

淡蓝小点技术系列：面向机器学习（深度学习、人工智能）的数学基础

淡蓝小点Bluedotdot

微信：[bluedotdot.cn](https://mp.weixin.qq.com/s?__biz=MzU4NjQyMjIwMA==&mid=2247484868&idx=1&sn=12345678901234567890123456789012)

2024 年 8 月 13 日

目录

1 项目说明

2 实变函数

3 概率统计



Draft版，未完成



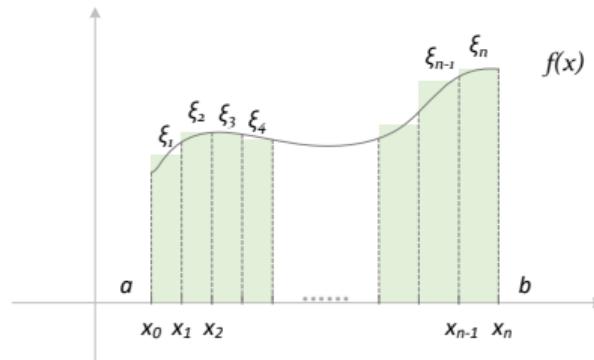
为什么需要勒贝格积分



实变函数：讨论以实数为变量的函数，但重点是将数学分析中的黎曼积分改造拓展为勒贝格积分¹

- 测度：[集合](#)、[测度](#)、[Lebesgue可测集](#)等
- 积分：[可测函数](#)、[Lebesgue积分](#)等

机器学习主要用到测度的概念，它是现代概率论的基础



¹ Lebesgue integral, 凡是黎曼可积的一定勒贝格可积



设函数 $f(x)$ 在闭区间 $[a, b]$ 上有定义，对闭区间 $[a, b]$ 作任意分割 $\Delta : a = x_0 < x_1 < \dots < x_n = b$ ，用 Δ_i 表示区间 $[x_{i-1}, x_i]$ 的长度。在每个子区间 $[x_{i-1}, x_i]$ 内任取一点 ξ_i ，作积分和（也称黎曼和）²

$$S_n(f, \Delta) = \sum_{i=1}^n f(\xi_i) \Delta_i$$

记 $\lambda = \max\{\Delta x_1, \Delta x_2, \dots, \Delta x_n\}$ ，若不论对区间 $[a, b]$ 作哪种划分，也不论如何选取每个 ξ_i ，只要当 $\lambda \rightarrow 0$ 时，积分和 $S_n(f, \Delta)$ 都趋于同一常数 I

$$\lim_{\lambda \rightarrow 0} S_n(f, \Delta) = \lim_{\lambda \rightarrow 0} \sum_{i=1}^n f(\xi_i) \Delta x_i = I$$

则称 $f(x)$ 在 $[a, b]$ 上可积，并把此极限 I 叫做函数 $f(x)$ 在 $[a, b]$ 上的定积分，记作 $\int_a^b f(x) dx$ ，即

$$\int_a^b f(x) dx = I = \lim_{\lambda \rightarrow 0} \sum_{i=1}^n f(\xi_i) \Delta x_i$$

此积分由德国数学家黎曼（Riemann）给出严格定义，因而一般称为黎曼积分，简称为R积分

²注意，这里的分割和区间内选值都是任意的



设函数 $f(x)$ 在闭区间 $[a, b]$ 上有定义，对闭区间 $[a, b]$ 作任意分割 $\Delta : a = x_0 < x_1 < \dots < x_n = b$ ，把 $f([x_{i-1}, x_i])$ 的上确界和下确界分别记作 M_i 和 m_i

$$M_i = \sup\{f(x) | x \in [x_{i-1}, x_i]\}, \quad m_i = \inf\{f(x) | x \in [x_{i-1}, x_i]\}$$

用 $\bar{S}(f, \Delta)$ 和 $\underline{S}(f, \Delta)$ 分别表示函数 $f(x)$ 关于分割 Δ 的达布上和 (upper Darboux sum) 和达布下和 (lower Darboux sum)，简称上和与下和

$$\bar{S}(f, \Delta) = \sum_{i=1}^n M_i \Delta x_i, \quad \underline{S}(f, \Delta) = \sum_{i=1}^n m_i \Delta x_i$$

若记 $\omega_i = M_i - m_i$ ，用 ω_i 表示 $f(x)$ 在 $[x_{i-1}, x_i]$ 上的振动幅度，并用 $\omega(f, \Delta)$ 表示 $f(x)$ 的振幅和。 $f(x)$ 在 $[a, b]$ 上所有达布上和的下确界称为 $f(x)$ 在 $[a, b]$ 上的达布上积分，达布下和的上确界称为达布下积分，简称上积分和下积分，记作

$$\int_a^b f(x) dx = \inf_{\Delta} \{\bar{S}(f, \Delta)\}, \quad \underline{\int_a^b f(x) dx} = \sup_{\Delta} (\underline{S}(f, \Delta))$$



很显然，无论作何种分割 Δ 、无论如何选取 ξ_i 总有

$$\sum_{i=1}^n m_i \Delta x_i \leq \sum_{i=1}^n f(\xi_i) \Delta x_i \leq \sum_{i=1}^n M_i \Delta x_i$$

若 $f(x)$ 在 $[a, b]$ 上达布可积，则一定有 $\underline{\int_a^b} f(x) dx = \overline{\int_a^b} f(x) dx$ 。因而可以证明当 $\lambda \rightarrow 0$ 时有

$$\underline{\int_a^b} f(x) dx \leq \lim_{\lambda \rightarrow 0} \sum_{i=1}^n f(\xi_i) \Delta x_i \leq \overline{\int_a^b} f(x) dx$$

这一结论告诉我们，黎曼积分和达布积分是等价的，它们是从不同角度定义同一事物。黎曼积分定义更直观，但达布积分定义更便于计算。



设函数 $f(x)$ 在 $[a, b]$ 上有界，则 f 在 $[a, b]$ 上黎曼可积的充要条件是 f 在 $[a, b]$ 上的不连续点组成的集合测度为零

黎曼积分的局限性

- 限制了可积函数范围³
- 极限积分运算换序的要求过高

$$\lim_{n \rightarrow \infty} \int_0^1 f_n(x) dx \neq \int_0^1 \lim_{n \rightarrow \infty} f_n(x) dx$$

- 黎曼可积函数空间不完备
- 黎曼积分下原函数与导函数满足下列关系式⁴，但实际上存在有界原函数其导函数存在且有界，但不可积

$$\int_a^b f(x) dx = F(b) - F(a)$$

³黎曼积分要求被积函数几乎处处连续，但还有很多处处不连续的函数也能求积分

⁴根据牛顿-莱尼兹公式，要使这一公式成立，原函数必须可导且导函数必须是可积的



设有函数 $D(x)$ 定义在 $[0, 1]$ 范围内，若 $x \in [0, 1]$ 是有理数点则 $D(x)$ 取值为 1 否则取值为 0，称 $D(x)$ 为狄利克雷 (Dirichlet) 函数

$$D(x) = \begin{cases} 1, & x \text{ 为有理数} \\ 0, & x \text{ 为无理数} \end{cases}$$

因为有理数的旁边总有无理数，无理数的旁边也总有有理数，所以无论对 $[0, 1]$ 作何种划分，对任意区间 $[x_{i-1}, x_i]$ ，总有

$$M_i = \sup\{D(x) | x \in [x_{i-1}, x_i]\} = 1, \quad m_i = \inf\{D(x) | x \in [x_{i-1}, x_i]\} = 0$$

因此任意区间的达布上和达布下和有

$$\bar{S}(f, \Delta) = \sum_{i=1}^n M_i \Delta x_i = \sum_{i=1}^n \Delta x_i = 1, \quad \underline{S}(f, \Delta) = \sum_{i=1}^n m_i \Delta x_i = 0$$

可见，在黎曼积分下此函数是不可积的，因为 $\bar{S}(f, \Delta) \neq \underline{S}(f, \Delta)$ 。但实际上，狄利克雷函数在 Lebesgue 积分上是可积的，积分结果等于 0



对称黎曼积分，只有当 $f(x)$ 在 $[a, b]$ 范围内一致连续时才能交换积分和极限的顺序，否则不可以

$$\lim_{n \rightarrow \infty} \int_0^1 f_n(x) dx = \int_0^1 \lim_{n \rightarrow \infty} f_n(x) dx, \quad \text{若 } f(x) \text{ 一致连续}$$

但实际上存在很多函数不是一致连续的，但仍然可交换极限与积分的顺序。例如 $f_n(x) = x^n (0 \leq x \leq 1)$ ，这是一个点收敛而非一致收敛函数，它收敛于

$$\lim_{n \rightarrow \infty} f_n(x) = f(x) = \begin{cases} 0, & 0 \leq x < 1 \\ 1, & x = 1 \end{cases}$$

但是对 $f(x)$ 有

$$\lim_{n \rightarrow \infty} \int_0^1 f_n(x) dx = \lim_{n \rightarrow \infty} \int_0^1 x^n dx = \lim_{n \rightarrow \infty} \frac{x^{n+1}}{n+1} \Big|_0^1 = \lim_{n \rightarrow \infty} \frac{1}{n+1} = 0$$

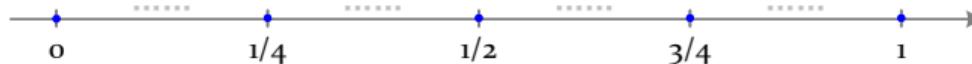
并且有

$$\int_0^1 \lim_{n \rightarrow \infty} f_n(x) dx = \int_0^1 f(x) dx = 0$$

即虽然 $f(x)$ 没有一致连续，但仍然可交换极限与积分的顺序



$[0, 1]$ 范围内二进有理数是 $\{0, 1, \frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{3}{8}, \frac{5}{8}, \frac{7}{8}, \dots\}$, 把这样的数列记作 $\{r_n\}$



定义函数列 $\{f_n(x)\}$

$$f_n(x) = \begin{cases} 1, & x = \{r_n\} \\ 0, & \text{其它} \end{cases}$$

很明显, 对于任意一个 $f_n(x)$ (n 有限), 它都是黎曼可积的, 因为它的不连续点都是有限的。但是该函数列的极限 $f(x) = \lim_{n \rightarrow \infty} f_n(x)$ 是一个类似于狄利克雷函数的函数, 即在每一点都不连续且任意区间内总有二进有理数也有非二进有理数, 因此它的达布上和不等于达布下和, 所以 $f(x)$ 不是黎曼可积的



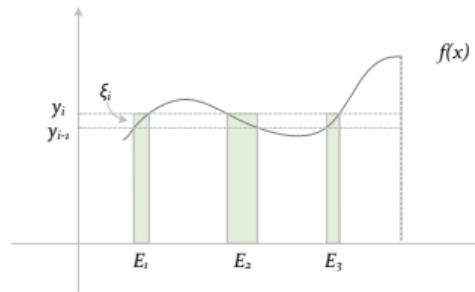
勒贝格积分的基本思想：从值域处作分割，例如第*i*个区间为 $[y_{i-1}, y_i]$ 。设所有取值在该区域内的 x 包括三部分 $\{E_1, E_2, E_3\}$ ，这三部分 x 的总测度为 $|E|$

$$|E| = |E_1| + |E_2| + |E_3|$$

从 $[y_{i-1}, y_i]$ 中任选一个值 ξ_i ，例如可令 $\xi_i = y_i$ ，那么 $f(x)$ 在这些区域围成的面积可近似表示为 S_i ，曲线围成的总面积 S 则有

$$S_i \approx \xi_i |E|, \quad S \approx S_\Delta = \sum_{i=1}^n \xi_i |E_i|$$

若 $\Delta y_i \rightarrow 0$ ，则 $S_\Delta \rightarrow S$ ，极限情况下二者相等





集合内容补充



设 A, B 都是 R^n 中非空点集，则定义 A 与 B 之间的距离为 $d(A, B) = \inf\{d(x, y) | x \in A, y \in B\}$ 。当 A 为单点集 $\{x\}$ 时，称 $d(x, B)$ 为点 x 与点集 B 之间的距离，即 $d(x, B) = d(\{x\}, B) = \inf\{d(x, y) | y \in B\}$

有以下两点注意

- 任意两个非空集合 A, B ，一定有 $d(A, B) \geq 0$
- 若 $A \cap B \neq \emptyset$ 则 $d(A, B) = 0$ ；但 $d(A, B) = 0$ 不一定有 $A \cap B \neq \emptyset$ ⁵

⁵例如 $A = (1, 2)$, $B = (2, 3)$ 时， $A \cap B = \emptyset$ 但 $d(A, B) = 0$



设 $A_1, A_2, A_3, \dots, A_n, \dots$ 是任意一列集合，记作集列 $\{A_n\}$ 。由该集列中属于无限多个集合的那些元素构成的集，称为此集列的上限集，记作 $\overline{\lim}_{n \rightarrow \infty} A_n$ 或 $\limsup_n A_n$ ；而由属于集列中从某个指标 $n_0(x)$ ⁶ 以后所有集 A_n 的那种元素 x 的全体（即附去有限多个集外的所有集都包含的元素）组成的集称为该列集的下限集，记作 $\underline{\lim}_{n \rightarrow \infty} A_n$ 或 $\liminf_n A_n$

$$\overline{\lim}_{n \rightarrow \infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m, \quad \underline{\lim}_{n \rightarrow \infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m$$

显然有

$$\bigcap_{n=1}^{\infty} A_n \subset \underline{\lim}_{n \rightarrow \infty} A_n \subset \overline{\lim}_{n \rightarrow \infty} A_n \subset \bigcup_{n=1}^{\infty} A_n$$

⁶这个指标不是固定的，与 x 有关



若对于集列 $\{A_n\}$, 上限集和下限集相等, 就说集列 $\{A_n\}$ 收敛, 并称 $A = \underline{\lim}_{n \rightarrow \infty} A_n = \bar{\lim}_{n \rightarrow \infty} A_n$ 是集列 $\{A_n\}$ 的极限, 记为 $A = \lim_{n \rightarrow \infty} A_n$

$$\bar{\lim}_{n \rightarrow \infty} A_n = \underline{\lim}_{n \rightarrow \infty} A_n$$



$$\bar{\lim}_{n \rightarrow \infty} A_n \longrightarrow \{ \text{●} , \text{●} \}$$

$$\underline{\lim}_{n \rightarrow \infty} A_n = \emptyset$$



$$\bar{\lim}_{n \rightarrow \infty} A_n = \underline{\lim}_{n \rightarrow \infty} A_n = \{ \text{●} \}$$

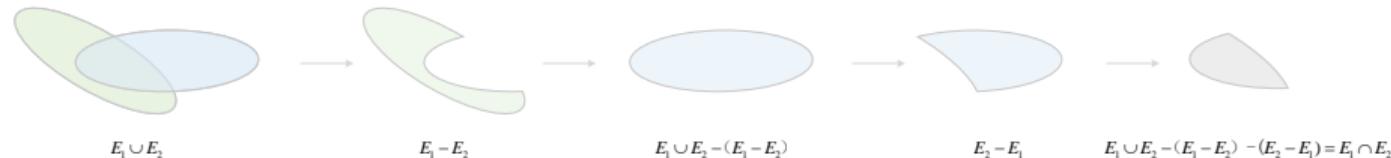


设 X 是给定的集合，它也被称为基本空间。以 X 的某些子集为元素构成的集称为 X 上的集族或集类或集合系，也简称为类。

设 X 是一个集合, R 是 X 上的集族, 如果对任何 $E_1, E_2 \in R$ 都有 $E_1 \cup E_2 \in R, E_1 - E_2 \in R$, 那么就称 R 是 X 上的环。特别地, 如果还有 $X \in R$ 就称 R 是 X 上的代数, 或也称为域⁷

环(或者代数或者域)上不仅对并、差运算封闭,还对交运算封闭,因为交运算可由并、差得到

$$E_1 \cap E_2 = (E_1 \cup E_2) - (E_1 - E_2) - (E_2 - E_1)$$



⁷代数的本质就是集合加集合元素上的操作



设 S 是由集 X 的某些子集所成的集类, 若对任何一列 $E_i \in S (i = 1, 2, 3, \dots)$ 都有 $\bigcup_{i=1}^{\infty} E_i \in S, E_1 - E_2 \in S$, 就称 S 是 X 上的 σ -环。如果 $X \in S$ 就称 S 是 X 上的 σ -代数或 σ -域⁸

对于可数集而言也有

$$\bigcap_{i=1}^{\infty} E_i = \bigcup_{i=1}^{\infty} E_i - \bigcup_{i=1}^{\infty} \left(\bigcup_{j=1}^{\infty} E_j - E_i \right)$$

设 E 是由集 X 的某些子集所成的集类, 那么必定有唯一的 σ -环 (σ -代数) S 使得⁹

- $E \subset S$
- 对于包含 E 的任何 σ -环 (σ -代数) S_1 都成立 $S \subset S_1$

⁸ 代数和 σ -代数的主要区别在于前者是对有限并交差封闭, 后者对可数并交差封闭

⁹ 此定理表示任意 σ -环或 σ -代数一定存在最小 σ -环或 σ -代数, 对于代数或环同样如此



任意多个开集的并集仍然是开集，有限多个开集的交集仍然是开集；任意多个闭集的交集仍是闭集，有限多个闭集的并集仍是闭集¹⁰

设 E 是 R^n 中的点集， $\{G_\lambda\}_{\lambda \in I}$ 是 R^n 中一族开集。如果对任意的 $P \in E$ 存在开集 $G_\lambda \in \{G_\lambda\}_{\lambda \in I}$ ，使 $P \in G_\lambda$ ，则称开集族 $\{G_\lambda\}_{\lambda \in I}$ 是 E 的一个开覆盖，简称覆盖

波雷尔 (Borel) 有限覆盖定理：设 F 是有界闭集， $\{G_\lambda\}_{\lambda \in I}$ 是 F 的任一开覆盖，则 $\{G_\lambda\}_{\lambda \in I}$ 中存在有限多个开集 G_1, G_2, \dots, G_m 同样覆盖 F ¹¹

林德略夫 (Lindelöf) 可数覆盖定理：设 E 是 R^n 中任一点集，一族开集 A 覆盖 E ，则 A 中存在可列多个（或有限多个）开集也覆盖 E

¹⁰ 无限多开集的交有可能是闭集，例如设 $A_n = (-\frac{1}{n}, \frac{1}{n})$ ， $\cap_{n=1}^{\infty} A_n = \{0\}$ ，这是一个闭集；任意多个闭集的并不一定是闭集，例如 $A_n = [-1 + \frac{1}{n}, 2 - \frac{1}{n}]$ ， $\cup_{n=1}^{\infty} A_n = (-1, 2)$ ，这是一个开集

¹¹ 波雷尔有限覆盖定理中 F 必须是“有界”且“闭”的，如果这两个条件不存在的话此定理也会改变



在实变函数或测度论中，波雷尔集（Borel set）是由实数集的开集通过交、并、补运算后生成的集合，它是由拓扑空间中的开集生成的最小 σ -代数。波雷尔集的构造过程

- ① 以所有的开集为构造基础
- ② 若一个集合在 $\mathcal{B}(R)$ 中，那么它的补集也在 $\mathcal{B}(R)$ 中
- ③ 若集列 $\{A_n\}$ 在 $\mathcal{B}(R)$ 中，那么其可数并 $\bigcup_{n=1}^{\infty} A_n$ 和可数交 $\bigcap_{n=1}^{\infty} A_n$ 也在 $\mathcal{B}(R)$ 中

所有开集、闭集、区间（包括 (a, b) , $(a, b]$, $[a, b)$, $[a, b]$ ）都是波雷尔集；它对补、可数并、可数交都具有封闭性。可测集、概率空间、随机变量、事件空间等等都是定义在波雷尔集之上的。波雷尔集实际上是非常“大”的，可以认为它包含了几乎所有的“好”集合¹²

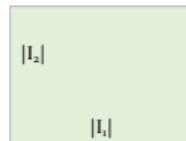
¹²如果在论文或资料中遇到了波雷尔集，只要要求不太严格，可直接忽视掉“波雷尔”三个字，直接把集合想像成最一般的集合



