笔记 · 《概率导论》

Yifeng Xu
中国科学院计算技术研究所
中国科学院大学

前言

这是《概率导论》[1] 一书的笔记，但不包括第 6 章与第 7 章关于随机过程的部分（《随机过程》相关内容将单独记录）. 另外，原书将离散随机变量和连续随机变量分别写在第 2 章和第 3 章中，导致许多概念需要重复书写，因此本笔记将这两章合并为了一章. 最后，本笔记附录 A 整理总结了常用随机变量及其期望、方差、矩母函数及性质，附录 B 整理了二元正态分布相关内容，附录 C 整理了正态分布的三个导出分布相关内容.

目录

1 样本空间与概率

1.1 概率模型

定义 1.1 (概率模型的基本构成)**.** 概率模型由样本空间和概率律构成：

- 样本空间 Ω：一个试验的所有可能结果的集合.
- 概率律：为试验结果的集合 A（称为事件）确定一个非负数 $\mathrm{P}(A)$，称为事件 A 的概率. 概率律需要满足概率公理.

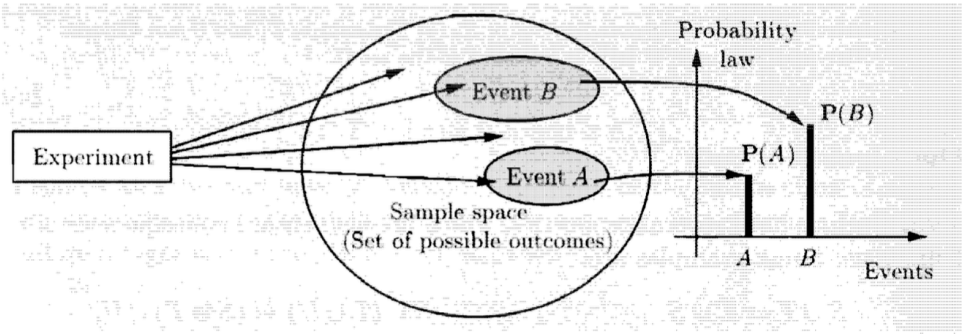

图 1: 概率模型的基本构成

定义 1.2 (概率公理)**.** 概率满足以下几条公理：

1. 非负性：对一切事件 A，满足 $\mathrm{P}(A) \geqslant 0$.
2. 归一化：整个样本空间为必然事件，即 $\mathrm{P}(\Omega) = 1$.
3. 可数可加性：设 $A_1, A_2, \ldots$ 是一列互不相容事件，则 $\mathrm{P}(A_1 \cup A_2 \cup \cdots) = \mathrm{P}(A_1) + \mathrm{P}(A_2) + \cdots$.

性质. 设 A, B, C 为事件，则由概率公理可以推导出如下性质：

- 空事件概率为 0，即 $\mathrm{P}(\varnothing) = 0$
- $A \subset B \implies \mathrm{P}(A) \leqslant \mathrm{P}(B)$
- $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$
- $\mathrm{P}(A \cup B) \leqslant \mathrm{P}(A) + \mathrm{P}(B)$
- $\mathrm{P}(A \cup B \cup C) = \mathrm{P}(A) + \mathrm{P}(A^c \cap B) + \mathrm{P}(A^c \cap B^c \cap C)$

1.2 条件概率

定义 1.3 (条件概率)**.** 设 A, B 为两个事件且 $\mathrm{P}(B) > 0$，定义 B 发生的条件下 A 发生的概率为：

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}, \quad \mathrm{P}(B) > 0$$

定理 1.1. 条件概率是一个公理化定义下的概率，从而概率的所有性质都适用于条件概率.

证明. 只需验证条件概率是否满足非负性、归一化和可数可加性即可.

1. 非负性：显然；
2. 归一化：

$$\mathrm{P}(\Omega|B) = \frac{\mathrm{P}(\Omega \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(B)}{\mathrm{P}(B)} = 1$$

3. 可数可加性：

$$\mathrm{P}\left(\bigcup_{i=1}^{n} A_i \middle| B\right) = \frac{\mathrm{P}\left(\left(\bigcup_{i=1}^{n} A_i\right) \cap B\right)}{\mathrm{P}(B)} = \frac{\mathrm{P}\left(\bigcup_{i=1}^{n} (A_i \cap B)\right)}{\mathrm{P}(B)}$$
$$= \frac{\sum_{i=1}^{n} \mathrm{P}(A_i \cap B)}{\mathrm{P}(B)} = \sum_{i=1}^{n} \mathrm{P}(A_i|B)$$

□

定理 1.2 (乘法公式)**.** 设 A, B 为两个事件，则由定义易知：

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

推论 1.3. 设 $A_1, A_2, \ldots, A_n$ 为 n 个事件，则：

$$\mathrm{P}(A_1 A_2 \cdots A_n) = \mathrm{P}(A_1)\mathrm{P}(A_2|A_1)\mathrm{P}(A_3|A_1A_2)\cdots\mathrm{P}(A_n|A_1A_2\cdots A_{n-1})$$

1.3 全概率公式和贝叶斯公式

定理 1.4 (全概率公式)**.** 设 $A_1, A_2, \cdots, A_n$ 互不相容，$B \subset A_1 \cup A_2 \cup \cdots \cup A_n$，则：

$$\mathrm{P}(B) = \sum_{i=1}^{n} \mathrm{P}(A_i)\mathrm{P}(B|A_i)$$

证明.

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

□

定理 1.5 (贝叶斯公式)**.** 设 $A_1, A_2, \cdots, A_n$ 互不相容且 $\mathrm{P}(A_i) > 0$，$B \subset A_1 \cup A_2 \cup \cdots \cup A_n$，则：

$$\mathrm{P}(A_i|B) = \frac{\mathrm{P}(A_i \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(A_i)\mathrm{P}(B|A_i)}{\sum\limits_{j=1}^{n} \mathrm{P}(A_j)\mathrm{P}(B|A_j)}$$

其中 $\mathrm{P}(A_i)$ 称作先验概率，$\mathrm{P}(A_i|B)$ 称作后验概率.

注解 (关于贝叶斯公式的理解). 视事件 A_i 是导致事件 B 发生的原因，我们对于事件 A_i 已有一个先验概率 $\mathrm{P}(A_i)$，现在事件 B 发生了，这必然给我们带了一定的信息，于是我们可以由此修正 A_i 发生的概率，得到 $\mathrm{P}(A_i|B)$，即后验概率.

1.4 独立性

定义 1.4 (独立)**.** 设 A, B 是两个事件，称 A 与 B 独立，若：

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$$

注释. 若 $\mathrm{P}(B) \neq 0$，则 A 与 B 独立等价于：

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

直观上，这说明事件 B 的发生与否并不给 A 带来信息，不改变 A 发生的概率.

注意. 事件的独立性常常不能直观地看出来. 例如，若事件 A 与事件 B 互不相容，并且 $\mathrm{P}(A) > 0, \mathrm{P}(B) > 0$，则它们一定不独立，因为 $\mathrm{P}(A \cap B) = 0 \neq \mathrm{P}(A)\mathrm{P}(B)$. 直观上，事件 B 发生意味着 A 一定没有发生，因此 B 的发生与否会给 A 带来信息.

定理 1.6. 设事件 A 与事件 B 独立，则 A 与 B^c 也独立.

证明. 由于 A 可写作两个不相容事件之并 $A = (A \cap B) \cup (A \cap B^c)$，故：

$$\mathrm{P}(A) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B^c)$$

于是：

$$\mathrm{P}(A \cap B^c) = \mathrm{P}(A) - \mathrm{P}(A \cap B) = \mathrm{P}(A) - \mathrm{P}(A)\mathrm{P}(B) = \mathrm{P}(A)(1 - \mathrm{P}(B)) = \mathrm{P}(A)\mathrm{P}(B^c)$$

□

定义 1.5 (条件独立)**.** 给定事件 C，称事件 A, B 在给定 C 下条件独立，若：

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

注释. A, B 条件独立并不能推出 A, B 独立，反之亦不成立.

定义 1.6 (一组事件的相互独立性)**.** 设 $A_1, A_2, \ldots, A_n$ 是一组事件，称它们相互独立，若：

$$\mathrm{P}\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} \mathrm{P}(A_i), \quad \forall S \subset \{1, 2, \ldots, n\}$$

注意 (两两独立与相互独立). 设 A_1, A_2, A_3 相互独立，则有：

$$\begin{aligned}
\mathrm{P}(A_1 \cap A_2) &= \mathrm{P}(A_1)\mathrm{P}(A_2) \\
\mathrm{P}(A_2 \cap A_3) &= \mathrm{P}(A_1)\mathrm{P}(A_3) \\
\mathrm{P}(A_1 \cap A_3) &= \mathrm{P}(A_1)\mathrm{P}(A_3) \\
\mathrm{P}(A_1 \cap A_2 \cap A_3) &= \mathrm{P}(A_1)\mathrm{P}(A_2)\mathrm{P}(A_3)
\end{aligned}$$

前三个式子称作两两独立. 但是第四个式子也非常重要，它并不是前三个式子的推论；反之，第四个式子也不能推出前三个式子.

例 1.1 (两两独立不能推出独立)**.** 设试验是抛掷两枚均匀的硬币，考虑事件：

$$H_1 = \{\text{第一次正面}\}, \quad H_2 = \{\text{第二次正面}\}, \quad D = \{\text{两次结果不同}\}$$

则易知 H_1 与 H_2 独立. 另外，

$$\mathrm{P}(D|H_1) = \frac{\mathrm{P}(D \cap H_1)}{\mathrm{P}(H_1)} = \frac{1/4}{1/2} = \frac{1}{2} = \mathrm{P}(D)$$

故 D 与 H_1 独立. 同理可知 D 与 H_2 独立，故 H_1, H_2, D 两两独立. 但是：

$$\mathrm{P}(H_1 \cap H_2 \cap D) = 0 \neq \mathrm{P}(H_1)\mathrm{P}(H_2)\mathrm{P}(D) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}$$

故 H_1, H_2, D 不是独立的.

2 随机变量

2.1 基本概念

定义 2.1 (随机变量)**.** 随机变量是试验结果的实值函数.

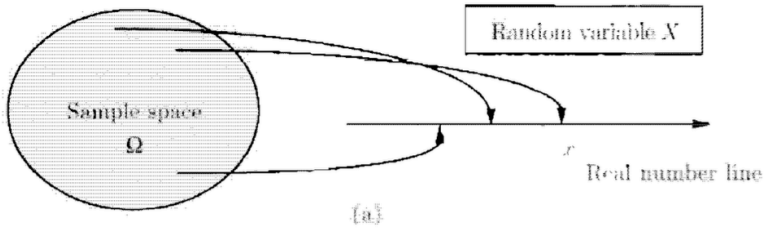

图 2: 随机变量示意图

注释. 一般用大写字母表示随机变量，小写字母表示其取值.

定义 2.2 (离散随机变量)**.** 若一个随机变量的值域为有限集或可数集，则称这个随机变量是离散的.

定义 2.3 (概率质量函数)**.** 定义随机变量 X 取值为 x 的概率为事件 $\{X = x\}$ 的概率，即所有与 x 对应的试验结果组成的事件的概率，记作 $p_X(x)$，即：

$$p_X(x) = \mathrm{P}(\{X = x\})$$

称 p_X 为 X 的概率质量函数 (PMF).

性质. PMF 满足非负性和归一化条件：

$$\sum_x p_X(x) = \sum_x \mathrm{P}(X = x) = 1$$

性质. 设 S 为任一 X 可能取值的集合，则：

$$\mathrm{P}(X \in S) = \sum_{x \in S} p_X(x)$$

例 2.1. 常见的离散随机变量包括伯努利、二项、几何和泊松随机变量等，详见附录 A.

定义 2.4 (连续随机变量，概率密度函数)**.** 对随机变量 X，若存在一个非负函数 f_X，使得：

$$\mathrm{P}(X \in B) = \int_B f_X(x)\mathrm{d}x$$

对实数轴的集合 B 都成立[1]，则称 X 为连续的随机变量，函数 f_X 称为概率密度函数 (PDF). 特别地，当 B 是一个区间时，有：

$$\mathrm{P}(a \leqslant X \leqslant b) = \int_a^b f_X(x)\mathrm{d}x$$

[1]本书只考虑黎曼积分，且 f_X 为有有限/可数个间断点的分段连续函数.

性质. PDF 满足非负性和归一化条件：

$$\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = \mathrm{P}(-\infty < X < \infty) = 1$$

性质. 对于充分小的 δ，有：

$$\mathrm{P}(x \leqslant X \leqslant x+\delta) = \int_{x}^{x+\delta} f_X(x)\mathrm{d}x \approx f(x)\cdot\delta$$

例 2.2. 常见的连续随机变量包括均匀、指数和正态随机变量等，详见附录 A.

2.2 分布函数

定义 2.5 (分布函数)**.** 设 X 是一个随机变量（离散或连续），定义其分布函数（CDF）F_X 为：

$$F_X(x) = \mathrm{P}(X \leqslant x) = \begin{cases} \sum\limits_{k\leqslant x} p_X(k), & X \text{ 离散} \\ \int_{-\infty}^{x} f_X(t)\mathrm{d}t, & X \text{ 连续} \end{cases}$$

注释. 分布函数统一刻画了离散和连续情形. 离散情形下的 PMF、连续情形下的 PDF 和一般情形下的 CDF 都是相应随机变量的概率律.

性质. 设 F_X 是随机变量 X 的分布函数，则：

- F_X 单调非减.
- $F_X(x) \to 0\ (x \to -\infty),; F_X(x) \to 1(x \to \infty)$.
- 当 X 是离散随机变量时，F_X 是阶梯函数.
- 当 X 是连续随机变量时，F_X 是连续函数.

定理 2.1 (分布列与分布函数)**.** 设 X 是离散随机变量且取整数值，则：

$$F_X(k) = \sum_{i=-\infty}^{k} p_X(i), \quad p_X(k) = F_X(k) - F_X(k-1)$$

定理 2.2 (概率密度函数与分布函数)**.** 设 X 是连续随机变量，则：

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t, \quad f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x}F_X(x)$$

第二个等式只在分布函数可微处成立.

2.3 期望和方差

定义 2.6 (期望/均值)**.** 随机变量 X 的期望定义为：

$$\mathbb{E}X = \begin{cases} \sum\limits_{x} x p_X(x), & X \text{ 离散} \\ \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x, & X \text{ 连续} \end{cases}$$

特别地，对于连续情形，若 f_X 不是绝对可积的，即 $\int_{-\infty}^{\infty}|x|f_X(x)\mathrm{d}x=\infty$，则称期望不存在.

定义 2.7 (矩，中心矩)**.** 定义随机变量 X 的 n 阶矩为 $\mathbb{E}[X^n]$，n 阶中心矩为 $\mathbb{E}[(X-\mathbb{E}X)^n]$.

定义 2.8 (方差，标准差)**.** 定义随机变量 X 的方差为其 2 阶中心矩，标准差为方差的平方根：

$$\mathrm{var}(X)=\mathbb{E}\left[(X-\mathbb{E}X)^2\right],\quad \sigma(X)=\sqrt{\mathrm{var}(X)}$$

定理 2.3 (随机变量函数的期望)**.** 设 X 是一随机变量，则 $Y=g(X)$ 的期望为：

$$\mathbb{E}Y=\mathbb{E}[g(X)]=\begin{cases}\displaystyle\sum_x g(x)p_X(x), & X\text{ 离散}\\ \displaystyle\int_{-\infty}^{+\infty} g(x)f_X(x)\mathrm{d}x, & X\text{ 连续}\end{cases}$$

因此，我们不必先求出 Y 的分布，只需知道 X 的分布就能求出 Y 的期望.

定理 2.4 (随机变量的线性函数的期望和方差)**.** 设 X 是一个随机变量，$Y=aX+b$，其中 a,b 为常数，则：

$$\mathbb{E}Y=a\mathbb{E}X+b,\quad \mathrm{var}(Y)=a^2\mathrm{var}(X)$$

证明. 仅对离散情形证明，连续情形类似.

$$\mathbb{E}Y=\sum_x(ax+b)p_X(x)=a\sum_x xp_X(x)+b\sum_x p_X(x)=a\mathbb{E}X+b$$
$$\mathrm{var}(Y)=\mathbb{E}[(Y-\mathbb{E}Y)^2]=\mathbb{E}[((aX+b)-(a\mathbb{E}X+b))^2]=a^2\mathbb{E}[(X-\mathbb{E}X)^2]=a^2\mathrm{var}(X)$$

□

定理 2.5 (用矩表达方差)**.** 设 X 是一个随机变量，则：

$$\mathrm{var}(X)=\mathbb{E}[X^2]-(\mathbb{E}X)^2$$

证明.

$$\begin{aligned}\mathrm{var}(X)&=\mathbb{E}\left[(X-\mathbb{E}X)^2\right]\\&=\mathbb{E}\left[X^2-2X\mathbb{E}X+(\mathbb{E}X)^2\right]\\&=\mathbb{E}[X^2]-2\mathbb{E}X\cdot\mathbb{E}X+(\mathbb{E}X)^2\\&=\mathbb{E}[X^2]-(\mathbb{E}X)^2\end{aligned}$$

□

例 2.3. 常见随机变量的期望和方差及其推导过程见附录 A.

