机器学习笔记

Huang Andong
andonghuang1991@gmail.com
https://github.com/GitHuang

September 3, 2017

Contents

Chapter 1

事件的概率

1.1 事件的运算, 条件概率和独立性

定义 1.1 (**事件的互斥和对立**)**.** 若两个事件 A,B 不能在一次试验中同时发生, 则称它们是互斥的. 如果一些事件中任意两个事件都是互斥的, 则称这些事件是两两互斥的. 互斥事件的另一个重要的情况是" 对立事件", 若 A 为一事件, 则事件 $B = \{A$不发生$\}$, 称为 A 的对立事件, 多记作 $\bar{A}$ 或 A^c.

定义 1.2 (**事件的和 (或称并)**)**.** 设有两个事件 A,B, 定义一个新事件 C 如下:

$$C = \{A\text{发生, 或}B\text{发生}\} = \{A, B\text{至少发生一个}\}$$

则称事件 C 为事件 A 与事件 B 的和, 记为

$$C = A + B$$

事件的和可以很自然的推广到很多个事件的情形. 设有若干个事件 $A_1, A_2, \cdots, A_n$. 它们的和 A 定义为事件

$$A = \{A_1, A_2, \cdots, A_n\text{至少发生一个}\}$$

且记为

$$A = A_1 + A_2 + \cdots + A_n = \sum_{i=1}^{n} A_i = \bigcup_{i=1}^{n} A_i \tag{1.1}$$

定义 1.3 (**概率的加法定理**)**.** 若干个**互斥事件**之和的概率, 等于各事件的概率之和, 即

$$P(A_1 + A_2 + \cdots) = P(A_1) + P(A_2) + \cdots \tag{1.2}$$

事件个数是可以有限的或无限的, 这个定理就称为概率的加法定理. 其重要条件是各事件必须**两两互斥**.

定义 1.4 (**事件的积 (或称交)**)**.** 设有两个事件 A,B, 定义一个新事件 C 如下:

$$C = \{A, B\text{都发生}\}$$

称为 A,B 的积或乘积, 并记为 AB. 多个事件 $A_1, A_2, \cdots$(有限或无限个都可以) 的积的定义类似:$A = \{A_1, A_2, \cdots \text{都发生}\}$, 记为

$$A = A_1 A_2 \cdots = \prod_{i=1}^{n} A_i \tag{1.3}$$

定义 1.5 (**事件的差**)**.** 设有两个事件 A,B, 两个事件的差记为 A-B, 定义为

$$A - B = \{A\text{发生}, B\text{不发生}\} = A\bar{B}$$

定理 1.6 (**事件的和差积运算**)**.** 事件也满足加法乘法交换律和结合律

$$A + B = B + A \tag{1.4}$$

$$AB = BA \tag{1.5}$$

$$(AB)C = A(BC) \tag{1.6}$$

$$A(B - C) = AB - AC \tag{1.7}$$

定义 1.7 (**条件概率**)**.** 设有两个事件 A,B, 而 $P(B) \neq 0$. 则" 在给定 B 发生的条件下 A 的条件概率", 记为 $P(A \mid B)$, 定义为

$$P(A \mid B) = P(AB)/P(B) \tag{1.8}$$

当 $P(B) = 0$ 时, 上述公式失去了意义. 但是在高等概率论中, 可以用更加高深的数学来考虑这种情况, 或者会用极限的概念来处理.

定义 1.8 (**事件的独立性**)**.** 设有两个事件 A,B,A 的无条件概率 $P(A)$ 与其在给定 B 发生之下的条件概率 $P(A \mid B)$, 一般是有差异的. 这反映了这两个事件之间存在着一些关联. 例如, 若 $P(A \mid B) > P(A)$, 则 B 的发生使 A 的发生的可能性增大了, 即 B 存进了 A 的发生.

反之, 若 $P(A \mid B) = P(A)$, 则 B 的发生与否对 A 的发生的可能性毫无影响. 这时, 在概率上就称 A,B 两事件独立, 根据 (1.8), 可知

$$P(AB) = P(A)P(B) \tag{1.9}$$

两个事件 A,B 弱满足 (1.9) 式, 则称 A,B**独立**. 式 (1.9) 称为**概率的乘法定理**

定义 1.9 (**多个独立事件判定公式**)**.** 设 $A_1, A_2, \cdots$ 为有限或无限个事件. 如果从其中取出有限个 $A_{i_1}, A_{i_2}, \cdots, A_{i_m}$, 都成立

$$P(A_{i_1} A_{i_2} \cdots A_{i_m}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_m}) \tag{1.10}$$

则称事件 $A_1, A_2, \cdots$ **相互独立**, 或简称独立. 这个定义与由条件概率出发的定义是等价的, 后者是说, 对任何互不相同的 $i_1, i_2, \cdots, i_m$, 有

$$P(A_{i_1} \mid A_{i_2} \cdots A_{i_m}) = P(A_{i_1}) \tag{1.11}$$

即任意事件 A_{i_1} 发生的可能性大小, 不受其他事件发生的影响. 这更加接近独立性的原本含义.

定理 1.10 (**概率乘法定理**). *若干个独立事件 $A_1, A_2, \cdots, A_n$ 之积的概率, 等于各事件概率的乘积:*

$$P(A_1 \cdots A_n) = P(A_1) \cdots P(A_n). \tag{1.12}$$

乘法定理的作用与加法定理一样, 把复杂事件的概率的计算归结为更简单的事件的概率的计算, 这当然要有条件: **相加是互斥, 相乘是独立**.

推论 1.11. *若一列事件 $A_1, A_2, \cdots, A_n$ 相互独立, 则将任一部分改为对立时, 所得事件仍相互独立.*

例如, 若 A_1, A_2, A_3 相互独立, 则 $\bar{A}_1, A_2, A_2$, 或 $\bar{A}_1, \bar{A}_2, A_3$ 也相互独立.

1.2 全概率公式与贝叶斯公式

定理 1.12 (**全概率公式**). *设 $B_1, B_2, \cdots$ 为有限或无限个事件, 它们两两互斥且在每次试验中至少发生一个. 用式表之, 即*

$$B_i B_j = \emptyset(\text{不可能事件}) \quad (i \neq j)$$
$$B_1 + B_2 + \cdots = \Omega(\text{必然事件})$$

有时, 把具有这些性质的一组事件称为一个"完备的事件群". 注意, 任一事件 B 及其对立事件构成一个完备的事件群. 现在考虑任一事件 A, 因 Ω 为必然事件, 有

$$A = A\Omega = AB_1 + AB_2 + \cdots$$

因 $B_1, B_2, \cdots$ 两两互斥, 显然 $AB_1, AB_2, \cdots$ 也两两互斥. 故依照加法定理 (1.2) 可得

$$P(A) = P(AB_1) + P(AB_2) + \cdots \tag{1.13}$$

再由条件概率的定义

$$P(AB_i) = P(B_i)P(A \mid B_i) \tag{1.14}$$

把式 (1.14) 代入式 (1.13) 可得

$$P(A) = P(B_1)P(A \mid B_1) + P(B_2)P(A \mid B_2) + \cdots \tag{1.15}$$

式 (1.15) 称为 **全概率公式**.

定理 1.13 (**贝叶斯公式**)**.** *在全概率公式的假定之下, 有*

$$P(B_i \mid A) = \frac{P(AB_i)}{P(A)} = \frac{P(B_i)P(A \mid B_i)}{\sum_j P(B_j)P(A \mid B_j)} \tag{1.16}$$

*这个公式就叫做***贝叶斯公式***, 是概率论中的一个著名的公式. 这个公式首先出现在英国学者贝叶斯 (1702 1761) 去世后的 1763 年的一项著作中.*

注 1.14. 如果我们把事件 A 看成是" 结果", 把诸事件 $B_1, B_2, \cdots$ 看成导致这个结果的可能的" 原因", 则可以形象地把全概率公式看成" 由原因推结果"; 而贝叶斯公式则恰好相反, 其作用在于" 由结果推原因": 现在有一个结果 A 发生了, 在众多可能的原因中, 到底哪一个导致了这个结果?

下面给出另外一个理解贝叶斯的途径, 其实就是我在知乎对贝叶斯相关问题的回答 https://www.zhihu.com/question/19725590/answer/201940379

Chapter 2

随机变量及概率分布

2.1 离散型随机变量的分布及重要例子

定义 2.1 (**离散随机变量的概率函数**)**.** 设 X 为离散型随机变量, 其全部可能值为 $\{a_1, a_2, \cdots\}$. 则

$$p_i = P(X = a_i) \quad (i = 1, 2, \cdots) \tag{2.1}$$

称为 X 的概率函数. 显然有

$$p_i \geq 0, \quad p_i + p_2 + \cdots = 1. \tag{2.2}$$

对于离散型随机变量, 用概率函数去表达其分布是最方便的. 也可以用下面定义的分布函数表示:

定义 2.2 (**分布函数**)**.** 设 X 为离散型随机变量, 则函数

$$P(X \leq x) = F(x) \quad (-\infty < x < \infty) \tag{2.3}$$

称为 X 的**分布函数**. 注意, 这里并未限定 X 为离散型的, 它对任何随机变量都有定义. 对离散型随机变量而言, 概率函数与分布函数在下述意义上是等价的, 即知道其一即可决定另一个. 事实上, 若知道概率函数 (2.1), 则

$$F(x) = P(X \leq x) = \sum_{\{i | a_i \leq x\}} p_i. \tag{2.4}$$

例 2.2.1 (**二项分布**)**.** 设某事件 A 在一次试验中发生的概率为 p. 现把这个试验独立地重复 n 次, 以 X 记 A 在这 n 此试验中发生的次数, 则 X 可取 0,1,$\cdots$,n 等值. 其概率分布为

$$p_i = b(i; n, p) = \binom{n}{i} p^i (1-p)^{n-i} \quad (i = 0, 1, \cdots, n) \tag{2.5}$$

X 所遵循的分布为二项分布, 并常常记为 $B(n, p)$. 以后, 当随机变量 X 服从某分布 F 时, 我们用 $X \sim F$ 来表达这一点. 例如 X 服从二项分布就记为 $X \sim B(n, p)$.

例 2.2.2 (泊松分布). *若随机变量 X 的可能取值为 0,1,2,$\cdots$, 且概率分布为*

$$P(X=i)=e^{-\lambda}\lambda^i/i! \tag{2.6}$$

则称 X 服从泊松分布, 常记为 $X\sim P(\lambda)$. 此处 $\lambda>0$ 是某一常数.(2.6) 式右边对 $i=0,1,\cdots$ 求和结果为 1, 可以从公式 $e^{\lambda}=\sum_{i=0}^{\infty}\lambda^i/i!$ 得出.

注 2.3. 泊松分布也是最重要的离散型分布之一, 它出现在当 X 表示在一定的事件或空间内出现的事件个数这种场合. 前面提到的在一定时间内某交通路口所发生的事故数, 是一个典型的例子. 这个分布产生的机制也可以通过这个例子来解释. 为了方便, 设所观察的这段事件为 [0,1). 取一个很大的自然数 n, 把时间段 [0,1) 分为等长的 n 段:

$$l_1=[0,\frac{1}{n}),\ l_2=[\frac{1}{n},\frac{2}{n}),\cdots,\ l_i=[\frac{i-1}{n},\frac{i}{n}),\cdots,\ l_n=[\frac{n-1}{n},1). \tag{2.7}$$

做几个假定:

(1) 在每段 l_i 内, 恰好发生一个事故的概率, 近似地与这段事件的长 $1/n$ 成证毕, 即可取为 λ/n. 又假定在 n 很大因而 $1/n$ 很小时, 在 l_i 这么短的事件内, 要发生两次或更多的事故是不可能的. 因此, 在 l_i 时段内不发生事故的概率为 $1-\lambda/n$.

(2) $l_1,\cdots,l_n$ 各段是否发生事故是独立的.

把在 [0,1) 时段内发生的事故数 X 视作在 n 个小时段 $l_1,\cdots,l_n$ 内有事故的时段数, 则按上述两条假定,X 应该服从二项分布 $B(n,\lambda/n)$. 于是

$$P(X=i)=\binom{n}{i}\left(\frac{\lambda}{n}\right)^i\left(1-\frac{\lambda}{n}\right)^{n-i} \tag{2.8}$$

让 $n\to\infty$ 时有

$$\binom{n}{i}\Big/n^i\to 1/i!,\quad \left(1-\frac{\lambda}{n}\right)^n\to e^{-\lambda} \tag{2.9}$$

因此得出了 (2.6) 式.

2.2 连续型随机变量的分布及重要例子

连续型随机变量的概率分布, 不能像离散型随机变量那种方法去描述. 原因在于, 这种变量的取值充满区间, 无法一一排出, 若指定一个值 a, 则变量 X 恰好是 a 的概率微乎其微, 只能取零. 刻画连续型随机变量的概率分布的一个方法, 是使用 (2.3) 所定义的概率分布函数. 但是, 在理论和实用上更方便而常用的方法, 是使用所谓的" 概率密度函数", 或简称密度函数.