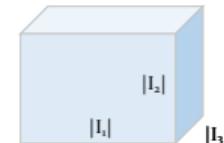
测度



测度：一维长度、二维面积、三维体积向 n 维空间集合的推广及抽象



$$\text{Area} = |I_1| * |I_2|$$



$$\text{Volume} = |I_1| * |I_2| * |I_3|$$

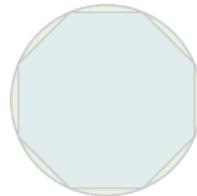
设有点集 $I = \{x|x = (x_1, x_2, \dots, x_n), a_i < x_i < b_i, i = 1, 2, \dots, n\}$ 为 R^n 中的开区间。类似地，若 $a_i < x_i \leq b_i$ 则称为左半开区间， $a_i \leq x_i < b_i$ 为右半开区间， $a_i \leq x_i \leq b_i$ 为闭区间。若 I 为区间¹³，则称 $|I|$ 为区间 I 的体积

$$|I| = \prod_{i=1}^n (b_i - a_i) = (b_1 - a_1) \times (b_2 - a_2) \times \cdots \times (b_n - a_n)$$

¹³无论是开区间、闭区间还是半开半闭区间



体积的概念仍然比较特殊¹⁴，还要进一步将体积的概念推广到一般点集上。在圆内接正多边形，此多边形面积总是小于圆的面积；在圆外接正多边形，此正多边形的面积总是大于圆的面积。当正 n 边形的 n 取值逐渐增大时，内外接正多边形的面积越来越接近于圆，极限情况下内外多边形面积相等且等于圆的面积。



圆内接正多边形



圆外接正多边形

结合三维空间经验，对于抽象的测度它也应具备如下性质

- 空集的测度等于0
- 任意集合的测度非负
- 两个不相交集合的测度等于两个集合各自测度的和

¹⁴ 它要求各个维度的取值在连续区间内且各个维度实际是相互独立的



定义 $\hat{R} = R \cup \{+\infty\} \cup \{-\infty\}$ 。设 E 是一个集族，若 μ 是从 E 到 \hat{R} 的映射，就说 μ 是集函数¹⁵。

设 R 是由集 X 的某些子集所构成的环， μ 是 R 上的集函数，如果 μ 具有如下性质则称集函数 μ 为环 R 上的测度， $\mu(E)$ 为集合 E 的测度

- $\mu(\emptyset) = 0$
- 非负性：对任何 $E \in R$ 有 $\mu(E) \geq 0$
- 可列可加性：对任意一列 $E_i \in R (i = 1, 2, 3, \dots)$ ，如果 $E_i \cap E_j = \emptyset (i \neq j)$ 且 $\bigcup_{i=1}^{\infty} E_i \in R$ 则必定有

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i)$$

可简单的令所有非空集合 $\mu(E_i) = \infty$ ，此时仍满足测度定义。但这样的测度是无意义的、平凡的。一般所指的测度都是非负、可列可加集函数，即至少有一部分集合的测度是有限的¹⁶。

¹⁵ 集函数以集合为自变量，以实数或 $+\infty$ 、 $-\infty$ 为值域

¹⁶ $\mu(E) = E$ 中元素的个数 ($E \in R$)，函数就是一个合法的测度



若 μ 是环 R 上的测度，那么它有如下一些性质

- 有限可加性：如果 $E_1, E_2, \dots, E_n \in R$ 且这些集合两两不相交，那么 $\mu\left(\bigcup_{i=1}^n E_i\right) = \sum_{i=1}^n \mu(E_i)$
- 单调性：如果 $E_1, E_2 \in R$ 且 $E_1 \subset E_2$ ，那么 $\mu(E_1) \leq \mu(E_2)$
- 可减性：如果 $E_1, E_2 \in R$ 且 $E_1 \subset E_2$ ，又如果 $\mu(E_1) < \infty$ 那么 $\mu(E_2 - E_1) = \mu(E_2) - \mu(E_1)$
- 次可列可加性：如果 $E_n (n = 1, 2, 3, \dots)$ 及 E 都属于 R 且 $E \subset \bigcup_{i=1}^{\infty} E_i$ ，那么 $\mu(E) \leq \sum_{i=1}^{\infty} \mu(E_i)$
- 如果 $E_n \in R (n = 1, 2, 3, \dots)$ 且 $E_1 \subset E_2 \subset E_3 \subset \dots$ ， $\bigcup_{n=1}^{\infty} E_n \in R$ ，那么 $\mu\left(\bigcup_{n=1}^{\infty} E_n\right) = \lim_{n \rightarrow \infty} \mu(E_n)$
- 如果 $E_n \in R (n = 1, 2, 3, \dots)$ 且 $E_1 \supset E_2 \supset E_3 \supset \dots$ ， $\bigcap_{n=1}^{\infty} E_n \in R$ ，而且至少有一个 E_n 使 $\mu(E_n) < \infty$ ，那么 $\mu\left(\bigcap_{n=1}^{\infty} E_n\right) = \lim_{n \rightarrow \infty} \mu(E_n)$



设 X 是基本空间, R 是 X 的环。下面引进一个包含 R 的 σ -环: $H(R)$ 表示 X 中能用 R 内一列元素加以覆盖的子集全体所构成的类¹⁷, 即

$$H(R) = \{E | E \subset X, \text{存在 } E_i \in R (i = 1, 2, 3, \dots) \text{ 使 } E \subset \bigcup_{i=1}^{\infty} E_i\}$$

$H(R)$ 实际上是 R 的一个外覆盖

如果 μ 是环 R 上的测度, 在 $H(R)$ 上作集函数 μ^* : 当 $E \in H(R)$ 时有下式, μ^* 称为由测度 μ 所引出的外测度¹⁸

$$\mu^*(E) = \inf \left\{ \sum_{i=1}^{\infty} \mu(E_i) | E_i \in R \text{ 且 } E \subset \bigcup_{i=1}^{\infty} E_i \right\}$$

¹⁷ 环和 σ -环的区别就在于前者只对集合的有限并满足封闭性, 后者对可数并满足封闭性, 此处定义就用到了 $\bigcup_{i=1}^{\infty} E_i$, 所以它是 σ -环

¹⁸ 外测度表示用一组集合覆盖另一个集合, 这一组集合测度和的下确界就是被覆盖集合的外测度



设 E 是 R^n 中任一点集，对每一覆盖 E 的可数个开区间 $I_1, I_2, \dots, I_n, \dots$ ， $\bigcup_{n=1}^{\infty} I_n \supset E$ ，作出它的体积总和 $\mu = \sum_{n=1}^{\infty} |I_n|$ ，显然 μ 是一个确定的非负数（ μ 也可以等于 $+\infty$ ）。所有这一切 μ 组成一个下方有界的数集（比如，0 就是它的一个下界），因此有惟一下确界，这个下确界（由 E 完全确定）称为 E 的勒贝格外测度，简称为 L 外测度或外测度，记为 $m^*(E)$ ¹⁹

$$m^*(E) = \inf \left\{ \sum_{i=1}^{\infty} |I_i| \mid E \subset \bigcup_{i=1}^{\infty} I_i, I_i \text{ 都是开区间} \right\}$$

外测度的基本性质

- ① $m^*(E) \geq 0$ ；当 E 为空集时 $m^*(E) = 0$
- ② 若 $A \subset B$ 则 $m^*(A) \leq m^*(B)$
- ③ $m^*\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} m^*(A_i)$
- ④ 若 A, B 之间的距离 $d(A, B) > 0$ ，则 $m^*(A \cup B) = m^*(A) + m^*(B)$

¹⁹ 许凤，《实变函数论》：勒贝格外测度是在勒贝格可测集上定义的外测度，后面会介绍

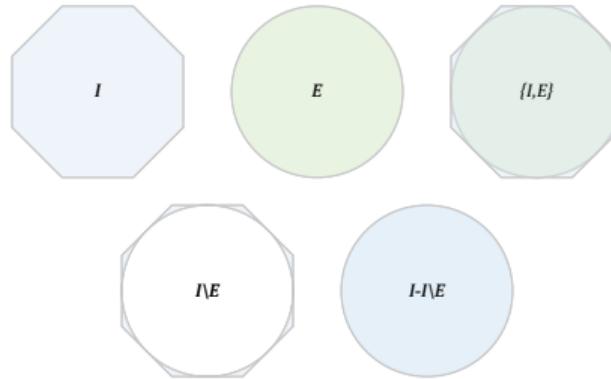


设 E 是 R^n 中有界点集²⁰, I 是任一包含 E 的开区间, 则称 $|I| - m^*(I \setminus E)$ 为 E 的内测度, 记为 $m_*(E)$

内测度的基本性质

- $m_*(E) \geq 0$
- $m_*(E) \leq m^*(E)$
- 内测度可变形为

$$\begin{aligned} m_*(E) &= |I| - m^*(I \setminus E) \\ &= |I| - \inf_{\forall E \subset \cup_{i=1}^{\infty} I_i} \sum_{i=1}^{\infty} |I_i| \\ &= \sup_{\forall E \subset \cup_{i=1}^{\infty} I_i} \{ |I| - \sum_{i=1}^{\infty} |I_i| \} \end{aligned}$$



²⁰注意, 这里内测度只针对有界集定义



我们希望当 $A \cap B = \emptyset$ 时有 $m^*(A \cup B) = m^*(A) + m^*(B)$ ，但是外测度的第4条性质要求 $d(A, B) > 0$ 时它们并的外测度才等于外测的和。特别注意，这里 $d(A, B) > 0$ 的要求不能替换为 $A \cap B = \emptyset$ ²¹。这跟我们的经验和预期是不相符的，这说明我们对测度的定义还不够好。

经研究发现，破坏外测度可加性的集合同样也破坏内测度可加性。这样的集合其测度从外逼近（外测度）和从内逼近（内测度）所得结果是不一样的，即总是有 $m^*(E) > m_*(E)$ 。称这样的集合为不可测集合。将这样的集合剔除后剩下的就都是可测集（它们测度的可加性符合我们的预期）。

²¹ 内测度也有类似的第4条性质：两个集合的距离大于0，那么这两个集合并的内测度等于它们内测度的和；它也面临同样的问题



设 E 是 R^n 中的**有界集合**，如果 $m^*(E) = m_*(E)$ ，则称 E 为**有界可测集**。此时称 E 的外（内）测度值为 E 的勒贝格测度，简称为 E 的测度，记作 $m(E)$ 即

$$m(E) = m^*(E) = m_*(E)$$

设 E 是 R^n 中的无界点集，若对任何开区间 I ， $E \cap I$ 都是有界可测集²²，则称 E 为无界可测集且 $m(E) = m^*(E)$

有界可测集和无界可测集统称为可测集。因为测度是用外测度定义的，所以外测度的性质测度全都具备。

²²因为开区间 I 是有界的，所以 $E \cap I$ 也是有界的，可利用对有界可测集的定义来定义无界可测集



前述可测集定义是由勒贝格给出的，它存在至少两个问题

- 有界集可测集和无界可测集的定义不一样
- 同时用到了外测度和内测度

卡拉泰屋独利（Caratheodory）给出了统一的可测集定义且只用到外测度²³：设 E 是 R^n 中的点集，如果对任意点集 T 都有下式，则称 E 是勒贝格可测集，简称可测集。此时称 E 的外测度 $m^*(E)$ 为 E 的测度，记为 $m(E)$

$$m^*(T) = m^*(T \cap E) + m^*(T \cap \complement E)$$

²³也称卡氏判定条件



点集 E 可测的充要条件

- ① $\complement E$ 为可测集
- ② 对任意 $A \subset E, B \subset \complement E$ 恒有 $m^*(A \cup B) = m^*(A) + m^*(B)$

设 E_1, E_2 为可测集，则

- ① $E_1 \cup E_2$ 为可测集，特别地当 $E_1 \cap E_2 = \emptyset$ 时，对任意点集 T 总有

$$m^*[T \cap (E_1 \cup E_2)] = m^*(T \cap E_1) + m^*(T \cap E_2)$$

- ② $E_1 \cap E_2$ 为可测集
- ③ $E_1 \setminus E_2$ 为可测集



可测集的其它一些性质

- 测度的可列²⁴可加性：设 $E_i(i = 1, 2, \dots)$ 都是可测集且 $E_i \cap E_j = \emptyset(i \neq j)$ ，则 $\bigcup_{i=1}^{\infty} E_i$ 可测，并且有

$$m\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} m(E_i)$$

- 设 $E_i(i = 1, 2, 3, \dots)$ 都是可测集，则 $\bigcup_{i=1}^{\infty} E_i$ 也是可测集
- 设 $E_i(i = 1, 2, 3, \dots)$ 都是可测集，则 $\bigcap_{i=1}^{\infty} E_i$ 也是可测集

常见的可测集

- 任何区间 I 都是可测集且 $m(I) = |I|$ ，任何开集、闭集都是可测集，任何波雷尔集都是可测集
- R^n 中任何可测集都可表示为至多可列个互不相交的有界可测集的并

²⁴ 可列就是可数



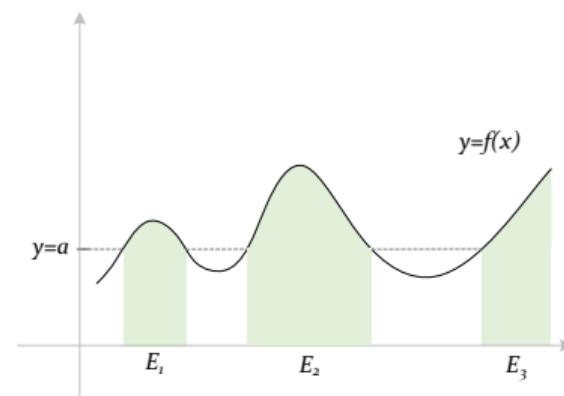
可测函数



设 $f(x)$ 是可测集 E 上定义的函数，如果对任何实数 a ，集合 $E[f(x) > a]$ 都是可测集，则称 $f(x)$ 是 E 上的勒贝格可测函数，简称可测函数

下列命题等价

- $f(x)$ 是 E 上的可测函数
- 对任何实数 a ， $E[f(x) \geq a]$ 是可测集
- 对任何实数 a ， $E[f(x) < a]$ 是可测集
- 对任何实数 a ， $E[f(x) \leq a]$ 是可测集





已知 $R^1 = (0, 1)$ 范围内的狄利克雷函数 $D(x)$, 在有理数上取值为1、无理数上取值为0

$$D(x) = \begin{cases} 1, & x \text{ 为有理数} \\ 0, & x \text{ 为无理数} \end{cases}$$

对任意实数 a 有

$$E[D(x) > a] = \begin{cases} \emptyset, & a \geq 1 \\ Q^{25}, & 0 \leq a < 1 \\ R^1, & a < 0 \end{cases}$$

因为 \emptyset, Q 和 R^1 都是可测集²⁶, 所以 $D(x)$ 是可测函数。有结论: 连续函数一定可测, 可测函数不一定连续, 甚至可以是处处不连续。

²⁵ Q 表示 $(0, 1)$ 范围内的有理数

²⁶ $(0, 1)$ 范围内有理数测度为0 无理数测度为1



可测函数的一些性质

- 若 $f(x)$ 是可测集 E 上的可测函数，而 $E_1 \subset E$ 为 E 的可测子集，则 $f(x)$ 在 E_1 上也是可测函数
- 若 $f(x)$ 在每个可测集 $E_i (i = 1, 2, \dots, m, m \text{有限或取} +\infty)$ 上都可测，则 $f(x)$ 在 $E = \bigcup_{i=1}^{\infty} E_i$ 上也可测
- 若 $\{f_n(x)\}$ 是可测集 E 上的可测函数列，则 $\sup_n\{f_n(x)\}$ 和 $\inf_n\{f_n(x)\}$ 都是 E 上可测函数

设有一个关于集合 E 上点 x 的命题 $P(x)$ ，如果有 E 的子集 N , $m(N) = 0$ (测度为0)，使 $P(x)$ 在 $E \setminus N$ 上恒成立，则称 $P(x)$ 在 E 上几乎处处成立，记作 $P(x)$ 成立a.e.于 E ²⁷

²⁷a.e.表示almost everywhere



设 $f(x)$ 和 $g(x)$ 都是 E 上的可测函数，则

- 对任何实数 c , $cf(x)$ 是 E 上的可测函数
- 当 $f(x) + g(x)$ 在 E 上几乎处处有意义时, $f(x) + g(x)$ 是 E 上的可测函数
- 当 $f(x) \cdot g(x)$ 在 E 上几乎处处有意义时, $f(x) \cdot g(x)$ 是 E 上的可测函数
- 当 $\frac{f(x)}{g(x)}$ 在 E 上几乎处处有意义时, $\frac{f(x)}{g(x)}$ 是 E 上的可测函数

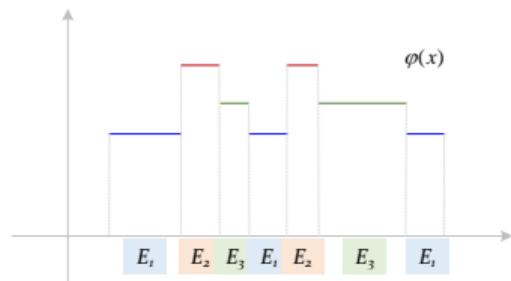
设 $\{f_n(x)\}$ 是 E 上的可测函数列并且 $\lim_{n \rightarrow \infty} f_n(x) = f(x)$ a.e.于 E , 则 $f(x)$ 也是 E 上的可测函数



设 E 是可测集, $\phi(x)$ 是 E 上的函数, 如果 E 可分解为有限个互不相交的可测子集的并: $E = \bigcup_{i=1}^n E_i$, 使 $\phi(x)$ 在每个 E_i 上都恒取某个常数值 C_i , 则称 $\phi(x)$ 是 E 上的简单函数

简单函数是可测函数

- 狄利克雷函数是简单函数
- 简单函数的和、差、积仍是简单函数
- 若 $f(x)$ 是 E 上可测函数, 则存在 E 上的简单函数列 $\{\phi_n(x)\}$,
使 $f(x) = \lim_{n \rightarrow \infty} \phi_n(x)$





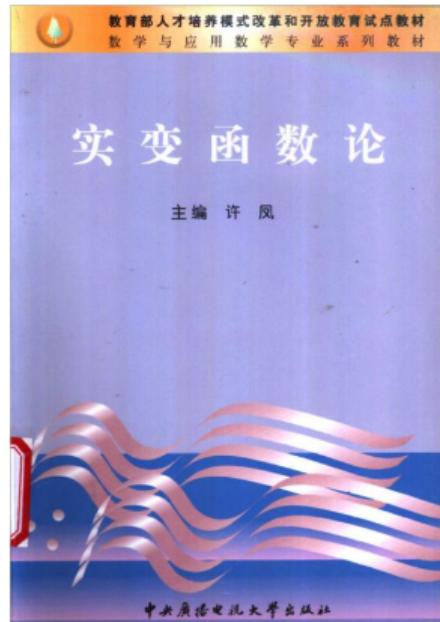
设 $\{f_n(x)\}$ 是点集 E 上一列几乎处处取有限值的可测函数, $f(x)$ 是 E 上几乎处处有限的可测函数, 如果对任意的 $\sigma > 0$ 有 $\lim_{n \rightarrow \infty} m(E[|f_n(x) - f(x)| \leq \sigma]) = 0$, 则称函数列 $\{f_n(x)\}$ 在 E 上依测度收敛于 $f(x)$, 记为 $f_n(x) \Rightarrow f(x)$

或表述为: 对任意 $\varepsilon > 0$ 和任意 $\sigma > 0$, 总存在自然数 N , 当 $n \geq N$ 时就有 $m(E[|f_n(x) - f(x)| \geq \sigma]) < \varepsilon$ 。在概率论中, 用概率函数 $p(x)$ 替代这里的测度函数 $m(E)$ 就得到了依概率收敛。

函数列的依测度收敛和几乎处处收敛的关系²⁸

- 里斯定理 (F. Riesz): 设在 E 上可测函数列 $\{f_n(x)\}$ 依测度收敛于 $f(x)$, 则存在 $\{f_n(x)\}$ 的子列 $\{f_{n_k}(x)\}$ 使得 $\lim_{k \rightarrow \infty} f_{n_k}(x) = f(x)$ a.e. 于 E
- 设函数列 $\{f_n(x)\}$ 在 E 上 a.e. 于可测函数 $f(x)$, 若 $m(E) < \infty$, 那么 $\{f_n(x)\}$ 在 E 上必然依测度收敛于 $f(x)$