2.4 联合分布与边缘分布

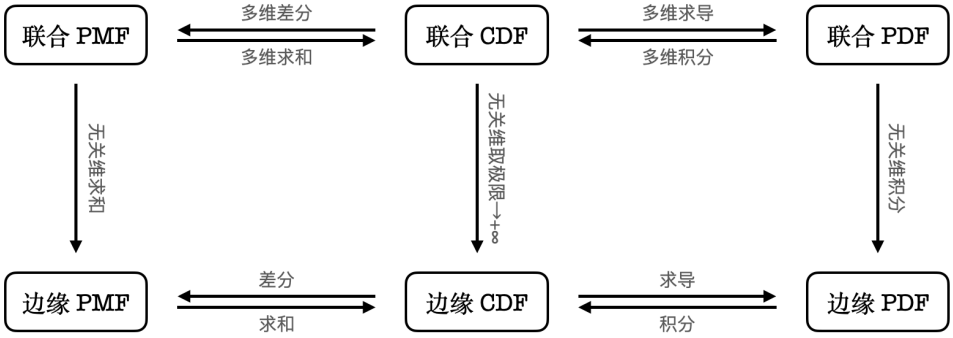

图 3: 联合分布与边缘分布的关系概览

定义 2.9 (联合概率质量函数)**.** 设 X, Y 是离散随机变量，定义 (X, Y) 取值 (x, y) 的概率为事件 $\{X = x, Y = y\}$ 的概率，记作 $p_{X,Y}(x, y)$，即：

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

称 $p_{X,Y}$ 为 X, Y 的联合概率质量函数.

定义 2.10 (联合概率密度函数)**.** 设 X, Y 是连续随机变量，若存在一个非负二元函数 $f_{X,Y}$ 使得：

$$\mathrm{P}((X, Y) \in B) = \iint_{(x,y)\in B} f_{X,Y}(x, y)\mathrm{d}x\mathrm{d}y$$

对平面上任意集合 B 成立，则称 $f_{X,Y}$ 为联合概率密度函数. 特别地，当 B 是一个矩形区域时有：

$$\mathrm{P}(a \leqslant X \leqslant b, c \leqslant Y \leqslant d) = \int_c^d \int_a^b f_{X,Y}(x, y)\mathrm{d}x\mathrm{d}y$$

性质. 联合 PDF 满足归一化条件：

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\mathrm{d}x\mathrm{d}y = 1$$

性质. 对于充分小的 δ，有：

$$\mathrm{P}(a \leqslant X \leqslant a + \delta, c \leqslant Y \leqslant c + \delta) = \int_c^{c+\delta} \int_a^{a+\delta} f_{X,Y}(x, y)\mathrm{d}x\mathrm{d}y \approx f_{X,Y}(a, c) \cdot \delta^2$$

定理 2.6 (边缘概率质量函数)**.** 设 X, Y 是离散随机变量且联合 PMF 为 $p_{X,Y}$，则：

$$p_X(x) = \sum_y p_{X,Y}(x, y), \quad p_Y(y) = \sum_x p_{X,Y}(x)$$

称 p_X 和 p_Y 为边缘概率质量函数.

定理 2.7 (边缘概率密度函数)**.** 设 X, Y 是连续随机变量且联合 PDF 为 $f_{X,Y}$，则：

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\mathrm{d}y, \quad f_Y(y) = \int_{-\infty}^{\infty} p_{X,Y}(x)\mathrm{d}x$$

称 f_X 和 f_Y 为边缘概率密度函数.

定义 2.11 (联合分布函数)**.** 设 X, Y 是两个随机变量（离散或连续），定义其联合分布函数为：

$$F_{X,Y}(x,y) = \mathrm{P}(X \leqslant x, Y \leqslant y)$$

定理 2.8 (联合概率密度函数与联合分布函数)**.** 若随机变量 X, Y 有联合概率密度函数 $f_{X,Y}$，则：

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(s,t)\mathrm{d}s\mathrm{d}t, \quad f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)$$

定理 2.9 (随机变量的二元函数的期望)**.** 设 X 和 Y 是随机变量，则 $Z = g(X,Y)$ 的期望为：

$$\mathbb{E}Z = \mathbb{E}[g(X,Y)] = \begin{cases} \displaystyle\sum_x \sum_y g(x,y) p_{X,Y}(x,y), & X, Y \text{ 离散} \\ \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y, & X, Y \text{ 连续} \end{cases}$$

定理 2.10 (随机变量的二元线性函数的期望)**.** 设 X 和 Y 是随机变量，a, b, c 为常数，则：

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}X + b\mathbb{E}Y + c$$

定理 2.11 (随机变量的多元线性函数的期望)**.** 设 $X_1, X_2, \ldots, X_n$ 是 n 个随机变量，$a_1, a_2, \ldots, a_n$ 是 n 个常数，则：

$$\mathbb{E}[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = a_1\mathbb{E}X_2 + a_2\mathbb{E}X_2 + \cdots + a_n\mathbb{E}X_n$$

2.5 条件

定义 2.12 (条件分布)**.** 设 X, Y 是离散随机变量，定义给定 $Y = y$ 下 X 的条件概率质量函数为：

$$p_{X|Y}(x|y) = \mathrm{P}(X = x|Y = y) = \frac{\mathrm{P}(X = x, Y = y)}{\mathrm{P}(Y = y)} = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

类似地，设 X, Y 是连续随机变量，定义给定 $Y = y$ 下 X 的条件概率密度函数为：

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

性质. 条件 PMF/PDF 满足归一化条件：

$$\sum_x p_{X|Y}(x|y) = 1, \quad X \text{ 离散}$$
$$\int_{-\infty}^{\infty} f_{X|Y}(x|y)\mathrm{d}x = 1, \quad X \text{ 连续}$$

定理 2.12 (条件分布与联合分布)**.** 设 X, Y 是随机变量，由定义可知：

$$p_{X,Y}(x,y) = p_Y(y)p_{X|Y}(x|y) = p_X(x)p_{Y|X}(y|x), \quad X \text{ 离散}$$
$$f_{X,Y}(x,y) = f_Y(y)f_{X|Y}(x|y) = f_X(x)f_{Y|X}(y|x), \quad X \text{ 连续}$$

定理 2.13 (条件分布与边缘分布)**.** 设 X, Y 是随机变量，则根据全概率公式，有：

$$p_X(x) = \sum_y p_{X|Y}(x|y)p_Y(y), \quad X \text{ 离散}$$
$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)\mathrm{d}y, \quad X \text{ 连续}$$

定理 2.14 (贝叶斯公式)**.** 设 X, Y 是随机变量，则有：

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_{Y|X}(y|x)p_X(x)}{\sum_x p_{Y|X}(y|x)p_X(x)}, \quad X, Y \text{ 离散}$$
$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y|x)f_X(x)\mathrm{d}x}, \quad X, Y \text{ 连续}$$

定义 2.13 (条件期望)**.** 设 X, Y 是随机变量，则给定 $Y = y$ 下 X 的条件期望为：

$$\mathbb{E}[X|Y=y] = \begin{cases} \sum_x x p_{X|Y}(x|y), & X \text{ 离散} \\ \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\mathrm{d}x, & X \text{ 连续} \end{cases}$$

对于随机变量的函数 $g(X)$，有：

$$\mathbb{E}[g(X)|Y=y] = \begin{cases} \sum_x g(x) p_{X|Y}(x|y), & X \text{ 离散} \\ \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y)\mathrm{d}x, & X \text{ 连续} \end{cases}$$

注意. $\mathbb{E}[X|Y=y]$ 是一个数，其值依赖于 y，因此 $\mathbb{E}[X|Y]$ 是关于随机变量 Y 的函数，从而也是一个随机变量.

定理 2.15 (全期望公式)**.** 设 X, Y 是随机变量（离散或连续）且 X 期望存在，则：

$$\mathbb{E}X = \mathbb{E}[\mathbb{E}[X|Y]] = \begin{cases} \sum_y \mathbb{E}[X|Y=y]p_Y(y), & Y \text{ 离散} \\ \int_{-\infty}^{\infty} \mathbb{E}[X|Y=y]f_Y(y)\mathrm{d}y, & Y \text{ 连续} \end{cases}$$

证明. 仅对离散情形证明，连续情形类似.

$$\begin{aligned} \mathbb{E}X &= \sum_x x p_X(x) \\ &= \sum_x x \sum_y p_{X|Y}(x|y)p_Y(y) \\ &= \sum_y p_Y(y) \sum_x x p_{X|Y}(x|y) \\ &= \sum_y \mathbb{E}[X|Y=y]p_Y(y) \end{aligned}$$

其中第二行应用了全概率公式. □

注解. 全期望公式常常是“反过来”使用的：当 $\mathbb{E}X$ 不好计算时，引入 Y 转而计算 $\mathbb{E}[\mathbb{E}[X|Y]]$.

定义 2.14 (条件方差)**.** 设 X, Y 是随机变量，则给定 $Y = y$ 下 X 的条件方差为：

$$\mathrm{var}(X|Y=y) = \mathbb{E}[(X - \mathbb{E}[X|Y=y])^2|Y=y]$$

注意. $\mathrm{var}(X|Y=y)$ 是一个数，其值依赖于 y，因此 $\mathrm{var}(X|Y)$ 是关于随机变量 Y 的函数，从而也是一个随机变量，并且有：

$$\mathrm{var}(X|Y) = \mathbb{E}[(X - \mathbb{E}[X|Y])^2|Y] = \mathbb{E}[X^2|Y] - (\mathbb{E}[X|Y])^2$$

定理 2.16 (全方差公式)**.** 设 X, Y 是随机变量且 X 方差存在，则：

$$\mathrm{var}(X) = \mathbb{E}[\mathrm{var}(X|Y)] + \mathrm{var}(\mathbb{E}[X|Y])$$

证明.

$$\begin{aligned}
\mathrm{var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}X)^2 \\
&= \mathbb{E}[\mathbb{E}[X^2|Y]] - (\mathbb{E}[\mathbb{E}[X|Y]])^2 \\
&= \mathbb{E}[\mathbb{E}[X^2|Y]] - \mathbb{E}[(\mathbb{E}[X|Y])^2] + \mathbb{E}[(\mathbb{E}[X|Y])^2] - (\mathbb{E}[\mathbb{E}[X|Y]])^2 \\
&= \mathbb{E}[\mathrm{var}(X|Y)] + \mathrm{var}(\mathbb{E}[X|Y])
\end{aligned}$$

□

2.6 独立性

定义 2.15 (独立)**.** 设 X, Y 是随机变量，称 X 与 Y 独立，若：

$$\begin{aligned}
p_{X,Y}(x,y) &= p_X(x)p_Y(y), \quad \forall x, y, \quad X, Y \text{ 离散} \\
f_{X,Y}(x,y) &= f_X(x)f_Y(y), \quad \forall x, y, \quad X, Y \text{ 连续}
\end{aligned}$$

注释. 离散情形下，若 $p_Y(y) > 0$，则 X 与 Y 独立等价于：

$$p_{X|Y}(x|y) = p_X(x),\ \forall y$$

直观上，这说明 Y 的取值不会给 X 的取值带来信息. 连续情形同理.

定义 2.16 (条件独立)**.** 设 X, Y, Z 是随机变量，给定 $Z = z$ 下（设 $p_Z(z) > 0$ 或 $f_Z(z) > 0$），称随机变量 X 与 Y 条件独立，若：

$$\begin{aligned}
p_{X,Y|Z}(x,y|z) &= p_{X|Z}(x|z)p_{Y|Z}(y|z), \quad \forall x, y, \quad X, Y \text{ 离散} \\
f_{X,Y|Z}(x,y|z) &= f_{X|Z}(x|z)f_{Y|Z}(y|z), \quad \forall x, y, \quad X, Y \text{ 连续}
\end{aligned}$$

定理 2.17. 若随机变量 X, Y 相互独立，则：

$$\mathbb{E}[XY] = \mathbb{E}X\mathbb{E}Y$$

进一步地，对任意函数 g, h，有：

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

证明. 仅对离散情形证明，连续情形类似.

$$\begin{aligned}\mathbb{E}[XY] &= \sum_x \sum_y xyp_{X,Y}(x,y) \\ &= \sum_x \sum_y xyp_X(x)p_Y(y) \\ &= \sum_x xp_X(x) \sum_y yp_Y(y) \\ &= \mathbb{E}X\mathbb{E}Y\end{aligned}$$

第二个式子类似可证. □

推论 2.18. 若随机变量 X, Y 独立，则对任意函数 g, h，有 $g(X)$ 与 $h(Y)$ 独立.

定理 2.19. 若随机变量 X, Y 相互独立，则：

$$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y)$$

证明. 令 $\tilde{X} = X - \mathbb{E}X$, $\tilde{Y} = Y - \mathbb{E}Y$，由于方差在加减常数后保持不变，所以：

$$\begin{aligned}\mathrm{var}(X+Y) &= \mathrm{var}(\tilde{X}+\tilde{Y}) \\ &= \mathbb{E}[(\tilde{X}+\tilde{Y})^2] \\ &= \mathbb{E}[\tilde{X}^2 + 2\tilde{X}\tilde{Y} + \tilde{Y}^2] \\ &= \mathbb{E}[\tilde{X}^2] + 2\mathbb{E}[\tilde{X}\tilde{Y}] + \mathbb{E}[\tilde{Y}^2] \\ &= \mathrm{var}X + \mathrm{var}Y\end{aligned}$$

其中利用了 $\mathbb{E}[\tilde{X}\tilde{Y}] = \mathbb{E}[\tilde{X}]\mathbb{E}[\tilde{Y}] = 0$. □

定义 2.17 (多个随机变量独立性)**.** 称随机变量 X, Y, Z 相互独立，若：

$$\begin{aligned}p_{X,Y,Z}(x,y,z) &= p_X(x)p_Y(y)p_Z(z), \quad \forall x,y,z, \quad X,Y,Z \text{ 离散} \\ f_{X,Y,Z}(x,y,z) &= f_X(x)f_Y(y)f_Z(z), \quad \forall x,y,z, \quad X,Y,Z \text{ 连续}\end{aligned}$$

定理 2.20 (多个独立随机变量和的方差)**.** 设 $X_1, X_2, \ldots, X_n$ 为相互独立的随机变量，则：

$$\mathrm{var}(X_1 + X_2 + \cdots + X_n) = \mathrm{var}(X_1) + \mathrm{var}(X_2) + \cdots + \mathrm{var}(X_n)$$

3 随机变量的深入内容

3.1 随机变量的函数的分布

定理 3.1 (随机变量的函数)**.** 设 X 是离散随机变量，则 $Y=g(X)$ 的 PMF 为：

$$p_Y(y)=\sum_{\{x|y=g(x)\}}p_X(x)$$

设 X 是连续随机变量，则 $Y=g(X)$ 的 CDF 为：

$$F_Y(y)=\mathrm{P}(Y\leqslant y)=\mathrm{P}(g(X)\leqslant y)=\int\limits_{\{x|g(x)\leqslant y\}}f_X(x)\mathrm{d}x$$

进而 PDF 为：

$$f_Y(y)=\frac{\mathrm{d}}{\mathrm{d}y}F_Y(y)$$

定理 3.2 (线性函数情形)**.** 设 X 是连续随机变量，$a,b\in\mathbb{R}$ 且 $a\neq 0$，设 $Y=aX+b$，则：

$$f_Y(y)=\frac{1}{|a|}f_X\left(\frac{y-b}{a}\right)$$

证明. 先求 Y 的 CDF：

$$F_Y(y)=\mathrm{P}(Y\leqslant y)=\mathrm{P}(aX+b\leqslant y)=\begin{cases}\mathrm{P}\left(X\leqslant\frac{y-b}{a}\right)=F_X\left(\frac{y-b}{a}\right), & a>0\\ \mathrm{P}\left(X\geqslant\frac{y-b}{a}\right)=1-F_X\left(\frac{y-b}{a}\right), & a<0\end{cases}$$

然后求导得到 Y 的 PDF：

$$f_Y(y)=\frac{\mathrm{d}F_Y(y)}{\mathrm{d}y}=\begin{cases}\frac{1}{a}f_X\left(\frac{y-b}{a}\right), & a>0\\ -\frac{1}{a}f_X\left(\frac{y-b}{a}\right), & a<0\end{cases}\quad=\quad\frac{1}{|a|}f_X\left(\frac{y-b}{a}\right)$$

□

例 3.1 (正态分布的线性变换仍然是正态分布)**.** 设 $X\sim N(\mu,\sigma^2)$，$a,b\in\mathbb{R}$ 且 $a\neq 0$，$Y=aX+b$，则根据定理 3.2，有：

$$\begin{aligned}f_Y(y)&=\frac{1}{|a|}f_X\left(\frac{y-b}{a}\right)\\&=\frac{1}{\sqrt{2\pi}\sigma|a|}\exp\left(-\frac{\left(\frac{y-b}{a}-\mu\right)^2}{2\sigma^2}\right)\\&=\frac{1}{\sqrt{2\pi}\sigma|a|}\exp\left(-\frac{(y-a\mu-b)^2}{2a^2\sigma^2}\right)\end{aligned}$$

所以，$Y=aX+b\sim N(a\sigma+b,a^2\sigma^2)$.