定义 2.4 (概率密度函数). 设连续型随机变量 X 有概率分布函数 $F(x)$, 则 $F(x)$ 的导数 $f(x)=F'(x)$ 称为 X 的概率密度函数.

连续型随机变量 X 的密度函数 $f(x)$ 都具有如下三条基本性质:

(1) $f(x) \geq 0$;

(2) $\int_{-\infty}^{\infty} f(x)dx = 1$;

(3) 对任何常数 $a < b$, 有

$$P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(x)dx \tag{2.10}$$

例 2.4.1 (**正态分布**)**.** 如果一个随机变量具有概率密度函数

$$f(x) = (\sqrt{2\pi}\sigma)^{-1} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad (-\infty < x < \infty) \tag{2.11}$$

则称 X 为正态随机变量, 并记为 $X \sim N(\mu, \sigma^2)$. 这里,N 为"Normal" 一词的首字母. 当 $\mu = 0, \sigma^2 = 1$ 时,*(2.11)* 成为

$$f(x) = \frac{\exp\left(-x^2/2\right)}{\sqrt{2\pi}} \tag{2.12}$$

它是正态分布 $N(0,1)$ 的密度函数, 称为标准正态分布. 在概率论著作中, 其密度函数和分布函数分别记为 $\varphi(x), \Phi(x)$.

2.3 多维随机变量 (随机向量)

2.3.1 离散型随机向量的分布

定义 2.5 (**n 维随机变量**)**.** 设 $X = (X_1, X_2, \cdots, X_n)$ 为一个 n 维向量, 其中每个分量, 即 $X_1, \cdots, X_n$, 都是一维随机变量, 则称 X 是一个 n 维随机向量或 n 维随机变量.

定义 2.6 (**离散型 n 维随机变量**)**.** 若一个随机向量 $X = (X_1, X_2, \cdots, X_n)$ 中每一个分量 X_i 都是一维离散型随机变量, 则称 X 为离散型的.

定义 2.7 (**随机向量的概率函数**)**.** 以 $\{a_{i1}, a_{i2}, \cdots\}$ 记 X_i 的全部可能值 $(i = 1, 2, \cdots)$, 则事件 $\{X_1 = a_{1j_1}, X_2 = a_{2j_2}, \cdots, X_n = a_{nj_n}\}$ 的概率

$$\begin{aligned} &p(j_1, j_2, \cdots, j_n) = p(X_1 = a_{1j_1},\ X_2 = a_{2j_2}, \cdots, X_n = a_{nj_n}) \\ &(j_1 = 1, 2, \cdots;\ j_2 = 1, 2, \cdots; \cdots;\ j_n = 1, 2, \cdots) \end{aligned} \tag{2.13}$$

称为随机向量 $X = (X_1, X_2, \cdots, X_n)$ 的概率函数或概率分布, 概率函数应该满足条件

$$p(j_1, j_2, \cdots, j_n) \geq 0, \quad \sum_{j_n} \cdots \sum_{j_2} \sum_{j_1} p(j_1, j_2, \cdots, j_n) = 1. \tag{2.14}$$

例 2.7.1 (**多项分布**). 多项分布是最重要的离散型多为分布. 设 $A_1, A_2, \cdots, A_n$ 是某一试验之下的完备事件群, 即事件 $A_1, A_2, \cdots, A_n$ 两两互斥, 其和为必然事件 (每次试验, 事件 $A_1, A_2, \cdots, A_n$ 必然发生一个且只发生一个). 分别以 $p_1, p_2, \cdots, p_n$ 记事件 $A_1, A_2, \cdots, A_n$ 的概率, 则

$$\sum_{i=1}^{n} p_i = 1, \quad p_i \geq 0 \tag{2.15}$$

现在将试验独立重复 N 次, 而以 X_i 记在这 N 此试验中事件 A_i 出现的次数 ($i = 1, 2, \cdots, n$), 则 $X = (X_1, X_2, \cdots, X_n)$ 为一个 n 维随机变量. 它取值的范围是:X_i 都是非负正数, 且

$$\sum_{i=1}^{n} X_i = N, \quad X_i \geq 0,\ X_i \in \mathbb{N} \tag{2.16}$$

X 的概率分布就叫做**多项分布**, 有时记作 $M(N; p_1, p_2, \cdots, p_n)$. 为了定出这个分布, 要计算事件

$$B = \{X_1 = k_1, \cdots, X_i = k_i, \cdots, X_n = k_n\}, \quad \sum_{i=1}^{n} k_i = N \tag{2.17}$$

的概率. 为了计算 $P(B)$ 的概率, 从 N 此试验的原始结果 $j_1, j_2, \cdots j_N$ 出发, 它表示第一次试验事件 A_{j_1} 发生, 第二次试验事件 A_{j_2} 发生等等. 为了使事件 B 发生, 在 $j_1, j_2, \cdots j_N$ 中有 k_1 个 1,k_2 个 2, 等等. 这种序列的数目, 等于把 N 个相异物体分成 n 维, 各堆依次有 $k_1, k_2, \cdots, k_n$ 件不同的分法. 而不同的分法总共有 $N!/(k_1!, \cdots, k_n!)$ 种. 其次, 由于独立性, 利用概率乘法定理可知, 每个适合上述条件的原始结果序列 $j_1, j_2, \cdots j_N$ 出现的概率应该为 $p_1^{k_1} p_2^{k_2} \cdots p_n^{k_n}$. 于是得到

$$\begin{gathered} P(X_1 = k_1, \cdots, X_i = k_i, \cdots, X_n = k_n) = \frac{N!}{k_1!, \cdots, k_n!} p_1^{k_1} p_2^{k_2} \cdots p_n^{k_n} \\ k_i \in \mathbb{N}, \quad \sum_{i=1}^{n} k_i = N \end{gathered} \tag{2.18}$$

式 (2.18) 就是多项分布. 名称的由来是因多项展开式

$$(x_1 + x_2 + \cdots + x_n)^N = \sum \frac{N!}{k_1!, \cdots, k_n!} x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n} \tag{2.19}$$

$\sum$ 表示求和的范围为 $k_i \in \mathbb{N}$, $k_1 + k_2 + \cdots + k_n = N$. 在 (2.19) 中令 $x_i = p_i$, 并利用 $p_1 + p_2 + \cdots + p_n = 1$, 得

$$\sum \frac{N!}{k_1!, \cdots, k_n!} p_1^{k_1} p_2^{k_2} \cdots p_n^{k_n} = 1. \tag{2.20}$$

说明分布 (2.18) 满足条件 (2.14).

2.3.2 连续型随机向量的分布

定义 2.8 (**连续型随机向量**). 设 $X = (X_1, \cdots, X_n)$ 是一个 n 维随机向量. 其取值可以视为 n 维欧式空间 $\mathbb{R}^n$ 中的一点. 如果 X 的取值能够充满 $\mathbb{R}^n$ 中某一区域, 则称它是连续性的.

与一维随机变量一样, 描述多维随机变量的概率分布, 最方便的是用概率密度函数, 为此, 我们引进一个记号:$X \in A$,$A \subset \mathbb{R}^n$.$\{X \in A\}$ 是一个随机事件, 因为做了试验后,X 的值就知道了, 因为也就能知道它是否落在了 A 内.

定义 2.9 (**连续型随机向量的概率密度函数**)**.** 若 $f(x_1, \cdots, x_n)$ 是定义在 $\mathbb{R}^n$ 上的非负函数, 使对 $\mathbb{R}^n$ 中的任何集合 A, 有

$$P(X \in A) = \int \cdots \int_A f(x_1, \cdots, x_n) dx_1 \cdots dx_n \tag{2.21}$$

则称 f 是 X 的概率密度函数. 如果把 A 取成全空间, 则 $\{X \in A\}$ 为必然事件, 其概率为 1. 因此有

$$\int \cdots \int_A f(x_1, \cdots, x_n) dx_1 \cdots dx_n = 1. \tag{2.22}$$

例 2.9.1 (**二维正态分布**)**.** 最重要的多维连续型分布是多维正态分布. 对二维的情况, 其概率密度函数有形式

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-a)^2}{\sigma_1^2} - \frac{2\rho(x_1-a)(x_2-b)}{\sigma_1\sigma_2} + \frac{(x_2-b)^2}{\sigma_2^2}\right)\right] \tag{2.23}$$

2.3.3 边缘分布

定义 2.10 (**边缘分布**)**.** 设 $X = (X_1, \cdots, X_n)$ 是一个 n 维随机向量.X 有一定的分布 F, 这是一个 n 维分布. 因为 X 的每个分量 X_i 都是一维随机变量, 故它们都有各自的分布 $F_i(i = 1, 2, \cdots)$, 这些都是一维分布, 称为随机向量 X 或其分布 F 的" 边缘分布". 对于离散的情况, 边缘分布就是把其他维度的概率加起来, 比如

$$P(X_1 = a_{1k}) = \sum_{j_2, \cdots, j_n} p(k, j_2, \cdots, j_n) \quad (k = 1, 2, \cdots) \tag{2.24}$$

实际上, 多项分布的边缘分布就是二项分布.

现在来考虑连续型随机向量的边缘分布. 先考虑二维情况, 设 $X = \{X_1, X_2\}$ 有概率密度函数 $f(x_1, x_2)$. 我们来证明: 这时 X_1 和 X_2 都具有概率密度函数. 为了证明这一点, 考虑 X_1 的分布函数 $F_1(x_1) = P(X_1 \leq x_1)$. 它可以写为 $P(X_1 \leq x_1, X_2 < \infty)$, 且

$$F_1(x_1) = P(X_1 \leq x_1) = \int_{-\infty}^{x_1} dt_1 \int_{-\infty}^{\infty} f(t_1, t_2) dt_2. \tag{2.25}$$

$\int_{-\infty}^{\infty} f(t_1, t_2) dt_2$ 是 t_1 的函数, 记之为 $f_1(t_1)$. 于是, 上式可写为

$$F_1(x_1) = \int_{-\infty}^{x_1} f_1(t_1) dt_1. \tag{2.26}$$

两边对 x_1 求导, 得到 X_1 的概率密度函数为

$$dF_1(x_1)/dx_1 = f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_2. \tag{2.27}$$

这不仅证明了 X_1 的密度函数存在, 而且还推出了其公会四. 同理求出 X_2 的密度函数为

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_1. \tag{2.28}$$

这个结果很容易推广到 n 维情况. 设 $X = (X_1, \cdots, X_n)$ 有概率密度函数 $f(x_1, \cdots, x_n)$, 为求得某分量 X_i 的概率密度函数, 只需要把 $f(x_1, \cdots, x_n)$ 中的 x_i 固定, 然后对其余变量在 $-\infty$ 到 ∞ 之间做定积分. 例如 X_1 的密度函数为

$$f_1(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \cdots, x_n)dx_2 \cdots dx_n \tag{2.29}$$

2.4 条件概率分布与随机变量的独立性

2.4.1 条件概率分布的概念

一个随机变量或向量 X 的条件概率分布, 就是在某种给定的条件之下 X 的概率分布. 它一般采取如下的形式: 设有两个随机变量或向量 X,Y, 在给定了 Y 取某个或某些值的条件下, 去求 X 的条件分布.