²⁸ 在一般情况下, 此二者互不包含, 但二者又有很紧密的联系





概率与统计

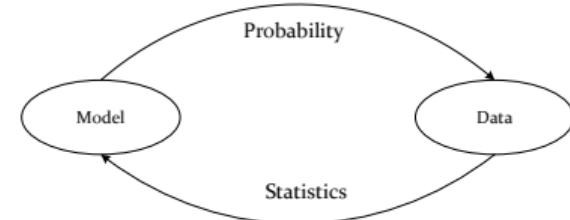


从机器学习的视角来看，概率与统计是两个正好相反的过程

- 概率：已知数据的分布信息，求 x 的概率（概率密度）或对 x 做预测
- 统计：基于已有观察数据推断总体的特性及规律

机器学习与概率统计

- 训练阶段对应统计，预测阶段对应概率
- 频率派和贝叶斯派之争主要发生在数理统计领域
- 推断（inference）在机器学习和概率统计中同词不同意





概率定义



自然界及社会上有一些现象具有确定性，还有一些现象在大量重复试验或观察下在总体上呈现特定规律，但每个具体事件又表现出随机性。我们把这种总体上的特定规律称为统计规律，把每一个具体的试验称为随机试验。

将随机试验记为 E ，虽然每次试验的结果不可预知，但试验有可能出现的所有结果是确定的，将所有结果组成的集合称为样本空间记为 $\Omega = \{\omega\}$ ， Ω 中的每个元素 ω （即随机试验的每种结果）称为样本点

- E : 抛一枚硬币，观察正面 H 、反面 T 出现的情况， $\Omega = \{H, T\}$
- E : 抛两枚硬币，观察正面 H 、反面 T 出现的情况， $\Omega = \{HH, HT, TH, TT\}$
- E : 抛有六面的骰子，观察得到的点数， $\Omega = \{1, 2, 3, 4, 5, 6\}$



很多时候在一次试验中，我们关心的不仅是某一种结果是否会出现，更多的是某一类结果是否会出现。例如抛投骰子出现的是奇数还是偶数？航班起飞延误时间是否超过3个小时以上？等等。将一类结果构成的样本点集合称为随机事件，简称为事件。如果某事件中只有一个样本点则称为基本事件；因为 Ω 中包含所有可能的试验结果，所以又称为必然事件；空集 \emptyset 不含任何事件，称为不可能事件。

抛掷一枚均匀的骰子的样本空间为 $\Omega = \{1, 2, 3, 4, 5, 6\}$ ，下面是一些可能的随机事件

- 得到6点： $A = \{6\}$
- 出现偶数点： $A = \{2, 4, 6\}$
- 点数小于等于3： $A = \{1, 2, 3\}$
- 点数大于6： $A = \{\emptyset\}$

若试验结果为 ω , $\omega \in A$ 则说事件A发生，否则说事件A未发生



一种定义事件发生概率的方法是利用事件发生的频率：设有样本空间 Ω ，做重复试验记录事件A发生的次数为 $N(A)$ ，那么事件A发生的概率为

$$P(A) = \lim_{N \rightarrow \infty} \frac{N(A)}{N}$$

这种定义方式虽然简单直观，但也存在严重缺陷

- 为什么可以用频率的极限来定义概率？如何证明它的合理性？
- 如何确保 $N \rightarrow \infty$ 时 $\lim \frac{N(A)}{N}$ 一定会收敛并且总是收敛到同一个值？



支持这种做法的人认为这是一种假设或公理（axiom），整个概率体系就建立在这样一个前提假设上。但还有很多人认为这种假设过去直接和强硬，我们应给定一些更基本、更简单、更一般的假设（或公理），然后去证明在这些公理下频率的极限的确趋于一个常数，这就是现代概率论的公理化方法。它的三个基本公理是

- ① 非负性：对任意事件 A 其概率值满足 $0 \leq P(A) \leq 1$
- ② 规范性：对于样本空间其概率值为 $P(\Omega) = 1$
- ③ 可列可加性：对任意不相容事件 A_1, A_2, \dots （即任意两事件不相交），有 $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$

回顾实变函数中的测度：设有基本空间 X ，由 X 的子集形成一个 σ -域 \mathcal{F} ，再在 \mathcal{F} 上定义一个测度 μ ，由此得到的三元组 (X, \mathcal{F}, μ) 称为测度空间；如果 $X = \Omega$ 并且令 $\mu = P$ 满足前面的三个公理，则 (Ω, \mathcal{F}, P) 三元组称为概率空间， P 称为概率测度， \mathcal{F} 中的元素就是事件， $P(A)$ 表示事件 A 的概率测度值，简称概率值²⁹

²⁹ 回忆前面测度所具有的性质，包括非负性和可列可加性，而概率化公理中也有非负性和可列可加性，概率空间只是多了一个 $P(\Omega) = 1$ 。



因为 σ -域本身就对交、并、补运算封闭，所以概率空间也对交、并、补封闭，所以概率运算满足如下结果

$$P(\bar{A}) = 1 - P(A), \quad P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

不同的概率空间可以定义不同的概率分布，设某硬币正面朝上的概率为0.5、反面朝上的概率为0.5，这是一个伯努利分布它对应的概率空间为

$$\{\Omega, \mathcal{F}, P\} : \Omega = \{H, T\}, \mathcal{F} = \{\{H, T\}, \emptyset, \{H\}, \{T\}\}, P = \{P(\{H\}) = 0.5, P(\{T\}) = 0.5\}$$

若另一个硬币正面朝上概率为0.3、反面朝上的概率为0.7，它对应另一个概率空间

$$\{\Omega, \mathcal{F}, P\} : \Omega = \{H, T\}, \mathcal{F} = \{\{H, T\}, \emptyset, \{H\}, \{T\}\}, P = \{P(\{H\}) = 0.3, P(\{T\}) = 0.7\}$$



不同的概率分布也可能来源于相同的概率空间：设有两个高斯分布 $N(x_1|\mu_1, \sigma_1^2), N(x_2|\mu_2, \sigma_2^2)$ ，假设有概率空间 $\{\Omega, \mathcal{F}, P\}$

$$\Omega = \{-\infty, +\infty\}, \quad \mathcal{F} = \mathcal{B}(\mathbb{R}), \quad P = \Phi(A) = \int_{-\infty}^{\beta} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

很明显此概率空间定义了一个标准正态分布 ε ，只需要对 ε 做线性映射就能得到 x_1, x_2 ，映射后它们仍对应相同的样本空间、事件空间、测度

$$x_1 = \mu_1 + \sigma_1 \varepsilon, \quad x_2 = \mu_2 + \sigma_2 \varepsilon$$

随机变量只是将 $\Omega \rightarrow \mathbb{R}$ ，对任意的 $x_i \sim (\alpha, \beta)$ 求其概率时应首先求 $E[\alpha < x_i < \beta]$ ，再求 $P(E[\alpha < x_i < \beta])$ ³⁰

³⁰ 也就是先基于 $E[\alpha < x_i < \beta]$ 得到对应的 $\varepsilon_1 < \varepsilon < \varepsilon_2$ ，再求 $P(\varepsilon_1 < \varepsilon < \varepsilon_2)$ ，同样的 $(\varepsilon_1, \varepsilon_2)$ 必定有相同的概率值



现在还有一个很大的问题：基于概率公理化如何证明或推导出 $P(A) = \lim_{n \rightarrow \infty} \frac{N(A)}{N}$ 。这需要用到极限定理，主要包括大数定律和中心极限定理。

本系列极限定理主要介绍以下内容

- ① 随机变量序列的两种收敛方式：依概率收敛和依分布收敛
- ② 两个不等式：马尔可夫不等式、切比雪夫不等式
- ③ 三条弱大数定律：切比雪夫大数定律、辛钦大数定律、伯努利大数定律，一条强大数定律
- ④ 两条中心极限定理：林德伯格-列维中心极限定理、棣莫弗-拉普拉斯中心极限定理³¹

³¹ Lindberg-Levi 也称独立同分布中心极限定理，就是我们一般所说的中心极限定理；De Moivre-Laplace 是局部中心极限定理。



设有随机变量序列 X_1, X_2, \dots , X 是一个随机变量或常数, 若对任意的 $\varepsilon > 0$ 有下式成立, 则称随机变量序列 $\{X_n\}$ 依概率收敛于 X ,

$$\lim_{n \rightarrow \infty} P\{|X_n - X| \geq \varepsilon\} = 0 \quad \text{或} \quad \lim_{n \rightarrow \infty} P\{|X_n - X| < \varepsilon\} = 1$$

也可记为

$$X_n \xrightarrow{P} X \quad \text{或} \quad \lim_{n \rightarrow \infty} P\{|X_n - X| \geq \varepsilon\} = 0$$

当 n 很大时, X_n 与 X 出现较大偏差的可能性很小, 或者说 X_n 与 X 有很大把握充分接近。这种“很大把握”大到发生的概率为 1



疑问：[发生概率为1不就是一定发生么？跟一般意义上的收敛有什么区别？](#)

回答：对于离散型事件，概率为1的确相当于一定发生，但对于连续型事件，发生概率为1并不代表一定发生，发生概率为0也并不代表一定不发生。严格来说 $P(A) = 1$ 这里的 P 是概率测度，若样本空间为 Ω （ Ω 一定是必然事件，因为它包含所有可能的情况）， $P(A) = 1$ 并不代表 $A = \Omega$ ，它只表示 A 和 Ω 有相同的测度

例如设 $L = (0, 1)$, $x_0 = \frac{1}{2}$, 从 L 中任取一点 x , 恰好有 $x = x_0$ 的概率为 $p(x = x_0) = \frac{1}{\infty} = 0$, 但这并不代表 x 不可能等于 x_0 ; 反过来, 所取 x 不等于 x_0 的概率为 $p(x \neq x_0) = \frac{L \setminus \{x_0\}}{L} = 1$, 虽然 $x \neq x_0$ 的概率为1但也并不代表会一定发生



设有随机变量序列 X_1, X_2, \dots , X 为一个随机变量, $F_n(x)$ 和 $F(x)$ 分别为 X_n 和 X 的分布函数, 如果在 $F(x)$ 连续点上均有下式成立, 则称 X_1, X_2, \dots 随机变量序列依分布收敛于 X

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

通常记为

$$X_n \xrightarrow{L} X \quad \text{或者} \quad X_n \xrightarrow{d} X$$

弱大数定律都是依概率收敛, 中心极限定理都是依分布收敛, 强大数定律是几乎必然收敛³²

³² 几乎必然收敛强于依概率收敛, 它蕴含了依概率收敛



设 X 为取非负值的随机变量，则对于任何常数 $a > 0$ ，下列不等式成立称为马尔可夫不等式

$$P(X \geq a) \leq \frac{E[X]}{a}$$

证明：对任意 $a > 0$ 令

$$I = \begin{cases} 1 & , \text{ 若 } X \geq a \\ 0 & , \text{ 其它} \end{cases}$$

若 $X \geq a$ 则 $\frac{X}{a} \geq 1$ ，所以 $I \leq \frac{X}{a}$ ；若 $X < a$ 则 $\frac{X}{a} = 0$ 且 $I = 0$ ，所以仍有 $I \leq \frac{X}{a}$ ，所以总有 $I \leq \frac{X}{a}$ 。在不等式两边同时求期望得

$$E[I] \leq \frac{1}{a} E[X]$$

又因为

$$E[I] = \{I = 0\}P\{I = 0\} + \{I = 1\}P\{I = 1\} = P(X \geq a)$$

马尔可夫不等式告诉我们：任意随机变量，只知道分布的期望值就能得到随机变量取值概率的上限（或下限）³³。

³³ 马尔可夫不等式仅适用于非负随机变量



设 X 是一随机变量，均值为 μ 方差为 σ^2 且都是有限的，则对任何 $k > 0$ 下列不等式成立称为切比雪夫不等式

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

证明：因为 $(X - \mu)^2$ 为非负随机变量，利用马尔可夫不等式 ($a = k^2$) 有

$$P((X - \mu)^2 \geq k^2) \leq \frac{E[(X - \mu)^2]}{k^2}$$

因为 $(X - \mu)^2 \geq k^2$ 等价于 $|X - \mu| \geq k$ ，所以上式等价于

$$P(|X - \mu| \geq k) \leq \frac{E[(X - \mu)^2]}{k^2} = \frac{\sigma^2}{k^2}$$

切比雪夫不等式告诉我们：任意随机变量，知道分布的期望方差就能得到随机变量取值概率的上限（或下限）³⁴。

³⁴但由切比雪夫不等式得到的上限往往离概率的真实取值差距较大，它更多的用于理论证明



设有随机变量序列 X_1, X_2, \dots 互不相关³⁵，若存在常数 c 使得 $D(X_i) = \sigma_i^2 \leq c < +\infty (i = 1, 2, \dots)$ ，则对任意 $\varepsilon > 0$ 有

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E(X_i)\right| < \varepsilon\right) = 1$$

证明：因为随机变量间两两不相关，所以根据期望和方差的性质有³⁶

$$E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E(X_i), \quad D\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n D(X_i) \leq \frac{c}{n}$$

根据前面的切比雪夫不等式有，对任意 $\varepsilon > 0$ 当 $n \rightarrow \infty$ 时有

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E(X_i)\right| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} D\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \leq \frac{c}{n\varepsilon^2} \rightarrow 0$$

³⁵ 切比雪夫大数定律只要求随机变量间互不相关，不要求它们独立同分布；互不相关通常指随机变量间线性无关

³⁶ 对任意两个随机变量都有 $E(X \pm Y) = E(X) \pm E(Y)$ ；对于两个线性无关的变量有 $D(X \pm Y) = D(X) + D(Y)$ ； $D(kX) = k^2 D(X)$



设有随机变量序列 X_1, X_2, \dots 相互独立同分布，若 $E[X_i] = \mu < +\infty$, $D(X_i) = \sigma^2 < +\infty (i = 1, 2, \dots)$ ，则对任意 $\varepsilon > 0$ 有

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| < \varepsilon\right) = 1$$

很明显，只需要令切比雪夫大数定律中所有随机变量有相同的期望 μ ，就能有 $\mu = \frac{1}{n} \sum_{i=1}^n E[X_i]$ ，对切比雪夫不等式做简单的替换就能得到辛钦大数定律。辛钦大数定律是切比雪夫大数定律的特殊情况，它不仅要求随机变量序列的期望存在且有限，还要求变量独立同分布³⁷。

³⁷ 最开始要求所有随机变量独立同分布且期望、方差均存在，因此该大数定律也被称为独立同分布大数定律；后来辛钦进一步证明仅期望存在、方差不存在时上述结论仍成立，因此该大数定律又称为辛钦大数定律



设有随机变量序列 X_1, X_2, \dots 相互独立同分布，并且 $X_i \sim B(1, p)$ ($i = 1, 2, \dots$)³⁸，则对任意 $\varepsilon > 0$ 都有

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - p\right| < \varepsilon\right) = 1$$

很明显，伯努利大数定律是辛钦大数定律的特例，它进一步要求随机变量服从伯努利分布。因为伯努利分布期望为 p ，所以 $\mu = \frac{1}{n} \sum_{i=1}^n E[X_i] = p$

若 X_1, X_2, \dots 是相互独立同分布的随机变量，并且 $E[X^k]$ 和 $D(X^k)$ 存在，在独立同分布大数定律作用下一定有

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i^k - E[X^k]\right| < \varepsilon\right) = 1$$

这正是后面要介绍的矩估计法的思想来源³⁹

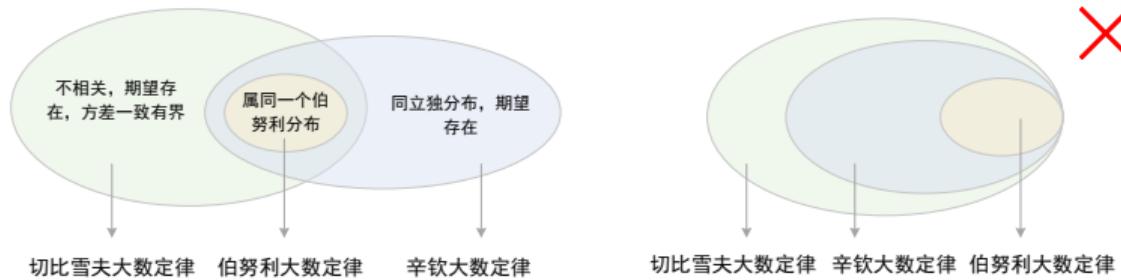
³⁸ $B(1, p)$ 表示伯努利分布，即只有事件发生不发生两种情况，发生的概率为 p

³⁹ 注意，若大数定律对 X 成立但不一定对任意 $g(X)$ 成立，需要 $g(X)$ 满足一些特定的性质，例如确定变量间仍是独立同分布的，不改变期望、方差的存在、一致有界性等



以上三条大数定律都属于弱大数定律⁴⁰

- 三条大数定律都体现的是 n 充分大时，事件发生的频率将依概率收敛到概率
- 证明了概率存在的客观性，也告诉我们在大量实验下频率的稳定性
- 切比雪夫只要求随机变量不相关，辛钦和伯努利要求随机变量独立同分布
- 切比雪夫要求随机变量期望存在方差一致有界，辛钦只要求期望存在，伯努利要求服从伯努利分布⁴¹



⁴⁰现实生活中用多次测量的均值作为物品重量测量值，其理论基础就是辛钦大数定律：多次测量时可用测量结果的算术平均值作为期望的近似；三条大数定律是从一般到特殊的关系，因此在历史上是先有伯努利大数定律，再有辛钦大数定律，最后才有切比雪夫大数定律

⁴¹伯努利分布的期望方差一定存在



列维-林德伯格中心极限定理：设随机变量序列 X_1, X_2, \dots 独立同分布， $E[X_i] = \mu, D(X_i) = \sigma^2$ 且 $0 < \sigma^2 < +\infty$ ，则对任意实数 a 有⁴²

$$\lim_{n \rightarrow \infty} P\left(\frac{\frac{1}{n} \sum_{i=1}^n X_i - \mu}{\sigma/\sqrt{n}} \leq a\right) = \Phi(a)$$

$\Phi(a)$ 前面介绍过，它表示标准正态分布函数。中心极限定理。

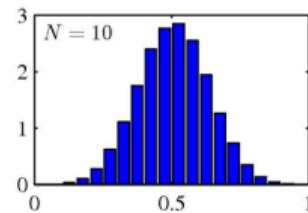
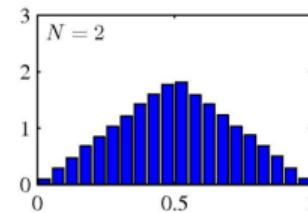
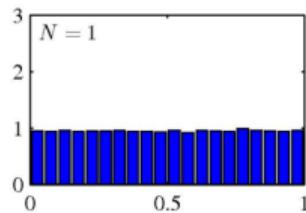
此定理的直观意义是：当 n 足够大时可近似地认为 $\sum_{i=1}^n X_i$ 服从正态分布，以下三种形式等价

$$\sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2), \quad \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad \frac{\frac{1}{n} \sum_{i=1}^n X_i - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

⁴² 强调方差不等于0，如果方差等于0那么随机变量永远只能等于某一个值，不具有随机性也就不满足中心极限定理；这里只要求随机变量独立同分布，不要求它服从某种特定的分布，换句话说此定理对任意分布的随机变量都成立



设 X 服从 $(0, 1)$ 间一致分布，将 $(0, 1)$ 作20等份转变为离散分布，每次采集 N 个样本后求 $S_N = \frac{1}{N} \sum_{i=1}^N x_i$ 。将第*i*次采集后得到的 S_N 记为 S_N^i 。很显然， S_N^i 也必然属于20等份中的某一份⁴³。



这一定理就是我们常说的中心极限定理，它有力的奠定了正态分布在概率统计中的中心地位，也揭示了自然界的神奇特性：任何不受有规律外力作用的事物，其整体上总是服从或近似服从正态分布。例如噪声或误差，由于噪声或误差是受多方面因素的共同作用造成的，如测量仪器的精准度、测量环境的温湿度、操作者技能水平、心理状态等等，但当这些因素迭加到一起时最终的噪声将服从正态分布⁴⁴。

⁴³ 此例来源于PRML的2.3节，但作者没有详细解释图中纵坐标的含义，应该对频数作了某种处理

⁴⁴ 如果发现噪声不服从正态分布，说明它正受某固定干扰源的干扰，应立即对设备进行排查检修



棣莫弗-拉普拉斯中心极限定理：设随机变量序列 X_1, X_2, \dots 独立同分布且都服从伯努利分布 $\mathcal{B}(1, p)$ ，则对任意实数 a 有

$$\lim_{n \rightarrow \infty} P\left(\frac{\frac{1}{n} \sum_{i=1}^n X_i - p}{\sqrt{p(1-p)/n}} \leq a\right) = \Phi(a)$$