定理 3.3 (单调函数情形)**.** 设 X 是连续随机变量，g 是严格单调的可逆函数，且其反函数 h 可微，

则 $Y=g(X)$ 在其支撑集 $\{y \mid f_Y(y)>0\}$ 上的 PDF 是：

$$f_Y(y)=f_X(h(y))\left|\frac{\mathrm{d}}{\mathrm{d}y}h(y)\right|$$

证明. 先求 Y 的 CDF：

$$F_Y(y)=\mathrm{P}(Y\leqslant y)=\begin{cases}\mathrm{P}(X\leqslant h(y))=F_X(h(y)), & g \text{ 单调递增}\\ \mathrm{P}(X\geqslant h(y))=1-F_X(h(y)), & g \text{ 单调递减}\end{cases}$$

于是求导得：

$$f_Y(y)=\begin{cases}f_X(h(y))h'(y), & g \text{ 单调递增}\\ -f_X(h(y))h'(y), & g \text{ 单调递减}\end{cases}\quad=\quad f_X(h(y))\left|\frac{\mathrm{d}}{\mathrm{d}y}h(y)\right|$$

□

定理 3.4 (随机变量的二元函数)**.** 设 X,Y 是离散随机变量，则 $Z=g(X,Y)$ 的 PMF 为：

$$p_Z(z)=\sum_{\{(x,y)|z=g(x,y)\}}p_{X,Y}(x,y)$$

设 X,Y 是连续随机变量，则 $Z=g(X,Y)$ 的 CDF 为：

$$F_Z(z)=\mathrm{P}(Z\leqslant z)=\mathrm{P}(g(X,Y)\leqslant z)=\iint_{\{(x,y)|g(x,y)\leqslant z\}}f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y$$

进而 PDF 为：

$$f_Z(z)=\frac{\mathrm{d}}{\mathrm{d}z}F_Z(z)$$

例 3.2 (瑞利分布)**.** 设 $X,Y\sim N(0,\sigma^2)$ 且相互独立，$R=\sqrt{X^2+Y^2}$，称 R 服从瑞利分布. 首先计算 CDF：

$$\begin{aligned}F_R(r)&=\mathrm{P}(R\leqslant r)=\mathrm{P}(X^2+Y^2\leqslant r^2)\\&=\iint_{\{(x,y)|x^2+y^2\leqslant r^2\}}p_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y\\&=\iint_{x^2+y^2\leqslant r^2}\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}}\mathrm{d}x\mathrm{d}y\\&=\frac{1}{2\pi\sigma^2}\int_0^{2\pi}\mathrm{d}\theta\int_0^r e^{-\frac{\rho^2}{2\sigma^2}}\rho\mathrm{d}\rho\\&=e^{-\frac{\rho^2}{2\sigma^2}}\Big|_r^0=1-e^{-\frac{r^2}{2\sigma^2}}\end{aligned}$$

故瑞利分布的 PDF 为：

$$f_R(r)=\frac{\mathrm{d}F_R(r)}{\mathrm{d}r}=\frac{r}{\sigma^2}e^{-\frac{r^2}{2\sigma^2}}$$

定理 3.5 (极值分布)**.** 设 X,Y 是独立的随机变量，$M=\max(X,Y),N=\min(X,Y)$，则：

$$F_M(z)=F_X(z)F_Y(z),\qquad F_N(z)=1-(1-F_X(z))(1-F_Y(z))$$

证明.

$$\begin{aligned}F_M(z)&=\mathrm{P}(M\leqslant z)=\mathrm{P}(X\leqslant z,Y\leqslant z)=\mathrm{P}(X\leqslant z)\mathrm{P}(Y\leqslant z)=F_X(z)F_Y(z)\\F_N(z)&=\mathrm{P}(N\leqslant z)=1-\mathrm{P}(N>z)=1-\mathrm{P}(X>z,Y>z)\\&=1-\mathrm{P}(X>z)\mathrm{P}(Y>z)=1-(1-F_X(z))(1-F_Y(z))\end{aligned}$$

□

推论 3.6. 设 $X_1,X_2,\cdots,X_n$ 是 n 个相互独立的随机变量，$M=\max\limits_{1\leqslant i\leqslant n}X_i$，$N=\min\limits_{1\leqslant i\leqslant n}X_i$，则：

$$F_M(z)=\prod_{i=1}^{n}F_{X_i}(z),\qquad F_N(z)=1-\prod_{i=1}^{n}(1-F_{X_i}(z))$$

定理 3.7 (独立随机变量之和——卷积)**.** 设 X,Y 是独立的离散随机变量，$Z=X+Y$，则：

$$p_Z(z)=\sum_x p_X(x)p_Y(z-x)$$

设 X,Y 是独立的连续随机变量，$Z=X+Y$，则：

$$f_Z(z)=\int_{-\infty}^{+\infty}f_X(x)f_Y(z-x)\mathrm{d}x$$

称上述两个式子为卷积（convolution）.

例 3.3 (独立正态随机变量之和仍服从正态分布)**.** 设 $X\sim N(\mu_x,\sigma_x^2)$，$Y\sim N(\mu_y,\sigma_y^2)$ 且相互独立，$Z=X+Y$，则根据定理 3.7，有：

$$f_Z(z)=\int_{-\infty}^{+\infty}\frac{1}{2\pi\sigma_x\sigma_y}e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}-\frac{(z-x-\mu_y)^2}{2\sigma_y^2}}\mathrm{d}x=\int_{-\infty}^{+\infty}\frac{1}{2\pi\sigma_x\sigma_y}e^{-\frac{1}{2}\left[\frac{u^2}{\sigma_x^2}+\frac{(v-u)^2}{\sigma_y^2}\right]}\mathrm{d}u$$

其中做了代换 $u=x-\mu_x$, $v=z-\mu_x-\mu_y$. 由于：

$$\frac{u^2}{\sigma_x^2}+\frac{(v-u)^2}{\sigma_y^2}=\frac{u^2}{\sigma_x^2}+\frac{u^2}{\sigma_y^2}+\frac{v^2}{\sigma_y^2}-\frac{2uv}{\sigma_y^2}=\left(\frac{\sqrt{\sigma_x^2+\sigma_y^2}}{\sigma_x\sigma_y}u-\frac{\sigma_x v}{\sigma_y\sqrt{\sigma_x^2+\sigma_y^2}}\right)^2+\frac{v^2}{\sigma_x^2+\sigma_y^2}$$

令 $t=\frac{\sqrt{\sigma_x^2+\sigma_y^2}}{\sigma_x\sigma_y}u-\frac{\sigma_x v}{\sigma_y\sqrt{\sigma_x^2+\sigma_y^2}}$，则：

$$f_Z(z)=\frac{1}{2\pi\sigma_x\sigma_y}\frac{\sigma_x\sigma_y}{\sqrt{\sigma_x^2+\sigma_y^2}}e^{\frac{v^2}{\sigma_x^2+\sigma_y^2}}\int_{-\infty}^{+\infty}e^{-\frac{1}{2}t^2}\mathrm{d}t=\frac{1}{\sqrt{2\pi}\sqrt{\sigma_x^2+\sigma_y^2}}e^{\frac{(z-\mu_x-\mu_y)^2}{\sigma_x^2+\sigma_y^2}}$$

所以 $Z\sim N(\mu_x+\mu_y,\sigma_x^2+\sigma_y^2)$.

例 3.4 (n 个独立正态随机变量之和)**.** 设 $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, n$ 且相互独立，则：

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

定理 3.8 (随机变量的多元函数（可逆函数情形）)**.** 设 $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ 是一个随机向量，T 是 $\mathbb{R}^n$ 上的一可逆映射，$\mathbf{Y} = T(\mathbf{X})$，则 $\mathbf{Y}$ 的 PDF 为：

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(T^{-1}(\mathbf{y}))|J|$$

其中，J 表示 $T^{-1} : \mathbf{y} \mapsto \mathbf{x}$，即 $\mathbf{x} = T^{-1}(\mathbf{y})$ 的雅各比行列式：

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}$$

证明. 设 $D \subset \mathbb{R}^n$ 是一个性质好的集合，则：

$$\begin{aligned} \mathrm{P}(\mathbf{Y} \in D) &= \mathrm{P}(T(\mathbf{X}) \in D) \\ &= \mathrm{P}(\mathbf{X} \in T^{-1}(D)) && \text{两边同时施以 } T^{-1} \\ &= \int_{T^{-1}(D)} f_{\mathbf{X}}(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \int_{D} f_{\mathbf{X}}(T^{-1}(y))|J|\mathrm{d}\mathbf{y} && \text{变量代换 } \mathbf{y} = T(\mathbf{x}) \end{aligned}$$

又：

$$\mathrm{P}(\mathbf{Y} \in D) = \int_{D} f_{\mathbf{Y}}(\mathbf{y})\mathrm{d}\mathbf{y}$$

根据 D 一定的任意性，可知：

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(T^{-1}(\mathbf{y}))|J|$$

□

3.2 协方差与相关系数

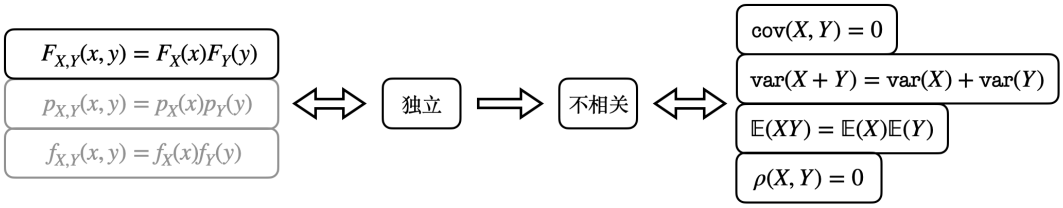

图 4: 随机变量的独立性与相关性示意图

定义 3.1 (协方差与相关系数)**.** 设 X, Y 是两个随机变量，定义协方差与相关系数分别为：

$$\mathrm{cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)], \quad \rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}X\mathrm{var}Y}}$$

当 $\mathrm{cov}(X,Y) = \rho(X,Y) = 0$ 时，称 X 和 Y 不相关.

性质. 从定义出发容易证明以下性质：

- $\mathrm{cov}(X,X) = \mathrm{var}(X)$
- $\mathrm{cov}(X, aY+b) = a \cdot \mathrm{cov}(X,Y)$
- $\mathrm{cov}(X, Y+Z) = \mathrm{cov}(X,Y) + \mathrm{cov}(X,Z)$

性质. $|\rho(X,Y)| \leqslant 1$.

证明. 对 $\forall t \in \mathbb{R}$，由于 $\mathrm{var}(Y-tX) = t^2\mathrm{var}X - 2t\mathrm{cov}(X,Y) + \mathrm{var}Y \geqslant 0$，所以：

$$\Delta = 4\mathrm{cov}^2(X,Y) - 4\mathrm{var}X\mathrm{var}Y \leqslant 0$$

即有 $|\rho(X,Y)| \leqslant 1$. □

定理 3.9. 设 X, Y 是两个随机变量，则：

$$\mathrm{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y$$

证明.

$$\begin{aligned}
\mathrm{cov}(X,Y) &= \mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)] \\
&= \mathbb{E}[XY - \mathbb{E}X\cdot Y - X\cdot\mathbb{E}Y + \mathbb{E}X\mathbb{E}Y] \\
&= \mathbb{E}[XY] - \mathbb{E}X\cdot\mathbb{E}Y - \mathbb{E}X\cdot\mathbb{E}Y + \mathbb{E}X\cdot\mathbb{E}Y \\
&= \mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y
\end{aligned}$$

□

推论 3.10 (独立与相关)**.** 设 X, Y 是两个随机变量，若 X, Y 独立，则 X, Y 不相关.

注意. 反之不成立，不相关不能推出独立.

例 3.5 (不相关且不独立)**.** 设随机变量 X, Y 以 $1/4$ 的概率取 $(1,0),(0,1),(-1,0),(0,-1)$，则 $\mathbb{E}X = \mathbb{E}Y = \mathbb{E}[XY] = 0$，于是 $\mathrm{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y = 0$，故 X, Y 不相关. 但是 $p_{X,Y}(1,0) = 1/4 \neq p_X(1)p_Y(0) = 1/4 \cdot 1/2 = 1/8$，故二者不独立. 直观上，$X$ 取非零值就要求 Y 取零值，因此不独立.

定理 3.11 (随机变量和的方差)**.** 设随机变量 X, Y 有有限的方差，则：

$$\mathrm{var}(X+Y) = \mathrm{var}X + 2\mathrm{cov}(X,Y) + \mathrm{var}Y$$

更一般的，设随机变量 $X_1, X_2, \ldots, X_n$ 有有限的方差，则：

$$\mathrm{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{var}(X_i) + \sum_{\{(i,j)|i\neq j\}} \mathrm{cov}(X_i, X_j)$$

证明. 设 $\tilde{X}_i = X_i - \mathbb{E}[X_i]$，则：

$$\begin{aligned}
\mathrm{var}\left(\sum_{i=1}^{n} X_i\right) &= \mathbb{E}\left[\left(\sum_{i=1}^{n} \tilde{X}_i^2\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{X}_i \tilde{X}_j\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}\left[\tilde{X}_i \tilde{X}_j\right] \\
&= \sum_{i=1}^{n} \mathbb{E}[\tilde{X}_i^2] + \sum_{\{(i,j)|i\neq j\}}^{n} \mathbb{E}\left[\tilde{X}_i \tilde{X}_j\right] \\
&= \sum_{i=1}^{n} \mathrm{var}(X_i) + \sum_{\{(i,j)|i\neq j\}} \mathrm{cov}(X_i, X_j)
\end{aligned}$$

□

推论 3.12. 设随机变量 X, Y 有有限的方差，a, b 为常数，则：

$$\mathrm{var}(aX + bY) = a^2 \mathrm{var}X + 2ab\mathrm{cov}(X, Y) + b^2 \mathrm{var}Y$$

或写作矩阵形式：

$$\mathrm{var}(aX + bY) = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} \mathrm{var}X & \mathrm{cov}(X,Y) \\ \mathrm{cov}(Y,X) & \mathrm{var}Y \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

3.3 再论条件期望与条件方差

定义 3.2 (条件期望作为估计量)**.** 设 X, Y 是随机变量，若将 Y 视作能提供关于 X 的信息的观测值，则可将条件期望视为给定 Y 下对 X 的估计，记作：

$$\hat{X} = \mathbb{E}[X|Y]$$

注意 $\hat{X}$ 是随机变量 Y 的函数. 进一步地，记估计误差为：

$$\tilde{X} = \hat{X} - X$$

则 $\tilde{X}$ 是随机变量 X, Y 的二元函数.

性质. 根据全期望公式易知，条件期望是无偏估计，即 $\mathbb{E}\hat{X} = \mathbb{E}X$，或 $\mathbb{E}\tilde{X} = 0$.

性质. $\mathbb{E}[\tilde{X}|Y] = 0$，即对任意 y，都有 $\mathbb{E}[\tilde{X}|Y = y] = 0$.

证明.

$$\mathbb{E}[\tilde{X}|Y] = \mathbb{E}[(\hat{X} - X)|Y] = \mathbb{E}[\hat{X}|Y] - \mathbb{E}[X|Y] = \hat{X} - \hat{X} = 0$$

□

性质. 估计量 $\hat{X}$ 与估计误差 $\tilde{X}$ 不相关.

证明.

$$\mathbb{E}[\hat{X}\tilde{X}] = \mathbb{E}[\mathbb{E}[\hat{X}\tilde{X}|Y]] = \mathbb{E}[\hat{X}\mathbb{E}[\tilde{X}|Y]] = 0 = \mathbb{E}\hat{X}\mathbb{E}\tilde{X}$$

□

性质. $\mathrm{var}(X) = \mathrm{var}(\hat{X}) + \mathrm{var}(\tilde{X})$.

证明. 由于 $\hat{X}$ 与 $\tilde{X}$ 不相关，故 $\mathrm{cov}(\hat{X}, \tilde{X}) = 0$，又 $X = \hat{X} + \tilde{X}$，故：

$$\mathrm{var}(X) = \mathrm{var}(\hat{X}) + 2\mathrm{cov}(\hat{X}, \tilde{X}) + \mathrm{var}(\tilde{X}) = \mathrm{var}(\hat{X}) + \mathrm{var}(\tilde{X})$$

□

3.4 矩母函数

定义 3.3 (矩母函数)**.** 设 X 是一个随机变量，定义其矩母函数为：

$$M_X(s) = \mathbb{E}[e^{sX}] = \begin{cases} \sum\limits_x e^{sx} p_X(x), & X \text{ 离散} \\ \int_{-\infty}^{\infty} e^{sx} f_X(x)\mathrm{d}x, & X \text{ 连续} \end{cases}$$

其定义域为使得 $\mathbb{E}[e^{sX}]$ 存在的 s. 在上下文清晰时可简记作 $M(s)$.