2.4.2 离散型随机变量的条件分布

这种情况比较简单. 设 (X_1, X_2) 为一个二维随机向量.X_1 的全部可能值为 $a_1, a_2, \cdots$;X_2 的全部可能值为 $b_1, b_2, \cdots$; 而 (X_1, X_2) 的联合分布概率为

$$p_{ij} = P(X_1 = a_i, X_2 = b_j) \quad (i, j = 1, 2, \cdots) \tag{2.30}$$

现在考虑 X_1 在给定 $X_2 = b_j$ 的条件下的条件分布, 那无非是要找条件概率 $P(X_1 = a \mid X_2 = b_j)$. 依照条件概率的定义, 有

$$\begin{aligned} P(X_1 = a \mid X_2 = b_j) &= P(X_1 = a_i, X_2 = b_j)/P(X_2 = b_j) \\ &= p_{ij}/P(X_2 = b_j) = p_{ij} \Big/ \sum_k p_{kj}, \quad (i = 1, 2, \cdots) \end{aligned} \tag{2.31}$$

类似有

$$P(X_2 = b_j \mid X_1 = a) = p_{ij} \Big/ \sum_k p_{ik}, \quad (j = 1, 2, \cdots) \tag{2.32}$$

2.4.3 连续型随机变量的条件分布

设二维随机向量 $X = (X_1, X_2)$ 有概率密度函数 $f(x_1, x_2)$. 我们先来考虑在限定 $a \le x_2 \le b$ 的条件下 X_1 的条件分布. 有

$$P(X_1 \le x_1, \mid a \le X_2 \le b) = P(X_1 \le x_1, a \le X_2 \le b) \big/ P(a \le X_2 \le b) \tag{2.33}$$

X_2 的边缘分布的密度函数 f_2 由 (2.28) 给出

$$P(X_1 \le x_1, a \le X_2 \le b) = \int_{-\infty}^{x_1} dt_1 \int_a^b f(t_1, t_2) dt_2 \tag{2.34}$$

$$P(a \le X_2 \le b) = \int_a^b f_2(t_2) dt_2 \tag{2.35}$$

由此得到

$$P(X_1 \le x_1, \mid a \le X_2 \le b) = \left(\int_{-\infty}^{x_1} dt_1 \int_a^b f(t_1, t_2) dt_2 \right) \Big/ \int_a^b f_2(t_2) dt_2 \tag{2.36}$$

这是 X_1 的条件分布函数. 对 x_1 求导数, 得到条件密度函数为

$$f_1(x_1 \mid a \le X_2 \le b) = \int_a^b f(x_1, t_2) dt_2 \Big/ \int_a^b f_2(t_2) dt_2 \tag{2.37}$$

更有兴趣的是 $a = b$ 的情况, 即在给定 X_2 等于一个值之下 X_1 的条件概率. 但是不能直接令式 (2.37)$a = b$, 可以使用极限运算获得

$$\begin{aligned} f_1(x_1 \mid x_2) &= f_1(x_1 \mid X_2 = x_2) \\ &= \lim_{h \to 0} f_1(x_1 \mid x_2 \le X_2 \le x_2 + h) \\ &= \lim_{h \to 0} \frac{1}{h} \int_{x_2}^{x_2+h} f(x_1, t_2) dt_2 \Big/ \left(\lim_{h \to 0} \frac{1}{h} \int_{x_2}^{x_2+h} f_2(t_2) dt_2 \right) \\ &= f(x_1, x_2) / f_2(x_2) \end{aligned} \tag{2.38}$$

上式可以写为

$$f(x_1, x_2) = f_2(x_2) f_1(x_1 \mid x_2) \tag{2.39}$$

就是说: 两个随机变量 X_1 和 X_2 的联合概率密度, 等于其中一个变量的概率密度乘以在给定这一个变量下另一个变量的条件概率密度. 这个公式相应于条件概率的公式 $P(AB) = P(B)P(A \mid B)$. 除了 (2.40) 外, 还有

$$f(x_1, x_2) = f_1(x_1) f_2(x_2 \mid x_1) \tag{2.40}$$

其中 f_1 为 x_1 的边缘密度, 而

$$f_2(x_2 \mid x_1) = f(x_1, x_2) / f_1(x_1) \tag{2.41}$$

则是在给定 $X_1 = x_1$ 的条件下 X_2 的条件密度. 这些公式反映的实质可以推广到任意多个变量的场合: 没有 n 维随机向量 $(X_1, \cdots, X_n)$, 其概率密度函数为 $f(x_1, \cdots, x_n)$. 则

$$f(x_1, \cdots, x_n) = g(x_1, \cdots, x_k) \cdot h(x_{k+1}, \cdots, x_n \mid x_1, \cdots, x_k) \tag{2.42}$$

其中 g 是 $(X_1, \cdots, X_k)$ 的概率密度, 而 h 是在给定 $X_1 = x_1, \cdots, X_k = x_k$ 的条件下,$X_{k+1}, \cdots, X_k$ 的条件概率密度.

2.4.4 随机变量的独立性

先考虑两个变量的情况, 设 (X_1, X_2) 为连续型.

联合分布	边缘分布	条件概率密度
$f(x_1, x_2)$	$f_1(x_1), f_2(x_2)$	$f_1(x_1 \mid x_2), f_2(x_2 \mid x_1)$

一般地,$f_1(x_1 \mid x_2)$ 是随 x_2 变化而变化的, 这反映了 X_1 与 X_2 在概率上有相依关系的事实, 即 X_1 的条件分布如何, 取决于另一个变量的值.

如果 $f_1(x_1 \mid x_2)$ 不依赖于 x_2, 而只是 x_1 的函数, 则表示 X_1 的分布情况与 X_2 的取值毫无关系. 因此 X_1 的无条件密度 $f_1(x_1)$ 就等于其条件密度 $f_1(x_1 \mid x_2)$, 这也可取为独立性的定义. 即

$$f_1(x_1) = f_1(x_1 \mid x_2) \tag{2.43}$$

$$f(x_1, x_2) = f_2(x_2) f_1(x_1 \mid x_2) = f_1(x_1) f_2(x_2) \tag{2.44}$$

即 (X_1, X_2) 的联合密度等于其各分量的密度之积. 下面我们推广到 n 为随机向量

定义 2.11 (**连续型 n 维随机向量独立性定义**)**.** 设 n 维随机向量 $(X_1, \cdots, X_n)$ 的联合密度函数为 $f(x_1, \cdots, x_n)$, 而 X_i 的边缘密度函数为 $f_i(x_i)(i = 1, \cdots, n)$. 如果

$$f(x_1, \cdots, x_n) = f_1(x_1) \cdots f_n(x_n) \tag{2.45}$$

就称随机变量 $X_1, \cdots, X_n$ 相互独立, 简称独立.

定理 2.12. *如果连续随机变量 $X_1, \cdots, X_n$ 相互独立, 则对任何 $a_i < b_i (i = 1, \cdots, n)$, 考察如下 n 个事件*

$$A_1 = \{a_1 \le X_1 \le b_1\}, \cdots, A_n = \{a_n \le X_n \le b_n\} \tag{2.46}$$

由上式定义的 n 个事件 $A_1, \cdots, A_n$ 也相互独立. 反之, 对任何 $a_i < b_i (i = 1, \cdots, n)$, 事件 $A_1, \cdots, A_n$ 独立, 则变量 $X_1, \cdots, X_n$ 相互独立.

定理 2.13. *若连续型随机向量 $X_1, \cdots, X_n$ 的概率密度函数 $f(x_1, \cdots, x_n)$ 可表为 n 个函数 $g_1, \cdots, g_n$ 之积, 其中 g_i 只依赖于 x_i, 即*

$$f(x_1, \cdots, x_n) = \prod_{i=1}^{n} g_i(x_i) \tag{2.47}$$

则 $X_1, \cdots, X_n$ 相互独立, 且 X_i 的边缘密度函数 $f_i(x_i)$ 与 $g_i(x_i)$ 只相差一个常数因子 (乘).

定理 2.14. *若连续型随机向量 $X_1, \cdots, X_n$ 相互独立, 而*

$$Y_1 = g_1(X_1, \cdots, X_m), \quad Y_2 = g_2(X_{m+1}, \cdots, X_n)$$

则 Y_1 和 Y_2 独立.

定义 2.15 (离散型随机变量相互独立). 设 $X_1, \cdots, X_n$ 都是离散型随机变量. 若对任何常数 $a_1, \cdots, a_n$ 都有

$$P(X_1 = a_1, \cdots, X_n = a_n) = P(X_1 = a_1) \cdots P(X_n = a_n)$$

则称 $X_1, \cdots, X_n$ 相互独立.

2.5 随机变量的函数的概率分布

在理论和应用上, 经常碰到这种情况: 已知某个或某些随机变量 $X_1, \cdots, X_n$ 的分布, 现另有一些随机变量 $Y_1, \cdots, Y_m$, 它们都是 $X_1, \cdots, X_n$ 的函数:

$$Y_i = g_i(X_1, \cdots, X_n) \quad (i = 1, \cdots, m) \tag{2.48}$$

要求 $Y_1, \cdots, Y_m$ 的概率分布.

2.5.1 离散型分布的情况

To be Continued...

2.5.2 连续型分布的情况

先考虑一个变量的情况. 设 X 有密度函数 $f(x)$. 设 $Y = g(x)$,g 是一个严格上升的函数. 又设 g 的导数 g' 存在. 由于 g 的严格上升性, 其反函数 $X = h(Y)$ 存在, 且 h 的导数 h' 也存在.

任取实数 y, 因为 g 严格上升, 有

$$P(Y \le y) = P(g(X) \le y) = P(X \le h(y)) = \int_{-\infty}^{h(y)} f(t)dt. \tag{2.49}$$

Y 的密度函数 $l(y)$ 即是这个表达式对 y 求导数, 有

$$l(y) = f(h(y))h'(y) \tag{2.50}$$

如果 $Y = g(X)$, 而 g 是严格递减的, 则 $\{g(X) \le y\}$ 相当于 $\{X \ge h(Y)\}$. 于是

$$P(Y \le y) = P(g(X) \le y) = P(X \ge h(y)) = \int_{h(y)}^{-\infty} f(t)dt. \tag{2.51}$$

对这个表达式对 y 求导数, 有

$$l(y) = -f(h(y))h^{'}(y) \tag{2.52}$$

而此时 $h^{'}(y)$ 也小于零, 故总有

$$l(y) = f(h(y))|h^{'}(y)| \tag{2.53}$$

例 2.15.1. 设 $Y = aX + b(a \neq 0)$, 则反函数为 $X = (Y - b)/a$. 由 *(2.53)* 式得出:$aX + b$ 的密度函数为

$$l(y) = f((y - b)/a)/|a| \tag{2.54}$$

若 $X \sim N(\mu, \sigma^2)$, 则根据正态分布的表达式可以算出 $aX+b$ 服从正态分布 $N(a\mu+b, a^2\sigma^2)$.

当 $Y = g(X)$ 而 g 不为严格单调时, 情况复杂了一些, 但并无原则困难. 我们考虑一个特例 $Y = X^2$. 仍以 f 记 X 的概率密度. 因 Y 非负, 有 $P(Y \leq y) = 0(y \leq 0)$, 对于 $y > 0$ 有

$$P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} f(t)dt \tag{2.55}$$

对 y 求导数, 得 Y 的密度函数 $l(y)$ 为

$$l(y) = \frac{1}{2\sqrt{y}}\left[f(\sqrt{y}) + f(-\sqrt{y})\right] \tag{2.56}$$

例 2.15.2. 若 $X \sim N(0, 1)$, 试求 $Y = X^2$ 的密度函数. 以 $f(x) = (\sqrt{2\pi})^{-1}e^{-x^2/2}$ 代入上式, 得

$$l(y) = \begin{cases} (\sqrt{2\pi})^{-1}e^{-y/2}, & y > 0 \\ 0, & y \leq 0. \end{cases} \tag{2.57}$$

现在考虑多个变量的函数的情况, 以两个为例. 设 (X_1, X_2) 的密度函数为 $f(x_1, x_2), Y_1, Y_2$ 都是 (X_1, X_2) 的函数:

$$Y_1 = g_1(X_1, X_2), \quad Y_2 = g_2(X_1, X_2) \tag{2.58}$$

要求 (Y_1, Y_2) 的概率密度函数 $l(y_1, y_2)$. 在此, 我们要假定式 (2.58) 是 (X_1, X_2) 到 (Y_1, Y_2) 的一一对应变换, 因而有逆变换

$$X_1 = h_1(Y_1, Y_2), \quad X_2 = h_2(Y_1, Y_2) \tag{2.59}$$

又假定 g_1, g_2 都有一阶连续偏导数. 这时, 逆变换 (2.59) 的函数 h_1, h_2 也有一阶连续偏导数, 且在一一对应变换的假定下, 雅克比行列式

$$\mathbf{J}(y_1, y_2) = \begin{vmatrix} \partial h_1/\partial y_1 & \partial h_1/\partial y_2 \\ \partial h_2/\partial y_1 & \partial h_2/\partial y_2 \end{vmatrix} \neq 0 \tag{2.60}$$

现在我们在 (Y_1, Y_2) 的平面内任取一个区域 A. 在变换 (2.59) 下, 这个区域映射到了 (X_1, X_2) 平面上的区域 B. 也就是说, 事件 $\{(Y_1, Y_2) \in A\}$ 等价于事件 $\{(X_1, X_2) \in B\}$. 考虑到 f 是 (X_1, X_2) 的密度函数, 有

$$P(Y_1, Y_2) \in A) = P((X_1, X_2) \in B) = \iint_B f(x_1, x_2) dx_1 dx_2 \tag{2.61}$$

使用重积分变量代换的公式, 在变换 (2.59) 下, 上式可变为

$$P(Y_1, Y_2) \in A) = \iint_A f\left(h_1(y_1, y_2), h_2(y_1, y_2)\right) \cdot |\mathbf{J}(y_1, y_2)| dy_1 dy_2 \tag{2.62}$$

于是 (Y_1, Y_2) 的概率密度函数为

$$l(y_1, y_2) = f\left(h_1(y_1, y_2), h_2(y_1, y_2)\right) \cdot |\mathbf{J}(y_1, y_2)| \tag{2.63}$$