此情况为前述中心极限定理的特例。因为伯努利分布的期望为 p ，方差为 $p(1 - p)$ ，只需要令前述公式中的 $\mu = p, \sigma = \sqrt{p(1 - p)}$ 即可得到此式。因为 $\sum_{i=1}^n X_i \sim \mathcal{B}(n, p)$ 服从二项分布，所以此定理表示二项分布的极限形式是正态分布⁴⁵。

⁴⁵ 在PRML的4.4节，PRML Page-by-page项目的4-047处就涉及到了中心极限定理



强大数定律⁴⁶：设有随机变量序列 X_1, X_2, \dots 相互独立同分布，所属分布的期望值为 $\mu = E[X]$ ，则下式以概率为1成立

$$P\left(\lim_{n \rightarrow \infty} \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| = 0\right) = 1$$

强大数定律之所以为“强”，是因为它发生的必然程度比“弱”大数定律更大。“弱”大数定律的数学形式是依概率收敛，“强”大数定律的数学形式是几乎必然收敛⁴⁷。依概率收敛表现的是“事件的概率极限”，而几乎必然收敛表现的是“极限事件的概率”。从数学上证明“强”比“弱”更强并不容易，但我们可以体会。

- 设 X_n 服从 $N(0, \frac{1}{n})$ ，因为随着 n 的增加 X_n 的方差变小直到趋近于 0，所以有 $X_n \xrightarrow{P} 0$
- 设 $X_n = \frac{1}{n}$ ，因为随着 n 的增加， X_n 将以确定的方式收敛到 0，所以有 $X_n \xrightarrow{\text{a.s.}} 0$

⁴⁶ 强大数定律可能是概率论中最有名的结论之一

⁴⁷ almost sure convergence，通常简写为 a.s.，它类似于 a.e.，只不过一个是概率论、测度论中的用语，一个是实变函数、数学分析中的用语



概率计算



从不同的角度对概率有不同的理解

- 事件发生的确信度：既有主观性又有客观性，但主观性也是基于客观性的
- 信息量：事件发生的概率值与含有的信息量是相反的，越是确定发生的事情信息量越小
- 事件集合的测度：数学观点，事件集合的测度大小

等可能概型 $\begin{cases} \text{有限样本空间：古典概型，抛硬币、扔骰子、摸小球等} \\ \text{样本空间是n维区域：几何概型，等待时间、坠落区域等} \end{cases}$



古典概型概率为: $P(A) = \frac{|A|}{|\Omega|}$, 几何概型概率为: $P(A) = \frac{m(A)}{m(\Omega)}$

在某事件 A 已发生的条件下, 另一事件 B 发生的概率记为 $P(B|A)$, 称作条件概率。条件概率的计算公式为⁴⁸

$$P(B|A) = \frac{P(A, B)}{P(A)}$$

条件概率也满足前面定义的概率公理化三条基本性质

- 非负性: 对任意事件 A, B , 总有 $P(B|A) \geq 0$
- 规范性: 在已知 A 发生的情况下, 下一次发生的事件总是属于 Ω , 所以有 $P(\Omega|A) = 1$
- 可列可加性: 若 A_1, A_2, \dots 互不相交, 则有 $P\left(\bigcup_{i=1}^{\infty} A_i|B\right) = \sum_{i=1}^{\infty} P(A_i|B)$

⁴⁸有一种观点是: 所有的概率都是条件概率



独立性和条件独立性是概率计算中最重要的性质

- 独立性： $P(A, B) = P(A)P(B) \iff A, B$ 相互独立
- 条件独立性： $P(A, B|C) = P(A|C)P(B|C) \iff A, B$ 关于条件 C 独立

独立性和条件独立性互不包含，独立推不出条件独立，条件独立更推不出独立。PRML第八章概率图模型本质上研究的就是随机变量间的条件独立性⁴⁹，通过寻找条件独立性简化概率计算。独立和条件独立常用以下符号表示。

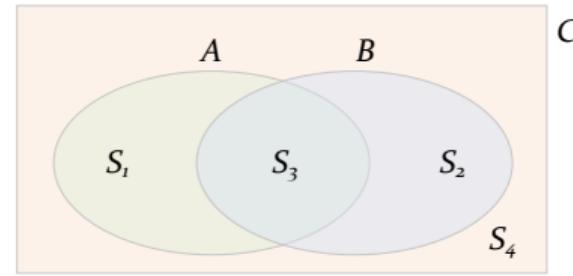
$$A \perp\!\!\!\perp B, \quad A \perp\!\!\!\perp B|C$$

⁴⁹需运用专门的方法和规则判定变量间是否具有条件独立性，详细内容请参考PRML第八章8.2节



和积公式是概率计算最基本的公式

- 和公式: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- 积公式: $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$



求证: 若 $A \perp\!\!\!\perp B$, 则 $\bar{A} \perp\!\!\!\perp B, \bar{A} \perp\!\!\!\perp \bar{B}$

$$P(\bar{A}, B) + P(A, B) = P(B)$$

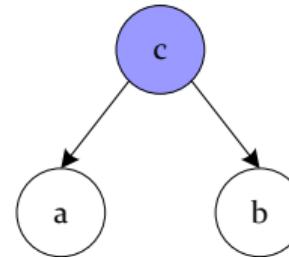
$$\begin{aligned} P(\bar{A}, B) &= P(B) - P(A, B) \\ &= P(B) - P(A)P(B) \\ &= P(\bar{A})P(B) \end{aligned}$$

$$P(\bar{A}, \bar{B}) + P(\bar{A}, B) = P(\bar{A})$$

$$\begin{aligned} P(\bar{A}, \bar{B}) &= P(\bar{A}) - P(\bar{A}, B) \\ &= P(\bar{A}) - P(\bar{A})P(B) \\ &= P(\bar{A})P(\bar{B}) \end{aligned}$$



若已知 $a \perp\!\!\!\perp b|c$, 求证 $\bar{a} \perp\!\!\!\perp b|c, a \perp\!\!\!\perp \bar{b}|c, \bar{a} \perp\!\!\!\perp \bar{b}|c$



$$p(\bar{a}, b|c) + p(a, b|c) = p(b|c)$$

$$\begin{aligned} p(\bar{a}, b|c) &= p(b|c) - p(a, b|c) \\ &= p(b|c) - p(a|c)p(b|c) \\ &= p(b|c)(1 - p(a|c)) \\ &= p(\bar{a}|c)p(b|c) \end{aligned}$$

$$p(\bar{a}, \bar{b}|c) + p(\bar{a}, b|c) = p(\bar{a}|c)$$

$$\begin{aligned} p(\bar{a}, \bar{b}|c) &= p(\bar{a}|c) - p(\bar{a}, b|c) \\ &= p(\bar{a}|c) - p(\bar{a}|c)p(b|c) \\ &= p(\bar{a}|c)(1 - p(b|c)) \\ &= p(\bar{a}|c)p(\bar{b}|c) \end{aligned}$$



证明两个条件概率公式

- 若已知 $a \perp\!\!\!\perp c|b$, 求证 $p(a|b, c) = p(a|b)$
- 求证无论在什么条件下都有 $p(a|b, c) = \frac{p(a, b|c)}{p(b|c)}$

$$\begin{aligned} p(a|b, c) &= \frac{p(a, b, c)}{p(b, c)} \\ &= \frac{p(a, c|b)p(b)}{p(c|b)p(b)} \\ &= \frac{p(a|b)p(c|b)}{p(c|b)} \\ &= p(a|b) \end{aligned}$$

$$\begin{aligned} p(a|b, c) &= \frac{p(a, b, c)}{p(b, c)} \\ &= \frac{p(a, b|c)p(c)}{p(b|c)p(c)} \\ &= \frac{p(a, b|c)}{p(b|c)} \end{aligned}$$

注意：若 $a \perp\!\!\!\perp c$ (非 $a \perp\!\!\!\perp c|b$) 不一定有 $p(a|b, c) = p(a|b)$



全概率公式：设样本空间为 Ω , B_1, B_2, \dots, B_n 为 Ω 的一个有限划分, 此时有

$$P(A) = \sum_{i=1}^n P(B_i)P(A|B_i)$$

贝叶斯公式：设 A_1, A_2, \dots, A_n 为样本空间 Ω 的一个完备事件组, 事件 B 有 $P(B) > 0$, 此时有

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

贝叶斯公式可能是机器学习中最重要的概率公式, 对于学习问题, 设问题的数据集为 X , 待学习参数为 θ 则有

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$



对于贝叶斯公式

- $P(\theta)$ 是先验, $P(X|\theta)$ 是似然, $P(\theta|X)$ 是后验, $P(X)$ 是正则化项⁵⁰
- “先”、“后”的判断标准是在 X 之前建模还是之后建模

条件概率公式和贝叶斯公式有什么区别?

- 从计算的角度没有区别, 因为 $P(X, Y) = P(X|Y)P(Y)$
- 建模思想上有本质区别, 条件概率中 X, Y 一般是同等地位变量, 贝叶斯公式中 θ 和 X 有明确的不同含义

$$P(Y|X) = \frac{P(X, Y)}{P(X)}, \quad P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

⁵⁰如果只是学习训练, 可以不求 $P(X)$; 如果是概率计算问题, 则需要求 $P(X)$: 先验和后验都是在对 θ 建模, 似然是在对 X 建模



设有事件 A ，它的优势比定义为 A 发生与否概率之比，表示事件发生可能性是不发生的多少倍。如果事件的优势比为 α ，意味着优势比为 $\alpha : 1$ 。若找到了关于事件新的证据 E ，可求 E 作为条件时新的优势比

$$\frac{P(A)}{P(\bar{A})} = \frac{P(A)}{1 - P(A)}, \quad \frac{P(A|E)}{P(\bar{A}|E)} = \frac{P(A)}{P(\bar{A})} \frac{P(E|A)}{P(E|\bar{A})}$$

设在同一数据集下建立两个不同的模型 M_1, M_0 ，用于模型选择的贝叶斯因子为⁵¹

$$\begin{aligned} BF_{1,0} &= \frac{P(D|M_1)}{P(D|M_0)} \\ &= \frac{P(M_1|D)P(D)/P(M_1)}{P(M_0|D)P(D)/P(M_0)} \\ &= \frac{P(M_1|D)}{P(M_0|D)} / \frac{P(M_1)}{P(M_0)} \\ &= \frac{P(M_1|D)}{P(M_0|D)} \quad (\text{无信息先验}) \end{aligned}$$

$BF_{1,0}$	说明
$BF_{1,0} < 1/100$	建议果断选择 M_0
$BF_{1,0} < 1/10$	强烈建议选择 M_0
$1/10 < BF_{1,0} < 1/3$	一般建议选择 M_0
$1/3 < BF_{1,0} < 1$	可以考虑选择 M_0

⁵¹ 例如 M_1 表示二次多项式模型， M_0 表示三次多项式模型；此处请参考PRML Page-by-page项目的3-031



贝叶斯观点



以下问题该如何求概率？

- 到本世纪末气温上升3°C及以上的概率有多大？
- X-Space下次成功发射并回收的概率有多大？
- 明天出门高速公路堵车的可能性有多大？

贝叶斯观点：某件事情是否发生与众多条件相关联，概率就是人对该事件发生与否的确信程度。如果某项相关条件发生改变，人对事物发生与否的判断也会发生改变。因此，概率是一个变量，它的取值随相关条件的变化而不断变化。

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$



频率派和贝叶斯派都认同似然 $P(X|\theta)$ 的意义，它们的争论主要在于

- 频率派认为概率值是客观存在且不变的，贝叶斯派认为概率值是主观感受且随对问题认识的加深而变化
- 频率派认为贝叶斯派对先验的选取是完全主观性的，这是不合理的

贝叶斯派的一个重要工作就是如何选择恰当的先验

- 根据现有知识选择⁵²
- 共轭先验
- 无信息先验⁵³
- 经验贝叶斯

⁵²这种方法允许使用者任意选择他认为合适的先验，我们不做介绍

⁵³无信息先验英文叫*non-informative*先验；其它先验在选择参数时都会更倾向于某一部分参数，因此也称为有信息先验（*informative*先验）



若参数 θ 的先验分布 $\pi(\theta)$ 属于分布族 \mathcal{F} ，若在样本 X 的作用下后验分布 $\pi(\theta|X)$ 仍属于 \mathcal{F} ，则称 \mathcal{F} 是一个共轭先验分布族。更具体的，若先验 $\pi(\theta)$ 在样本 X 的作用下与后验 $\pi(\theta|X)$ 有相同的函数形式，则称此时先验和后验是共轭分布。

共轭先验的优缺点

- 优点1：计算较为简便
- 优点2：具有较好的可解释性
- 缺点1：缺少合理性，不一定是参数真正服从的分布
- 缺点2：很多分布找不到共轭先验



设 x 服从伯努利分布，因此有 $\text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}$ ，它的共轭先验为Beta分布⁵⁴

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \propto \mu^{a-1}(1-\mu)^{b-1}$$

假设数据集中有 m 次事件发生， l 次事件不发生，那么此时其后验为

$$\begin{aligned} p(\mu|X) &\propto p(X|\mu)p(\mu) \\ &\propto \mu^m(1-\mu)^l \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\ &\propto \mu^{m+a-1}(1-\mu)^{l+b-1} \end{aligned}$$

此时后验和先验的函数形式相同， m, l 分别表示试验中事件发生、不发生的次数，先验中的参数 a, b 则可以解释为先验中事件发生、不发生的次数⁵⁵

⁵⁴ $\Gamma(t) = \int_0^\infty \mu^{t-1} e^{-\mu} d\mu$ ，其中 $t > 0$

⁵⁵ 可以这样理解 a, b ，但注意 a, b 不一定要取整数



高斯分布有两个参数：均值 μ 和方差 σ^2 ，它的共轭先验分三种情况⁵⁶

- 方差已知时 μ 的共轭先验：高斯分布
- 均值已知时 σ^{-1} 的共轭先验：Gamma分布
- 均值、方差均未知时的共轭先验：高斯-Gamma分布

设数据集 X 全都服从高斯分布， σ^2 取值已知， μ 的先验 $p(\mu)$ 服从高斯分布

$$p(X|\mu) = \prod_{n=1}^N p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2\right\}, \quad p(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right\}$$

此时后验为

$$\begin{aligned} p(\mu|X) &\propto p(X|\mu)p(\mu) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right\} \end{aligned}$$

⁵⁶ 多维高斯分布是均值 $\boldsymbol{\mu}$ 和协方差阵 $\boldsymbol{\Sigma}$ ，是类似的，这里以一维高斯分布为例；这里只介绍第一种情况，其它情况请参考PRML Page-by-page项目
第二章



对于一个一般的高斯分布，其密度函数指数项部分为

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = -\frac{1}{2} \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} + \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + C$$

对后验分布指数项部分变形有

$$-\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{(\mu - \mu_0)^2}{2\sigma_0^2} = -\frac{\mu^2}{2} \left(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \right) + \mu \left(\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma^2} \sum_{n=1}^N x_n \right) + C$$

所以对应有⁵⁷

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \rightarrow \sigma_N^2 = \frac{\sigma^2 \sigma_0^2}{N \sigma_0^2 + \sigma^2}$$

$$\frac{1}{\sigma_N^2} \mu_N = \left(\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma^2} \sum_{n=1}^N x_n \right) \rightarrow \mu_N = \frac{\sigma^2}{N \sigma_0^2 + \sigma^2} \mu_0 + \frac{N \sigma_0^2}{N \sigma_0^2 + \sigma^2} \frac{1}{N} \sum_{n=1}^N x_n$$

⁵⁷当 $N \rightarrow 0$ 时有 $\sigma_N^2 \rightarrow \sigma_0^2, \mu_N \rightarrow \mu_0$ ；当 $N \rightarrow \infty$ 时有 $\sigma_N^2 \rightarrow 0, \mu_N \rightarrow \mu_{ML}$ ，此时表示 μ 的取值越发稳定，几乎只取某一个确定值



指数族分布的一般形式如下所示，它是目前唯一已知一定有共轭先验的分布族⁵⁸

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\}$$

其参数 $\boldsymbol{\eta}$ 的共轭先验 $p(\boldsymbol{\eta}|\chi, \nu)$ 一般形式为

$$p(\boldsymbol{\eta}|\chi, \nu) = f(\chi, \nu)g(\boldsymbol{\eta})^\nu \exp\{\nu \boldsymbol{\eta}^T \chi\}$$

得到的后验 $p(\boldsymbol{\eta}|\mathbf{X}, \chi, \nu)$ 一般形式为

$$p(\boldsymbol{\eta}|\mathbf{X}, \chi, \nu) \propto g(\boldsymbol{\eta})^{\nu+N} \exp\{\boldsymbol{\eta}^T \left(\sum_{n=1}^N \mathbf{u}(\mathbf{x}_n) + \nu \chi \right)\}$$

⁵⁸ $\boldsymbol{\eta}$ 称为自然参数，它直接决定了分布的函数形式，相当于高斯分布中的均值和方差； $\mathbf{u}(\mathbf{x})$ 是分布的充分统计量； $h(\mathbf{x})$ 是缩放因子， $g(\boldsymbol{\eta})$ 是正则化因子，它确保该函数积分为1；关于指数族分布更详细介绍请参考PRML Page-by-page项目的2-051~2-059



贝叶斯派提出，在实在不知道参数先验时可选择一种特殊的先验，这种先验下无论 θ 取分布中的哪个值，它对后验的影响都是一样的。贝叶斯把这种“对 θ 取值一无所知”的说法在当年直接表示为 θ 服从 $U(0, 1)$ ，这种假设称为“贝叶斯假设”⁵⁹。

但Fisher对此批评道

- 对 θ 取值一无所知和知道任何 θ 对后验影响都一样，是两件不同的事
- 既然任意 θ 对后验的影响是一样的，那任意 $\frac{1}{\theta}$ 、 θ^2 以至 $g(\theta)$ 对后验的影响也应该是一样的

虽然一致分布有时候可以达到无信息先验的目的，但很多时候也并不是无信息先验。另外，根据概率密度函数映射变换可知，Fisher批评的第二点是很明显的⁶⁰。也就是说贝叶斯假设并不总是成立的，不同情况下的无信息先验需要具体讨论。

⁵⁹ 关于无信息先验更多介绍请参考PRML Page-by-page项目的2-057

⁶⁰ PRML公式1.27式，其推导过程请参考PRML Page-by-page项目1-020；从 θ 到 $g(\theta)$ 的变换称为再参数化（reparameterization），这是一种常用的技巧，常用于难以直接对待估参数求导数的情况，在淡蓝小点技术系列介绍VAE时曾介绍过



以一致分布为例，如果用一致分布作某些参数的先验，它可能根本就不是一个合法的概率密度，因为它的积分很可能大于1，我们把这种先验称为反常先验。例如高斯分布的均值 μ ，因为 μ 的取值是 $(-\infty, +\infty)$ ，所以若用 $p(\mu) = c$ 作先验，那么总有 $\int_{-\infty}^{+\infty} c d\mu = \infty$ 。但是反常先验仍可以作为有效先验使用，只要由它得到的后验是合法⁶¹。

设 x 服从正态分布 $N(x|\theta, \sigma^2)$ ，令 $p(\theta) = c, c > 0, \theta \in (-\infty, +\infty)$ ，很明显 $p(\theta)$ 不是一个合法的概率密度。下面验证它是一个反常先验。根据贝叶斯公式有

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x, \theta)}{\int_{-\infty}^{+\infty} p(x|\theta)p(\theta)d\theta}$$

设有 n 个数据，则似然 $p(x|\theta)$ 为

$$p(x|\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i - \theta)^2\right)$$

⁶¹improper prior, 有些地方也称为广义先验；若 $\pi(\theta)$ 是一个广义先验，则任意 $c\pi(\theta)$ 也是一个广义先验



因而有

$$\begin{aligned}
 p(x) &= \int_{-\infty}^{+\infty} p(x|\theta)p(\theta)d\theta \\
 &= \int_{-\infty}^{+\infty} \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i - \theta)^2\right) c d\theta \\
 &= \dots \\
 &= \frac{cn^{-1/2}}{(2\pi)^{n-1/2}\sigma^{n-1}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 + \frac{n}{2\sigma^2} \bar{x}^2\right) \quad (\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i)
 \end{aligned}$$

同时还有⁶²

$$p(x, \theta) = \frac{c \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 + \frac{n}{2\sigma^2} \bar{x}^2\right)}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{n}{2\sigma^2}(\theta - \bar{x})^2\right)$$