定理 3.13 (矩母函数计算矩)**.** 设随机变量 X 的矩母函数为 $M(s)$，则：

$$\left.\frac{\mathrm{d}^n}{\mathrm{d}s^n} M(s)\right|_{s=0} = \mathbb{E}[X^n]$$

证明. 假设积分（期望）与微分可交换，则：

$$\frac{\mathrm{d}^n}{\mathrm{d}s^n} M(s) = \frac{\mathrm{d}^n}{\mathrm{d}s^n}\mathbb{E}[e^{sX}] = \mathbb{E}\left[\frac{\mathrm{d}^n}{\mathrm{d}s^n} e^{sX}\right] = \mathbb{E}[X^n e^{sX}]$$

代入 $s = 0$ 得：

$$\left.\frac{\mathrm{d}^n}{\mathrm{d}s^n} M(s)\right|_{s=0} = \mathbb{E}[X^n]$$

□

定理 3.14 (矩母函数与分布)**.** 若随机变量 X 的矩母函数 $M_X(s)$ 满足：存在一个正数 a，使得对 $\forall s \in [-a, a]$，$M_X(s)$ 都是有限的，则矩母函数 $M_X(s)$ 唯一决定 X 的分布函数.

定理 3.15 (独立随机变量之和)**.** 设 X, Y 是独立随机变量，$Z = X + Y$，则：

$$M_Z(s) = M_X(s)M_Y(s)$$

证明.

$$M_Z(s) = \mathbb{E}[e^{sZ}] = \mathbb{E}[e^{s(X+Y)}] = \mathbb{E}[e^{sX}e^{sY}] = \mathbb{E}[e^{sX}]\mathbb{E}[e^{sY}] = M_X(s)M_Y(s)$$

□

推论 3.16. 设 $X_1, X_2, \ldots, X_n$ 是独立随机变量，$Z = X_1 + X_2 + \cdots + X_n$，则：

$$M_Z(s) = M_{X_1}(s)M_{X_2}(s)\cdots M_{X_n}(s)$$

定义 3.4 (多元矩母函数)**.** 设 $X_1, X_2, \ldots, X_n$ 是随机变量，定义它们的多元矩母函数为：

$$M_{X_1,X_2,\ldots,X_n}(s_1, s_2, \ldots, s_n) = \mathbb{E}[e^{s_1X_1+s_2X_2+\cdots+s_nX_n}]$$

3.5 随机个随机变量之和

定理 3.17 (随机个随机变量之和的期望、方差与矩母函数)**.** 设 N 是取正整数值的随机变量，$X_1, X_2, \ldots$ 是独立同分布的随机变量，且 $N, X_1, X_2, \ldots$ 相互独立（即这些随机变量的任意有限子集都是独立的）. 设 $Y = X_1 + X_2 + \cdots + X_N$，则：

$$\mathbb{E}Y = \mathbb{E}N\mathbb{E}X$$
$$\mathrm{var}(Y) = \mathbb{E}N\mathrm{var}(X) + (\mathbb{E}X)^2\mathrm{var}(N)$$
$$M_Y(s) = \sum_{n=1}^{\infty}(M_X(s))^n p_N(n)$$

其中 $\mathbb{E}X, \mathrm{var}(X), M_X(s)$ 表示各 X_i 的期望、方差和矩母函数.

证明. 给定正整数 n，随机变量 $X_1 + X_2 + \cdots + X_n$ 与 N 独立，故与事件 $\{N = n\}$ 独立，故：

$$\begin{aligned}\mathbb{E}[Y|N = n] &= \mathbb{E}[X_1 + X_2 + \cdots + X_N|N = n] \\ &= \mathbb{E}[X_1 + X_2 + \cdots + X_n|N = n] \\ &= \mathbb{E}[X_1 + X_2 + \cdots + X_n] \\ &= n\mathbb{E}X\end{aligned}$$

这对于任意非负整数 n 都成立，因此：

$$\mathbb{E}[Y|N] = N\mathbb{E}X$$

于是根据全期望公式，有：

$$\mathbb{E}Y = \mathbb{E}[\mathbb{E}[Y|N]] = \mathbb{E}[N\mathbb{E}X] = \mathbb{E}N\mathbb{E}X$$

类似地，给定正整数 n，有：

$$\begin{aligned}\mathrm{var}(Y|N=n) &= \mathrm{var}(X_1+X_2+\cdots+X_N|N=n)\\ &= \mathrm{var}(X_1+X_2+\cdots+X_n|N=n)\\ &= \mathrm{var}(X_1+X_2+\cdots+X_n)\\ &= n\mathrm{var}(X)\end{aligned}$$

这对任意正整数 n 都成立，因此：

$$\mathrm{var}(Y|N) = N\mathrm{var}(X)$$

于是根据全方差公式，有：

$$\begin{aligned}\mathrm{var}(Y) &= \mathbb{E}[\mathrm{var}(Y|N)] + \mathrm{var}(\mathbb{E}[Y|N])\\ &= \mathbb{E}[N\mathrm{var}(X)] + \mathrm{var}(N\mathbb{E}X)\\ &= \mathbb{E}N\mathrm{var}(X) + (\mathbb{E}X)^2\mathrm{var}(N)\end{aligned}$$

类似地，给定正整数 n，有：

$$\begin{aligned}\mathbb{E}[e^{sY}|N=n] &= \mathbb{E}[e^{s(X_1+X_2+\cdots+X_N}|N=n]\\ &= \mathbb{E}[e^{s(X_1+X_2+\cdots+X_n}|N=n]\\ &= \mathbb{E}[e^{s(X_1+X_2+\cdots+X_n}]\\ &= \mathbb{E}[e^{sX_1}]\mathbb{E}[e^{sX_2}]\cdots\mathbb{E}[e^{sX_n}\\ &= (M_X(s))^n\end{aligned}$$

上式对任意正整数 n 都成立，因此：

$$\mathbb{E}[e^{sY}|N] = (M_X(s))^N$$

于是根据全期望公式，有：

$$M_Y(s) = \mathbb{E}[e^{sY}] = \mathbb{E}[\mathbb{E}[e^{sY}|N]] = \mathbb{E}[(M_X(s))^N] = \sum_{n=1}^{\infty}(M_X(s))^n p_N(n)$$

□

注释. 对比 $M_Y(s)$ 与 $M_N(s)$：

$$\begin{aligned}M_Y(s) &= \sum_{n=1}^{\infty}(M_X(s))^n p_N(n)\\ M_N(s) &= \sum_{n=1}^{\infty}(e^s)^n p_N(n)\end{aligned}$$

可以看见 $M_Y(s)$ 就是将 $M_N(s)$ 中的函数 e^s 替换为 X_i 的矩母函数 $M_X(s)$.

4 极限理论

给定一列独立同分布随机变量 $X_1, X_2, \ldots$，定义前 n 项和 $S_n = X_1 + X_2 + \cdots + X_n$，本章的极限理论研究 S_n 及其相关变量在 $n \to \infty$ 时的极限性质.

4.1 马尔可夫不等式与切比雪夫不等式

定理 4.1 (马尔可夫不等式)**.** 设随机变量 X **只取非负值**，则对任意 $a > 0$，有：

$$\mathbb{P}(X \geqslant a) \leqslant \frac{\mathbb{E}X}{a}$$

注释. 粗略来讲，马尔可夫不等式指出，一个非负随机变量如果均值很小，那么该随机变量取大值的概率也很小.

证明. 这里假设 X 是连续随机变量，离散类似.

$$\mathbb{E}X = \int_0^{+\infty} x f_X(x)\mathrm{d}x \geqslant \int_a^{+\infty} x f_X(x)\mathrm{d}x \geqslant a\int_a^{+\infty} f_X(x)\mathrm{d}x = a \cdot \mathbb{P}(X \geqslant a)$$

□

例 4.1. 设 $X \sim U(0,4)$，则 $\mathbb{E}X = 2$，由马尔可夫不等式可得：

$$\mathrm{P}(X \geqslant 2) \leqslant \frac{2}{2} = 1, \quad \mathrm{P}(X \geqslant 3) \leqslant \frac{2}{3}, \quad \mathrm{P}(X \geqslant 4) \leqslant \frac{2}{4} = \frac{1}{2}$$

而真实概率是：

$$\mathrm{P}(X \geqslant 2) = \frac{1}{2}, \quad \mathrm{P}(X \geqslant 3) = \frac{1}{4}, \quad \mathrm{P}(X \geqslant 4) = 0$$

可见由马尔可夫不等式给出的上界非常的粗糙.

定理 4.2 (切比雪夫不等式)**.** 设随机变量 X 的均值为 μ，方差为 σ^2，则对任意 $c > 0$，有：

$$\mathbb{P}(|X - \mu| \geqslant c) \leqslant \frac{\sigma^2}{c^2}$$

注释. 粗略来讲，切比雪夫不等式指出，如果一个随机变量的方差非常小，那么该随机变量取远离均值的概率也非常小.

证明. 利用马尔可夫不等式，

$$\mathbb{P}(|X - \mu| \geqslant c) = \mathbb{P}((X-\mu)^2 \geqslant c^2) \leqslant \frac{\mathbb{E}[(X-\mu)^2]}{c^2} = \frac{\sigma^2}{c^2}$$

□

相比马尔可夫不等式，切比雪夫不等式利用了随机变量的方差的信息，因此会更准确. 不过均值和方差也仅仅是粗略描述了随机变量的性质，因此它给出的上界可能依旧距离精确概率较远.

例 4.2. 仍然考虑例 4.1，即 $X \sim U(0,4)$, $\mathbb{E}X = 2$, $\mathrm{var}X = \frac{4}{3}$，则根据切比雪夫不等式有：

$$\mathrm{P}(|X-2| \geqslant 1) \leqslant \frac{4}{3}$$

由于概率一定小于等于 1，所以这个不等式并没有带来任何信息.

例 4.3. 设 $X \sim E(1)$，则 $\mathbb{E}X = \mathrm{var}X = 1$，对任意 $c > 2$，应用切比雪夫不等式可得：

$$\mathrm{P}(X \geqslant c) = \mathrm{P}(X - 1 \geqslant c - 1) = \mathrm{P}(|X-1| \geqslant c-1) \leqslant \frac{1}{(c-1)^2}$$

而真实概率是 $\mathrm{P}(X \geqslant c) = e^{-c}$，可见切比雪夫不等式给出的上界比较保守.

4.2 弱大数定律

定义 4.1 (依概率收敛)**.** 设 $X_1, X_2, \ldots$ 是随机变量序列，$a \in \mathbb{R}$，若对任意 $\epsilon > 0$，都有：

$$\lim_{n\to\infty} \mathrm{P}(|X_n - a| \geqslant \epsilon) = 0$$

则称 $\{X_n\}$ 依概率收敛到 a.

注释. 用 ϵ-N 语言描述上述定义中的数列极限，则依概率收敛可以等价叙述为：对任意 $\epsilon > 0$ 和 $\delta > 0$，存在 $N \in \mathbb{N}$，当 $n \geqslant N$ 时，都有：

$$\mathrm{P}(|X_n - a| \geqslant \epsilon) \leqslant \delta$$

其中 ϵ 称作精度，δ 称作置信水平.

定理 4.3 (弱大数定律)**.** 设 $X_1, X_2, \ldots$ 是独立同分布的随机变量序列，其公共分布均值为 μ，则样本均值 $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ 依概率收敛到 μ，即对任意 $\epsilon > 0$，有：

$$\lim_{n\to\infty} \mathbb{P}(|M_n - \mu| \geqslant \epsilon) = 0$$

证明. 这里仅对方差有界的情形进行证明，方差无界时弱大数定律依然成立，但是证明较为精巧. 考虑 M_n 的均值和方差：

$$\mathbb{E}M_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}X_i = \mu$$
$$\mathrm{var}(M_n) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(X_i) = \frac{\sigma^2}{n}$$

于是根据切比雪夫不等式，对任意 $\epsilon > 0$，有：

$$\mathbb{P}(|M_n - \mu| \geqslant \epsilon) \leqslant \frac{\sigma^2}{n\epsilon^2} \to 0 \quad (n \to \infty)$$

□

注释. 一般情形的弱大数定理称为辛钦大数定律，而方差有界的情形称之为切比雪夫大数定律. 更特殊的，对于 $X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} B(1,p)$ 的情形而言，我们称之为伯努利大数定律：

$$\lim_{n\to\infty} \mathbb{P}(|M_n - p| \geqslant \epsilon) = 0$$

4.3 中心极限定理

定理 4.4 (中心极限定理)**.** 设 $X_1, X_2, \ldots$ 是独立同分布的随机变量序列，序列的每一项的均值为 μ，方差为 σ^2，则对 $\forall x \in \mathbb{R}$，有：

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \leqslant x \right) = \Phi(x)$$

其中 $\Phi(x)$ 表示标准正态分布的 CDF.

注释. 粗略来讲，中心极限定理表明，在大样本的情况下，独立同分布的随机变量序列的样本均值的标准化结果服从标准正态分布.

注释. 这个一般情形的定理被称作林德伯格-莱维中心极限定理.

注释. 对于 $X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} B(1,p)$ 的特殊情形而言，我们称之为棣莫弗-拉普拉斯中心极限定理：

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} \leqslant x \right) = \Phi(x)$$

4.4 强大数定律

定义 4.2 (以概率 1 收敛/几乎必然收敛)**.** 设 $X_1, X_2, \ldots$ 是随机变量序列，$a \in \mathbb{R}$，若：

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

则称 $\{X_n\}$ 以概率 1 收敛到 a 或几乎必然收敛到 a.

定理 4.5 (强大数定律)**.** 设 $X_1, X_2, \ldots$ 是均值为 μ 的独立同分布的随机变量序列，则样本均值 $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ 以概率 1 收敛到 μ，即：

$$\mathrm{P}\left(\lim_{n\to\infty} M_n = \mu \right) = 1$$

注释. 弱大数定律是指 M_n 显著性偏离 μ 的事件的概率在 $n \to \infty$ 时趋于 0，但是对任意有限的 n，这个概率可以是正的. 所以可以想象的是，在 M_n 这个无穷的序列中，常常有 M_n 显著偏离 μ. 而强大数定律则进一步告诉我们，这样的显著偏离事件只能发生有限次.

5 贝叶斯统计推断

5.1 贝叶斯推断与后验分布

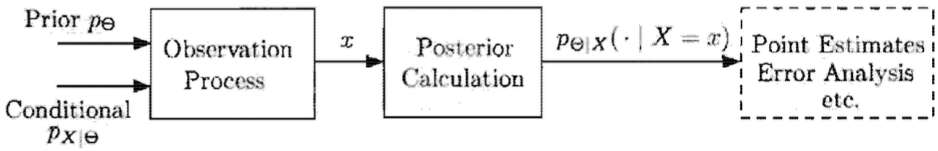

图 5: 贝叶斯推断模型

记感兴趣的未知量为 Θ，并视其为一个随机变量或随机变量的有限集合. 我们的目标是基于观测到相关随机变量的值 $X = (X_1, \ldots, X_n)$ 来提取 Θ 的信息. 假定我们已知先验分布 p_Θ 或 f_Θ，以及条件分布 $p_{X|\Theta}$ 或 $f_{X|\Theta}$，则贝叶斯推断问题由 Θ 的后验分布 $p_{\Theta|X}$ 或 $f_{\Theta|X}$ 完全决定. 后验分布可根据贝叶斯公式计算.

例 5.1 (正态随机变量公共均值的推断)**.** 设随机变量观测值 $X = (X_1, \ldots, X_n)$ 具有相同的未知均值. 假设在给定均值的条件下，X_i 是正态的，且相互独立，方差分别为 $\sigma_1^2, \ldots, \sigma_n^2$. 又设 X_i 的公共均值为随机变量 Θ，且先验分布为正态分布 $N(x_0, \sigma_0^2)$. 那么：

$$f_\Theta(\theta) = c_1 \cdot \exp\left(-\frac{(\theta - x_0)^2}{2\sigma_0^2}\right)$$
$$f_{X|\Theta}(x|\theta) = c_2 \cdot \exp\left(-\frac{(x_1 - \theta)^2}{2\sigma_1^2}\right) \cdots \exp\left(-\frac{(x_n - \theta)^2}{2\sigma_n^2}\right)$$

其中 c_1, c_2 为常数. 根据贝叶斯公式，有：

$$f_{\Theta|X}(\theta|x) \propto f_\Theta(\theta) f_{X|\Theta}(x|\theta) \propto \exp\left(-\sum_{i=0}^{n} \frac{(\theta - x_i)^2}{2\sigma_i^2}\right) \propto \exp\left(-\frac{(\theta - m)^2}{2v}\right)$$

其中根据配方可得：

$$m = \frac{\sum_{i=0}^n x_i/\sigma_i^2}{\sum_{i=0}^n 1/\sigma_i^2}, \quad v = \frac{1}{\sum_{i=0}^n 1/\sigma_i^2}$$

于是后验概率就是以 m 为均值、v 为方差的正态分布.