我们考虑一个重要的线性变换

$$Y_1 = a_{11} X_1 + a_{12} X_2, \quad Y_2 = a_{21} X_1 + a_{22} X_2. \tag{2.64}$$

假定逆变换存在

$$X_1 = b_{11} Y_1 + b_{12} Y_2, \quad X_2 = b_{21} Y_1 + b_{22} Y_2. \tag{2.65}$$

此变换的雅克比行列式为常数

$$\mathbf{J}(y_1, y_2) = (a_{11} a_{12} - a_{21} a_{22})^{-1} \tag{2.66}$$

因此有

$$l(y_1, y_2) = f\left(b_{11} y_1 + b_{12} y_2, b_{21} y_1 + b_{22} y_2\right) \cdot |b_{11} b_{12} - b_{21} b_{22}|. \tag{2.67}$$

以上所说可完全平行地推广到 n 个变量的情形: 设 $(X_1, \cdots, X_n)$ 有密度函数 $f(x_1, \cdots, x_n)$ 而

$$Y_i = g_i(X_1, \cdots, X_n) \quad (i = 1, \cdots, n) \tag{2.68}$$

构成 $(X_1, \cdots, X_n)$ 到 $(Y_1, \cdots, Y_n)$ 的一一变换, 其逆变换为

$$X_i = h_i(Y_i, \cdots, Y_n) \quad (i = 1, \cdots, n) \tag{2.69}$$

此变换的雅可比行列式为

$$\mathbf{J}(y_1, \cdots, y_n) = \begin{vmatrix} \partial h_1/\partial y_1 & \cdots & \partial h_1/\partial y_n \\ \vdots & & \vdots \\ \partial h_n/\partial y_1 & \cdots & \partial h_n/\partial y_n \end{vmatrix} \tag{2.70}$$

则 $(Y_1, \cdots, Y_n)$ 的密度函数为

$$l(y_1, \cdots, y_n) = f\left(h_1(y_1, \cdots, y_n), \cdots, h_n(y_1, \cdots, y_n)\right) \cdot |\mathbf{J}(y_1, \cdots, y_n)| \tag{2.71}$$

在测度论中, 概率其实就是一种测度, 和面积分, 体积分是一个概念. 上述公式其实就是多维微积分的变量替换公式, 从一个域映射到另外一个域.

2.5.3 随机变量和的密度函数

设随机变量 (X_1, X_2) 的联合密度函数为 $f(x_1, x_2)$, 要求 $Y = X_1 + X_2$ 的密度函数
一个办法是考虑事件

$$\{Y \le y\} = \{X_1 + X_2 \le y\} \tag{2.72}$$

它所对应的 (X_1, X_2) 坐标平面上的集合 B, 就是直线 $x_1 + x_2 = y$ 的下方那个部分. 按密度函数的定义

$$P(Y \le y) = P(X_1 + X_2 \le y) = \iint_B f(x_1, x_2) dx_1 dx_2. \tag{2.73}$$

$$B = \{(x_1, x_2) \mid x_1 + x_2 \le y\} \tag{2.74}$$

将重积分化成累次积分, 得

$$P(Y \le y) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{y-x_1} f(x_1, x_2) dx_2 \right) dx_1 \tag{2.75}$$

对 y 求导, 即得 Y 的密度函数

$$l(y) = \int_{-\infty}^{\infty} f(x_1, y - x_1) dx_1 = \int_{-\infty}^{\infty} f(x, y - x) dx \tag{2.76}$$

作变量 $t = y - x$, 再把积分变量 t 换回到 x, 也得到 (其实就是 x_1, x_2 积分先后的问题)

$$l(y) = \int_{-\infty}^{\infty} f(y - x, x) dx \tag{2.77}$$

如果 X_1, X_2 独立, 则有 $f(x_1, x_2) = f(x_1) f(x_2)$. 此时式 (2.76,2.77) 有形式

$$l(y) = \int_{-\infty}^{\infty} f_1(x) f_2(y - x) dx = \int_{-\infty}^{\infty} f_1(y - x) f_2(x) dx \tag{2.78}$$

这个方法在数学上有些不足的地方是要在积分号下求导数, 这在理论上是有条件的, 实分析里的勒贝格积分部分导出了相应的条件. 另一个方法是配上另一个函数, 例如 $Z = X_1$, 则

$$Y = X_1 + X_2, \quad Z = X_1 \tag{2.79}$$

构成 (X_1, X_2) 到 (Y, Z) 的一一对应变换. 逆变换为

$$X_1 = Z, \quad X_2 = Y - Z \tag{2.80}$$

雅可比行列式为 -1, 绝对值为 1, 故 (Y, Z) 的联合密度函数为 $f(z, y - z)$. 然后 Y 的密度函数为边缘分布, 对 z 积分.

例 2.15.3. *若 $X_1, \cdots, X_n$ 相互独立, 分别服从正态分布 $N(\mu_1, \sigma_1^2) \cdots N(\mu_n, \sigma_n^2)$, 则 $X_1 + \cdots + X_n$ 服从正态分布 $N(\mu_1 + \cdots + \mu_n, \sigma_1^2 + \cdots + \sigma_n^2)$.*

定义 2.16 (Γ 函数)**.** Γ 函数, 也读作 Gamma 函数, 通过如下积分来定义

$$\Gamma(x)=\int_0^\infty e^{-t}t^{x-1}dt\quad(x>0). \tag{2.81}$$

Γ 函数有重要的递归公式

$$\Gamma(x+1)=x\Gamma(x) \tag{2.82}$$

Γ 函数的整数点值

$$\Gamma(n)=(n-1)! \tag{2.83}$$

$$\Gamma(n/2)=1\cdot 3\cdots 5\cdots(n-2)\cdot 2^{-(n-1)/2}\sqrt{\pi}. \tag{2.84}$$

定义 2.17 (B 函数)**.** B 函数, 也读作 Beta 函数, 通过如下积分来定义

$$B(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}dt\quad(x>0,y>0). \tag{2.85}$$

Γ 函数和 B 函数之间的关系为

$$B(x,y)=\Gamma(x)\Gamma(y)/\Gamma(x+y) \tag{2.86}$$

定义 2.18 (**独立同分布**)**.** 若 $X_1,\cdots,X_n$ 独立同分布, 则简记为 iid(independently identically distributed)

定义 2.19 (**自由度为 n 的皮尔逊卡方分布**)**.** 函数

$$k_n(x)=\begin{cases}\frac{1}{\Gamma(\frac{n}{2})2^{n/2}}e^{-x^2/2}x^{(n-2)/2}, & x>0\\ 0 & x\le 0.\end{cases} \tag{2.87}$$

是概率密度函数, 叫做自由度为 n 的皮尔逊卡方分布, 常记为 χ^2_n.

例 2.19.1. *若 $X_1,\cdots,X_n$ 独立同分布,iid.,$\sim N(0,1)$. 则 $Y=X_1^2+\cdots+X_n^2$ 服从自由度为 n 的卡方分布 χ^2_n.*

2.5.4 随机变量商的密度函数

设 (X_1,X_2) 有密度函数 $f(x_1,x_2),Y=X_2/X_1$, 要求 Y 的密度函数. 为简单计, 限制 X_1 只取正值的情况.

事件 $\{Y\le y\}=\{X_2/X_1\le y\}$ 可以写成 $\{X_2\le X_1y\}$, 由于 $X_1>0$, 故

$$P(Y\le y)=\iint\limits_B f(x_1,x_2)dx_1dx_2=\int_0^\infty\left[\int_{-\infty}^{x_1y}f(x_1,x_2)dx_2\right]dx_1. \tag{2.88}$$

对 y 求导, 可以得到 Y 的密度函数为

$$l(y)=\int_0^\infty x_1f(x_1,x_1y)dx_1 \tag{2.89}$$

若 X_1,X_2 独立, 则 $f(x_1,x_2)=f_1(x_1)f_2(x_2)$, 而上式称为

$$l(y)=\int_0^\infty x_1f_1(x_1)f_2(x_1y)dx_1 \tag{2.90}$$

Chapter 3

随机变量的数字特征

3.1 数学期望 (均值) 与中位数

3.1.1 数学期望的定义

定义 3.1 (**数学期望**)**.** 设随机变量 X 只取有限个可能值 $a_1, \cdots, a_m$, 其概率分布为 $P(X = a_i) = p_i (i = 1, \cdots, m)$. 则 X 的数学期望, 记为 $E(X)^*$ 或 EX, 定义为

$$E(X) = \sum_{i=1}^{m} a_i p_i \tag{3.1}$$

定义 3.2. 如果随机变量 X 只取有限个可能值 $a_1, \cdots$, 其概率分布为 $P(X = a_i) = p_i (i = 1, \cdots)$.

$$\sum_{i=1}^{\infty} |a_i| p_i < \infty \tag{3.2}$$

则称

$$E(X) = \sum_{i=1}^{\infty} a_i p_i \tag{3.3}$$

为 X 的数学期望.

定义 3.3. 设 X 有概率密度函数 $f(x)$. 如果

$$\int_{-\infty}^{\infty} |x| f(x) dx < \infty \tag{3.4}$$

则称

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx \tag{3.5}$$

为 X 的数学期望.

例 3.3.1. 设 X 服从泊松分布 $P(\lambda)$, 则

$$E(X) = \sum_{i=0}^{\infty} i \frac{\lambda^i}{i!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda \tag{3.6}$$

这解释了珀松分布 $P(\lambda)$ 中参数 λ 的含义, 拿以前的例子来说,λ 就是在所制定的事件段内发生事故的平均次数.

3.1.2 数学期望的性质

定理 3.4. 若干个随机变量之和的期望等于各变量的期望之和

$$E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) \tag{3.7}$$

定理 3.5. 若干个独立随机变量之积的期望等于各变量的期望之积, 即

$$E(\prod_{i=1}^{n} X_i) = \prod_{i=1}^{n} E(X_i) \tag{3.8}$$

定理 3.6 (随机变量函数的期望). 设随机变量 X 为离散型, 有分布 $P(X = a_i) = p_i (i = 1, 2, \cdots)$ 或者为连续型, 有概率密度函数 $f(x)$. 则

$$E(g(X)) = \sum_i g(a_i)p_i \quad (\text{当} \sum_i |g(a_i)|p_i < \infty) \tag{3.9}$$

或

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx \quad (\text{当} \sum_i \int_{-\infty}^{\infty} |g(x)|f(x)dx < \infty) \tag{3.10}$$

3.1.3 条件数学期望 (条件均值)

与条件分布的定义相似, 随机变量 Y 的条件数学期望就是它在给定的某种附加条件之下的数学期望. 对统计学来说, 最重要的情况是: 在给定了某些其他随机变量 $X, Z, \cdots$ 的值 $x, z, \cdots$ 的条件下 Y 的条件期望, 记为 $E(Y \mid X = x, Z = z, \cdots)$. 以只有一个变量 X 为例, 就是 $E(Y \mid X = x)$, 在不引起误解时, 也可以记为 $E(Y \mid x)$.

$$E(Y \mid x) = \int_{-\infty}^{\infty} yf(y \mid x)dy \tag{3.11}$$

如果说条件分布是变量 X 与 Y 的相依关系在概率上的完全刻画, 那么条件期望则是在一个很重要的方面刻画了两者的关系, 它反映了随着 X 的取值 x 的变化 Y 的平均变化的情况如何, 而这常常是研究者所关系的主要内容. 例如, 随着人的身高 X 的变化, 具有身高 x 的那些人的平均体重变化情况如何. 在统计学上, 常把条件期望 $E(Y \mid x)$ 作为 x 的函数, 称为 Y 对 X 的" 回归函数".

例 3.6.1. 条件期望的一个最重要的例子是 (X, Y) 服从二维正态分布 $N(a, b, \sigma_1^2, \sigma_2^2, \rho)$. 根据前面的案例, 在给定 $X = x$ 时的条件分布为正态分布

$$N(b + \rho\sigma_2\sigma_1^{-1}(x - a),\ \sigma_2^2(1 - \rho^2))$$

因为正态分布 $N(\mu,\sigma^2)$ 的期望是 μ, 故有

$$E(Y \mid x) = b + \rho\sigma_2\sigma_1^{-1}(x-a) \tag{3.12}$$

它是 x 的线性函数. 如果 $\rho > 0$, 则 $E(Y \mid x)$ 随着 x 的增加而增加, 即 Y 平均来说, 有随着 X 的增长而增长的趋势, 也就是所谓的"正相关"; 如果 $\rho < 0$, 则负相关; 如果 $\rho = 0$, 则不相关,X, Y 相互独立.