⁶²篇幅所限，推导的具体过程请参考PRML Page-by-page项目的2-057



此时后验 $p(\theta|x)$ 为

$$\begin{aligned}
 p(\theta|x) &= \frac{p(x, \theta)}{p(x)} \\
 &= \frac{\frac{c \exp(-\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 + \frac{n}{2\sigma^2} \bar{x}^2)}{\sigma^n (2\pi)^{n/2}} \exp(-\frac{n}{2\sigma^2} (\theta - \bar{x})^2)}{\frac{cn^{-1/2}}{(2\pi)^{n-1/2}\sigma^{n-1}} \exp(-\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 + \frac{n}{2\sigma^2} \bar{x}^2)} \\
 &= \frac{n^{\frac{1}{2}}}{\sqrt{2\pi}\sigma} \exp(-\frac{n}{2\sigma^2} (\theta - \bar{x})^2) \\
 &= \frac{1}{\sqrt{2\pi}(\frac{\sigma^2}{n})^{1/2}} \exp(-\frac{1}{2(\frac{\sigma}{\sqrt{n}})^2} (\theta - \bar{x})^2)
 \end{aligned}$$

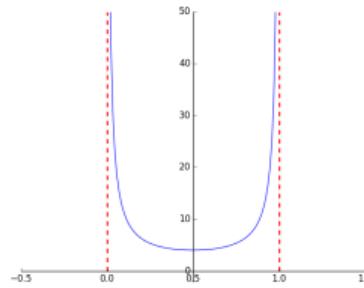
很明显， $p(\theta|x) \sim N(\theta|\bar{x}, \frac{\sigma^2}{n})$ ，只要观察数据量 $n \geq 1$ 它就是一个合法的概率密度。



很显然，并非所有的反常先验都可以得到合法的后验分布。设有伯努利分布如下，并为其指定反常先验 $p(\theta)$ ⁶³

$$p(x|\theta) = \theta^x(1-\theta)^{1-x}, \quad p(\theta) = \frac{1}{\theta(1-\theta)}, 0 \leq \theta \leq 1$$

$$\begin{aligned} \int_0^1 \frac{1}{\theta(1-\theta)} d\theta &> \int_0^1 \frac{1}{\theta} d\theta \\ &> \ln \theta \Big|_0^1 \\ &> \ln 1 - \ln 0 \\ &> +\infty \end{aligned}$$



⁶³这一反常先验又称“Haldane prior”，其实就是特殊的Beta分布Beta(0,0)



此时后验分布为

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta) \\ &\propto \theta^x(1-\theta)^{1-x}\theta^{-1}(1-\theta)^{-1} \\ &\propto \theta^{x-1}(1-\theta)^{-x} \end{aligned}$$

假设试验中 $x = 0$ (即事件不发生), 此时后验分布为

$$p(\theta|x=0) \propto \frac{1}{\theta}, \quad (\text{前面已证明此分布积分为 } +\infty)$$

此例表示, 一个反常先验导致一个反常后验, 所以此先验是不可用的。



一个概率分布可能有多个参数，不同参数对概率分布的控制作用不同，根据参数在分布中的不同地位，介绍三类无信息先验

- 位置参数的无信息先验：形如 $p(x|\mu) = f(x - \mu)$ 的参数 μ 是分布的位置参数
- 尺度参数的无信息先验：形如 $p(x|\sigma) = \frac{1}{\sigma}f(\frac{x}{\sigma})$ 的参数 σ 是分布的尺度参数
- 一般参数的无信息先验：**Jeffreys先验**

位置参数对于分布而言具有平移不变性，例如假设 x 平移成 $\hat{x} = x + c$ ，只需要将 μ 平移成 $\hat{\mu} = \mu + c$ 就有

$$p(\hat{x}|\hat{\mu}) = f(\hat{x} - \hat{\mu}) = f(x + c - \mu - c) = f(x - \mu) = p(x|\mu)$$

因为 μ 可以被任意平移，要使 $p(\mu)$ 是无信息的，它意味着对于任意一个固定长度的区间无论它被移动到哪儿，其先验的密度积分不变

$$\int_A^B p(\mu) d\mu = \int_{A-c}^{B-c} p(\mu) d\mu = \int_A^B p(\mu - c) d\mu$$



因此前述积分式对任意区间 (A, B) 成立，所以有

$$p(\mu) = p(\mu - c) = \text{const}$$

或者从概率密度函数映射变换的角度看

$$p(\mu) = p(\mu - c) \left| \frac{d(\mu - c)}{d\mu} \right| = p(\mu - c) = \text{const}$$

所以，位置参数的无信息先验一定是一致分布，并且该一致分布很有可能是反常的。高斯分布的均值是其位置参数，另外我们在前面推导过均值 μ 的共轭先验仍然是高斯分布，其后验为

$$\mu_n = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\mu_{ML}, \quad \frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

如果先验 $p(\mu) \sim N(\mu|\mu_0, \sigma_0^2)$, $\sigma_0^2 \rightarrow \infty$ 则 $p(\mu)$ 趋向于 $(-\infty, +\infty)$ 内的一致分布，理论上它就是无信息先验。实际上，如果 $\sigma_0^2 \rightarrow \infty$ 则上式中 $\mu_n \rightarrow \mu_{ML}$, $\sigma_n^2 \rightarrow \frac{\sigma^2}{N}$ ，可见后验分布只受数据集影响、不受先验影响。这与无信息先验的初衷是一致的。



如果概率密度函数对于参数 σ 具有如下形态，则说 σ 是该分布的尺度参数⁶⁴。

$$p(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$$

尺度参数对于概率分布具有缩放不变性，例如令 $\hat{x} = cx$ （将 x 扩大 c 倍），那么 $p(\hat{x}|\hat{\sigma})$ 在 (cA, cB) 范围内的积分等于 $p(x|\sigma)$ 在 (A, B) 范围内的积分。根据概率密度函数映射变换公式⁶⁵

$$p_{\hat{x}}(\hat{x}|\hat{\sigma}) = p_x(g(\hat{x})) \left| \frac{dg(\hat{x})}{dx} \right| = \frac{1}{\sigma} f\left(\frac{\frac{1}{c}\hat{x}}{\sigma}\right) \frac{1}{c} = \frac{1}{c\sigma} f\left(\frac{\hat{x}}{c\sigma}\right) = \frac{1}{\hat{\sigma}} f\left(\frac{\hat{x}}{\hat{\sigma}}\right)$$

对 $p(\hat{x}|\hat{\sigma})$ 在 (cA, cB) 范围内积分则有

$$\int_{cA}^{cB} p(\hat{x}|\hat{\sigma}) d\hat{x} = \int_{cA}^{cB} \frac{1}{c\sigma} f\left(\frac{\hat{x}}{c\sigma}\right) d\hat{x} = \int_{cA}^{cB} \frac{1}{\sigma} f\left(\frac{\frac{1}{c}\hat{x}}{\sigma}\right) d\frac{1}{c}\hat{x}$$

⁶⁴ scale parameter, 也称刻度参数

⁶⁵ PRML中的公式1.27，推导请参考PRML Page-by-page项目的1-020



可将积分变量看作是 $\frac{1}{c}\hat{x}$, $\frac{1}{c}\hat{x} \in (cA, cB)$ 。令 $x = \frac{1}{c}\hat{x}$ 则 $x \in (A, B)$, 所以有

$$\int_{cA}^{cB} p(\hat{x}|\hat{\sigma}) d\hat{x} = \int_{cA}^{cB} \frac{1}{c\sigma} f\left(\frac{\hat{x}}{c\sigma}\right) d\hat{x} = \int_A^B \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right) dx$$

σ 的无信息先验 $p(\sigma)$ 也应具有类似性质, 即对任意的 $\sigma \in (A, B)$, 若 $\hat{\sigma} = c\sigma \in (cA, cB)$, 则 $p(\hat{\sigma})$ 在 (cA, cB) 内的积分应等于 $p(\sigma)$ 在 (A, B) 内的积分

$$\int_A^B p(\sigma) d\sigma = \int_{cA}^{cB} p(\hat{\sigma}) d\hat{\sigma} = \int_A^B cp(c\sigma) d\sigma$$

这相当于在任意一点处都有

$$cp(c\sigma) = p(\sigma) \longrightarrow p(c\sigma) = \frac{1}{c}p(\sigma)$$

这跟前面已推导的密度函数的性质是类似的⁶⁶

$$p_{\hat{x}}(\hat{x}|\hat{\sigma}) = \frac{1}{c\sigma} f\left(\frac{\hat{x}}{c\sigma}\right) = \frac{1}{c\sigma} f\left(\frac{cx}{c\sigma}\right) = \frac{1}{c} \left(\frac{1}{\sigma} f\left(\frac{x}{\sigma}\right) \right) = \frac{1}{c} p(x|\sigma)$$

⁶⁶对于尺度参数概率密度而言, 将 x 缩放为 cx 后, cx 处的概率密度将变成原来 x 处密度的 $\frac{1}{c}$

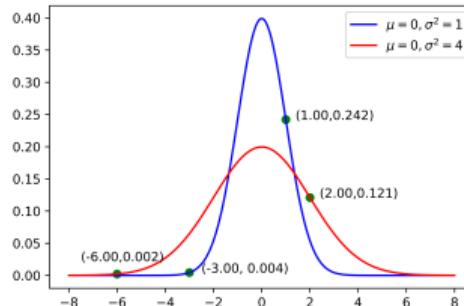


举例，假设高斯分布的均值 μ 已知，那么有⁶⁷

$$p(x|\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right), \quad f\left(\frac{x}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

若令 $\hat{x} = 2x, \hat{\mu} = 2\mu, \hat{\sigma} = 2\sigma$ （即将所有的变量、参数都放大2倍）则有

$$\begin{aligned} p(\hat{x}|\hat{\sigma}) &= \frac{1}{\hat{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\hat{x}-\hat{\mu})^2}{2\hat{\sigma}^2}\right) \\ &= \frac{1}{2\sigma} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(2x-2\mu)^2}{2(2\sigma)^2}\right) \\ &= \frac{1}{2} \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right) \\ &= \frac{1}{2} p(x|\sigma) \end{aligned}$$



⁶⁷注意高斯分布的尺度参数是 σ 而非 σ^2



通过分析已知，尺度参数的先验当 σ 放大 c 倍变成 $c\sigma$ 后， $p(c\sigma)$ 却是原密度值 $p(\sigma)$ 的 $\frac{1}{c}$ 。也就是说，尺度参数的无信息先验的特性应是：当 σ 变大时 $p(\sigma)$ 反而减小，当 σ 减小时 $p(\sigma)$ 反而变大，即

$$p(\sigma) \propto \frac{1}{\sigma}$$

凡是具有这种数学关系的密度函数，都可以作为尺度参数的无信息先验。例如，假设就令 $p(\sigma) = \frac{1}{\sigma}$ ，最开始令 $(a, b) = (1, 10)$ ，然后将变量放大10倍、100倍，则相应的取值范围也变成 $(10, 100), (100, 1000)$ ⁶⁸

$$\int_1^{10} p(\sigma) d\sigma = \int_1^{10} \frac{1}{\sigma} d\sigma = \ln \sigma \Big|_1^{10} = \ln 10 - \ln 1 \approx 2.303 - 0 = 2.303$$

$$\int_{10}^{100} p(\hat{\sigma}) d\hat{\sigma} = \int_{10}^{100} \frac{1}{10\sigma} d10\sigma = \int_{10}^{100} \frac{1}{\sigma} d\sigma = \ln 100 - \ln 10 \approx 4.605 - 2.303 = 2.302$$

$$\int_{100}^{1000} p(\hat{\sigma}) d\hat{\sigma} = \int_{100}^{1000} \frac{1}{100\sigma} d100\sigma = \int_{100}^{1000} \frac{1}{\sigma} d\sigma = \ln 1000 - \ln 100 \approx 6.908 - 4.605 = 2.303$$

⁶⁸以下结果应精确相等，误差来源于计算精度受限



Jeffrey先验是一类适用面更广泛的无信息先验，它有两个特点

- 位置参数的Jeffrey先验是一致分布，尺度参数的Jeffrey先验其密度值与参数值呈反比
- 它在再参数化上表现出特殊的不变性（invariant）⁶⁹

Jeffrey先验不是某种固定形式的先验，它是根据变量似然函数按一定方法计算得到的先验。设参数为 θ ，用 $\pi_J(\theta)$ 表示 θ 的Jeffrey先验，那么有

$$\pi_J(\theta) = c I(\theta)^{1/2} \propto I(\theta)^{1/2}$$

$I(\theta)$ 表示样本中关于 θ 的费歇尔信息量（Fisher Information），它的定义是

$$I(\theta) = -E_{p(x|\theta)} \left[\frac{d^2 \ln p(x|\theta)}{d\theta^2} \right] = - \int \frac{\partial^2}{\partial \theta^2} \ln p(x|\theta) p(x|\theta) dx$$

⁶⁹这种不变性或者也称“共变性”（co-variance）；关于共变性此处不介绍，请参考PRML Page-by-page 2-058+



先来验证平移参数、尺度参数的Jeffrey先验都是前面讨论过的无信息先验。假设 $p(x)$ 服从高斯分布，其均值 μ 的Jeffrey先验为

$$\ln p(x|\mu, \sigma) = \ln \frac{1}{\sqrt{2\pi}} - \ln \sigma - \frac{(x-\mu)^2}{2\sigma^2} \rightarrow \frac{d}{d\mu} \ln p(x|\mu, \sigma) = \frac{x-\mu}{\sigma^2}, \frac{d^2}{d\mu^2} \ln p(x|\mu, \sigma) = -\frac{1}{\sigma^2}$$

所以

$$\pi_J(\mu) \propto I(\mu)^{1/2} = \left\{ - \int \left(-\frac{1}{\sigma^2} \right) p(x|\mu, \sigma) dx \right\}^{1/2} = \frac{1}{\sigma} = C$$

这与前面我们分析的，位置参数的无信息先验应服从一致分布是相吻合的。



再来验证尺度参数的Jeffrey先验是无信息的。仍然假设 $p(x)$ 服从高斯分布，那么

$$\frac{d}{d\sigma} \ln p(x|\mu, \sigma) = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}, \quad \frac{d^2}{d\sigma^2} \ln p(x|\mu, \sigma) = \frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4}$$

所以有

$$\begin{aligned}\pi_J(\sigma) \propto I(\sigma)^{1/2} &= \left\{ - \int \left(\frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4} \right) p(x|\mu, \sigma) dx \right\}^{1/2} \\ &= \left\{ -\frac{1}{\sigma^2} + \frac{3}{\sigma^4} E[(x-\mu)^2] \right\}^{1/2} \\ &= \frac{1}{\sigma}\end{aligned}$$

这与我们前面所分析的尺度参数的无信息先验是一致的



经验贝叶斯为参数 θ 的先验引入超参数 η ，即假设 θ 的先验是 $p(\theta|\eta)$ 。在一般的贝叶斯中，对先验的选择（相当于这里的 η ）具有主观性甚至随意性。经验贝叶斯则根据数据集 X 来选择 η ⁷⁰。

$$\hat{\eta} = \arg \max_{\eta} \int p(X|\theta)p(\theta|\eta)d\theta = \arg \max_{\eta} p(X|\eta)$$

PRML中3.5.1以高斯分布为例，介绍了如何利用经验贝叶斯求超参数 α, β 。由于整个过程较为冗长繁琐，此处不展开介绍。可参考PRML的3.5.1部分或PRML Page-by-page项目的3-043~3-045。

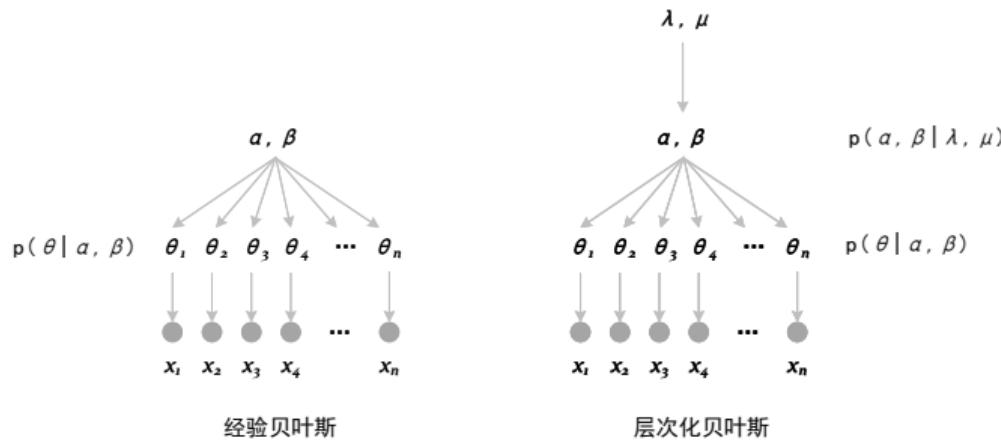
⁷⁰ 先验应在获得数据集之前确定，因此严格来说经验贝叶斯（empirical bayes）并不是真正意义上的贝叶斯，它介于经典统计学和贝叶斯统计之间；这里介绍的经验贝叶斯主要来自于PRML的3.5节，也称第二型极大似然估计ML-II、部分贝叶斯（partial bayes）、证据近似；更多详细介绍请参考PRML Page-by-page项目的3-040



层次化贝叶斯是为先验的超参数再次建立概率分布及控制参数

$$\text{似然: } p(x|\theta, \alpha, \beta, \lambda, \mu), \quad \text{先验: } p(\theta, \alpha, \beta, \lambda, \mu) = p(\theta|\alpha, \beta)p(\alpha, \beta|\lambda, \mu)$$

理论上可以继续为 λ, μ 引入先验及控制参数，但层次过多时会使得计算非常复杂且模型准确性不一定继续提升





随机变量及其描述



设 X 是基本空间， \mathcal{B} 是其子集 σ -代数，称二元组 (X, \mathcal{B}) 为可测空间， \mathcal{B} 中的元素为可测集。设 (Ω, \mathcal{F}) 和 (X, \mathcal{B}) 为可测空间，映射 $\xi : X \mapsto Y$ 称为可测映射，如果 $\xi^{-1}(\mathcal{B}) \subset \mathcal{F}$ ⁷¹。

概率空间 (Ω, \mathcal{F}, P) 到可测空间 (X, \mathcal{B}) 的可测变换 ξ 称为定义在 (Ω, \mathcal{F}, P) 上取值为 (X, \mathcal{B}) 的随机元， ξ 的导出测度 $P\xi^{-1}$ 称为它的分布。

用 \mathbb{R} 表示全体实数，用 \mathcal{R} 表示 \mathbb{R} 中形如 $(-\infty, x]$ 的集合构成的 σ -代数： $\mathcal{R} = \sigma(\{(-\infty, x] : x \in \mathbb{R}\})$ 。定义在概率空间 (Ω, \mathcal{F}, P) 上取值于 $(\mathbb{R}, \mathcal{R})$ 的随机元 ξ 称为这个概率空间上的随机变量⁷²； $F(x) = (P\xi^{-1})(-\infty, x] = P(\xi \leq x), x \in \mathbb{R}$ 称为随机变量 ξ 的分布函数⁷³。

⁷¹像空间的 σ -代数其原像也是 σ -代数

⁷²随机变量本质上是一个可测映射

⁷³随机变量的作用是将样本空间转换为一个无量纲的数集



设 Ω 是随机试验的样本空间，若对于每个样本点 $\omega \in \Omega$ ，都有唯一实数 $\xi(\omega)$ 与之对应且对于任意实数 x 都有确定的概率 $P\{\xi(\omega) \leq x\}$ 与之对应，则称 $\xi(\omega)$ 为随机变量，简记为 ξ 。

随机变量不同于普通意义上的变量，它具有两个特点

- 它是由随机试验结果所决定的量（是从 Ω 而非 \mathcal{F} 到实数的映射）
- 随机变量取各值的可能性大小有确定的统计规律性（主要由概率空间中的 P 决定）

将一枚均匀硬币连续抛两次，用 H, T 分别表示正面、反面朝上，则样本空间为 $\Omega = \{HH, HT, TH, TT\}$ 。用 ξ 表示两次抛掷中正面出现的次数，它是一个随机变量

$$\xi(TT) = 0, \quad \xi(TH) = 1, \quad \xi(HT) = 1, \quad \xi(HH) = 2$$



设抛投一个有六面的骰子，则其样本空间为 $\Omega = \{1\text{点}, 2\text{点}, 3\text{点}, 4\text{点}, 5\text{点}, 6\text{点}\}$ 。用 ξ 表示每次抛投得到的点数，它是一个随机变量

$$\xi(1\text{点}) = 1, \quad \xi(2\text{点}) = 2, \quad \xi(3\text{点}) = 3, \quad \xi(4\text{点}) = 4, \quad \xi(5\text{点}) = 5, \quad \xi(6\text{点}) = 6$$