定义 5.1 (共轭分布)**.** 在贝叶斯推断中，若先验分布与后验分布是同一个分布族，则先验分布与后验分布被称为共轭分布.

注释. 例 5.1 显示正态分布与其自身是共轭分布. 这并不是一个普遍的情形，除了正态分布以外，其他常见的与自身共轭的分布有伯努利分布和二项分布.

5.2 点估计，假设检验，最大后验概率准则

点估计 给定 X 的观测值 x，贝叶斯推断给出了后验分布 $p_{\Theta|X}(\theta|x)$ 或 $f_{\Theta|X}(\theta|x)$. 后验分布包含了 x 提供的所有信息，但有时我们希望得到一个数而不是一个概率分布，这就是点估计. 具体而言，点估计指从 X 的观测值中提取一个数 $\hat{\theta} = g(x)$ 作为对 Θ 的估计的方法.

定义 5.2 (估计量，估计值)**.** 设 g 是一个关于 X 的函数，则称随机变量 $\hat{\Theta} = g(X)$ 为估计量. 当观测到 X 的值 x 后，则称 $\hat{\Theta}$ 的取值 $\hat{\theta} = g(x)$ 为估计值. 不同的函数 g 引出不同的估计量.

定义 5.3 (最大后验概率估计量)**.** 给定 X 的观测值 x，选择使得后验概率最大的 θ 作为估计值，即：

$$\hat{\theta} = \max_{\theta} p_{\Theta|X}(\theta|x) \text{ 或 } \max_{\theta} f_{\Theta|X}(\theta|x)$$

定义 5.4 (条件期望估计量)**.** 给定 X 的观测值 x，选择条件期望作为估计值，即：

$$\hat{\theta} = \mathbb{E}[\Theta|X = x]$$

注释. 在下一节中将看到，条件期望估计量其实就是最小均方估计量.

例 5.2 (正态随机变量公共均值的估计量)**.** 在例 5.1 中，我们得到了正态随机变量公共均值 Θ 的后验分布是以 m 为均值、v 为方差的正态分布，其中：

$$m = \frac{\sum_{i=0}^{n} x_i/\sigma_i^2}{\sum_{i=0}^{n} 1/\sigma_i^2}, \quad v = \frac{1}{\sum_{i=0}^{n} 1/\sigma_i^2}$$

由于正态分布的概率密度函数在均值处取最大值，所以最大后验概率估计为：

$$\hat{\theta} = m$$

而正态分布的均值参数正是其期望，因此条件期望估计也为：

$$\hat{\theta} = \mathbb{E}[\Theta|X = x] = m$$

因此在该模型中，最大后验概率估计和条件期望估计恰好相同.

假设检验 在一个假设检验问题中，Θ 取 $\theta_1, \ldots, \theta_m$ 中的一个值，其中 m 是一个较小的整数（常常是 $m = 2$）. 称事件 $H_i = \{\Theta = \theta_i\}$ 为第 i 个假设. 观测到 X 的取值 x 后，我们希望依据某种准则选出一个最“合理”的假设.

定义 5.5 (假设检验的最大后验概率准则)**.** 给定观测值 x，最大后验概率准则选择使得后验概率 $\mathrm{P}(\Theta = \theta_i|X = x)$ 最大的假设 H_i：

$$\arg\max_i \ \mathrm{P}(\Theta = \theta_i|X = x)$$

根据贝叶斯公式，这等价于：

$$\arg\max_i \ p_\Theta(\theta_i)p_{X|\Theta}(x|\theta_i), \quad X \text{ 离散}$$
$$\arg\max_i \ p_\Theta(\theta_i)f_{X|\Theta}(x|\theta_i), \quad X \text{ 连续}$$

注释. 对任意观测值 x，最大后验概率准则是错误率最小的决策准则.

5.3 贝叶斯最小均方估计

定理 5.1 (最小均方估计——无观测值情形)**.** 考虑在没有观测值的情况下，用常数 $\hat{\theta}$ 去估计 Θ，则使得均方误差 $\mathbb{E}[(\Theta-\hat{\theta})^2]$ 最小的估计为：

$$\hat{\theta}=\mathbb{E}\Theta$$

换句话说，有：

$$\mathbb{E}[(\Theta-\mathbb{E}\Theta)^2]\leqslant\mathbb{E}[(\Theta-\hat{\theta})^2],\quad\forall\hat{\theta}$$

证明. 对任何估计 $\hat{\theta}$，有均方误差：

$$\mathbb{E}[(\Theta-\hat{\theta})^2]=\mathrm{var}(\Theta-\hat{\theta})+(\mathbb{E}[\Theta-\hat{\theta}])^2=\mathrm{var}(\Theta)+(\mathbb{E}\Theta-\hat{\theta})^2$$

由于 $\mathrm{var}(\Theta)$ 与 $\hat{\theta}$ 无关，故当 $(\mathbb{E}\Theta-\hat{\theta})^2$ 最小时均方误差最小，也即 $\hat{\theta}=\mathbb{E}\Theta$. □

定理 5.2 (最小均方估计——给定观测值情形)**.** 设给定观测值 x, 则使得均方误差 $\mathbb{E}[(\Theta-\hat{\theta})^2|X=x]$ 最小的估计为条件期望估计：

$$\hat{\theta}=\mathbb{E}[\Theta|X=x]$$

换句话说，有：

$$\mathbb{E}[(\Theta-\mathbb{E}[\Theta|X=x])^2|X=x]\leqslant\mathbb{E}[(\Theta-\hat{\theta})^2|X=x],\quad\forall\hat{\theta}$$

证明. 在定理 5.1 的证明中加入 $X=x$ 的条件即可. □

定理 5.3 (最小均方估计——总体情形)**.** 总体上，设估计量为 $\hat{\Theta}$，则使得均方误差 $\mathbb{E}[(\Theta-\hat{\Theta})^2]$ 最小的估计量为：

$$\hat{\Theta}=\mathbb{E}[\Theta|X]$$

换句话说，有：

$$\mathbb{E}[(\Theta-\mathbb{E}[\Theta|X])^2]\leqslant\mathbb{E}[(\Theta-\hat{\Theta})^2],\quad\forall\hat{\Theta}=g(X)$$

证明. 对于任意给定 X 的取值 x，$\hat{\theta}=g(x)$ 是一个数，因此：

$$\mathbb{E}[(\Theta-\mathbb{E}[\Theta|X=x])^2|X=x]\leqslant\mathbb{E}[(\Theta-g(x))^2|X=x]$$

根据 x 的任意性，有：

$$\mathbb{E}[(\Theta-\mathbb{E}[\Theta|X])^2|X]\leqslant\mathbb{E}[(\Theta-g(X))^2|X]$$

根据全期望公式，两边取期望得：

$$\mathbb{E}[(\Theta-\mathbb{E}[\Theta|X])^2]\leqslant\mathbb{E}[(\Theta-g(X))^2]$$

□

性质. 将最小均方估计和估计误差分别记为：

$$\hat{\Theta} = \mathbb{E}[\Theta|X], \quad \tilde{\Theta} = \hat{\Theta} - \Theta$$

在 3.3 节中已经推导了一些性质，这里列举如下：

- $\tilde{\Theta}$ 是无偏的，它的条件期望和非条件期望都是 0: $\mathbb{E}[\tilde{\Theta}] = 0$, $\mathbb{E}[\tilde{\Theta}|X = x] = 0, \forall x$
- 估计误差 $\tilde{\Theta}$ 与估计量 $\hat{\Theta}$ 不相关：$\mathrm{var}(\hat{\Theta}, \tilde{\Theta}) = 0$
- Θ 的方差可以分解为：$\mathrm{var}(\Theta) = \mathrm{var}(\hat{\Theta}) + \mathrm{var}(\tilde{\Theta})$

定理 5.4 (推广到多观测情形)**.** 设有多个观测量 $X_1, \ldots, X_n$，则最小均方估计为：

$$\hat{\Theta} = \mathbb{E}[\Theta|X_1, \ldots, X_n]$$

换句话说，有：

$$\mathbb{E}[(\Theta - \mathbb{E}[\Theta|X_1, \ldots, X_n])^2] \leqslant \mathbb{E}[(\Theta - \hat{\Theta})^2], \quad \forall \hat{\Theta} = g(X_1, \ldots, X_n)$$

5.4 贝叶斯线性最小均方估计

定义 5.6 (线性估计量)**.** 限定 g 是关于 X 的线性函数 $g(X) = aX + b$，称随机变量：

$$\hat{\Theta} = g(X) = aX + b$$

为线性估计量. 进一步地，若有多个观测量 $X_1, \ldots, X_n$，则线性估计量的形式为：

$$\hat{\Theta} = a_1 X_1 + \cdots + a_n X_n + b$$

定理 5.5 (线性最小均方估计)**.** 基于 X 的 Θ 的线性最小均方估计是：

$$\hat{\Theta} = \mathbb{E}\Theta + \frac{\mathrm{cov}(\Theta, X)}{\mathrm{var}(X)}(X - \mathbb{E}X) = \mathbb{E}\Theta + \rho\frac{\sigma_\Theta}{\sigma_X}(X - \mathbb{E}X)$$

其中 $\rho = \dfrac{\mathrm{cov}(\Theta, X)}{\sigma_\Theta \sigma_X}$ 是相关系数，此时所得均方误差为：

$$\mathbb{E}[(\Theta - aX - b)^2] = (1 - \rho^2)\sigma_\Theta^2$$

证明. 假设固定 a，则问题等价于选择常数 b 来估计随机变量 $\Theta - aX$. 根据之前的讨论，最优解为：

$$b = \mathbb{E}[\Theta - aX] = \mathbb{E}\Theta - a\mathbb{E}X$$

因此问题转化为：

$$\min_a \ \mathbb{E}[(\Theta - aX - \mathbb{E}[\Theta - aX])^2] = \mathrm{var}(\Theta - aX)$$

打开方差：

$$\mathrm{var}(\Theta - aX) = \mathrm{var}(\Theta) + \mathrm{var}(aX) - 2\mathrm{cov}(\Theta, aX) = \sigma_\Theta^2 + a^2\sigma_X^2 - 2a\rho\sigma_\Theta\sigma_X$$

这是关于 a 的二次函数，最小值在顶点处取得：

$$a = \frac{\rho\sigma_\Theta\sigma_X}{\sigma_X^2} = \rho\frac{\sigma_\Theta}{\sigma_X}$$

因此线性最小均方估计为：

$$\hat{\Theta} = aX + b = aX + \mathbb{E}\Theta - a\mathbb{E}X = \mathbb{E}\Theta + \rho\frac{\sigma_\Theta}{\sigma_X}(X - \mathbb{E}X)$$

且估计误差为：

$$\mathrm{var}(\Theta - aX) = \sigma_\Theta^2 + a^2\sigma_X^2 - 2a\rho\sigma_\Theta\sigma_X = \sigma_\Theta^2 + \rho^2\sigma_\Theta^2 - 2\rho^2\sigma_\Theta^2 = (1-\rho^2)\sigma_\Theta^2$$

□

注解. 直观上，估计量以 $\mathbb{E}\Theta$ 为基础，通过 $X - \mathbb{E}X$ 的取值来调整. 例如，不妨假设 $\rho > 0$，则当观测到比 $\mathbb{E}X$ 更大的 X 取值后，我们对 Θ 的估计也就相应地提高. 另外，当 $|\rho|$ 接近 1 时，X 和 Θ 高度相关，了解 X 将帮助我们准确地估计 Θ，因此均方误差较小.

例 5.3 (正态随机变量公共均值的估计量-续)**.** 在例 5.2 中，我们得到正态随机变量公共均值的最小均方估计（条件期望估计）为：

$$\hat{\theta} = \mathbb{E}[\Theta|X = x] = m = \frac{\sum_{i=0}^n x_i/\sigma_i^2}{\sum_{i=0}^n 1/\sigma_i^2}$$

进一步，注意到上式是观测值 $x_1, \ldots, x_n$ 的线性组合，因此它其实也是线性最小均方估计. 也即，在该模型中，最大后验概率估计、最小均方估计和线性最小均方估计恰好都是相同的.

6 经典统计推断

经典统计推断认为未知参数 θ 是确定的，观测 X 是随机的，根据 θ 取值的不同服从 $p_X(x;\theta)$ 或 $f_X(x;\theta)$. 因此，我们将同时处理多个候选模型，每个模型对应 θ 的一个可能的取值.

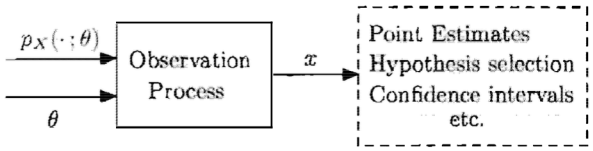

图 6: 经典推断模型

6.1 经典参数估计

定义 6.1 (估计量，估计值)**.** 给定观测 $X=(X_1,\dots,X_n)$，估计量指形如 $\hat{\Theta}_n=g(X)$ 的随机变量，估计量的取值称为估计值. 注意，由于 X 的分布依赖于参数 θ，因而 $\hat{\Theta}_n$ 的分布也依赖于 θ. 记 $\hat{\Theta}_n$ 的期望为 $\mathbb{E}_\theta[\hat{\Theta}_n]$，方差为 $\mathrm{var}_\theta(\hat{\Theta}_n)$.

定义 6.2 (估计误差，偏差)**.** 设 $\hat{\Theta}_n$ 是未知参数 θ 的一个估计量，定义估计误差为 $\tilde{\Theta}_n=\hat{\Theta}_n-\theta$，并定义偏差为估计误差的期望 $b_\theta(\tilde{\Theta}_n)=\mathbb{E}_\theta[\hat{\Theta}_n]-\theta$.

定义 6.3 (无偏，渐近无偏)**.** 若 $\mathbb{E}_\theta[\hat{\Theta}_n]=\theta$ 对 θ 所有可能的取值都成立，则称 $\hat{\Theta}_n$ 是无偏的；若 $\lim_{n\to\infty}\mathbb{E}_\theta[\hat{\Theta}_n]=\theta$ 对 θ 所有可能的取值都成立，则称 $\hat{\Theta}_n$ 是渐近无偏的.

定义 6.4 (相合)**.** 若对 θ 所有可能的取值，序列 $\hat{\Theta}_n$ 依概率收敛到参数 θ 的真值，则称 $\hat{\Theta}_n$ 是 θ 的相合估计序列.

性质. 均方误差、偏差和 $\hat{\Theta}_n$ 的方差有关系：

$$\mathbb{E}_\theta[\tilde{\Theta}_n^2]=b_\theta^2(\hat{\Theta}_n)+\mathrm{var}_\theta(\hat{\Theta}_n)$$

> 注解. 这个公式很重要，等式右侧的两项体现了偏差与方差之间的权衡.

定义 6.5 (极大似然估计，似然函数)**.** 设观测向量 $X=(X_1,\dots,X_n)$ 的联合 PMF/PDF 为 $p_X(x;\theta)$ 或 $f_X(x;\theta)$，其中 $x=(x_1,\dots,x_n)$ 为 X 的观测值. 极大似然估计是使得 $p_X(x;\theta)$ 或 $f_X(x;\theta)$ 达到最大的参数值，即：

$$\hat{\theta}_n=\arg\max_\theta\ p_X(x;\theta)\ \text{或}\ \arg\max_\theta\ f_X(x;\theta)$$

称 $p_X(x;\theta)$ 或 $f_X(x;\theta)$ 为似然函数，称 $\ln p_X(x;\theta)$ 或 $\ln f_X(x;\theta)$ 为对数似然函数.

> 注解 (极大似然估计与最大后验概率估计). 在贝叶斯最大后验概率估计中，估计的选择是使得 $p_\Theta(\theta)p_{X|\Theta}(x|\theta)$ 达到最大的 θ，其中 $p_\Theta(\theta)$ 是未知参数 θ 的先验分布. 因此，最大似然估计可以解释为具有均匀先验分布的最大后验概率估计.

定理 6.1 (极大似然估计的不变原理)**.** 设 $\hat{\Theta}_n$ 是 θ 的极大似然估计，那么对于任意关于 θ 的一一映射函数 h，$h(\hat{\Theta})$ 是 $\zeta=h(\theta)$ 的极大似然估计.

例 6.1 (随机变量均值的估计). 设观测值 $X_1, \ldots, X_n$ 独立同分布，均值为未知参数 θ，定义样本均值为:

$$M_n = \frac{X_1 + \cdots + X_n}{n}$$

由于 $\mathbb{E}_\theta[M_n] = \mathbb{E}_\theta[X] = \theta$，故样本均值是 θ 的无偏估计. 进一步地，根据弱大数定律，M_n 依概率收敛到 θ，因此具有相合性.