从条件数学期望的概念, 可得出求通常的 (无条件的) 数学期望的一个重要公式. 这个公式与计算概率的全概率公式相当. 回想全概率公式 $P(A) = \sum_i P(B_i)P(A \mid B_i)$, 它可以理解为通过事件 A 的条件概率 $P(A \mid B_i)$ 去计算无条件概率 $P(A)$. 更确切的说 $P(A)$ 就是条件概率 $P(A \mid B_i)$ 的某种加权平均, 权重即为 B_i 的概率. 以此类推, 变量 Y 的 (无条件) 期望应等于其条件期望 $E(Y \mid x)$ 对 x 取加权平均,x 的权与变量 X 在点 x 的概率密度 $f_1(x)$ 称比例

$$E(Y) = \int_{-\infty}^{\infty} E(Y \mid X) f_1(x)dx. \tag{3.13}$$

上式可以用例外一种记法. 记 $g(x) = E(Y \mid x)$, 则上式为

$$E(Y) = \int_{-\infty}^{\infty} g(x) f_1(x)dx = E[E(Y \mid X)]. \tag{3.14}$$

如果 X 为 n 维随机向量 $(X_1, \cdots, X_n)$, 有概率密度 $f(x_1, \cdots, x_n)$, 则上式扩展为

$$E(Y) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} E(Y \mid x_1, \cdots, x_n) f(x_1, \cdots, x_n) \tag{3.15}$$

假设 X 为一维离散变量, 有分布

$$P(X_i = a_i) = p_i \quad (i = 1, 2 \cdots) \tag{3.16}$$

则

$$E(Y) = \sum_{i=1}^{\infty} p_i E(Y \mid a_i). \tag{3.17}$$

3.1.4 中位数

定义 3.7. 设连续型随机变量 X 的分布函数为 $F(x)$, 则满足条件

$$P(X \le m) = F(m) = 1/2 \tag{3.18}$$

的数 m 称为 X 或分布 F 的中位数.

3.2 方差与矩

3.2.1 方差和标准差

定义 3.8. 设 X 为随机变量, 分布为 F, 则

$$Var(X) = E(X - EX)^2 \tag{3.19}$$

称为 X 的方差, 其平方根 $\sqrt{Var(X)}$ 为 X 的标准差. 暂记 $EX = a$, 由于 $(X-a)^2 = X^2 - 2aX + a^2$, 故有

$$Var(X) = E(X - EX)^2 = E(X^2) - 2aE(X) + a^2 = E(X^2) - (EX)^2. \tag{3.20}$$

定理 3.9. *(1) 常数的方差为 0;*

(2) 若 c 为常数, 则 $Var(X + c) = Var(X)$;

(3) 若 c 为常数, 则 $Var(cX) = c^2 Var(X)$.

定理 3.10. 独立随机变量之和的方差等于各变量的方差之和, 即

$$Var(X_1 + \cdots + X_n) = Var(X_1) + \cdots + Var(X_n) \tag{3.21}$$

3.3 协方差与相关系数

在多维随机向量中, 最有兴趣的数字特征是反映分量之间的关系的那种量, 其中最重要的, 是下面要讨论的协方差和相关系数. 我么记

$$E(X) = m_1, \quad E(Y) = m_2, \quad Var(X) = \sigma_1^2, \quad Var(Y) = \sigma_2^2.$$

定理 3.11. 称 $E[(X - m_1)(Y - m_2)]$ 为 X, Y 的协方差, 并记为 $Cov(X, Y)^*$.

协方差的一些简单性质

$$Cov(c_1X + c_2, c_3Y + c_4) = c_1c_3Cov(X, Y). \tag{3.22}$$

$$Cov(X, Y) = E(XY) - m_1m_2. \tag{3.23}$$

定理 3.12. 下面定理包含了协方差的重要性质:

(1) 若 X, Y 独立, 则 $Cov(X, Y) = E(XY) - m_1m_2 = 0$;

(2) $[Cov(X, Y)]^2 \le \sigma_1^2\sigma_2^2$. 等号当且仅当 X, Y 之间有严格的线性关系 (即存在常数 a, b, 使得 $Y = a + bX$) 时成立.

Proof. 只证明第二个. 如果把 X, Y 看作是两列高维向量, 则 $Cov(X, Y)$ 的形式非常像是两列质心为原点的向量的内积, 因此可以借助柯西不等式来证明. 或者考虑如下事实

$$E[t(X - m_1) + (Y - m_2)]^2 = \sigma_1^2 t^2 + 2Cov(X, Y)t + \sigma_2^2. \tag{3.24}$$

然后根据等式左边非负, 记得上述性质 2. 事实上, 在泛函分析里面, 涉及到函数的正交定义, 和 Cov 的概念类似. □

定义 3.13 (**相关系数**)**.** 称 $Cov(X, Y)/(\sigma_1\sigma_2)$ 为 X, Y 的相关系数, 并记为 $Corr(X, Y)^*$. 这个概念可以类比于两个向量的夹角 $(\mathbf{a} \cdot \mathbf{b}/(|\mathbf{a}| \cdot |\mathbf{b}|))$.

对于离散的随机变量, 相关系数可以理解为两个向量的夹角, 对应于柯西不等式. 对于连续的随机变量, 相关系数可以理解为函数的" 相关性", 对应于 Hölder 不等式

离散 Cauchy 不等式.

$$\sum_{k=1}^{n} a_k b_k \le \left(\sum_{k=1}^{n} a_k^2\right)^{1/2} \left(\sum_{k=1}^{n} b_k^2\right)^{1/2} \tag{3.25}$$

Hölder 不等式

$$\left(\int_E |f(x)g(x)|d\mu\right)^2 \le \int_E |f(x)|^2 d\mu \int_E |g(x)|^2 d\mu \tag{3.26}$$

定理 3.14. *相关系数的一些性质:*

(1) 若 X, Y 独立, 则 $Corr(X, Y) = 0$. 这一点并不以外, 关键利用到了若干个独立随机变量之积的期望等于各变量的期望之积这个性质. 另外一点, 随机变量独立, 可以对应理解为向量之间的正交.

(2) $-1 \le Corr(X, Y) \le 1$, 或者 $|Corr(X, Y)| \le 1$. 这个性质来源于 $[Cov(X, Y)]^2 \le \sigma_1^2\sigma_2^2$.

对上面这个定理, 有两点需要注释

(1) 当 $Corr(X, Y) = 0$ 时, 称 X, Y 不相关. 而上面性质 (1) 说明独立则一定不相关, 但是不相关却不一定独立. 独立的判定是需要看 $f(x, y) = f_1(x)f_2(y)$ 是否成立, 即联合密度是否等于其边缘密度之积.

(2) 相关系数也常常称为" 线性相关系数". 这是因为, 相关系数并不是刻画了 X, Y 的" 一般" 关系的程度, 而只是" 线性" 关系程度, 可以直接理解为就是两个向量的夹角. 即使 X, Y 有某种严格的函数关系但非线性关系,$|Corr(X, Y)|$ 不仅不必为 1, 还可以为 0.

(3) " 线性相关" 含义还可以从最小二乘法的角度去理解. 设有两个随机变量 X, Y, 现在想用 X 的某一线性函数 $a + bX$ 来逼近 Y, 问要选择怎样的常数 a, b, 才能使逼近的程度最高? 这个逼近成都, 我们就用最小二乘的观点来衡量, 要使 $E[(Y - a - bX)^2]$ 达到最小.

(4) 当 (X,Y) 为二维正态分布时, 由 $Corr(X,Y)=0$ 能够推导出 X,Y 独立. 当 (X,Y) 服从二维正态分布 $N(a,b,\sigma_1^2,\sigma_2^2,\rho)$, 则可以证明 $Corr(X,Y)=\rho$, 而当 $\rho=0$, 很容易证明 X,Y 相互独立.

3.4 大数定理和中心极限定理

先来介绍下随机变量. 就用投硬币这个例子好了. 假设只用同一个硬币, 出现正面的概率为 p, 那么投硬币出现正面或者是反面, 这个是随机的. 我们构造一个随机变量 $\{Z_n=$ 出现正面的次数/投币次数$n\}$.Z_n 对于任意给定的 n, 这明显是一个随机变量. 如果 n 从 1 到 N, 这是一个随机变量序列. 然后呢, 从 Z_1 到 Z_N, 该随机变量序列里的每个元素, 都是一个随机变量. 当正整数 n 趋向于无穷大, 我们说,Z_n 收敛于 Z, 这个 Z 可以是一个随机变量, 也可以是一个实数.

现在我们进行 n 次投币试验, 事件 A 定义为某一次试验硬币的正面朝上. 并且定义随机变量 X_n, 其中

$$X_n=\begin{cases}1 & \text{第 n 次硬币正面朝上}\\ 0 & \text{第 n 次硬币正面朝下.}\end{cases} \tag{3.27}$$

则 X_n 为随机变量. 另外定义新的变量 $\overline{X}_n$

$$\overline{X}_n=p_n=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{3.28}$$

若 $\overline{X}_n$ 随着 $n\to\infty$ 也趋于一个常数, 则就是说频率趋于概率时, 就称为大数定理.

3.4.1 大数定理

定理 3.15 (大数定理). *设 $X_1,X_2,\cdots,X_n,\cdots$ 是独立同分布的随机变量 (iid.), 记它们的公共均值为 a. 又设它们的方差存在并记为 σ^2, 记 $\overline{X}_n=\frac{1}{n}\sum_{k=1}^{n}X_k$. 则对任意给定的 $\varepsilon>0$, 有*

$$\lim_{n\to\infty}P(|\overline{X}_n-a|\geq\varepsilon)=0. \tag{3.29}$$

定理 3.16 (马尔科夫不等式). *若 Y 为只取非负值的随机变量, 则对任给常数 $\varepsilon>0$, 有*

$$P(Y\geq\varepsilon)\leq E(Y)/\varepsilon \tag{3.30}$$

Proof. 设 Y 为连续型变量, 密度函数为 $f(y)$. 因为 Y 只取非负值, 则当 $y<0$ 时 $f(y)=0$, 故

$$E(Y)=\int_0^{\infty}yf(y)dy\geq\int_{\varepsilon}^{\infty}yf(y)dy. \tag{3.31}$$

因为在 $[\varepsilon,\infty)$ 内总有 $y\geq\varepsilon$, 且

$$P(Y\geq\varepsilon)=\int_{\varepsilon}^{\infty}f(y)dy \tag{3.32}$$

故有

$$E(Y) \geq \int_{\varepsilon}^{\infty} yf(y)dy \geq \int_{\varepsilon}^{\infty} \varepsilon f(y)dy = \varepsilon P(Y \geq \varepsilon) \tag{3.33}$$

□

不等式 (3.30) 的一个重要案例为契比雪夫不等式.

定理 3.17 (**契比雪夫不等式**)**.** *若 $Var(Y)$ 存在, 则*

$$P(|Y - EY| \geq \varepsilon) \leq Var(Y)/\varepsilon^2. \tag{3.34}$$

此式的证明很简单, 只需要在式 (3.30) 式中以 $[Y - EY]^2$ 代替 Y, 以 ε^2 代替 ε 就可以了.

下面我们来证明大数定理 (3.15). 利用契比雪夫不等式, 并且注意到

$$E(\overline{X}_n) = \sum_{i=1}^{n} E(X_i)/n = na/n = a \tag{3.35}$$

故有

$$P(|\overline{X}_n - a| \geq \varepsilon) \leq Var(\overline{X}_n)/\varepsilon^2. \tag{3.36}$$

而由于 X_i 独立同分布, 且 $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, 有

$$Var(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} n\sigma^2 = \sigma^2/n. \tag{3.37}$$

因此有

$$P(|\overline{X}_n - a| \geq \varepsilon) \leq \sigma^2/(n\varepsilon^2) \to 0 \quad (\text{当} n \to \infty). \tag{3.38}$$

3.4.2 中心极限定理

在概率上, 习惯于把和的分布收敛于正态分布的那一类定理都叫做" 中心极限定理".

定义 3.18 (**中心极限定理**)**.** 设 $X_1, X_2, \cdots, X_n, \cdots$ 为独立同分布的随机变量,$E(X_i) = a, Var(X_i) = \sigma^2 (0 < \sigma^2 < \infty)$. 则对任何实数 x 有

$$\lim_{n\to\infty} P\left(\frac{1}{\sqrt{n\sigma}}\left(\sum_{i=1}^{n} X_i - na\right) \leq x\right) = \Phi(x) \tag{3.39}$$

这里 $\Phi(x)$ 是标准正态分布 $N(0,1)$, 即

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2} dt. \tag{3.40}$$

显然,$\frac{1}{\sqrt{n\sigma}}\left(\sum_{i=1}^{n} X_i - na\right)$ 就是 $\sum_{i=1}^{n} X_i$ 的标准化.

定理 3.19 (棣莫弗 -拉普拉斯定理). 设 $X_1, X_2, \cdots, X_n, \cdots$ 为独立同分布的随机变量,X_i 的分布是

$$P(X_i) = p, \quad P(X_i = 0) = 1 - p \quad (0 < p < 1) \tag{3.41}$$

则对任何实数 x 有

$$\lim_{n\to\infty} P\left(\frac{1}{\sqrt{np(1-p)}}\left(\sum_{i=1}^{n} X_i - np\right) \leq x\right) = \Phi(x) \tag{3.42}$$

这个定理讲的是用正态分布去逼近一个二项分布.