随机变量的两种不同看待角度⁷⁴

- 特指某种映射规则，既不随机也不变量
- 相同的试验结果具有不确定性、随机性

⁷⁴从数学定义或应用的角度看待随机变量



设 Ω 为样本空间， ξ 是 Ω 上的随机变量，若 ξ 的值域为有限集或可数集，那么称 ξ 为离散型随机变量。

若离散随机变量 ξ 的取值为 $x_1, x_2, \dots, x_n, \dots$ ，称相对应的概率 $P(\xi = x_i) = p_i$ 为随机变量 ξ 的分布律（或分布列、概率函数）⁷⁵，分布列可用表格表示。注意， p_i 要满足概率公理化，例如非负性、规范性。

ξ	x_1	x_2	\cdots	x_n	\cdots
概率	p_1	p_2	\cdots	p_n	\cdots

离散型随机变量及其分布列定义有两点注意

- ξ 的值域是有限或可数的， Ω 本身可以是连续的、离散的甚至混合的
- 映射结果也可能有无穷多，若是无穷多只要求其可数

⁷⁵ p_i 有时也称为 x_i 的概率质量



离散变量的分布函数由不同情况的概率值累加得到

$$F(\xi) = \begin{cases} 0, & \xi < -1 \\ 0.2, & -1 \leq \xi < 0 \\ 0.6, & 0 \leq \xi < 2 \\ 1, & \xi \geq 2 \end{cases}$$

ξ	-1	0	2
概率	0.2	0.4	0.4

了解一个随机变量的分布列或分布函数，相当于了解该随机变量的全部信息

- 已知一个分布列可得其分布函数，已知一个分布函数也可得其分布列，二者等价
- 任意离散变量都一定对应有分布列或分布函数且是唯一的



设 Ω 为样本空间， ξ 是 Ω 上的随机变量，若 $\xi(\omega) = x$ ， $F(x)$ 是 ξ 的分布函数，若存在非负函数 $f(x)$ 使得下式成立，则称 ξ 是连续型随机变量， $f(x)$ 称为 ξ 的概率密度函数⁷⁶

$$F(x) = \int_{-\infty}^x f(t)dt$$

连续型随机变量有下列特性

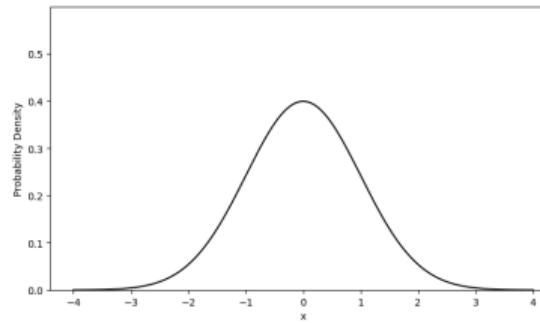
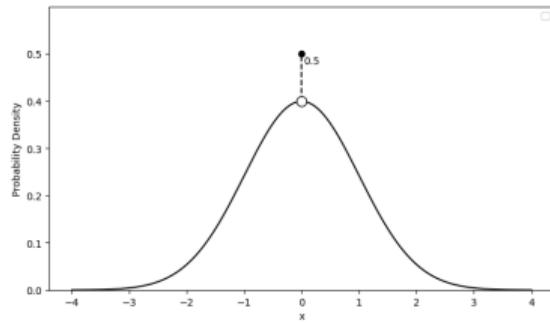
- $F(x)$ 是连续函数，在 $f(x)$ 的连续点处有 $F'(x) = f(x)$
- $f(x)$ 的积分为1但 $f(x)$ 本身的取值完全有可能大于1⁷⁷
- 对任意一点都有概率值 $P(x) = 0$
- 知道密度函数就几乎相当于了解了概率分布的全部信息

⁷⁶ 连续型随机变量的取值充满了某个区间或某些区间的并，在这个区间里有无穷不可数个实数；密度函数也要满足概率公理化，包括 $f(x) \geq 0$ 和 $\int_{-\infty}^{+\infty} f(x)dx = 1$

⁷⁷ 密度值大于1与总积分等于1不矛盾，因为大于1的点的集合测度很小



理论上，同一个分布的概率密度函数并不唯一，若 $f(x)$ 和 $g(x)$ 在某些测度为零的集合上取值不同，它们仍可表示同一个概率分布，因为它们对应同一个累积分布函数。





定义一个概率分布有多种不同的表示方法

- 累积分布函数（分布列）
- 概率密度函数
- 矩母函数
- 特征函数

实数范围内分布函数和特征函数一定存在且唯一，密度函数可能不唯一也可能不存在，矩母函数可能不存在⁷⁸。若 $f(x)$ 为密度函数， $y = g(x)$ ， y 的密度函数 $f_y(y)$ 就是概率密度函数的映射变换⁷⁹。

⁷⁸ 康托分布只有累积分布函数没有密度函数，柯西分布没有矩母函数

⁷⁹ 这一问题已介绍多次，请参考PRML Page-by-page项目中的1-020或PRML重点问题选讲Chapter1



矩母函数（moment-generating function）也称矩生成函数、动差生成函数，它可以表示一个概率分布，但前提是该分布的矩存在⁸⁰。设随机变量 X 的矩母函数为 $M_X(t)$ ，离散型和连续型变量下 $M_X(t)$ 定义如下

$$M_X(t) = E[e^{tX}] = \sum_X e^{tX} p(X) \quad (\text{离散}), \quad M_X(t) = E[e^{tX}] = \int_{-\infty}^{+\infty} e^{tX} p(X) dX \quad (\text{连续})$$

矩母函数的两个重要用途

- 可以更加简单的求出分布的矩⁸¹
- 独立随机变量之和的矩母函数等于随机变量矩母函数之积

⁸⁰用矩母函数表示分布，其本质依赖于用矩表示某个分布，由于矩并非总是存在（积分可能发散）所以矩母函数也并非总是存在；矩对分布的表示是否唯一，这一问题的答案并非十分简洁明了，但在不严格的情况下，我们可以认为矩对分布的表示是唯一的

⁸¹这也是它被称为矩母函数的重要原因



已知 e^x 的泰勒展开式为

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots$$

那么 e^{tX} 可展开为

$$\begin{aligned} e^{tX} &= 1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \cdots + \frac{(tX)^n}{n!} + \cdots \\ &= 1 + tX + t^2 \frac{X^2}{2!} + t^3 \frac{X^3}{3!} + \cdots + t^n \frac{X^n}{n!} + \cdots \end{aligned}$$

矩母函数 $M_X(t)$ 可展开为

$$M_X(t) = E[e^{tX}] = 1 + tE[X] + \frac{t^2}{2!} E[X^2] + \cdots + \frac{t^n}{n!} E[X^n] + \cdots$$



求 $M_X(t)$ 关于 t 的一阶导、二阶导直至 n 阶导，并令 $t = 0$ 后有

$$\frac{dM_X(t)}{dt} \Big|_{t=0} = E[X], \quad \frac{d^2M_X(t)}{dt^2} \Big|_{t=0} = E[X^2], \quad \dots, \quad \frac{d^nM_X(t)}{dt^n} \Big|_{t=0} = E[X^n]$$

只需要求 $M_X(t)$ 关于 t 的 n 阶导并令 $t = 0$ 就可以得到分布的 n 阶矩，这可以大大简少求矩的计算量和难度。

设 X 服从高斯分布 $N(X|\mu, \sigma^2)$ ，它的矩母函数则为

$$\begin{aligned} M_X(t) &= E[e^{tX}] = \int_{-\infty}^{+\infty} e^{tX} p(X) dX \\ &= \int_{-\infty}^{+\infty} e^{tX} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right) dX \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \exp\left(tX - \frac{(X-\mu)^2}{2\sigma^2}\right) dX \end{aligned}$$



重点关注指数项里面的部分，将其变换成平方项

$$\begin{aligned} tX - \frac{(X - \mu)^2}{2\sigma^2} &= tX - \frac{X^2 - 2\mu X + \mu^2}{2\sigma^2} \\ &= -\frac{X^2 - 2\mu X - 2t\sigma^2 X + \mu^2}{2\sigma^2} \\ &= -\frac{X^2 - 2(\mu + t\sigma^2)X + (\mu + t\sigma^2)^2 - (\mu + t\sigma^2)^2 + \mu^2}{2\sigma^2} \\ &= -\frac{(X - (\mu + t\sigma^2))^2}{2\sigma^2} + \frac{(\mu + t\sigma^2)^2 - \mu^2}{2\sigma^2} \end{aligned}$$

所以有

$$M_X(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{(\mu + t\sigma^2)^2 - \mu^2}{2\sigma^2}\right\} \int_{-\infty}^{+\infty} \exp\left\{-\frac{(X - (\mu + t\sigma^2))^2}{2\sigma^2}\right\} dX$$



很明显，积分部分是一个不带正则化项的高斯密度函数，其均值 $\tilde{\mu} = \mu + t\sigma^2$ ，它的积分为正则化项部分的倒数

$$\int_{-\infty}^{+\infty} \exp\left(-\frac{(X - \tilde{\mu})^2}{2\sigma^2}\right) dX = \sqrt{2\pi}\sigma$$

所以最终有

$$\begin{aligned} M_X(t) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{(\mu + t\sigma^2)^2 - \mu^2}{2\sigma^2}\right\} \sqrt{2\pi}\sigma \\ &= \exp\left\{\frac{(\mu + t\sigma^2)^2 - \mu^2}{2\sigma^2}\right\} \\ &= \exp\left\{\frac{1}{2}\sigma^2t^2 + \mu t\right\} \end{aligned}$$

验证用 $M_X(t)$ 求高斯分布的一阶矩和二阶矩⁸²

$$\frac{dM_X(t)}{dt}\Big|_{t=0} = (\sigma^2 t + \mu) \exp\left\{\frac{1}{2}\sigma^2 t^2 + \mu t\right\}\Big|_{t=0} = \mu$$

$$\frac{d^2M_X(t)}{dt^2}\Big|_{t=0} = \sigma^2 \exp\left\{\frac{1}{2}\sigma^2 t^2 + \mu t\right\} + (\sigma^2 t + \mu)^2 \exp\left\{\frac{1}{2}\sigma^2 t^2 + \mu t\right\}\Big|_{t=0} = \mu^2 + \sigma^2$$

⁸²已知对高斯分布有 $E[x] = \mu, \sigma^2 = E[x^2] - E^2[x]$



假设某电磁干扰器上有两个独立的噪声生发源，各自服从高斯分布 $x_1 \sim \mathcal{N}(x_1|\mu_1, \sigma_1^2)$, $x_2 \sim \mathcal{N}(x_2|\mu_2, \sigma_2^2)$ ，求接收方采集到的电磁信号 $z = x_1 + x_2$ 所服从的分布。利用高斯卷积知识可以很快知道 $z \sim \mathcal{N}(z|\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ ⁸³，现在用矩母函数来推导 z 的分布，会更加简单。因为有

$$M_{x_1}(t) = E[e^{tx_1}], \quad M_{x_2}(t) = E[e^{tx_2}]$$

所以

$$M_z(t) = E[e^{tz}] = E[e^{t(x_1+x_2)}] = E[e^{tx_1} e^{tx_2}] = E[e^{tx_1}]E[e^{tx_2}] = M_{x_1}(t)M_{x_2}(t)$$

刚刚已推导 $M_X(t) = \exp\{\frac{1}{2}\sigma^2 t^2 + \mu t\}$ ，所以此时有

$$M_z(t) = \exp\{\frac{1}{2}\sigma_1^2 t^2 + \mu_1 t\} \exp\{\frac{1}{2}\sigma_2^2 t^2 + \mu_2 t\} = \exp\{\frac{1}{2}(\sigma_1^2 + \sigma_2^2)t^2 + (\mu_1 + \mu_2)t\}$$

那么很容易的就知道 $M_z(t)$ 服从高斯分布（因为它矩母函数的形式和高斯分布矩母函数形式完全一样）且

$$\mu_z = \mu_1 + \mu_2, \quad \sigma_z^2 = \sigma_1^2 + \sigma_2^2$$

⁸³ 详细过程请参考PRML Page-by-page项目的2-030



特征函数（characteristic function）在形式上类似于矩母函数，但它是在复指数上定义的。设随机变量 X 的特征函数为 $\phi_X(t)$ 为

$$\phi_X(t) = E[e^{itX}] = \sum_X e^{itX} p(X) \quad (\text{离散}), \quad \phi_X(t) = E[e^{itX}] = \int_{-\infty}^{+\infty} e^{itX} p(X) dX \quad (\text{连续})$$

特征函数和矩母函数主要异同点

- 独立变量之和的特征函数等于特征函数之积： $\phi_Z(t) = E[e^{itz}] = E[e^{it(x_1+x_2)}] = E[e^{itz_1}]E[e^{itz_2}] = \phi_{x_1}(t)\phi_{x_2}(t)$
- 任意一个概率分布特征函数总是存在，但矩母函数不一定总是存在⁸⁴
- 概率分布的特征函数总是唯一的⁸⁵
- 都可以通过求 n 阶导得到分布的 n 阶矩
- 矩母函数类似于密度函数的拉普拉斯变换，特征函数类似于密度函数的傅里叶变换⁸⁶
- 特征函数在处理周期性、振荡性问题时更有效

⁸⁴因为 e^{itX} 总是有界的而 e^{tX} 可能是无界的

⁸⁵一般情况下我们可认为若矩母函数存在也是唯一的，但详细讨论时还要分情况

⁸⁶离散型分布一样



已知欧拉公式为 $e^{ix} = \cos x + i \sin x$, 因此对于特征函数有

$$e^{itX} = \cos(tX) + i \sin(tX)$$

已知复数 $z = a + ib$ 的模为 $|z| = \sqrt{a^2 + b^2}$, 所以复指数的模为

$$|e^{itX}| = \sqrt{\cos^2(tX) + \sin^2(tX)} = 1$$

对于离散型变量有

$$|\phi_X(t)| = \left| \sum_X e^{itX} p(X) \right| \leq \sum_X |e^{itX}| p(X) = \sum_X p(X) = 1$$

对于连续型变量有

$$|\phi_X(t)| = \left| \int_{-\infty}^{+\infty} e^{itX} p(X) dX \right| \leq \int_{-\infty}^{+\infty} |e^{itX}| p(X) dX = \int_{-\infty}^{+\infty} p(X) dX = 1$$

对于任意随机变量 X 和实数 t , $E[e^{itX}]$ 总是存在并有限, 因此特征函数 $\phi_X(t)$ 总是存在



对于唯一性这里只证明若两个随机变量的特征函数相同，则它们有相同的密度函数⁸⁷。设对于随机变量 X, Y 有 $\phi_X(t) = \phi_Y(t)$

$$\phi_X(t) = \int_{-\infty}^{+\infty} e^{itX} p(X) dX = \int_{-\infty}^{+\infty} e^{itY} p(Y) dY = \phi_Y(t)$$

已知对任意函数 $f(x)$ 的傅里叶变换、逆傅里叶变换为

$$\hat{f}(k) = \int_{-\infty}^{+\infty} f(x) e^{-ikx} dx \text{ (傅里叶变换)}, \quad f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(k) e^{ikx} dk \text{ (逆傅里叶变换)}$$

变换和逆变换是互逆的过程，对 $f(x)$ 做傅里叶变换能得到 $\hat{f}(k)$

$$\begin{aligned} \int_{-\infty}^{+\infty} f(x) e^{-ikx} dx &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left\{ \int_{-\infty}^{+\infty} \hat{f}(k') e^{ik'x} dk' \right\} e^{-ikx} dx \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(k') \left\{ \int_{-\infty}^{+\infty} e^{ix(k'-k)} dx \right\} dk' \end{aligned}$$

⁸⁷严格来说这里应该进一步证明两个变量的分布函数相同，但它涉及到更多傅里叶变换及狄拉克函数知识，此处从略；这里仅以连续型变量为例证明



已知狄拉克函数可表示为⁸⁸

$$\delta(x - \alpha) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{ik(x-\alpha)} dk$$

所以有

$$\begin{aligned} \int_{-\infty}^{+\infty} f(x) e^{-ikx} dx &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(k') \left\{ \int_{-\infty}^{+\infty} e^{ix(k'-k)} dx \right\} dk' \\ &= \int_{-\infty}^{+\infty} \hat{f}(k') \delta(k' - k) dk' \\ &= \hat{f}(k) \end{aligned}$$

因为有 $\phi_X(t) = \phi_Y(t)$, 所以应有

$$\int_{-\infty}^{+\infty} \phi_X(t) e^{-ikt} dt = \int_{-\infty}^{+\infty} \phi_Y(t) e^{-ikt} dt \longrightarrow p(X) = p(Y)$$

⁸⁸此公式此处不证明；狄拉克函数的选值作用请参考PRML Page-by-page项目的4-050



用特征函数求矩的思路和用矩母函数求矩的思路基本一致。对 e^{itX} 作泰勒展开有

$$\begin{aligned} e^{itX} &= 1 + itX + \frac{(itX)^2}{2!} + \frac{(itX)^3}{3!} + \cdots + \frac{(itX)^n}{n!} + \cdots \\ &= 1 + itX + \frac{(it)^2}{2!} X^2 + \frac{(it)^3}{3!} X^3 + \cdots + \frac{(it)^n}{n!} X^n + \cdots \end{aligned}$$

所以特征函数可表示为

$$\phi_X(t) = E[e^{itX}] = 1 + itE[X] + \frac{(it)^2}{2!} E[X^2] + \frac{(it)^3}{3!} E[X^3] + \cdots + \frac{(it)^n}{n!} E[X^n] + \cdots$$

对特征函数求导

$$\frac{d^n \phi_X(t)}{dt^n} \Big|_{t=0} = i^n E[X^n] \rightarrow \frac{d\phi_X(t)}{dt} \Big|_{t=0} = iE[X], \quad \frac{d^2 \phi_X(t)}{dt^2} \Big|_{t=0} = i^2 E[X^2], \quad \frac{d^3 \phi_X(t)}{dt^3} \Big|_{t=0} = i^3 E[X^3]$$



已知任意函数 $f(x)$ 的傅里叶变换、逆傅里叶变换为

$$\hat{f}(k) = \int_{-\infty}^{+\infty} f(x) e^{-ikx} dx \text{ (傅里叶变换)}, \quad f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(k) e^{ikx} dk \text{ (逆傅里叶变换)}$$

所以特征函数 $\phi_X(t) = \int_{-\infty}^{+\infty} e^{itX} p(X) dX$ 可看作是密度函数 $p(X)$ 的逆傅里叶变换⁸⁹

已知任意函数 $f(x)$ 的拉普拉斯变换、拉普拉斯逆变换为

$$\hat{f}(k) = \int_{-\infty}^{+\infty} f(x) e^{-kx} dx \text{ (拉普拉斯变换)}, \quad f(x) = \frac{1}{2\pi i} \int_{\gamma-i\infty}^{\gamma+i\infty} \hat{f}(k) e^{kx} dk$$

所以矩母函数 $M_X(t) = \int_{-\infty}^{+\infty} e^{tX} p(X) dX$ 类似于密度函数 $p(X)$ 的拉普拉斯变换，只差了一个负号

以上结论告诉我们，矩母函数、特征函数相当于对密度函数做**时域、频域变换**

⁸⁹当然也可以看作是傅里叶变换，只不过是基于-的



设随机变量 X 服从正态分布 $N(X|\mu, \sigma^2)$, 求其特征函数

$$\begin{aligned}\phi_X(t) &= \int_{-\infty}^{+\infty} e^{itX} p(X) dX \\ &= \int_{-\infty}^{+\infty} e^{itX} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dX\end{aligned}$$

重点关注指数项部分

$$\begin{aligned}itX - \frac{(X-\mu)^2}{2\sigma^2} &= -\frac{X^2 - 2\mu X + \mu^2 - 2i\sigma^2 t X}{2\sigma^2} \\ &= -\frac{X^2 - 2(\mu + i\sigma^2 t)X + (\mu + i\sigma^2 t)^2 - (\mu + i\sigma^2 t)^2 + \mu^2}{2\sigma^2} \\ &= -\frac{\{X - (\mu + i\sigma^2 t)\}^2}{2\sigma^2} + \frac{2i\sigma^2 t \mu - \sigma^4 t^2}{2\sigma^2}\end{aligned}$$