例 6.2 (随机变量方差的估计). 设观测值 $X_1, \ldots, X_n$ 独立同分布，方差为未知参数 v，则有两个常用估计量:

$$\bar{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - M_n)^2, \qquad \hat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2$$

可以证明 $\bar{S}^2$ 是有偏但渐近无偏的，$\hat{S}^2$ 则是无偏的.

证明. 注意到 $\mathbb{E}_{(\theta,v)}[M_n] = \theta$, $\mathbb{E}_{(\theta,v)}[X_i^2] = v + \theta^2$, $\mathbb{E}_{(\theta,v)}[M_n^2] = \frac{v}{n} + \theta^2$，于是:

$$\begin{aligned}
\mathbb{E}_{(\theta,v)}[\bar{S}_n^2] &= \mathbb{E}_{(\theta,v)}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - M_n)^2\right] \\
&= \mathbb{E}_{(\theta,v)}\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \frac{2M_n}{n}\sum_{i=1}^{n}X_i + M_n^2\right] \\
&= \mathbb{E}_{(\theta,v)}\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2M_n^2 + M_n^2\right] \\
&= v + \theta^2 - \left(\frac{v}{n} + \theta^2\right) \\
&= \frac{n-1}{n}v
\end{aligned}$$

由于 $n \to \infty$ 时 $\mathbb{E}_{(\theta,v)}[\bar{S}_n^2] \to v$，故 $\bar{S}_n^2$ 是渐近无偏的. 又:

$$\mathbb{E}_{(\theta,v)}[\hat{S}_n^2] = \mathbb{E}_{(\theta,v)}[\frac{n}{n-1}\bar{S}_n^2] = v$$

故 $\hat{S}_n^2$ 是无偏的. □

区间估计 在前文中，无论是贝叶斯推断中的最大后验概率估计、最小均方估计，还是经典推断中的极大似然估计，这些方法都是点估计，即给出一个数作为估计量. 但有时我们还想建立一个置信区间，这就是区间估计. 具体而言，我们首先固定一个**置信度** $1-\alpha$，其中 α 是一个很小的数. 然后用一个略小的估计量 $\hat{\Theta}_n^-$ 和一个略大的估计量 $\hat{\Theta}_n^+$ 代替点估计量 $\hat{\Theta}_n$，使得:

$$\mathrm{P}_\theta(\hat{\Theta}_n^- \leqslant \theta \leqslant \hat{\Theta}_n^+) \geqslant 1 - \alpha$$

对 θ 所有可能的取值都成立. 称 $[\hat{\Theta}_n^-, \hat{\Theta}_n^+]$ 为**置信区间**.

例 6.3 (正态随机变量公共均值的区间估计)**.** 设观测量 $X_1, \ldots, X_n$ 独立同分布于 $N(\theta, v)$，其中均值 θ 未知，方差 v 已知. 根据例 3.4 的结论（独立正态随机变量之和仍是正态的），可知样本均值估计量:

$$\hat{\Theta}_n = \frac{X_1 + \cdots + X_n}{n} \sim N\left(\theta, \frac{v}{n}\right)$$

取 $\alpha = 0.05$，即置信度 0.95，查正态分布表可得 $\Phi(1.96) = 0.975 = 1 - \alpha/2$，于是:

$$\mathrm{P}_\theta\left(\left|\frac{\hat{\Theta}_n - \theta}{\sqrt{v/n}}\right| \leqslant 1.96\right) = \mathrm{P}_\theta\left(\hat{\Theta}_n - 1.96\sqrt{\frac{v}{n}} \leqslant \theta \leqslant \hat{\Theta}_n + 1.96\sqrt{\frac{v}{n}}\right) = 0.95$$

这说明:

$$\left[\hat{\Theta}_n - 1.96\sqrt{\frac{v}{n}},\ \hat{\Theta}_n + 1.96\sqrt{\frac{v}{n}}\right]$$

是 95% 置信区间.

例 6.4 (基于方差近似估计量的区间估计)**.** 设观测量 $X_1, \ldots, X_n$ 独立同分布于 $N(\theta, v)$，其中均值 θ 和方差 v 都是未知的. 考虑用方差的无偏估计量:

$$\hat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\Theta}_n)^2$$

代替方差，则根据例 6.3 和中心极限定理，可构造一个近似的 $1 - \alpha$ 置信区间:

$$\left[\hat{\Theta}_n - z\frac{\hat{S}_n}{\sqrt{n}},\ \hat{\Theta}_n + z\frac{\hat{S}_n}{\sqrt{n}}\right]$$

其中 z 满足 $\Phi(z) = 1 - \alpha/2$. 例如，近似 95% 置信区间为:

$$\left[\hat{\Theta}_n - 1.96\frac{\hat{S}_n}{\sqrt{n}},\ \hat{\Theta}_n + 1.96\frac{\hat{S}_n}{\sqrt{n}}\right]$$

事实上，当用 $\hat{S}_n^2$ 代替方差后，随机变量:

$$T_n = \frac{\sqrt{n}(\hat{\Theta}_n - \theta)}{\hat{S}_n}$$

服从自由度为 $n-1$ 的 t 分布（详见附录 C）. 因此，设 t 分布的分布函数为 $\Psi_{n-1}(z)$，则更精确的置信区间中，z 应满足 $\Psi_{n-1}(z) = 1 - \alpha/2$.

定义 6.6 (充分统计量)**.** 给定观测 $X = (X_1, \ldots, X_n)$，设随机变量 $T = q(X)$ 为关于 X 的标量或向量函数，若 X 在给定 T 下的条件分布不依赖于 θ，即:

$$\mathrm{P}_\theta(X \in D | T = t), \quad \forall t$$

对所有 θ 都是一样的，则称 T 为 θ 的充分统计量.

定理 6.2 (因子分解定理)**.** $T = q(X)$ 是 θ 的充分统计量当且仅当似然函数 $p_X(x;\theta)$ 或 $f_X(x;\theta)$ 可以写成 $r(q(x), \theta)s(x)$ 的形式，其中 r, s 是两个函数.

证明. 假设 X 是离散随机变量，连续情形类似. 首先证明充分性. 假设 $p_X(x;\theta) = r(q(x),\theta)s(x)$，取定 $T=t$，则对所有 $q(x)\neq t$ 的 x，容易知道：

$$\mathrm{P}_\theta(X=x|T=t)=\frac{\mathrm{P}_\Theta(X=x)\mathrm{P}_\theta(T=t|X=x)}{\mathrm{P}_\theta(T=t)}=0$$

而对于满足 $q(x)=t$ 的 x，有：

$$\begin{aligned}\mathrm{P}_\theta(X=x|T=t)&=\frac{\mathrm{P}_\theta(X=x,T=t)}{\mathrm{P}_\theta(T=t)}=\frac{\mathrm{P}_\theta(X=x)}{\mathrm{P}_\theta(T=t)}=\frac{\mathrm{P}_\theta(X=x)}{\sum_z \mathrm{P}_\theta(X=z,T=t)}\\&=\frac{\mathrm{P}_\theta(X=x)}{\sum_{\{z|q(z)=t\}}\mathrm{P}_\theta(X=z)}=\frac{r(q(x),\theta)s(x)}{\sum_{\{z|q(z)=t\}}r(q(z),\theta)s(z)}\\&=\frac{r(t,\theta)s(x)}{r(t,\theta)\sum_{\{z|q(z)=t\}}s(z)}=\frac{s(x)}{\sum_{\{z|q(z)=t\}}s(z)}\end{aligned}$$

可见对所有 x，$\mathrm{P}_\theta(X=x|T=t)$ 都不依赖于 θ，故 T 是充分统计量.
然后证明必要性. 设 $T=q(X)$ 是 θ 的充分统计量，则对任意 x，有：

$$p_X(x;\theta)=\sum_t \mathrm{P}_\theta(X=x|T=t)\mathrm{P}_\theta(T=t)=\mathrm{P}_\theta(X=x|T=q(x))\mathrm{P}_\theta(T=q(x))$$

右式两部分中，前者与 θ 无关，故可写作 $s(x)$ 的形式，而后者为 $r(q(x),\theta)$ 的形式. □

定理 6.3. 若 $q(X)$ 是 θ 的充分统计量，则对 θ 的任何函数 h，$q(X)$ 都是 $\zeta=h(\theta)$ 的充分统计量.

证明. 对 $\zeta=h(\theta)$ 有：

$$\mathrm{P}_\zeta(X\in D|T=t)=\mathrm{P}_\theta(X\in D|T=t)$$

所以 $\mathrm{P}_\zeta(X\in D|T=t)$ 对所有 ζ 都是一样的. □

定理 6.4. 若 $q(X)$ 是 θ 的充分统计量，则 θ 的极大似然估计可以写作 $\hat{\Theta}_n=\phi(q(X))$.

证明. 根据因子分解定理 6.2，$p_X(x;\theta)=r(q(x),\theta)s(x)$，故极大似然估计为：

$$\hat{\theta}_n=\arg\max_\theta\ p_X(x;\theta)=\begin{cases}\arg\max\limits_\theta\ r(q(x),\theta), & s(x)>0\\ \arg\min\limits_\theta\ r(q(x),\theta), & s(x)<0\end{cases}$$

即 $\hat{\theta}_n$ 只通过 $\phi(q(x))$ 依赖于 x，故 $\hat{\Theta}_n=\phi(q(X))$. □

注解. 该定理说明充分统计量抓住了由 X 提供的关于 θ 的所有核心信息.

例 6.5 (伯努利分布). 设 $X_1,\ldots,X_n$ 独立同分布于参数为 θ 的伯努利分布，则：

$$q(X)=\sum_{i=1}^n X_i$$

是 θ 的充分统计量. 这是因为似然函数为:

$$p_X(x;\theta)=\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}=\theta^{q(x)}(1-\theta)^{n-q(x)}$$

可以分解为 $r(q(x),\theta)=\theta^{q(x)}(1-\theta)^{n-q(x)}$ 与常函数 $s(x)=1$ 的乘积.

例 6.6 (泊松分布). 设 $X_1,\dots,X_n$ 独立同分布于参数为 θ 的泊松分布, 则:

$$q(X)=\sum_{i=1}^n X_i$$

是 θ 的充分统计量. 这是因为似然函数为:

$$p_X(x;\theta)=\prod_{i=1}^n p_{X_i}(x_i;\theta)=e^{-\theta}\prod_{i=1}^n\frac{\theta^{x_i}}{x_i!}=e^{-\theta}\frac{\theta^{q(x)}}{\prod_{i=1}^n x_i!}$$

可以分解为 $r(q(x),\theta)=e^{-\theta}\theta^{q(x)}$ 与 $s(x)=1/\prod_{i=1}^n x_i!$ 的乘积.

例 6.7 (正态分布). 设 $X_1,\dots,X_n$ 独立同分布于 $N(\mu,\sigma^2)$, 则:

- 若 σ^2 已知, 则 $q(X)=\sum_{i=1}^n X_i$ 是 μ 的充分统计量.
- 若 μ 已知, 则 $q(X)=\sum_{i=1}^n(X_i-\mu)^2$ 是 σ^2 的充分统计量.
- 若 μ,σ^2 都未知, 则 $q(X)=(\sum_{i=1}^n X_i,\sum_{i=1}^n X_i^2)$ 是 (μ,σ^2) 的充分统计量.

定理 6.5 (Rao-Blackwell). 给定观测 $X=(X_1,\dots,X_n)$, 设 $T=q(X)$ 是参数 θ 的充分估计量, $g(X)$ 是 θ 的一个估计量. 则:

1. $\mathbb{E}_\theta[g(X)|T]$ 对所有 θ 都一样, 因而可以去掉下标 θ, 将:

   $$\hat{g}(X)=\mathbb{E}[g(X)|T]$$

   视作 θ 的一个新估计量, 它只通过 T 依赖于 X.
2. 新估计量 $\hat{g}(X)$ 与原估计量 $g(X)$ 偏差相等.
3. 对满足 $\mathrm{var}_\theta(g(X))<\infty$ 的 θ, 有均方误差:

   $$\mathbb{E}_\theta[(\hat{g}(X)-\theta)^2]\leqslant\mathbb{E}_\theta[(g(X)-\theta)^2]$$

   进一步地, 给定 θ, 此不等式是严格的当且仅当:

   $$\mathbb{E}_\theta[\mathrm{var}(g(X)|T)]>0$$

证明. 我们逐条证明:

1. 由于 T 是充分估计量, 所以条件分布 $\mathrm{P}_\theta(X=x|T=t)$ 不依赖于 θ, 自然 $\mathbb{E}_\theta[g(X)|T]$ 不依赖于 θ.

2. 根据全期望公式，有：

$$\mathbb{E}_\theta[\hat{g}(X)] = \mathbb{E}[\mathbb{E}[g(X)|T]] = \mathbb{E}_\theta[g(X)]$$

因而偏差相等：

$$\mathbb{E}_\theta[\hat{g}(x)] - \theta = \mathbb{E}_\theta[g(x)] - \theta$$

3. 对于固定的 θ，记 $\hat{g}(X)$ 和 $g(X)$ 的偏差为 b_θ. 根据全方差公式，有：

$$\begin{aligned}
\mathbb{E}_\theta[(g(X)-\theta)^2] &= \mathrm{var}_\theta(g(X)) + b_\theta^2 \\
&= \mathbb{E}_\theta[\mathrm{var}(g(X)|T)] + \mathrm{var}_\theta(\mathbb{E}[g(X)|T]) + b_\theta^2 \\
&= \mathbb{E}_\theta[\mathrm{var}(g(X)|T)] + \mathrm{var}_\theta(\hat{g}(X)) + b_\theta^2 \\
&= \mathbb{E}_\theta[\mathrm{var}(g(X)|T)] + \mathbb{E}_\theta[(\hat{g}(X)-\theta)^2] \\
&\geqslant \mathbb{E}_\theta[(\hat{g}(X)-\theta)^2]
\end{aligned}$$

且不等式是严格的当且仅当 $\mathbb{E}_\theta[\mathrm{var}(g(X)|T)] > 0$.

□

注解. Rao-Blackwell 定理说明，一个一般的估计量可以改进为一个只依赖于充分统计量的估计量，新的估计量偏差与原估计量相同，但均方误差更小.

例 6.8. 设 $X_1, \ldots, X_n$ 是 $[0, \theta]$ 上独立同分布的均匀随机变量.

(a) 证明 $T = \max_{i=1,\ldots,n} X_i$ 是充分统计量.
(b) 证明 $g(X) = (2/n)\sum_{i=1}^n X_i$ 是无偏估计.
(c) 找出估计量 $\hat{g}(X) = \mathbb{E}[g(X)|T]$ 的形式，计算并比较 $\mathbb{E}_\theta[(\hat{g}(X)-\theta)^2]$ 和 $\mathbb{E}_\theta[(g(X)-\theta)^2]$.

证明.

(a) 由于似然函数为：

$$f_X(x_1, \ldots, x_n; \theta) = \begin{cases} 1/\theta^n, & 0 \leqslant \max_{i=1,\ldots,n} x_i \leqslant \theta \\ 0, & \text{otherwise} \end{cases}$$

只通过 $\max_{i=1,\ldots,n} x_i$ 依赖于 x，根据因子分解定理可知其为充分统计量.

(b) 由于：

$$\mathbb{E}_\theta[g(X)] = \frac{2}{n}\sum_{i=1}^n \mathbb{E}_\theta[X_i] = \frac{2}{n}\sum_{i=1}^n \frac{\theta}{2} = \theta$$

故 $g(X)$ 是无偏估计.

(c) 固定 $T = t$，则 $X_1, \ldots, X_n$ 中有一个等于 t，其余观测服从 $[0, t]$ 上的均匀分布，因此：

$$\mathbb{E}[g(X)|T=t] = \frac{2}{n}\mathbb{E}_\theta\left[\sum_{i=1}^n X_i | T=t\right] = \frac{2}{n}\left(t + (n-1)\frac{t}{2}\right) = \frac{n+1}{n}t$$

所以：

$$\hat{g}(X) = \mathbb{E}[g(X)|T] = \frac{n+1}{n}T$$

下面计算均方误差. 首先计算一阶矩：

$$\mathbb{E}_\theta[\hat{g}(X)] = \mathbb{E}_\theta[\mathbb{E}[g(X)|T]] = \mathbb{E}_\theta[g(X)] = \theta$$

可以看出 $\hat{g}(X)$ 也是无偏估计，这符合 Rao-Blackwell 定理的第二点. 为了计算二阶矩，需要首先确定 T 的分布：

$$F_T(t;\theta) = \mathrm{P}_\theta(T \leqslant t) = \left(\frac{t}{\theta}\right)^n \implies f_T(t;\theta) = \frac{nt^{n-1}}{\theta^n}, \quad t \in [0,\theta]$$

于是：

$$\mathbb{E}_\theta[(\hat{g}(X))^2] = \left(\frac{n+1}{n}\right)^2 \mathbb{E}_\theta\left[T^2\right] = \left(\frac{n+1}{n}\right)^2 \int_0^\theta t^2 \frac{nt^{n-1}}{\theta^n}\mathrm{d}t = \frac{(n+1)^2}{n(n+2)}\theta^2$$

由于 $\hat{g}(X)$ 是无偏估计，所以均方误差就是方差：

$$\mathbb{E}_\theta[(\hat{g}(X)-\theta)^2] = \mathbb{E}_\theta[(\hat{g}(X))^2] - \theta^2 = \frac{(n+1)^2}{n(n+2)}\theta^2 - \theta^2 = \frac{\theta^2}{n(n+2)}$$

同理，$g(X)$ 均方误差也是其方差，且：

$$\mathbb{E}_\theta[(g(X)-\theta)^2] = \mathrm{var}_\theta(g(X)) = \frac{4}{n^2}\sum_{i=1}^n \mathrm{var}_\theta(X_i) = \frac{4}{n^2}\sum_{i=1}^n \frac{\theta^2}{12} = \frac{\theta^2}{3n}$$

由于 $n(n+2) \geqslant 3n,\ \forall n > 0$，所以：

$$\mathbb{E}_\theta[(\hat{g}(X)-\theta)^2] \leqslant \mathbb{E}_\theta[(g(X)-\theta)^2]$$

即我们得到了一个方差更小的无偏估计. 这符合 Rao-Blackwell 定理的第三点.