Chapter 4

参数估计

4.1 数理统计基本概念

定义 4.1 (**总体**)**.** 总体是指与所研究的问题有关的对象的全体所构成的集合.

定义 4.2 (**样本**)**.** 样本是按一定的规则从总体中抽取出的一部分个体, 且每个个体被抽取的概率是相等的.

定义 4.3 (**统计量**)**.** 完全由样本所决定的量叫做统计量.

定义 4.4 (**样本方差**)**.** 方差反映了散布程度, 其表达式为

$$S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1) \tag{4.1}$$

平均因子为 $n-1$, 而不是 n, 其深层次的原因是, 每个点的散步程度, 其实是相对于均值 $\overline{X}$ 而言, 所以, 当均值确定后, 那么散射的自由度就确定了为 $n-1$.

定义 4.5 (**样本原点矩**)**.** 设 $X_1, \cdots, X_n$ 为样本,k 为正整数, 则

$$a_k = (X_1^k + \cdots + X_n^k)/n \tag{4.2}$$

称为 k 阶样本原点矩.

定义 4.6 (**样本中心矩**)**.** 设 $X_1, \cdots, X_n$ 为样本,k 为正整数, 则

$$m_k = \sum_{i=1}^{n} (X_i - \overline{X})^k/n \tag{4.3}$$

称为 k 阶样本中心矩.

4.2 矩估计, 极大似然估计和贝叶斯估计

4.2.1 矩估计法

矩估计是 K· 皮尔逊在 19 世纪末到 20 世纪初的一系列文章中引进的. 设总体分布为 $f(x,\theta_1,\cdots,\theta_k)$, 则它的矩 (原点矩和中心矩都可以, 此处以原点矩为例)

$$a_m=\int_{-\infty}^{\infty}x^m f(x,\theta_1,\cdots,\theta_k)dx \quad (\text{或}\sum_k f(x,\theta_1,\cdots,\theta_k)) \tag{4.4}$$

依赖于 $\theta_1,\cdots,\theta_k$. 另一方面, 至少在样本大小 n 较大时,a_m 又应接近于样本原点矩. 于是

$$a_m=a_m(\theta_1,\cdots,\theta_k)\approx a_m=\sum_{i=1}^{n}X_i^m/n. \tag{4.5}$$

取 $m=1,\cdots,k$, 并将上面的近似式改成等式, 就得到一个方程组:

$$a_m(\theta_1,\cdots,\theta_k)\approx a_m \quad (m=1,\cdots,k) \tag{4.6}$$

解上述方程组, 得到根 $\hat{\theta}_i=\hat{\theta}_i(X_1,\cdots,X_n)$ $(i=1,\cdots,k)$, 就是以 $\hat{\theta}_i$ 作为 θ_i 的估计. 如果要估计的是 $\theta_1,\cdots,\theta_k$ 的某函数 $g(\theta_1,\cdots,\theta_k)$, 则用 $\hat{g}=(X_1,\cdots,X_n)=g(\theta_1,\cdots,\theta_k)$ 去估计它. 这样定出的估计量就叫做矩估计.

4.2.2 极大似然估计法

设总体有分布 $f(x,\theta_1,\cdots,\theta_k),X_1,\cdots,X_n$ 为自这个总体中抽出的样本, 则样本 $(X_1,\cdots,X_n)$ 的分布 (即其概率密度函数或概率函数) 为

$$f(x_1,\theta_1,\cdots,\theta_k)f(x_2,\theta_1,\cdots,\theta_k)\cdots f(x_n,\theta_1,\cdots,\theta_k) \tag{4.7}$$

记为 $L(x_1,\cdots,x_n;\theta_1,\cdots,\theta_k)$. 固定 $\theta_1,\cdots,\theta_k$, 而看做 $x_1,\cdots,x_n$ 的函数时,L 是一个概率密度函数或概率函数. 可以这样理解: 若

$$L(Y_1,\cdots,Y_n;\theta_1,\cdots,\theta_k)>L(X_1,\cdots,X_n;\theta_1,\cdots,\theta_k) \tag{4.8}$$

则在观察时出现 $(Y_1,\cdots,Y_n)$ 这个点的概率要比出现 $(X_1,\cdots,X_n)$ 的可能性要大. 把这件事反过来说, 可以这样想: 当已观察到 $X_1,\cdots,X_n$ 时, 若

$$L(X_1,\cdots,X_n;\theta_1',\cdots,\theta_k')>L(X_1,\cdots,X_n;\theta_1'',\cdots,\theta_k'') \tag{4.9}$$

则被估计的参数 $(\theta_1,\cdots,\theta_k)$ 是 $(\theta_1',\cdots,\theta_k')$ 的概率要比 $\theta_1'',\cdots,\theta_k''$ 的大.

当 $X_1,\cdots,X_n$ 固定而把 L 看作是 $\theta_1,\cdots,\theta_k$ 的函数时, 它称为**" 似然函数"**. 可根据上述分析得到理解: 这个函数对不同的 $(\theta_1,\cdots,\theta_k)$ 的取值反映了在观察结果 $(X_1,\cdots,X_n)$ 已知的条件下,$(\theta_1,\cdots,\theta_k)$ 的各种值的" 似然程度".

上述分析很自然的导致如下的方法: 应该用似然程度最大的那个点 $\theta_1^*, \cdots, \theta_k^*$, 即满足条件

$$L(X_1, \cdots, X_n; \theta_1^*, \cdots, \theta_k^*) = \max_{\theta_1, \cdots, \theta_k} L(X_1, \cdots, X_n; \theta_1, \cdots, \theta_k) \tag{4.10}$$

这个估计 $(\theta_1^*, \cdots, \theta_k^*)$ 称为 $(\theta_1, \cdots, \theta_k)$ 的**极大似然估计**. 为了简化计算, 通常我们会构造如下的函数

$$\ln L = \sum_{i=1}^{n} \ln f(X_i, \cdots, X_n; \theta_1, \cdots, \theta_k) \tag{4.11}$$

为了使得 L 最大, 也就是 $\ln L$ 最大, 因此可以建立**似然方程组**

$$\frac{\partial \ln L}{\partial \theta_i} = 0 \quad (i = 1, \cdots, k). \tag{4.12}$$

例 4.6.1. 设 $X_1, \cdots, X_n$ 是从正态总体 $N(\mu, \sigma^2)$ 中抽出的样本, 则似然函数为

$$L = \prod_{i=1}^{n} \left[(\sqrt{2\pi\sigma^2})^{-1} \exp\left(-\frac{1}{2\sigma^2}(X_i - \mu)^2 \right) \right] \tag{4.13}$$

因此

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2. \tag{4.14}$$

求方程组 (4.12)(把 σ^2 作为一个整体看):

$$\begin{cases} \frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2}(X_i - \mu) = 0, \\ \frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2 = 0. \end{cases} \tag{4.15}$$

于是解得

$$\mu^* = \sum_{i=1}^{n} X_i/n = \overline{X} \tag{4.16}$$

$$(\sigma^*)^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/n = m_2. \tag{4.17}$$

我们可以看见,μ, σ^2 的最大似然估计与其矩估计完全一致.

4.2.3 贝叶斯法

设总体有概率密度 $f(X, \theta)$(或概率函数, 若总体分布为离散的), 从这个总体中抽取样本 $X_1, \cdots, X_n$, 则这组样本的密度为 $f(X_1, \theta) \cdots f(X_n, \theta)$. 它可视为在给定 θ 值时 $(X_1, \cdots, X_n)$ 的密度, 则 $(\theta, X_1, \cdots, X_n)$ 的联合密度为

$$h(\theta) f(X_1, \theta) \cdots f(X_n, \theta) \tag{4.18}$$

由此算出 $(X_1, \cdots, X_n)$ 的边缘密度为

$$P(X_1, \cdots, X_n) = \int h(\theta) f(X_1, \theta) \cdots f(X_n, \theta) d\theta \tag{4.19}$$

积分的范围, 要看参数 θ 的范围而定. 因此根据贝叶斯公式, 可以得到在给定 $X_1, \cdots, X_n$ 的条件下,θ 的条件密度为

$$h(\theta \mid X_1, \cdots, X_n) = h(\theta) f(X_1, \theta) \cdots f(X_n, \theta) / p(X_1, \cdots, X_n). \tag{4.20}$$

按照贝叶斯学派的观点, 这个条件密度代表了我们现在 (即取得样本 $X_1, \cdots, X_n$ 的条件下) 对 θ 的知识, 它综合了 θ 的先验信息 (以 $h(\theta)$ 反映) 与由样本带来的信息. 通常把 (4.20) 称为 θ 的后验密度, 因为它是在做了试验以后才取得的.

如果把上述过程和我们在第一章中讲过的贝叶斯公式相比, 就可以理解, 现在我们所做的, 可以说不过是把贝叶斯公式加以连续化而已. 看如下表所示

Table 4.1: 离散贝叶斯和连续贝叶斯估计

	问题	先验知识	当前知识	后验 (现在) 知识
贝叶斯公式	事件 $B_1, \cdots, B_n$ 中哪一个发生了?	$P(B_1), \cdots, P(B_n)$	事件 A 发生了	$P(B_1 \mid A), \cdots, P(B_n \mid A)$
此处的问题	$\theta =$?	$h(\theta)$	样本 $X_1, \cdots, X_n$	后验密度式 (4.20)

例 4.6.2. *设 $X_1, \cdots, X_n$ 是正态分布总体 $N(\theta, 1)$ 中抽出的样本. 为估计 θ, 给出 θ 的先验分布为正态分布 $N(\mu, \sigma^2)$(当然,μ, σ 都已知). 求 θ 的贝叶斯估计.*

在本例中

$$h(\theta) = (2\pi\sigma)^{-1} \exp\left[-\frac{1}{2\sigma^2}(\theta - \mu)^2\right] \tag{4.21}$$

$$f(x, \theta) = (2\pi)^{-1} \exp\left[-\frac{1}{2}(x - \theta)^2\right] \tag{4.22}$$

因此由公式 *(4.20)* 知,θ 的后验密度为

$$\begin{aligned} h(\theta \mid X_1, \cdots, X_n) &= h(\theta) f(X_1, \theta) \cdots f(X_n, \theta) / p(X_1, \cdots, X_n) \\ &= \exp\left[-\frac{1}{2\sigma^2}(\theta - \mu)^2 - \sum_{i=1}^{n} \frac{1}{2}(X_i - \theta)^2\right] \Big/ I. \end{aligned} \tag{4.23}$$

其中 I 是一个与 θ 无关而只和 μ, σ, X_i 有关的数. 把上式指数项目中的 θ 合并在一起得到

$$-\frac{1}{2\sigma^2}(\theta - \mu)^2 - \sum_{i=1}^{n} \frac{1}{2}(X_i - \theta)^2 = -\frac{1}{2\eta^2}(\theta - t)^2 + J \tag{4.24}$$

其中

$$t = (n\overline{X} + \mu/\sigma^2)/(n + 1/\sigma^2) \tag{4.25}$$

$$\eta^2 = 1/(n + 1/\sigma^2) \tag{4.26}$$

而 J 与 θ 无关, 把 *(4.24)* 代入 *(4.23)* 得到

$$h(\theta \mid X_1, \cdots, X_n) = I_1 \exp\left[-\frac{1}{2\eta^2}(\theta - t)^2\right] \tag{4.27}$$

这里 $I_1 = Ie^J$ 与 θ 无关。I_1 不用直接算, 因为 $h(\theta \mid X_1, \cdots, X_n)$ 作为 θ 的函数是一个概率密度函数, 必须满足条件

$$\int_{-\infty}^{\infty} h(\theta \mid X_1, \cdots, X_n)d\theta = 1 \tag{4.28}$$

这就决定了 $I_1 = (2\pi\eta)^{-1}$. 因此 θ 的后验概率就是正态分布 $N(t, \eta^2)$, 其均值 t 就是 θ 的贝叶斯估计 $\hat{\theta}$

$$\hat{\theta} = t = \frac{n}{n + 1/\sigma^2}\overline{X} + \frac{1/\sigma^2}{n + 1/\sigma^2}\mu \tag{4.29}$$

当样本信息 X_i 和先验信息 $N(\mu, \sigma^2)$ 都存在时,θ 的估计为两者的折中. 当样本数 n 很大时, 样本信息多, 那么 $\overline{X}$ 的权重应该更大. 对于 μ 而言,σ^2 越大, 表示先验信息更不准确 *(θ 在 μ 周围散布范围很大)*, 反之,σ^2 越小, 表示根据先验信息, 有很大把握 θ 在 μ 附近.