所以特征函数有

$$\begin{aligned}\phi_X(t) &= \int_{-\infty}^{+\infty} e^{itX} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dX \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{2i\sigma^2 t\mu - \sigma^4 t^2}{2\sigma^2}\right) \int_{-\infty}^{+\infty} \exp\left\{-\frac{(X-(\mu+i\sigma^2 t))^2}{2\sigma^2}\right\} dX \\ &= \exp(it\mu - \frac{1}{2}\sigma^2 t^2)\end{aligned}$$

特征函数形式简洁优雅，而且包含了分布的参数 μ, σ^2



数字特征



用于描述随机变量分布特征的数字称为数字特征⁹⁰

- 难以确切获得分布函数或密度函数，甚至难以知晓变量服从的分布类型
- 有些情况只关心随机变量的部分特征
- 能更直观、更突出的展现分布特征
- 计算更简单、代价更小

两类数字特征

- 单一随机变量数字特征：期望、方差、中位数、众数、条件期望、条件方差、偏度、峰度等
- 两随机变量相互关系数字特征：相关系数、协方差、协方差阵等

⁹⁰如随机变量的平均值、分布波动程度、最大值最小值等；购买商品时更关心产品的平均寿命而非产品寿命服从的分布



期望：随机变量在所有可能取值上的加权均值，它反映了随机变量的平均或中心趋势⁹¹

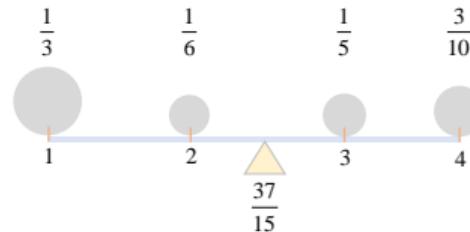
$$\text{离散型: } E[X] = \sum_x xp(x), \quad \text{连续型: } E[X] = \int xp(x)dx$$

物理上期望表示随机变量分布的概率质量中心。设有随机变量及其期望如下所示

$$p(x=1) = \frac{1}{3}, p(x=2) = \frac{1}{6}, p(x=3) = \frac{1}{5}, p(x=4) = \frac{3}{10}, \quad E[X] = \sum_x xp(x) = \frac{37}{15}$$

倘若在 x 处放置质量为 $p(x)$ 的小球，那么在 $E[X]$ 处设置支撑点两边质量刚好平衡

$$\begin{aligned} \text{Left: } & \frac{1}{3}(\frac{37}{15} - 1) + \frac{1}{6}(\frac{37}{15} - 2) = \frac{51}{90} \\ \text{Right: } & \frac{1}{5}(3 - \frac{37}{15}) + \frac{3}{10}(4 - \frac{37}{15}) = \frac{51}{90} \end{aligned}$$



⁹¹ 中心趋势（central tendency）描述数据集中位于某一位置的统计特征，它指的是某一类量而某一个量，如均值、中位数、众数都属于中心趋势



在求期望之前应先确保期望存在：若级数 $\sum_{i=1}^{\infty} x_i p(x_i)$ 绝对收敛则期望存在，否则期望不存在⁹²

- 绝对收敛蕴含条件收敛，反之不成立
- 确保期望值与变量的取值顺序无关

设有离散型随机变量 X ，取值为 $x_k = (-1)^k \frac{2^k}{k}$ ($k = 1, 2, \dots$)，对应的概率为 $p_k = \frac{1}{2^k}$ 。可验证这是一个合法的概率分布

$$\sum_{k=1}^{\infty} p_k = \sum_{k=1}^{\infty} \frac{1}{2^k} = \lim_{n \rightarrow \infty} \frac{1}{2} \frac{1 - (\frac{1}{2})^n}{1 - \frac{1}{2}} = 1$$

并且它条件收敛⁹³

$$\sum_{k=1}^{\infty} x_k p_k = \sum_{k=1}^{\infty} (-1)^k \frac{2^k}{k} \frac{1}{2^k} = \sum_{k=1}^{\infty} (-1)^k \frac{1}{k} = -\ln 2$$

但它并非绝对收敛，因为已知调和级数发散，所以该分布期望不存在

$$\sum_{k=1}^{\infty} |x_k| p_k = \sum_{k=1}^{\infty} \frac{1}{k} \rightarrow \text{发散}$$

⁹² 绝对收敛和条件收敛的关系及意义请参见本课题高等数学部分 [\(Jump to\)](#)

⁹³ 交错调和级数收敛，此处不证明，可参考本课题高等数学部分 [\(Jump to\)](#)



方差：反映了随机变量在其中心位置（期望）附近的散布程度，其定义如下

$$D(X) = E[(X - E[X])^2] = E[X^2] - E^2[X]$$

证明如下

$$\begin{aligned} D(X) &= E[(X - E[X])^2] \\ &= E[X^2 - 2XE[X] + E^2[X]] \\ &= E[X^2] - 2E[XE[X]] + E^2[X] \\ &= E[X^2] - 2E^2[X] + E^2[X] \\ &= E[X^2] - E^2[X] \end{aligned}$$

$E[X - \mu]$, $E[|X - \mu|]$ 理论上也能反应分布的散布程度，但前者存在数值抵消、后者不便计算。 $\sqrt{D(X)}$ 也称为标准差，理论上随机变量、期望、标准差三者量纲相同。



期望、方差通常被认为是常数，设有 $Y = g(X)$ 对于期望方差有

$$E[Y] = \int g(X)p(X)dX, \quad D(Y) = \int (g(X) - E[g(X)])^2 p(X)dX$$

期望具有如下性质

- 若 c 为常数，则 $E[c] = c$
- 若 k, c 为常数，则 $E[kX + c] = kE[X] + c$
- 若 $E[X]$ 和 $E[Y]$ 均存在，则 $E[X + Y] = E[X] + E[Y]$
- 若 X, Y 相互独立则 $E[XY] = E[X]E[Y]$

方差具有如下性质

- 若 c 为常数，则 $D(c) = 0$
- 若 k, c 为常数，则 $D(kX + c) = k^2D(X)$
- $D(X \pm Y) = D(X) + D(Y) \pm 2E\{[X - E[X]][Y - E[Y]]\}$
- 若 X, Y 相互独立则 $D(X \pm Y) = D(X) \pm D(Y)$



若任意随机变量 X 其 $E[X], D(X)$ 存在，令

$$X_* = X - E[X], \quad X^* = \frac{X - E[X]}{\sqrt{D(X)}}$$

X_* 称为 X 的中心化随机变量， X^* 称为 X 的标准化随机变量⁹⁴

- X_* 中心点在原点，分布既不偏左也不偏右期望为0
- X^* 不仅对 X 做了平移，还对分布做了压缩，使其分布既不疏也不密

⁹⁴深度学习中的各类正则化（normalization）就是基于这一思想



深度学习中的Batch-Normalization能有效防止网络的“内部协同漂移”⁹⁵。设有数据集 $D = \{x_1, x_2, \dots, x_n\}$, 那么

$$\mu_D = \frac{1}{n} \sum_{i=1}^n x_i, \quad \sigma_D^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_D)^2$$

再令

$$\hat{x}_i = \frac{x_i - \mu_D}{\sqrt{\sigma_D^2 + \epsilon}}, \quad y_i = \gamma \hat{x}_i + \beta$$

ϵ 是取值非常小的常量（而非服从高斯分布的变量），用于防止 σ_D^2 为零或接近于零，确保计算的稳定性。 γ 和 β 是待学习参数，它同网络其它参数（如 w ）一起被学习。它们存在的意义是恢复数据原有的特性。例如若激活函数是logistic函数，它原本是一个非线性函数但它在0附近却非常接近于线性。如果直接用 \hat{x}_i 作下一层的输入会使激活函数丧失非线性⁹⁶。

⁹⁵ Internal Covariate Shift，在神经网络中上一层的输出是下一层的输入，但在学习过程中由于网络的参数在不断变化，因此上一层输出值的特性（如范围、均值、方差等）也在不断变换，这会使得下游神经元的学习更加困难。这种情况称为内部协同漂移。Batch-Normalization可减小这一问题对模型训练的影响。

⁹⁶ 简单来说，BN包括四步：1.求均值；2.求方差；3.标准化；4.缩放和移动



对于随机变量 X : 若存在常数 c 使得 $P(X \leq c) \geq 1/2$ 且 $P(X \geq c) \geq 1/2$ 则称 c 为 X 的中位数。具体地，对于离散型变量和连续型变量中位数定义略有不同⁹⁷

- 离散型变量: 若一共有 n 种取值⁹⁸且每种取值概率相等, 对所有元素排序后

$$c_{\text{median}} = \begin{cases} x_{\frac{n}{2}+1} & (n \text{ 为奇数}) \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & (n \text{ 为偶数}) \end{cases}$$

- 连续型变量: 若 X 的分布函数为 F 则

$$c_{\text{median}} = F^{-1}(1/2)$$

设有离散型随机变量如下所示, 根据定义 $P(X \leq 5) \geq 0.5$ 且 $P(X \geq 5) \geq 0.5$, 所以其中位数为 $c_{\text{median}} = 5$

x	1	2	3	4	5	6	7	8
$p(x)$	0.1	0.1	0.15	0.05	0.2	0.2	0.15	0.05

⁹⁷若 $P(X \leq c) \geq p$ 且 $P(X \geq c) \geq 1 - p$ 称 c 为 p 分位数; 中位数就是二分位数, 常见的还有三分位数、四分位数, 但有两个三分位数、三个四分位数

⁹⁸离散型也包括可数种取值的情况, 这里仅指有限元素情况, 无限元素情况类似于连续型变量

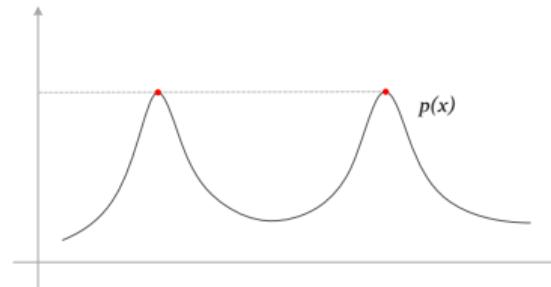


对于随机变量 X , 众数是其概率或概率密度最大的值(连续变量概率密度曲线的最高点)

$$P(X = c_{mode}) \geq P(X = x) \text{ (离散型), } p(c_{mode}) \geq p(x) \text{ (连续型)}$$

对任意分布, 期望是唯一的, 中位数、众数不一定是唯一的。设有随机变量 X 的分布函数如下, 则 $x \in [0, 1]$ 范围内的点都是该分布的中位数。若有随机变量 Y 的密度函数如下, 则它有两个众数⁹⁹。

$$F(X) = \begin{cases} 0, & \text{if } x < 0 \\ 0.5, & \text{if } 0 \leq x \leq 1 \\ 1, & \text{if } x > 1 \end{cases}$$



⁹⁹高斯分布是完美对称分布, 因为它的均值、中位数、众数为同一个值



设随机变量 X 的均值为 μ , 中位数为 c

- 若 m 使得 $E[(X - m)^2]$ 取最小值, 则一定有 $m = \mu$
- 若 m 使得 $E[|X - m|]$ 取最小值, 则一定有 $m = c$

已知根据方差定义及方差性质有

$$D(X - m) = E[(X - m)^2] - E^2[(X - m)], \quad D(X - m) = D(X)$$

所以有

$$E[(X - m)^2] = D(X) + E^2[(X - m)]$$

根据期望性质有

$$E[(X - m)] = E[X] - m = \mu - m \quad \rightarrow \quad E[(X - m)^2] = D(X) + (\mu - m)^2$$

因为 $D(X) \geq 0$, 所以要使 $E[(X - m)^2]$ 取值最小只能是 $m = \mu$



设 $L(m) = E[|X-m|] = \int |x-m|p(x)dx$, 要求使 $L(m)$ 取值最小的 m 只需要找到使 $\frac{dL(m)}{dm} = 0$ 的点, 但由于 $|x-m|$ 在 m 处不可导, 因此要分情况讨论

$$|x-m| = \begin{cases} x-m, & \text{if } x > m \\ m-x, & \text{if } x \leq m \end{cases} \rightarrow L(m) = \int_{-\infty}^m (m-x)p(x)dx + \int_m^{+\infty} (x-m)p(x)dx$$

因此

$$\frac{dL(m)}{dm} = \frac{d}{dm} \int_{-\infty}^m (m-x)p(x)dx + \frac{d}{dm} \int_m^{+\infty} (x-m)p(x)dx$$

根据莱不尼兹积分法则

$$\frac{d}{dx} \left(\int_{a(x)}^{b(x)} f(x, t)dt \right) = f(x, b(x)) \frac{db(x)}{dx} - f(x, a(x)) \frac{da(x)}{dx} + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} f(x, t)dt$$



设 x 的分布函数为 $F(x)$, 基于莱不尼兹积分法则有

$$\begin{aligned}\frac{d}{dm} \int_{-\infty}^m (m-x)p(x)dx &= (m-m)p(m)\frac{dm}{dm} - (m+\infty)p(-\infty)\frac{d(-\infty)}{dm} + \int_{-\infty}^m \frac{d}{dm}(m-x)p(x)dx \\ &= \int_{-\infty}^m p(x)dx = F(m)\end{aligned}$$

类似地, 另外一部分有

$$\begin{aligned}\frac{d}{dm} \int_m^{+\infty} (x-m)p(x)dx &= (m-\infty)p(\infty)\frac{d(+\infty)}{dm} - (m-m)p(m)\frac{dm}{dm} + \int_m^{+\infty} \frac{d}{dm}(x-m)p(x)dx \\ &= - \int_m^{+\infty} p(x)dx = -(1-F(m)) = F(m)-1\end{aligned}$$

所以

$$\frac{dL(m)}{dm} = 2F(m)-1 \longrightarrow 2F(m)-1=0 \longrightarrow F(m_{median})=\frac{1}{2}$$



“矩”借用了物理中力矩的概念，设 X 为随机变量， X 的原点矩和中心矩分别定义如下¹⁰⁰

$$E[X^k] \text{ (k阶原点矩)}, \quad E[(X - E[X])^k] \text{ (k阶中心矩)}$$

要求 k 阶矩，应首先证明 k 阶矩存在即 $E[X^k] < \infty$ 或者 $E[(X - E[X])^k] < \infty$ ¹⁰¹。

设 X 和 Y 为两个随机变量，若对任意正整数 n 都有 $E[X^n] = E[Y^n]$ ，则 X 和 Y 表示同一个分布。虽然用矩可以唯一地表示分布，但它具有如下问题

- 矩的存在性：并非所有的分布有所有阶矩。有些分布没有高阶矩，有些分布没有任何矩。
- 分布重建性：即使各阶矩都存在，但要利用矩来重建整个分布通常仍然比较困难。

但矩也并非无用，可以用矩来估计模型中参数的值，这类方法被称为“**矩估计**”是点估计中的重要一类。通常我们求正态分布的均值、方差就是用的矩估计¹⁰²。

¹⁰⁰ 原点矩、中心矩其命名分别来源于物理中的质量中心和惯性矩；很明显，均值就是一阶原点矩，方差就是二阶中心矩

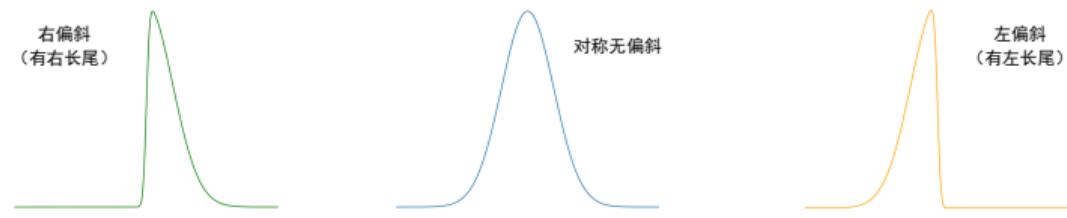
¹⁰¹ 理论上要通过积分求矩，但前面介绍过利用矩母函数求导也可求矩

¹⁰² 关于参数估计，在本课题的统计部分有详细介绍



正态分布是完美对称分布，偏度（Skewness）量化一个分布对称性缺失程度，或者说量化分布的偏斜程度

$$\text{Skew}(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$



数值上，偏度既可以看作是 X 标准化之后的三阶原点矩，也可以看作是 X 三阶中心矩的标准化¹⁰³

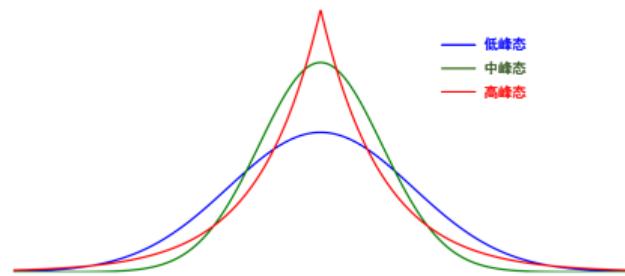
- 右偏也称正偏：金融中表示有较多较大正收益；左偏也称负偏：金融中表示负收益较多，亏损风险较高
- 方差只表示波度大小，偏度能表示偏移方向。医疗、生命科学、金融等学科中偏度重要性甚至可能高过方差

¹⁰³ $\text{Skew}(x) = E[(x-\mu)^3]/\sigma^3$



峰度（Kurtosis）用于量化分布的陡峭平缓程度，标准正态分布的峰度值等于3称为中峰态（Mesokurtic），若峰度比正态分布小称为低峰态（Platykurtic），若峰度比正态分布大称为高峰态（Leptokurtic）

$$\text{Kurt}(X) = E\left[\left(\frac{X - \mu}{\sigma}\right)^4\right]$$
$$\text{Kurt}(X) = E\left[\left(\frac{X - \mu}{\sigma}\right)^4\right] - 3$$



峰度主要随机变量的四阶矩决定。标准正态分布的峰度值为3，因此有时候也定义超额峰度（excess kurtosis）

- 高峰度峰值较尖同时尾部较厚，低峰度恰好相反
- 高峰度时极端值出现可能性较大



设 x 服从标准正态分布 $N(0, 1)$, 求其峰度。因为标准正态分布是对称的, 所以根据定义有

$$E[x^4] = \int_{-\infty}^{+\infty} x^4 f(x) dx \rightarrow E[x^4] = 2 \int_0^{+\infty} x^4 \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) dx$$

令 $u = \frac{x^2}{2}$, 因而有 $4u^2 = x^4$ 且两边求微分 $du = xdx$, 利用换元法有¹⁰⁴

$$E[X^4] = 2 \int_0^{\infty} 4u^2 \frac{1}{\sqrt{2\pi}} \exp(-u) \frac{1}{\sqrt{2u}} du = \frac{4}{\sqrt{\pi}} \int_0^{\infty} u^{\frac{3}{2}} \exp(-u) du$$

已知伽玛函数是一般意义上的阶乘且有以下性质

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt, \quad \Gamma(z+1) = z\Gamma(z) \quad \text{且} \quad \Gamma(\frac{1}{2}) = \sqrt{\pi}$$

所以

$$E[X^4] = \frac{4}{\sqrt{\pi}} \Gamma(\frac{5}{2}) = \frac{4}{\sqrt{\pi}} \frac{3}{2} \Gamma(\frac{3}{2}) = \frac{4}{\sqrt{\pi}} \frac{3}{2} \frac{1}{2} \Gamma(\frac{1}{2}) = \frac{4}{\sqrt{\pi}} \frac{3}{2} \frac{1}{2} \sqrt{\pi} = 3$$

¹⁰⁴ 此时积分范围为 $(0, +\infty)$, 所以 $x = \sqrt{2u}$

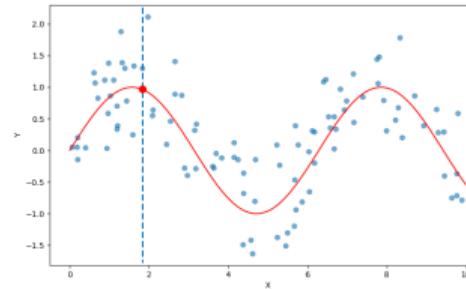
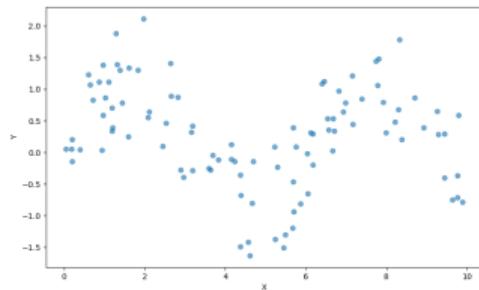