□

6.2 线性回归

线性回归为两个或多个变量建立线性模型，可以由**最小二乘法**完成而不需要任何概率上的解释，但也可以在各种概率框架下进行解释.

一元线性回归 设有 n 个数据对 $(x_i, y_i),\ i = 1,\ldots,n$，建立线性模型：

$$y \approx \theta_0 + \theta_1 x$$

其中 θ_0, θ_1 是未知的待估计参数. 特别地，给定参数的估计 $\hat{\theta}_0, \hat{\theta}_1$，则模型对 x_i 的估计为：

$$\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_i$$

其与真实值 y_i 之间的差异称作残差：

$$\tilde{y}_i = y_i - \hat{y}_i$$

线性回归的优化目标是最小化残差的平方和：；

$$\min_{\theta_0,\theta_1} \quad \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)^2$$

这是关于 θ_0, θ_1 的二次函数，容易求得最优估计为：

$$\hat{\theta}_1 = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$

其中，

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

概率框架解释 · 极大似然 对于每个 x_i，将 y_i 视作随机变量 Y_i 的一个取值. 设 Y_i 的模型为：

$$Y_i = \theta_0 + \theta_1 x_i + W_i,\ i = 1, \ldots, n$$

其中 W_i 是均值为零、方差为 σ^2 的正态独立同分布随机变量，因而 Y_i 也是独立的正态随机变量且 $Y_i \sim N(\theta_0 + \theta_1 x_i, \sigma^2)$. 于是似然函数为：

$$f_Y(y;\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^2}\right\}$$

当似然函数最大时，指数部分达到最大，即残差平方和最小. 因此最小二乘法可以视为 Y 的期望具有线性结构的正态模型中参数 θ_0, θ_1 的极大似然估计.

概率框架解释 · 近似贝叶斯线性最小均方估计 将 x_i, y_i 都视作随机变量 X_i, Y_i 的取值，设不同 (X_i, Y_i) 之间是独立同分布的，但是它们的二维联合分布未知. 考虑服从同一分布的另一随机变量对 (X_0, Y_0)，假设观测到 X_0 并希望用线性估计量 $\hat{Y}_0 = \theta_0 + \theta_1 X_0$ 来估计 Y_0. 根据定理 5.5，Y_0 的线性最小均方估计量为：

$$\hat{Y}_0 = \mathbb{E}Y_0 + \frac{\mathrm{cov}(Y_0, X_0)}{\mathrm{var}(X_0)}(X_0 - \mathbb{E}X_0)$$

也即：

$$\theta_1 = \frac{\mathrm{cov}(Y_0, X_0)}{\mathrm{var}(X_0)}, \quad \theta_0 = \mathbb{E}Y_0 - \theta_1 \mathbb{E}X_0$$

由于我们不知道 (X_0, Y_0) 的分布，因此作估计：

$$\mathbb{E}X_0 \approx \bar{x}, \quad \mathbb{E}Y_0 \approx \bar{y}, \quad \mathrm{cov}(Y_0, X_0) \approx \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \quad \mathrm{var}(X_0) \approx \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

代入即可得到最小二乘公式.

概率框架解释 · 近似贝叶斯最小均方估计 仍然假设 (X_i, Y_i) 独立同分布，并附加新的假设——数据满足模型：

$$Y_i = \theta_0 + \theta_1 X_i + W_i$$

其中 W_i 是独立同分布的零均值噪声项且与 X_i 独立. 根据定理 5.3 可知，Y_0 的最小均方估计量为：

$$\mathbb{E}[Y_0|X_0] = \mathbb{E}[\theta_0 + \theta_1 X_0 + W_0|X_0] = \theta_0 + \theta_1 X_0$$

换句话说，$\theta_0 + \theta_1 X_0$ 是所有关于 X_0 的函数 $g(X_0)$ 中使得均方误差最小的，于是：

$$\theta_0, \theta_1 = \arg\min_{\theta_0', \theta_1'} \mathbb{E}[(Y_0 - \theta_0' - \theta_1' X_0)^2]$$

注意到，根据弱大数定律，当 $n \to \infty$ 时，上式是：

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \theta_0' - \theta_1' X_i)^2$$

的极限，而该式正是最小二乘法的优化目标.

贝叶斯线性回归 将 $x_1, \ldots, x_n$ 视为给定的数，将 $(y_1, \ldots, y_n)$ 视为随机向量 $Y = (Y_1, \ldots, Y_n)$ 的观测值，且假设有线性关系：

$$Y_i = \Theta_0 + \Theta_1 x_i + W_i$$

其中 $\Theta = (\Theta_0, \Theta_1)$ 是待估计的参数（视为随机变量）. 假设 $\Theta_0, \Theta_1, W_1, \ldots, W_n$ 是相互独立的正态随机变量. $W_1, \ldots, W_n$ 均值为零，方差已知为 σ^2. Θ_0, Θ_1 的均值为零，方差分别为 σ_0^2, σ_1^2. 则根据贝叶斯公式，后验概率密度函数为：

$$\begin{aligned} f_{\Theta|Y}(\theta_0, \theta_1|y_1, \ldots, y_n) &\propto f_\Theta(\theta_0, \theta_1) f_{Y|\Theta}(y_1, \ldots, y_n|\theta_0, \theta_1) \\ &\propto \exp\left(-\frac{\theta_0^2}{2\sigma_0^2}\right) \cdot \exp\left(-\frac{\theta_1^2}{2\sigma_1^2}\right) \prod_{i=1}^{n} \cdot \exp\left(-\frac{y_i^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\left(\frac{\theta_0^2}{2\sigma_0^2} + \frac{\theta_1^2}{2\sigma_1^2} + \sum_{i=1}^{n}\frac{y_i^2}{2\sigma^2}\right)\right) \end{aligned}$$

因此，最大后验概率估计就是：

$$\hat{\theta}_0, \hat{\theta}_1 = \arg\min_{\theta_0, \theta_1} \left(\frac{\theta_0^2}{2\sigma_0^2} + \frac{\theta_1^2}{2\sigma_1^2} + \sum_{i=1}^{n}\frac{y_i^2}{2\sigma^2}\right)$$

求导并令导数为零，可解得：

$$\hat{\theta}_1 = \frac{\sigma_1^2}{\sigma^2 + \sigma_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \quad \hat{\theta}_0 = \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}(\bar{y} - \hat{\theta}_1 \bar{x})$$

多元线性回归 一元线性回归可以直接扩展到多元情形. 例如，设数据由三元组 (x_i, y_i, z_i) 组成，建立线性模型：

$$y \approx \theta_0 + \theta_1 x + \theta_2 z$$

则多元线性回归最小化残差的平方和：

$$\min_{\theta_0,\theta_1,\theta_2} \quad \sum_{i=1}^{n}(y_i-\theta_0-\theta_1x_i-\theta_2z_i)^2$$

特别地，z 可以是 x 的函数，例如 $z=x^2$. 尽管二次函数不是线性的，但是未知参数 θ_j 与随机变量 Y_i 是线性关系，所以这仍然属于线性回归的范畴. 一般地，考虑如下形式的模型：

$$y \approx \theta_0 + \sum_{j=1}^{m}\theta_j h_j(x)$$

则回归问题的优化目标为：

$$\min_{\theta_0,\ldots,\theta_m} \quad \sum_{i=1}^{n}\left(y_i-\theta_0-\sum_{j=1}^{m}\theta_j h_j(x_i)\right)^2$$

这样的最小化问题都有现存的公式.

6.3 简单假设检验

在经典统计推断中，我们不再有先验概率的假设，因此假设检验问题可以看作是 θ 只有两个取值的推断问题. 我们一般用 H_0 表示**原假设**，H_1 表示**备择假设**.

设观测随机变量 $X=(X_1,\ldots,X_n)$，记 $\mathrm{P}(X\in A;H_j)$ 表示假设 H_j 成立时 X 属于 A 的概率. 类似地，记 $p_X(x;H_j)$ 或 $f_X(x;H_j)$ 表示假设 H_j 成立时的 PMF 或 PDF. 我们希望找到一个决策准则将观测值 x 映射到其中一个假设上去. 任何一个决策准则都将样本空间划分为了两个区域——拒绝域与接受域. 当观测数据落入拒绝域时，称假设 H_0 被拒绝，反之则被接受.

一个决策准则有两种错误. 设该决策准则决定的拒绝域为 R，第一类错误指拒绝了 H_0 但它实际上是正确的，此类错误发生概率为 $\alpha(R)=\mathrm{P}(x\in R;H_0)$；第二类错误指接受了 H_0 但它实际上是错误的，此类错误发生概率为 $\beta(R)=\mathrm{P}(x\notin R;H_1)$.

定义 6.7 (似然比)**.** 定义似然比为：

$$L(x)=\frac{p_X(x;H_1)}{p_X(x;H_0)} \quad 或 \quad L(x)=\frac{f_X(x;H_1)}{f_X(x;H_0)}$$

基于似然比的概念，我们可以得到这样的决策准则：选择一个临界值 $\xi\in(0,\infty)$，当观测值 x 满足 $L(x)>\xi$ 时，拒绝原假设 H_0. 特别地，当 $\xi=1$ 时，我们总是选择似然最大的假设，此时正好对应极大似然准则.

似然比检验　首先确定错误拒绝的概率 α，然后选择一个常数 ξ，使得错误拒绝的概率为 α：

$$\mathrm{P}(L(x)>\xi;H_0)=\alpha$$

那么拒绝域就是 $R=\{x\mid L(x)>\xi\}$. 当观测值落入拒绝域中时，即 $L(x)>\xi$ 时，我们拒绝 H_0.

引理 6.6 (Neyman-Pearson 引理)**.** 考虑似然比检验中一个确定的 ξ，从而有犯错概率：

$$\mathrm{P}(L(X)>\xi;H_0)=\alpha, \quad \mathrm{P}(L(X)\leqslant\xi;H_1)=\beta$$

假设还有其他检验，拒绝域为 R，使得错误拒绝的概率一样或更小：

$$\mathrm{P}(X \in R; H_0) \leqslant \alpha$$

那么这个检验错误接受的概率一样或更大：

$$\mathrm{P}(X \notin R; H_1) \geqslant \beta$$

当 $\mathrm{P}(X \in R; H_0) < \alpha$ 严格成立时，$\mathrm{P}(X \notin R; H_1) > \beta$ 严格成立.

6.4 显著性检验

在实际情况中，假设检验问题并不总是只有两个特定的选择，因而简单假设检验的方法可能不再使用. 本节介绍更一般的问题和解决方法.

假设观测 $X = (X_1, \ldots, X_n)$ 服从由参数 θ 决定的 PMF/PDF，其中 θ 在给定集合 $\mathcal{M}$ 中取值. 我们称“断言 θ 的真值为 $\theta_0 \in \mathcal{M}$”为原假设，记作 H_0，而相应的 $\theta \neq \theta_0$ 为备择假设，记作 H_1.

显著性检验 基于观测 $X_1, \ldots, X_n$，对假设 $H_0 : \theta = \theta_0$ 做统计检验：

1. 选择统计量 $S = h(X_1, \ldots, X_n)$，其中 $h : \mathbb{R}^n \to \mathbb{R}$.
2. 确定拒绝域形状：拒绝域 R 为 S 的取值的子集，当 S 落入 R 时就拒绝 H_0；这个过程通常涉及一个常数 ξ，称作临界值；
3. 选择显著水平 α；
4. 选择临界值 ξ 使得错误拒绝的概率等于或约等于 α；此时拒绝域 R 就完全确定了；
5. 得到 $X_1, \ldots, X_n$ 的观测值 $x_1, \ldots, x_n$，计算统计值 $s = h(x_1, \ldots, x_n)$，若 s 落入 R 则拒绝 H_0.

注释. 很多时候，统计学家并不选择显著水平 α 和临界值 ξ，而是计算统计值 s 后，直接汇报 p-值：

$$p\text{-值} = \min\{\alpha \mid H_0 \text{在显著水平 } \alpha \text{ 下被拒绝}\}$$

A 常见随机变量

本节总结常见的随机变量及其期望、方差、矩母函数和性质，其中离散随机变量包括伯努利、二项、泊松、几何和超几何随机变量，连续随机变量包括均匀、指数和正态随机变量.

伯努利随机变量

- 记号：$X \sim B(1,p)$
- 实例：抛掷一枚硬币，硬币向上概率为 p，X 为是否向上.
- PMF：$p_X(k)=\begin{cases} p, & k=1 \\ 1-p, & k=0 \end{cases}$
- 期望：$\mathbb{E}X=p$
- 方差：$\mathrm{var}X=p(1-p)$
- 矩母函数：$M_X(s)=1-p+pe^s$

矩母函数的推导.

$$M_X(s)=\mathbb{E}[e^{sX}]=(1-p)\cdot e^0+p\cdot e^s=1-p+pe^s$$

□

二项随机变量

- 记号：$X \sim B(n,p)$
- 实例：抛掷 n 枚硬币，每枚硬币向上概率均为 p，X 为向上次数.
- PMF：$p_X(k)=\binom{n}{k}p^k(1-p)^{n-k}, \quad k=0,1,2,\ldots,n$
- 期望：$\mathbb{E}X=np$
- 方差：$\mathrm{var}X=np(1-p)$
- 矩母函数：$M_X(s)=(1-p+pe^s)^n$

期望的推导.

$$\begin{aligned}\mathbb{E}X&=\sum_{k=0}^{n}k\binom{n}{k}p^k(1-p)^{n-k}=\sum_{k=1}^{n}n\binom{n-1}{k-1}p^k(1-p)^{n-k}\\&=np\sum_{k=0}^{n-1}\binom{n-1}{k}p^k(1-p)^{n-1-k}=np(p+1-p)^{n-1}=np\end{aligned}$$

其中用到了恒等式 $\binom{n}{k}=\binom{n-1}{k-1}\frac{n}{k}$ 和二项式定理. □

二阶矩的推导.

$$\begin{aligned}\mathbb{E}X^2 &= \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} nk \binom{n-1}{k-1} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^{n} n \binom{n-1}{k-1} p^k (1-p)^{n-k} + \sum_{k=1}^{n} n(k-1) \binom{n-1}{k-1} p^k (1-p)^{n-k} \\ &= np + np \sum_{k=0}^{n-1} k \binom{n-1}{k} p^k (1-p)^{n-1-k} = np + np(n-1)p\end{aligned}$$

□

方差的推导.

$$\mathrm{var}X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = np + n(n-1)p^2 - n^2p^2 = np - np^2 = np(1-p)$$

□

矩母函数的推导.

$$M_X(s) = \mathbb{E}[e^{sX}] = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} e^{sk} = (1-p+pe^s)^n$$

□

泊松随机变量

- 记号：$X \sim P(\lambda)$
- 实例：一个城市一天中发生车祸次数.
- PMF：$p_X(k) = e^{-\lambda} \dfrac{\lambda^k}{k!}, \quad k = 0, 1, \ldots$
- 期望：$\mathbb{E}X = \lambda$
- 方差：$\mathrm{var}X = \lambda$
- 矩母函数：$M_X(s) = e^{\lambda(e^s - 1)}$
- 性质：取 $\lambda = np$，则当 $n \to \infty$ 时，泊松分布近似二项分布.