4.3 点估计的优良性准则

4.3.1 估计量的无偏性

设某统计总计的分布包含位置参数 $\theta_1, \cdots, \theta_k, X_1, \cdots, X_n$ 是从该总体中抽出的样本, 要估计 $g(\theta_1, \cdots, \theta_k)$.$g$ 为已知函数. 设 $\hat{g}(X_1, \cdots, X_n)$ 为估计量, 如果对于任何可能的 $\theta_1, \cdots, \theta_k$ 有

$$E_{\theta_1, \cdots, \theta_k}[\hat{g}(X_1, \cdots, X_n)] = g(\theta_1, \cdots, \theta_k) \tag{4.30}$$

则称 $\hat{g}$ 是 $g(\theta_1, \cdots, \theta_k)$ 的一个**无偏估计量**. 无偏估计有两层含义: 一是没有系统偏差, 只存在随机误差. 另一个含义需要结合大数定理理解. 设想每天把这个估计量 $\hat{g}(X_1, \cdots, X_n)$ 用一次, 第 i 天的样本记为 $\hat{g}(X_1^i, \cdots, X_n^i)(i = 1, \cdots, N)$. 则按照大数定理, 当 $N \to \infty$ 时, 各次估计值的平均, 即 $\sum_{i=1}^{N} \hat{g}(X_1^i, \cdots, X_n^i)/N$ 依概率收敛到被估计的值 $g(\theta_1, \cdots, \theta_k)$. 所以若估计量有无偏性, 则在大量次数使用取平均时, 能以接近于 100% 的概率无限接近被估计的量.

例 4.6.3. *样本方差 S^2 是总体分布方差 σ^2 的无偏估计.*

$$S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n - 1) \tag{4.31}$$

为了证明这一点, 以 a 记总体分布均值, 即 $E(X_i) = a$. 也有 $E(\overline{X}) = a$, 把 $X_i - \overline{X}$ 写成 $(X_i - a) - (\overline{X} - a)$, 则有

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\overline{X})^2 &= \sum_{i=1}^{n}\left[(X_i-a)-(\overline{X}-a)\right]^2 \\ &= \sum_{i=1}^{n}(X_i-a)^2-2(\overline{X}-a)\sum_{i=1}^{n}(X_i-a)+n(\overline{X}-a)^2 \\ &= \sum_{i=1}^{n}(X_i-a)^2-2(\overline{X}-a)\cdot n(\overline{X}-a)+n(\overline{X}-a)^2 \\ &= \sum_{i=1}^{n}(X_i-a)^2-n(\overline{X}-a)^2\end{aligned} \tag{4.32}$$

因为 $a=E(X_i)=E(\overline{X})$,

$$E(X_i-a)^2=Var(X_i)=\sigma^2 \quad (i=1,\cdots,n) \tag{4.33}$$

$$E(\overline{X}-a)^2=Var(\overline{X})=\sum_{i=1}^{n}Var(X_i)/n^2=n\sigma^2/n=\sigma^2/n \tag{4.34}$$

于是得到

$$E(S^2)=\frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i-\overline{X})^2\right]=\frac{1}{n-1}(n\sigma^2-n\cdot\sigma^2/n)=\sigma^2. \tag{4.35}$$

这就说明了 S^2 是 σ^2 的无偏估计. 另外一点,$\sum_{i=1}^{n}(X_i-\overline{X})^2$ 的自由度为 $n-1$, 这正好是正确的除数.

例 4.6.4. 下面我们来探讨下总体分布的标准差的无偏估计. 我们想要的标准差的无偏估计表达式为

$$E_{unbiasd}=E(\sqrt{(S^2)})=ES\neq\sqrt{E(S^2)}=\sqrt{\sigma^2}=\sigma \tag{4.36}$$

这里回顾方差定义式 *(3.19)* 以及公式 *(3.20)*

$$Var(Y)=E(Y-EY)^2 \tag{4.37}$$

$$Var(Y)=E(Y-EY)^2=E(Y^2)-2aE(Y)+a^2=E(Y^2)-(EY)^2. \tag{4.38}$$

这里我们定义新的随机变量 $Y=S$, 真实的标准差 $E(S)=\sigma$, 把 Y 替换成 S, 上式可得

$$Var(S)=E(S^2)-(ES)^2. \tag{4.39}$$

进一步有

$$E(S^2)=\sigma^2=Var(S)+(ES)^2 \tag{4.40}$$

易知 $ES<\sqrt{(E(S^2))}=\sigma$. 即如果用 S 去估计 σ, 总是系统性的偏低. 实际上上述不等式来源于 *Jensens' Inequality [1]*

$$ES=\frac{1}{n}\sum_{i=1}^{n}S_n\leq\sqrt{\frac{1}{n}\sum_{i=1}^{n}S_n^2}=\sqrt{E(S^2)}=\sigma \tag{4.41}$$

在一些情况下, 可以通过简单的调整来达到无偏估计. 办法是把 S 乘上一个大于 *1*, 与样本大小 n 有关的因子 c_n, 得到 c_nS. 适当选择 c_n, 可以使 $E(c_nS) = c_nE(S) = \sigma$.

定义 4.7 (矩母函数). 矩母函数又称为动差生成函数 (moment-generating function), 其定义为

$$M_X(t) = E(e^{tX}), \quad t \in \mathbb{R} \tag{4.42}$$

前提是这个期望存在. 如果随机变量 X 具有连续概率密度函数 $f(x)$, 则其矩母函数为

$$\begin{aligned} M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\,\mathrm{d}x &= \int_{-\infty}^{\infty} \left(1 + tx + \frac{t^2x^2}{2!} + \cdots\right) f(x)\,\mathrm{d}x \\ &= 1 + tm_1 + \frac{t^2 m_2}{2!} + \cdots \end{aligned} \tag{4.43}$$

只要动差生成函数在 $t = 0$ 周围的开区间存在, 第 n 个矩为:

$$\mathrm{E}\left(X^n\right) = M_X^{(n)}(0) = \left.\frac{\mathrm{d}^n M_X(t)}{\mathrm{d}t^n}\right|_{t=0} \tag{4.44}$$

如果动差生成函数在这个区间内是有限的, 则它唯一决定了一个概率分布. 一些其它在概率论中常见的积分变换也与动差生成函数有关, 包括特征函数以及概率生成函数.

定理 4.8. 设 $X_1, \cdots, X_n$ 为独立同分布随机变量, 且 $X_i \sim N(\mu, \sigma^2)$. 其中 $\bar{X} = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i$ 为 n 个样本的均值. 而 $S^2 = \dfrac{1}{n-1}\sum\limits_{i=1}^{n}(X_i - \bar{X})^2$ 为 n 个样本的采样方差. 那么有

(1) $\bar{X}$ 和 S^2 相互独立.

(2) $\dfrac{(n-1)S^2}{\sigma^2} = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$

Proof. 第一个证明很难, 这里省略其过程. 下面证明命题 (2). 我们定义一个新的变量 W

$$\begin{aligned} W &= \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{(X_i - \bar{X}) + (\bar{X} - \mu)}{\sigma}\right)^2 \\ &= \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 + \sum_{i=1}^{n}\left(\frac{\bar{X} - \mu}{\sigma}\right)^2 + 2\left(\frac{\bar{X} - \mu}{\sigma}\right)\sum_{i=1}^{n}(X_i - \bar{X}) \\ &= \sum_{i=1}^{n}\frac{(X_i - \bar{X})^2}{\sigma^2} + \frac{n(\bar{X} - \mu)^2}{\sigma^2} \end{aligned} \tag{4.45}$$

根据采样方程的定义有

$$(n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{4.46}$$

因此 W 可以写成

$$W = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} + \frac{n(\bar{X} - \mu)^2}{\sigma^2} \tag{4.47}$$

$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$ 为 n 个 χ_1 卡方分布之和, 根据前面章节的结论, 其分布为 $\chi(n)$,W 的矩母函数为

$$M_W(t)=(1-2t)^{-n/2},\quad t<\frac{1}{2} \tag{4.48}$$

由于

$$Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1) \tag{4.49}$$

可知 Z^2 服从自由度为 1 的卡方分布

$$Z^2=\frac{n(\bar{X}-\mu)^2}{\sigma^2}\sim\chi^2(1) \tag{4.50}$$

因此 Z^2 的的矩母函数为

$$M_{Z^2}(t)=(1-2t)^{-1/2},\quad t<\frac{1}{2} \tag{4.51}$$

根据定义有

$$M_W(t)=E(e^{tW})=E\left[e^{t((n-1)S^2/\sigma^2+Z^2)}\right]=M_{(n-1)S^2/\sigma^2}(t)\cdot M_{Z^2}(t) \tag{4.52}$$

上述公式利用到了 $\bar{X}$ 和 S^2 相互独立这个结论. 联立 (4.48,4.51,4.52) 可得

$$(1-2t)^{-n/2}=M_{(n-1)S^2/\sigma^2}(t)\cdot(1-2t)^{-1/2} \tag{4.53}$$

$$M_{(n-1)S^2/\sigma^2}(t)=(1-2t)^{-n/2}\cdot(1-2t)^{1/2}=(1-2t)^{-(n-1)/2},\quad t<\frac{1}{2} \tag{4.54}$$

故 $(n-1)S^2/\sigma^2$ 的矩母函数为自由度 $n-1$ 的卡方分布, 即

$$\frac{(n-1)S^2}{\sigma^2}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}\sim\chi^2_{(n-1)} \tag{4.55}$$

□

例 4.8.1 (**标准差的无偏估计**)**.** 上面的定理证明已经有了

$$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2_{n-1} \tag{4.56}$$

而 χ^2_k 的概率分布为

$$p(x)=\frac{(1/2)^{k/2}}{\Gamma(k/2)}x^{k/2-1}e^{-x/2} \tag{4.57}$$

由此可得

$$\begin{aligned}E(s)&=\sqrt{\frac{\sigma^2}{n-1}}E\left(\sqrt{\frac{s^2(n-1)}{\sigma^2}}\right)\\&=\sqrt{\frac{\sigma^2}{n-1}}\int_0^\infty\sqrt{x}\frac{(1/2)^{(n-1)/2}}{\Gamma((n-1)/2)}x^{((n-1)/2)-1}e^{-x/2}\,dx\end{aligned} \tag{4.58}$$

对上式变形

$$\begin{aligned} E(s) &= \sqrt{\frac{\sigma^2}{n-1}} \int_0^\infty \frac{(1/2)^{(n-1)/2}}{\Gamma(\frac{n-1}{2})} x^{(n/2)-1} e^{-x/2}\ dx \\ &= \sqrt{\frac{\sigma^2}{n-1}} \cdot \frac{\Gamma(n/2)}{\Gamma(\frac{n-1}{2})} \int_0^\infty \frac{(1/2)^{(n-1)/2}}{\Gamma(n/2)} x^{(n/2)-1} e^{-x/2}\ dx \\ &= \sqrt{\frac{\sigma^2}{n-1}} \cdot \frac{\Gamma(n/2)}{\Gamma(\frac{n-1}{2})} \cdot \frac{(1/2)^{(n-1)/2}}{(1/2)^{n/2}} \underbrace{\int_0^\infty \frac{(1/2)^{n/2}}{\Gamma(n/2)} x^{(n/2)-1} e^{-x/2}\ dx}_{\chi^2_n \text{ density}} \end{aligned} \tag{4.59}$$

现在我们知道最后一项积分为 *1*, 简化得到

$$E(s) = \sigma \cdot \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma(n/2)}{\Gamma(\frac{n-1}{2})} \tag{4.60}$$

4.3.2 最小方差无偏估计

一个参数往往有不止一个无偏估计, 从这些众多的无偏估计中, 我们想要挑出那个最优的. 这牵涉到两个问题: 一是为优良性制定一个准则, 而是在已定的准则下, 如何去找到最优者. 下面做一点点初步介绍.

均方误差

设 $X_1, \cdots, X_n$ 是从某一带参数 θ 的总体中抽出的样本, 要估计 θ, 若我们采用量 $\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n)$, 我们把

$$M_{\hat{\theta}}(\theta) = E_\theta \left[\hat{\theta}(X_1, \cdots, X_n) - \theta \right] \tag{4.61}$$

作为 $\hat{\theta}$ 的误差大小从整体角度的一个衡量. 这个量越小, 就表示 $\hat{\theta}$ 的误差平均来说越小, 也即越优.

4.4 区间估计

设 $X_1, \cdots, X_n$ 是从总体中抽出的样本. 所谓的 θ 的**区间估计**, 就是以满足条件 $\hat{\theta}_1(X_1, \cdots, X_n) \le \hat{\theta}_2(X_1, \cdots, X_n)$ 的两个统计量 $\hat{\theta}_1, \hat{\theta}_2$ 为端点的区间 $[\hat{\theta}_1, \hat{\theta}_2]$. 一旦有了样本 $X_1, \cdots, X_n$, 就把 θ 估计在区间 $[\hat{\theta}_1(X_1, \cdots, X_n), \hat{\theta}_2(X_1, \cdots, X_n)]$ 之内. 不难理解, 这里有两个要求:

(1) θ 要尽大可能性落在区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 内, 也就是说, 概率

$$P_\theta \left(\hat{\theta}_1(X_1, \cdots, X_n) \le \theta \le \hat{\theta}_2(X_1, \cdots, X_n) \right) \tag{4.62}$$

要尽可能大;

(2) 估计的精密度要尽可能高, 比如区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 要尽可能小.