期望的推导.

$$\mathbb{E}X = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} = \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda$$

其中用到了 e^x 的泰勒展开. □

二阶矩的推导.

$$\begin{aligned}\mathbb{E}X^2 &= \sum_{k=0}^{\infty} k^2 e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=1}^{\infty} k\frac{\lambda^k}{(k-1)!} \\ &= e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!} + e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^k}{(k-2)!} \\ &= e^{-\lambda}\lambda e^{\lambda} + e^{-\lambda}\lambda^2 e^{\lambda} = \lambda + \lambda^2\end{aligned}$$

□

方差的推导.

$$\mathrm{var}X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \lambda + \lambda^2 - \lambda^2 = \lambda$$

□

矩母函数的推导.

$$M_X(s) = \mathbb{E}[e^{sX}] = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!}e^{sk} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^s)^k}{k!} = e^{\lambda(e^s-1)}$$

□

证明泊松分布近似二项分布. 取 $\lambda = np$，则：

$$\begin{aligned}\lim_{n\to\infty}\binom{n}{k}p^k(1-p)^{nk} &= \lim_{n\to\infty}\frac{n^{\underline{k}}}{k!}\cdot\frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!}\lim_{n\to\infty}\frac{n^{\underline{k}}}{n^k}\cdot\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!}\cdot 1\cdot\lim_{n\to\infty}\left[\left(1-\frac{\lambda}{n}\right)^{-\frac{n}{\lambda}}\right]^{-\frac{\lambda(n-k)}{n}} \\ &= \frac{\lambda^k}{k!}e^{-\lambda}\end{aligned}$$

□

几何随机变量

- 记号：$X \sim G(p)$
- 实例：抛掷一枚硬币直至向上，硬币向上概率为 p，X 为抛掷次数.
- PMF：$p_X(k) = (1-p)^{k-1}p, \quad k = 1, 2, \ldots$
- 期望：$\mathbb{E}X = \dfrac{1}{p}$
- 方差：$\mathrm{var}X = \dfrac{1-p}{p^2}$

- 矩母函数：$M_X(s) = \dfrac{pe^s}{1-(1-p)e^s}$
- 无记忆性：$\mathrm{P}(X > n+m | X > n) = \mathrm{P}(X > m)$

期望的推导. 设：

$$f(x) = \sum_{k=1}^{\infty} kx^{k-1} = \sum_{k=1}^{\infty} (x^k)' = \left(\sum_{k=1}^{\infty} x^k \right)' = \left(\frac{x}{1-x} \right)' = \frac{1}{(1-x)^2}$$

则：

$$\mathbb{E}X = \sum_{k=1}^{\infty} k(1-p)^{k-1}p = pf(1-p) = \frac{1}{p}$$

□

二阶矩的推导. 设：

$$g(x) = \sum_{k=1}^{\infty} k^2 x^{k-1} = \sum_{k=1}^{\infty} k(x^k)' = \left(\sum_{k=1}^{\infty} kx^k \right)' = (xf(x))' = \left(\frac{x}{(1-x)^2} \right)' = \frac{1+x}{(1-x)^3}$$

则：

$$\mathbb{E}X^2 = \sum_{k=1}^{\infty} k^2(1-p)^{k-1}p = pg(1-p) = p \cdot \frac{2-p}{p^3} = \frac{2-p}{p^2}$$

□

方差的推导.

$$\mathrm{var}X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$$

□

矩母函数的推导.

$$M_X(s) = \mathbb{E}[e^{sX}] = \sum_{k=1}^{\infty} (1-p)^{k-1} pe^{sk} = pe^s \sum_{k=0}^{\infty} ((1-p)e^s)^k = \frac{pe^s}{1-(1-p)e^s}$$

□

证明无记忆性. 首先计算尾概率：

$$\mathrm{P}(X > n) = \sum_{k=n+1}^{\infty} (1-p)^{k-1}p = p\frac{(1-p)^n}{1-(1-p)} = (1-p)^n$$

于是：

$$\mathrm{P}(X > n+m | X > n) = \frac{P(X > n+m)}{P(X > n)} = \frac{(1-p)^{n+m}}{(1-p)^n} = (1-p)^m = \mathrm{P}(X > m)$$

□

超几何随机变量

- 实例：一盒内有 N 个球，其中 M 个白球 $N-M$ 个黑球，从中无放回地取 n 个球且每次取球独立，X 表示取出白球个数.
- PMF：$p_X(k)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}},\quad k=0,1,\ldots,n$
- 期望：$\mathbb{E}X=\frac{nM}{N}$
- 方差：$\mathrm{var}X=\frac{nM}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}$
- 性质：当 $N\to\infty$ 时，超几何分布近似二项分布.

期望的推导.

$$\mathbb{E}X=\sum_k k\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}=\frac{1}{\binom{N}{n}}\sum_k M\binom{M-1}{k-1}\binom{N-M}{n-k}=\frac{M}{\binom{N}{n}}\binom{N-1}{n-1}=\frac{nM}{N}$$

其中用到了恒等式 $\binom{N}{n}=\binom{N-1}{n-1}\frac{N}{n}$ 和范德蒙德卷积式 $\sum_k\binom{r}{k}\binom{s}{n-k}=\binom{r+s}{n}$. □

二阶矩的推导.

$$\begin{aligned}\mathbb{E}X^2&=\sum_k k^2\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}=\frac{M}{\binom{N}{n}}\sum_k k\binom{M-1}{k-1}\binom{N-M}{n-k}\\&=\frac{M}{\binom{N}{n}}\sum_k\binom{M-1}{k-1}\binom{N-M}{n-k}+\frac{M}{\binom{N}{n}}\sum_k(k-1)\binom{M-1}{k-1}\binom{N-M}{n-k}\\&=\frac{M}{\binom{N}{n}}\binom{N-1}{n-1}+\frac{M}{\binom{N}{n}}(M-1)\binom{N-2}{n-2}=\frac{nM}{N}+\frac{M(M-1)n(n-1)}{N(N-1)}\end{aligned}$$

□

方差的推导.

$$\mathrm{var}X=\mathbb{E}X^2-(\mathbb{E}X)^2=\frac{nM}{N}+\frac{M(M-1)n(n-1)}{N(N-1)}-\frac{n^2M^2}{N^2}=\frac{nM}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}$$

□

均匀随机变量

- 记号：$X\sim U(a,b)$
- PDF：$f_X(x)=\frac{1}{b-a},\quad a\leqslant x\leqslant b$
- 期望：$\mathbb{E}X=\frac{a+b}{2}$
- 方差：$\mathrm{var}X=\frac{(b-a)^2}{12}$

- 矩母函数：$M_X(s)=\dfrac{1}{b-a}\cdot\dfrac{e^{sb}-e^{sa}}{s}$

矩母函数的推导.

$$M_X(s)=\mathbb{E}[e^{sX}]=\int_a^b\frac{e^{sx}}{b-a}\mathrm{d}x=\frac{1}{b-a}\cdot\frac{1}{s}\int_{a/s}^{b/s}e^{sx}\mathrm{d}(sx)=\frac{1}{b-a}\cdot\frac{e^{sb}-e^{sa}}{s}$$

□

指数随机变量

- 记号：$X\sim E(\lambda)$
- PDF：$f_X(x)=\lambda e^{-\lambda x},\quad x\geqslant 0$
- 期望：$\mathbb{E}X=\dfrac{1}{\lambda}$
- 方差：$\mathrm{var}X=\dfrac{1}{\lambda^2}$
- 矩母函数：$M_X(s)=\dfrac{\lambda}{\lambda-s}\ (s<\lambda)$
- 无记忆性：$\mathrm{P}(X>x+y|X>x)=\mathrm{P}(X>y)$

期望的推导.

$$\mathbb{E}X=\int_0^{+\infty}x\lambda e^{-\lambda x}\mathrm{d}x=-\int_0^{+\infty}x\mathrm{d}e^{-\lambda x}=\int_0^{+\infty}e^{-\lambda x}\mathrm{d}x=-\frac{1}{\lambda}\,e^{-\lambda x}\Big|_0^{+\infty}=\frac{1}{\lambda}$$

□

二阶矩的推导.

$$\mathbb{E}X^2=\int_0^{+\infty}x^2\lambda e^{-\lambda x}\mathrm{d}x=-\int_0^{+\infty}x^2\mathrm{d}e^{-\lambda x}=2\int_0^{+\infty}xe^{-\lambda x}\mathrm{d}x=\frac{2}{\lambda^2}$$

□

方差的推导.

$$\mathrm{var}X=\mathbb{E}X^2-(\mathbb{E}X)^2=\frac{2}{\lambda^2}-\frac{1}{\lambda^2}=\frac{1}{\lambda^2}$$

□

矩母函数的推导.

$$M_X(s)=\mathbb{E}[e^{sX}]=\int_0^{\infty}\lambda e^{-\lambda x}e^{sx}\mathrm{d}x=\lambda\int_0^{\infty}e^{-(\lambda-s)x}\mathrm{d}x=\frac{\lambda}{\lambda-s}\quad(s<\lambda)$$

□

证明无记忆性. 首先计算尾概率:

$$\mathrm{P}(X>x)=\int_x^{+\infty}\lambda e^{-\lambda t}\mathrm{d}t=e^{-\lambda t}\Big|_{+\infty}^{x}=e^{-\lambda x}$$

于是:

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

□

正态随机变量

- 记号: $X\sim N(\mu,\sigma^2)$
- PDF: $f_X(x)=\dfrac{1}{\sqrt{2\pi}\sigma}\exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right)$
- 期望: $\mathbb{E}X=\mu$
- 方差: $\mathrm{var}X=\sigma^2$
- 矩母函数: $M_X(s)=\exp\left(\dfrac{\sigma^2s^2}{2}+\mu s\right)$

证明标准正态分布的归一性. 由于:

$$\begin{aligned}\left[\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x\right]^2&=\frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\exp\left(-\frac{x^2+y^2}{2}\right)\mathrm{d}x\mathrm{d}y\\&=\frac{1}{2\pi}\int_0^{2\pi}\mathrm{d}\theta\int_0^{+\infty}\exp\left(-\frac{r^2}{2}\right)r\mathrm{d}r\\&=\int_0^{+\infty}\exp\left(-\frac{r^2}{2}\right)\mathrm{d}\left(\frac{r^2}{2}\right)\\&=\exp\left(-\frac{r^2}{2}\right)\Big|_{+\infty}^{0}=1\end{aligned}$$

故:

$$\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x=1$$

□

B 二元正态分布

定义 B.1 (二元正态分布)**.** 若随机变量 X, Y 有如下联合概率密度函数:

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

则称 X, Y 服从参数为 $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$ 的二元正态分布.

注释. 矩阵形式: 设 $\mathbf{X} = \begin{pmatrix} X \\ Y \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, 则:

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

这一形式可以推广到多元正态分布.

定理 B.1 (二元正态分布的密度分解)**.** 对定义式进行变形可以得到:

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

注意到, 前一部分是 $N(\mu_1, \sigma_1)$ 的概率密度函数, 后一部分是 $N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1), \sigma_2^2(1-\rho^2)\right)$ 的概率密度函数. 又由于:

$$f_{X,Y}(x,y) = f_X(x) f_{Y|X}(y|x)$$

所以事实上后一部分是就是 $f_{Y|X}(y|x)$.

定理 B.2 (二元正态分布的边缘分布)**.** 根据密度分解容易知道, 二元正态分布的边缘分布仍是正态分布, 且 $X \sim N(\mu_1, \sigma_1^2), \quad Y \sim N(\mu_2, \sigma_2^2)$.

定理 B.3 (二元正态分布的协方差与相关系数)**.** 运用密度分解, 可以计算:

$$\begin{aligned}
\mathrm{cov}(X,Y) &= \mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)] \\
&= \iint_{\mathbb{R}^2} (x-\mu_1)(y-\mu_2) f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{+\infty} (x-\mu_1) f_X(x)\mathrm{d}x \int_{-\infty}^{+\infty} (y-\mu_2) f_{Y|X}(y|x)\mathrm{d}y \\
&= \int_{-\infty}^{+\infty} (x-\mu_1) f_X(x)\mathrm{d}x \left[\int_{-\infty}^{+\infty} y f_{Y|X}(y|x) - \mu_2\right] \\
&= \int_{-\infty}^{+\infty} (x-\mu_1) f_X(x)\mathrm{d}x \Big[\mathbb{E}[Y|X=x] - \mu_2\Big] \\
&= \int_{-\infty}^{+\infty} (x-\mu_1) f_X(x)\mathrm{d}x \Big[\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)\Big] \\
&= \rho\frac{\sigma_2}{\sigma_1}\int_{-\infty}^{+\infty} (x-\mu_1)^2 f_X(x)\mathrm{d}x \\
&= \rho\frac{\sigma_2}{\sigma_1}\mathrm{var}X \\
&= \rho\sigma_1\sigma_2
\end{aligned}$$

由此可得相关系数：

$$\rho(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}X}\sqrt{\operatorname{var}Y}}=\rho$$

也即二元正态分布定义中的参数 ρ 就是其相关系数.

定理 B.4 (二元正态分布的独立性)**.** 设 (X,Y) 服从二元正态分布，则 X,Y 独立当且仅当 $\rho=0$.

证明. 由于 X,Y 独立蕴含着 X,Y 不相关，而后者等价于相关系数 $\rho=0$，所以独立 $\Longrightarrow$ $\rho=0$. 又设 $\rho=0$，则：

$$p_{X,Y}(x,y)=\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)\cdot\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(y-\mu_2)^2}{2\sigma_2^2}\right)=p_X(x)p_Y(y)$$

所以 $\rho=0\implies$ 独立. □

推论 B.5. 对于二元正态分布而言，独立和不相关是等价的.

C 正态分布的三个导出分布

定义 C.1 (χ^2 分布)**.** 设 $X_1, X_2, \cdots, X_n$ 为 n 个独立的服从 $N(0,1)$ 的随机变量，则称

$$Z = \sum_{i=1}^{n} X_i^2$$

的分布为自由度为 n 的 χ^2 分布，记作 $Z \sim \chi^2(n)$.

期望与方差:

$$\mathbb{E}Z = n, \quad \mathrm{var}Z = 2n$$

证明. 由于

$$\mathbb{E}X_i^2 = \mathrm{var}X_i + (\mathbb{E}X_i)^2 = 1 + 0 = 1$$

故

$$\mathbb{E}Z = \mathbb{E}\left[\sum_{i=1}^{n} X_i^2\right] = \sum_{i=1}^{n} \mathbb{E}X_i^2 = n$$

又由于

$$\begin{aligned}
\mathbb{E}X_i^4 &= \int_{-\infty}^{+\infty} x^4 \varphi(x) \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^4 e^{-\frac{x^2}{2}} \mathrm{d}x = -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^3 \mathrm{d}e^{-\frac{x^2}{2}} \\
&= \frac{3}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2 e^{-\frac{x^2}{2}} \mathrm{d}x = -\frac{3}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x \mathrm{d}e^{-\frac{x^2}{2}} \\
&= \frac{3}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} \mathrm{d}x = 3
\end{aligned}$$

故

$$\mathrm{var}X_i^2 = \mathbb{E}X_i^4 - (\mathbb{E}X_i^2)^2 = 3 - 1 = 2$$

故

$$\mathrm{var}Z = \mathrm{var}\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} \mathrm{var}X_i = 2n$$

□

定义 C.2 (t 分布)**.** 设 $X \sim N(0,1)$，$Y \sim \chi^2(n)$，X, Y 相互独立，则称

$$t = \frac{X}{\sqrt{Y/n}}$$

的分布为自由度为 n 的 t 分布，记作 $t \sim t(n)$.

定义 C.3 (F 分布)**.** 设 $X \sim \chi^2(n), Y \sim \chi^2(m)$，$X, Y$ 独立，则称

$$Z = \frac{X/n}{Y/m}$$

的分布为自由度为 n, m 的 F 分布，记作 $Z \sim F(n,m)$.

定理 C.1. 设 $X_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$，则样本均值服从期望相同、方差更小的正态分布：

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

定理 C.2. 设 $X_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$，有样本均值的标准化：

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

定理 C.3. 设 $X_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$，$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 为样本方差，则：

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

定理 C.4. 设 $X_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$，$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 为样本方差，则 $\bar{X}$ 与 S^2 独立.

定理 C.5. 设 $X_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$，$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 为样本方差，则：

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t(n-1)$$

证明. 根据定理 C.2 和定理 C.3 可知；

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

于是，根据 t 分布的定义有：

$$\frac{\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t(n-1)$$

□

定理 C.6. 设 $X_i \overset{\text{i.i.d.}}{\sim} N(\mu_1, \sigma_1^2)$，$i = 1, 2, \cdots, n$，样本方差为 S_1^2；$Y_i \overset{\text{i.i.d.}}{\sim} N(\mu_2, \sigma_2^2)$，$i = 1, 2, \cdots, m$，样本方差为 S_2^2，且 X_i, Y_i 相互独立，则：

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n-1, m-1)$$

证明. 由定理 C.3 知：

$$\frac{(n-1)S_1^2}{\sigma_1^2} \sim \chi^2(n-1), \quad \frac{(m-1)S_2^2}{\sigma_2^2} \sim \chi^2(m-1)$$

于是根据 F 分布的定义有：

$$\frac{\frac{(n-1)S_1^2}{\sigma_1^2}/(n-1)}{\frac{(m-1)S_2^2}{\sigma_2^2}/(m-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n-1, m-1)$$

□

参考文献

[1] Dimitri Bertsekas and John N Tsitsiklis. 概率导论（第 *2* 版 · 修订版）. 人民邮电出版社, 2016.