定义 4.9. 给定一个很小的数 $\alpha > 0$. 如果对参数 θ 的任何值, 概率 (4.62) 式都等于 $1-\alpha$, 则称区间估计 $[\hat{\theta}_1, \hat{\theta}_2]$ 的置信系数为 $1-\alpha$.

区间估计也常称为置信区间, 字面上的意思是: 对该区间能包含未知参数 θ 可置信到何种程度. 比如要估计一个服从正态分布的随机变量的均值, 前面讲过的点估计就是直接用 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 来估计均值. 而区间估计则是给出一个均值的置信区间, 而非单个值.

Chapter 5

多元正态分布

5.1 多元正态分布定义

多维正态分布也称为 Multivariate Normal Distribution(MVN). 我们回顾下一维的情况 (2.11)

$$f(x) = (\sqrt{2\pi}\sigma)^{-1} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad (-\infty < x < \infty) \tag{5.1}$$

其中 μ 为分布的均值,σ^2 为分布的方差. 而 $p-$ 维度的概率密度为

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{5.2}$$

其中 $\boldsymbol{\mu}$ 有 p 个独立的参数,$\boldsymbol{\Sigma}$ 有 $\frac{1}{2}p(p+1)$ 个独立的参数. 我们记

$$\boldsymbol{x} \sim \boldsymbol{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{5.3}$$

为 $p-$ 维度的正态分布概率密度. 其中

$$E(\boldsymbol{X}) = \boldsymbol{\mu} = [E(X_1), \cdots, E(X_p)]^T \tag{5.4}$$

$$\begin{aligned} Cov(\boldsymbol{X}) = \boldsymbol{\Sigma} &= E\left[(\boldsymbol{X} - E(\boldsymbol{X})) \cdot (\boldsymbol{X} - E(\boldsymbol{X}))^T\right] \\ &= \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1p}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2p}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1}^2 & \sigma_{p2}^2 & \cdots & \sigma_{pp}^2 \end{bmatrix} \end{aligned} \tag{5.5}$$

5.2 多元正态分布的基本性质

如果 $\boldsymbol{x}(p \times 1)$ 为均值是 $\boldsymbol{\mu}$, 协方差 $\boldsymbol{\Sigma}$ 的 MVN. 那么有

1. 任何 $\boldsymbol{x}$ 的线性组合还是 MVN. 设 $\boldsymbol{A}(q\times p)$,$\boldsymbol{c}(q\times 1)$, 线性映射 $\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}+\boldsymbol{c}$, 则

$$\boldsymbol{y}\sim \boldsymbol{N}_q(\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y) \tag{5.6}$$

其中 $\boldsymbol{\mu}_y=\boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{c}$, 且 $\boldsymbol{\Sigma}_y=\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T$. 这里简单证明下

Proof. 由 $\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}+\boldsymbol{c}$ 可得

$$\boldsymbol{y}-(\boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{c})=\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu}) \tag{5.7}$$

另外

$$\begin{aligned}-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\cdot\boldsymbol{A}^T(\boldsymbol{A}^T)^{-1}\cdot\boldsymbol{\Sigma}^{-1}\cdot\boldsymbol{A}^{-1}\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu})\\ &= -\frac{1}{2}\left[(\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu}))^T(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)^{-1}\cdot\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu})\right]\\ &= -\frac{1}{2}\left[((\boldsymbol{y}-\boldsymbol{\mu}_y))^T(\boldsymbol{\Sigma}_y)^{-1}\cdot(\boldsymbol{y}-\boldsymbol{\mu}_y)\right]\end{aligned} \tag{5.8}$$

□

2. 任何 $\boldsymbol{x}$ 的子集都是 MVN 分布;

3. 如果 MVN 的变量不相关, 则它们相互独立. 特别地, 若 $\sigma_{ij}=0, i\neq j$, 则 x_i,x_j 独立;

4. MVN 的条件分布也是 MVN.

5.3 多元正态分布的条件分布

定义 $\boldsymbol{X}=[\boldsymbol{X_1},\boldsymbol{X_2}]^T$, 其维度为 $[q,p-q]^T$, 均值为 $\boldsymbol{\mu}=[\boldsymbol{\mu}_1,\boldsymbol{\mu}_2]^T$, 协方差为

$$\Sigma=\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}$$

的 MVN. 则在给定 $\boldsymbol{X_1}=\boldsymbol{x_1}$ 的条件下,$\boldsymbol{X_2}$ 的概率密度为 MVN, 其均值和协方差为

$$E(\boldsymbol{X}_2\mid\boldsymbol{X}_1=\boldsymbol{x}_1)=\boldsymbol{\mu}_2+\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}_1) \tag{5.9}$$

$$Cov(\boldsymbol{X}_2\mid\boldsymbol{X}_1=\boldsymbol{x}_1)=\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \tag{5.10}$$

在这里, 大写字母表示随机变量, 小写字母代表某个特定的值, 希望不要混淆. 下面我们来证明式 (5.9,5.10).

Proof. 最直接的方式计算出条件概率 $f_2(\boldsymbol{X}_2\mid\boldsymbol{X}_1=\boldsymbol{x}_1)$, 然后再根据条件期望 (3.11) 和方差的定义来计算. 但是这样会比较麻烦. 之前我们有定理说多元正态分布的条件分布也是多

元正态分布, 因此我们只需要计算相应的条件期望以及方差. 引入一个与 $\boldsymbol{X_1}$ 不相关的随机变量 $\boldsymbol{Z} = \boldsymbol{X}_2 + \boldsymbol{A}\boldsymbol{X}_1$, 其中 $\boldsymbol{A} = -\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}$. 现在证明 $\boldsymbol{X_1}, \boldsymbol{Z}$ 不相关.

$$\begin{aligned} Cov(\boldsymbol{Z}, \boldsymbol{X}_1) &= Cov(\boldsymbol{X}_2 + \boldsymbol{A}\boldsymbol{X}_1,\ \boldsymbol{X}_1) \\ &= Cov(\boldsymbol{X}_2,\ \boldsymbol{X}_1) + Cov(\boldsymbol{A}\boldsymbol{X}_1,\ \boldsymbol{X}_1) \\ &= \boldsymbol{\Sigma}_{21} + \boldsymbol{A} \cdot Cov(\boldsymbol{X}_1,\ \boldsymbol{X}_1) \\ &= \boldsymbol{\Sigma}_{21} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} \cdot \boldsymbol{\Sigma}_{11} = 0 \end{aligned} \tag{5.11}$$

上述证明用到的一些协方差的一些运算性质, 可以通过其定义得出. 上式可以知道, 随机变量 $\boldsymbol{X_1}, \boldsymbol{Z}$ 不相关, 且相互独立, 且显然有 $E(\boldsymbol{Z}) = \boldsymbol{\mu}_2 + \boldsymbol{A}\boldsymbol{\mu}_1$. 因此有

$$\begin{aligned} E(\boldsymbol{X}_2 \mid \boldsymbol{X}_1 = \boldsymbol{x}_1) = E(\boldsymbol{X}_2 \mid \boldsymbol{x}_1) &= E(\boldsymbol{Z} - \boldsymbol{A}\boldsymbol{X}_1 \mid \boldsymbol{x}_1) \\ &= E(\boldsymbol{Z} \mid \boldsymbol{x}_1) - E(\boldsymbol{A}\boldsymbol{X}_1 \mid \boldsymbol{x}_1) \\ &= E(\boldsymbol{Z}) - \boldsymbol{A}\boldsymbol{x}_1 \\ &= \boldsymbol{\mu}_2 + \boldsymbol{A}\boldsymbol{\mu}_1 - \boldsymbol{A}\boldsymbol{x}_1 \\ &= \boldsymbol{\mu}_2 + \boldsymbol{A}(\boldsymbol{\mu}_1 - \boldsymbol{x}_1) = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{A}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1) \end{aligned} \tag{5.12}$$

上述公式需要一点点解释,$E(\boldsymbol{A}\boldsymbol{X}_1 \mid \boldsymbol{x}_1)$ 这个条件概率的意思是当随机变量 $\boldsymbol{X}_1 = \boldsymbol{x}_1$ 时 $\boldsymbol{A}\boldsymbol{X}_1$ 的期望, 显然, 此时随机变量 $\boldsymbol{X}_1$ 为定值 $\boldsymbol{x}_1$, 均值也就是其本身了. 现在我们证明条件协方差.

$$\begin{aligned} &Cov(\boldsymbol{X}_2 \mid \boldsymbol{x}_1) = Cov(\boldsymbol{Z} - \boldsymbol{A}\boldsymbol{X}_1 \mid \boldsymbol{x}_1) \\ &= Cov(\boldsymbol{Z} - \boldsymbol{A}\boldsymbol{x}_1 \mid \boldsymbol{x}_1) \\ &= Cov(\boldsymbol{Z} \mid \boldsymbol{x}_1) = Cov(\boldsymbol{Z}) \\ &= Cov(\boldsymbol{X}_2 + \boldsymbol{A}\boldsymbol{X}_1) \\ &= Cov(\boldsymbol{X}_2, \boldsymbol{X}_2) + \boldsymbol{A}Cov(\boldsymbol{X}_1, \boldsymbol{X}_2) + Cov(\boldsymbol{X}_2, \boldsymbol{X}_1)\boldsymbol{A}^T + \boldsymbol{A}Cov(\boldsymbol{X}_1, \boldsymbol{X}_1)\boldsymbol{A}^T \\ &= \boldsymbol{\Sigma}_{22} + \boldsymbol{A}\boldsymbol{\Sigma}_{12} + \boldsymbol{\Sigma}_{21}\boldsymbol{A}^T + \boldsymbol{A}\boldsymbol{\Sigma}_{11}\boldsymbol{A}^T \\ &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_{21}(\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1})^T + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{11}(\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1})^T \\ &= \boldsymbol{\Sigma}_{22} - 2\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{21} \\ &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}. \end{aligned} \tag{5.13}$$

上述证明利用到了 $\boldsymbol{\Sigma}_{ij}$ 为对称阵列, 且对称矩阵的逆也是对称矩阵这两个事实, 因此有 $(\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1})^T = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{21}$. □

Chapter 6

Appendix

6.1 常见泰勒展开函数

一个实函数的泰勒级数为 [2]

$$f(x)=\sum_{n=0}^{\infty}\frac{f^n(a)}{n!}(x-a)^n \tag{6.1}$$

指数函数泰勒级数展开式为

$$e^x=\sum_{n=0}^{\infty}\frac{x^n}{n!}=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\cdots\quad(x\in\mathbb{R}) \tag{6.2}$$

自然对数泰勒级数展开式为

$$\ln(1-x)=-\sum_{n=0}^{\infty}\frac{x^n}{n}=-x-\frac{x^2}{2}-\frac{x^3}{3}+\cdots\quad(|x|<1). \tag{6.3}$$

$$\ln(1+x)=\sum_{n=0}^{\infty}(-1)^{n+1}\frac{x^n}{n}=x-\frac{x^2}{2}+\frac{x^3}{3}-\cdots\quad(|x|<1). \tag{6.4}$$

等比级数 (Geometric series) 的泰勒展开式为

$$\frac{1}{x-1}=\sum_{n=0}^{\infty}x^n\quad(|x|<1). \tag{6.5}$$

$$\frac{1}{(x-1)^2}=\sum_{n=1}^{\infty}vx^{n-1}\quad(|x|<1). \tag{6.6}$$

$$\frac{1}{(x-1)^3}=\sum_{n=2}^{\infty}\frac{(n-1)n}{2}x^{n-2}\quad(|x|<1). \tag{6.7}$$

二项级数 (Binomial series) 的泰勒展开式为

$$(1+x)^\alpha=\sum_{n=0}^{\infty}\binom{\alpha}{n}x^n \tag{6.8}$$

其中

$$\binom{\alpha}{n} = \prod_{k=1}^{n} \frac{\alpha - k + 1}{k} = \frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!}. \tag{6.9}$$

三角级数展开式为

$$\sin(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots \tag{6.10}$$

$$\cos(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots \tag{6.11}$$

Bibliography

[1] Wikipedia, "Jensen's inequality — wikipedia, the free encyclopedia," 2017, [Online; accessed 29-July-2017]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Jensen%27s_inequality&oldid=768419847

[2] ——, "Taylor series — wikipedia, the free encyclopedia," 2017. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Taylor_series&oldid=790017852