设有随机变量 X, Y , 则已知 $X = x$ 时 Y 的条件期望 $E[Y|X = x]$ 如下所示, 条件期望和期望的区别在于概率密度变成了条件概率密度

$$E[Y|X = x] = \int_{-\infty}^{+\infty} y p_{y|x}(y|x) dy$$

类似的条件方差 $D(Y|X = x)$ 定义如下¹⁰⁵

$$D(Y|X = x) = \int_{-\infty}^{+\infty} [y - E[Y|X = x]]^2 p_{y|x}(y|x) dy$$



¹⁰⁵注意式子中的 x 是一个特定值, 而 y 是变量

条件期望和条件方差的常用性质¹⁰⁶

- 若 X, Y 相互独立则 $E[Y|X] = E[Y]$
- 对任意函数 h 有 $E[h(X) \cdot Y|X] = h(X)E[Y|X]$
- 对任意 Y_1, Y_2 有 $E[Y_1 + Y_2|X] = E[Y_1|X] + E[Y_2|X]$
- 总是有 $E[E[Y|X]] = E[Y]$
- $D(Y|X) = E[Y^2|X] - E^2[Y|X]$
- $D(Y) = E[D(Y|X)] + D(E[Y|X])$

根据条件方差定义有

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

¹⁰⁶ $E[E[Y|X]] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} yp(y|x)dydp(x)dx = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} yp(x,y)dxdy = \int_{-\infty}^{+\infty} yp(y)dy = E[Y]$; 关于条件方差的第二条性质，只需要按定义展开后做变形化简就可得到，此处不再证明



协方差是表现两个变量相互关系的数字特征，设有 X, Y 两个随机变量，则它们的协方差 $\text{Cov}(X, Y)$ 定义为

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] \iff \text{Cov}(X, Y) = E[XY] - E[X]E[Y]$$

证明

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - XE[Y] - YE[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

很明显

$$\text{Cov}(X, X) = E[X^2] - E^2[X] = D(X)$$



令 $Z = (X - E[X])(Y - E[Y])$, 那么 $\text{Cov}(X, Y) = E[Z]$

- 若 $\text{Cov}(X, Y) > 0$ 说明 $\{Z > 0\}$ 可能性较大, 相当于 $\{X > E[X]\} \& \{Y > E[Y]\}$ 或 $\{X < E[X]\} \& \{Y < E[Y]\}$ 可能性较大
- 若 $\text{Cov}(X, Y) < 0$ 说明 $\{Z < 0\}$ 可能性较大, 相当于 $\{X > E[X]\} \& \{Y < E[Y]\}$ 或 $\{X < E[X]\} \& \{Y > E[Y]\}$ 可能性较大

因而 $\text{Cov}(X, Y)$ 反映了随机变量 X, Y 之间的协同变化关系

协方差的基本性质

- $\text{Cov}(X, c) = 0$
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- $\text{Cov}(kX, lY) = kl\text{Cov}(X, Y)$
- $\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y)$



设 (X, Y) 联合分布的概率密度函数如下所示，求 $\text{Cov}(X, Y)$

$$p(x, y) = \begin{cases} \frac{1}{\pi}, & x^2 + y^2 < 1 \\ 0, & \text{其它} \end{cases}$$

先求 $p_X(x)$ 及 $E[X]$

$$p_X(x) = \begin{cases} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi} dy = \frac{1}{\pi} \frac{x^2}{2} \Big|_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} = \frac{2}{\pi} \sqrt{1-x^2}, & -1 < x < 1 \\ 0, & \text{其它} \end{cases}$$

所以¹⁰⁷

$$E[X] = \int_{-1}^1 x \cdot \frac{2}{\pi} \sqrt{1-x^2} dx = \frac{2}{\pi} \int_{-1}^1 x \sqrt{1-x^2} dx = 0$$

同理有

$$p_Y(y) = \begin{cases} \frac{2}{\pi} \sqrt{1-y^2}, & -1 < y < 1 \\ 0, & \text{其它} \end{cases}, \quad E[Y] = 0$$

¹⁰⁷ $x \sqrt{1-x^2}$ 是奇函数且积分区间对称，所以积分为0



又因为有¹⁰⁸

$$E[XY] = \iint_{x^2+y^2 \leq 1} xy \frac{1}{\pi} dx dy = \frac{1}{\pi} \int_{-1}^1 y dy \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} x dx = 0$$

所以

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = 0$$

协方差对相关性的表现在数值上不稳定。例如设 X, Y 分别表示新生儿的身高体重，若 X, Y 的单位为厘米和克， $\text{Cov}(X, Y)$ 的取值会比单位为米、千克时大100000倍 (10^6)，但实际上 X, Y 的相关性是一样的。解决办法是将 X, Y 标准化后再求协方差，这一结果就是相关系数。

¹⁰⁸很明显 x 是奇函数且积分域对称



设有随机变量 X, Y 且它们各自的期望方差存在，将 X, Y 标准化

$$X^* = \frac{X - E[X]}{\sqrt{D(X)}}, \quad Y^* = \frac{Y - E[Y]}{\sqrt{D(Y)}}$$

X, Y 的相关系数 $\rho(X, Y)$ 定义为

$$\rho(X, Y) = E[X^* Y^*] = E\left[\frac{X - E[X]}{\sqrt{D(X)}} \frac{Y - E[Y]}{\sqrt{D(Y)}}\right] = \frac{E[(X - E[X])(Y - E[Y])]}{\sqrt{D(X)} \sqrt{D(Y)}} = \frac{\text{Cov}(X, Y)}{\sqrt{D(X)} \sqrt{D(Y)}}$$

相关系数 $\rho(X, Y)$ 的重要特性

- $-1 \leq \rho(X, Y) \leq 1$
- $\rho(X, Y) = 1$ 或 $\rho(X, Y) = -1$ 时分别表示 X, Y 有确定的正、负线性关系



要证明这两点需要用到柯西-施瓦茨不等式，设 f, g 在 $[a, b]$ 内连续则

$$\left(\int_a^b f(t)g(t)dt \right)^2 \leq \int_a^b f^2(t)dt \int_a^b g^2(t)dt$$

在概率论中该不等式的形式为

$$\{E[(X - E[X])(Y - E[Y])]\}^2 \leq E[(X - E[X])^2]E[(Y - E[Y])^2]$$

所以有

$$\text{Cov}^2(X, Y) \leq D(X)D(Y)$$

$$|\text{Cov}(X, Y)| \leq \sqrt{D(X)} \sqrt{D(Y)}$$



$$\begin{aligned} -\sqrt{D(X)} \sqrt{D(Y)} &\leq \text{Cov}(X, Y) \leq \sqrt{D(X)} \sqrt{D(Y)} \\ -1 &\leq \frac{\text{Cov}(X, Y)}{\sqrt{D(X)} \sqrt{D(Y)}} \leq 1 \end{aligned}$$

柯西-施瓦茨不等式等号成立的条件是 $X - E[X]$ 和 $Y - E[Y]$ 之间存在确定的线性关系¹⁰⁹，即

$$X - E[X] = a(Y - E[Y]) \quad \text{或} \quad Y - E[Y] = b(X - E[X])$$

这两种情况都相当于

$$X = kY + c$$

¹⁰⁹这一点已经被证明



验证充分性：即若 $X - E[X] = a(Y - E[Y])$ 那么 $\rho(X, Y) = \pm 1$ 。很明显此时有

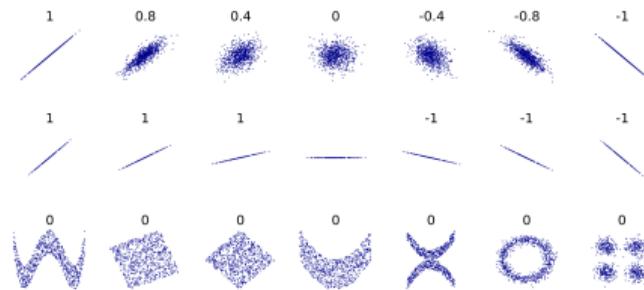
$$\begin{aligned}\{E[(X - E[X])(Y - E[Y])]\}^2 &= \{E[a(X - E[X])^2]\}^2 \\ &= a^2 E^2[(X - E[X])^2] \\ &= E[(X - E[X])^2] E^2[(a(X - E[X]))^2] \\ &= E[(X - E[X])^2] E[(Y - E[Y])^2]\end{aligned}$$

验证必要性：即若 $\rho(X, Y) = \pm 1$ 则 $X - E[X] = a(Y - E[Y])$ 。很明显，当 $\rho(X, Y) = \pm 1$ 时柯西-施瓦茨不等式等号成立，那么一定有 $X - E[X] = a(Y - E[Y])$

当 $\rho(X, Y) = \pm 1$ 时，表示 X, Y 具有正、负线性关系；若 $\rho(X, Y) = 0$ 表示 X, Y 是两个不线性相关（一般简称不相关）的随机变量，并不是表示它们完全独立； $\rho(X, Y)$ 越接近 ± 1 表示 X, Y 的线性相关性越强。



相关系数只是描述 X, Y 相互关系的一个数字特征，它并不能完全概括 X, Y 之间的关系； X, Y 独立表示从整体上两个变量无关，所以独立一定无关但无关不一定独立。一个例外是：若 X, Y 均服从[高斯分布](#)，则不相关等同于独立。



下面四个命题是等价的¹¹⁰

- (1) $\rho(X, Y) = 0$
- (2) $\text{Cov}(X, Y) = 0$
- (3) $E[XY] = E[X]E[Y]$
- (4) $D(X \pm Y) = D(X) + D(Y)$

¹¹⁰若 $\text{Cov}(X, Y) = 0$ 根据定义有 $0 = E[XY] - E[X]E[Y]$ ，其它几项均可根据定义证明



设 X, Y 是随机变量， k, l 是正整数，联合原点矩、联合中心矩的定义如下

$$E[X^k Y^l] \text{ ((k,l)阶联合原点矩)} \quad E[(X - E[X])^k (Y - E[Y])^l] \text{ ((k,l)阶联合中心矩)}$$

联合矩在机器学习中用途广泛，例如可以用来提取图像的全局特征¹¹¹。设用 (x, y) 表示图像的坐标， $f(x, y)$ 表示该处的图像强度（RGB值），图像的 (k, l) 阶矩阵定义如下

$$M^{k,l} = \sum_x \sum_y x^k y^l f(x, y)$$

可令 $F = (M^{1,1}, M^{2,0}, M^{0,2}, M^{2,1}, M^{1,2}, M^{2,2})^T$ 作为图像的特征向量

¹¹¹因为分布的矩是唯一的，所以可用作特征



重要概率分布



设有 (a, b) 范围内的一致分布 $U(a, b)$, 其密度函数为 $f(x) = \frac{1}{b-a}$, $a < x < b$, 期望、方差、矩母函数分别为

$$E[X] = \int_a^b x \frac{1}{b-a} dx = \frac{1}{b-a} \frac{x^2}{2} \Big|_a^b = \frac{1}{2(b-a)} (b^2 - a^2) = \frac{a+b}{2}$$

要求 $D[X]$ 先求 $E[X^2]$

$$E[X^2] = \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{b-a} \frac{x^3}{3} \Big|_a^b = \frac{1}{3(b-a)} (b^3 - a^3) = \frac{b^2 + ab + a^2}{3}$$

所以

$$\begin{aligned} D[X] &= E[X^2] - E^2[X] = \frac{b^2 + ab + a^2}{3} - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{4b^2 + 4ab + 4a^2 - 3a^2 - 6ab - 3b^2}{12} = \frac{(b-a)^2}{12} \end{aligned}$$



已知矩母函数定义为 $M_X(t) = E[e^{tX}]$, 所以有

$$\begin{aligned} M_X(t) &= \int_a^b e^{tx} \frac{1}{b-a} dx = \frac{1}{b-a} \frac{e^{tx}}{t} \Big|_a^b \\ &= \frac{1}{b-a} \left(\frac{e^{bt}}{t} - \frac{e^{at}}{t} \right) \\ &= \frac{e^{bt} - e^{at}}{t(b-a)} \end{aligned}$$

一些重要用途及性质

- 可作为分布位置参数的无信息先验，即使是反常先验很多时候也能正常使用
- 是所有采样方法的基础¹¹²

¹¹²例如可基于一致分布按Muller-Box方法采样得到服从高斯分布样本，具体方法请参见PRML Page-by-page的11-010



在概率论的早期研究发展中，伯努利分布曾起到重要作用。设随机实验只有事件发生、不发生两种结果，用 $\{x = 1\}$ 表示事件发生概率为 p , $\{x = 0\}$ 表示事件不发生概率为 $1 - p$, 分布的概率质量函数可写成

$$P(x) = p^x(1-p)^{1-x}, x \in \{0, 1\}, p \in (0, 1)$$

分布期望为

$$E[X] = \sum_x x p(x) = \{x = 1\}p + \{x = 0\}(1 - p) = p$$

要求方差先求 $E[X^2]$

$$E[X^2] = \sum_x x^2 p(x) = \{x = 1\}^2 p + \{x = 0\}^2 (1 - p) = p$$

所以

$$D(X) = E[X^2] - E^2[X] = p - p^2 = p(1 - p)$$

其矩母函数为

$$M_X(t) = E[e^{tX}] = \sum_x e^{tx} p(x) = e^{t\{x=1\}}p + e^{t\{x=0\}}(1 - p) = e^t p + 1 - p$$



伯努利分布只做一次实验，如果连续做 n 次实验则称为二项分布（Binomial distribution）。若做 n 次二元实验，有 x 次事件发生 $n - x$ 次不发生，其概率为

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

其期望为

$$\begin{aligned} E[X] &= \sum_x x \binom{n}{x} p^x (1-p)^{n-x} = \sum_x x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= np \sum_x \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x} = np \end{aligned}$$

倒数第二步时令 $k = n - 1, u = x - 1$

$$\sum_x \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x} = \sum_u \frac{k!}{u!(k-u)!} p^u (1-p)^{k-u} = \binom{k}{u} p^u (1-p)^{k-u} = 1$$



先求 $E[X^2]$

$$\begin{aligned} E[X^2] &= \sum_x x^2 \binom{n}{x} p^x (1-p)^{n-x} = \sum_x [x(x-1) + x] \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_x x(x-1) \binom{n}{x} p^x (1-p)^{n-x} + \sum_x x \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_x x(x-1) \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} + np \\ &= n(n-1)p^2 \sum_x \frac{(n-2)!}{(x-2)!(n-x)!} p^{x-2} (1-p)^{n-x} + np \\ &= n(n-1)p^2 \sum_u \frac{k!}{u!(k-u)!} p^u (1-p)^{k-u} + np \\ &= n(n-1)p^2 + np \end{aligned}$$

所以方差为

$$D(X) = E[X^2] - E^2[X] = n(n-1)p^2 + np - n^2p^2 = np - np^2 = np(1-p)$$



其矩母函数为

$$M_X(t) = E[e^{tX}] = \sum_x e^{tx} \binom{n}{x} p^x (1-p)^{n-x} = \sum_x \binom{n}{x} (e^t p)^x (1-p)^{n-x}$$

根据二项定理有

$$(a+b)^n = \sum_k \binom{n}{k} a^k b^{n-k}$$

所以

$$M_X(t) = \sum_x \binom{n}{x} (e^t p)^x (1-p)^{n-x} = (pe^t + 1 - p)^n$$

二项分布在深度学习中的应用¹¹³

- 若神经元采用logistic激活函数，其输出服从二项分布
- 交叉熵损失函数就是二项分布去掉排列数后取对数

¹¹³ $\sigma(z) = P(z=1|x)$ 可看作 z 事件发生的概率，当有多个独立同分布的数据项时神经元的输出就服从二项分布；交叉熵损失函数 $L = -\frac{1}{m} \sum_{i=1}^m [t_i \log y_i + (1-t_i) \log(1-y_i)]$



泊松分布是一种离散分布，它的参数为 λ ($\lambda > 0$)，设 $X = 0, 1, 2, \dots$ ，其概率质量函数为

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

求期望

$$E[X] = \sum_x x \frac{\lambda^x}{x!} e^{-\lambda} = \lambda \sum_x \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!} = \lambda \sum_j \frac{\lambda^k e^{-\lambda}}{k!} = \lambda$$

求 $E[X^2]$ 有

$$E[X^2] = \sum_x x^2 \frac{\lambda^x}{x!} e^{-\lambda} = \lambda \sum_x x \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!} = \lambda \left\{ \sum_x (x-1) \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!} + \sum_x \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!} \right\} = \lambda(\lambda + 1)$$

所以

$$D(X) = E[X^2] - E^2[X] = \lambda^2 + \lambda - \lambda^2 = \lambda$$



其矩母函数为

$$M_X(t) = E[e^{tX}] = \sum_x e^{tx} \frac{\lambda^x}{x!} e^{-\lambda} = e^{-\lambda} \sum_x e^{tx} \frac{\lambda^x}{x!} = e^{-\lambda} \sum_x \frac{(\lambda e^t)^x}{x!}$$

已知 e^x 的泰勒展开式为

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots$$

所以

$$e^{\lambda e^t} = 1 + \lambda e^t + \frac{(\lambda e^t)^2}{2!} + \frac{(\lambda e^t)^3}{3!} + \frac{(\lambda e^t)^n}{n!} + \cdots = \sum_x \frac{(\lambda e^t)^x}{x!}$$

所以

$$M_X(t) = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}$$



在 n 重伯努利试验中（二项分布），若事件发生的概率为 p ，若 n 取值很大而 p 较小且 $np = \lambda$ 取值适中¹¹⁴，那么二项分布的概率值约等于泊松分布

$$\lim_{n \rightarrow +\infty} \binom{n}{k} p^k (1-p)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}, \quad \lambda = np$$

泊松分布是一种常用概率分布（需提前确定参数 λ 的值）

- 某一时段内网站的点击量
- 某公交站上车人数数量
- 书本上印刷错误数

¹¹⁴此条件很重要



设 X 为随机变量，指数分布的概率密度函数如下所示

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0, \quad \lambda > 0$$

概率分布函数

$$F(x) = \int_0^x \lambda e^{-\lambda x} dx = -e^{-\lambda x} \Big|_0^x = 1 - e^{-\lambda x}$$

指数分布具有重要特性——无记忆性，即 $P(X > s + t | X > s) = P(X > t)$ ，证明如下

$$P(X > s + t | X > s) = \frac{P(X > s + t)}{P(X > s)} = \frac{1 - P(X \leq s + t)}{1 - P(X \leq s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t)$$



指数分布常用于对电子元器件寿命或可靠性建模。例如，已知某元件寿命服从指数分布，根据无记忆性，刚出厂的元件稳定使用 t 小时的概率等于使用 s 小时后再稳定使用 t 小时的概率

- 为什么旧产品比新产品更容易坏？
- 还有一些其它分布也具有无记忆性，如几何分布
- 注意区分指数分布和指数族分布
- $E[X] = \frac{1}{\lambda}$, $D(X) = \frac{1}{\lambda^2}$, $M_X(t) = \frac{\lambda}{\lambda+t} (t < \lambda)$
- 利用指数函数的非负性可以构建任意概率分布



高斯分布是最重要的概率分布，它是完美对称分布（均值、众数、中位数是同一个点），其密度函数为¹¹⁵

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\}$$

设 x, y 是两个独立的高斯变量且 $x \sim \mathcal{N}(x|\mu_x, \sigma_x^2), y \sim \mathcal{N}(y|\mu_y, \sigma_y^2)$ ，则 $z = x + y$ 是 x, y 的卷积， z 也服从高斯分布

$$p(z) \sim \mathcal{N}(z|\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

$z = x - y$ 可看作 $z = x + (-y)$ 所以 z 也服从高斯分布

$$p(z) \sim \mathcal{N}(z|\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2)$$

¹¹⁵最重要没有之一，无论从任何角度去看它都非常重要；它的均值为 μ 方差为 σ^2 矩母函数为 $\exp\{\frac{1}{2}\sigma^2 t^2 + \mu t\}$ ；关于高斯分布请参考PRML重点问题选讲（总结了高斯分布的14条重要性质）的Chapter2或PRML Page-by-page项目的2-050



高斯分布经仿射变换后仍服从高斯分布：设 $x \sim N(x|\mu, \sigma^2)$ ，若 $y = kx + b$ 则 y 服从高斯分布

$$p(y) \sim N(y|k\mu + b, k^2\sigma^2)$$

对于多元变量若 $\mathbf{x} \sim N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ ，则有

$$\mathbf{y} \sim N(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$$

高斯分布其它重要性质¹¹⁶

- 对于多维高斯分布，两个变量不相关等同于两个变量相互独立
- 两个高斯变量的商可以写成类似于高斯分布的形式，但不一定是合法的高斯变量
- 若 $f_1(x), f_2(x)$ 是两个不同的高斯密度函数，则 $f_1(x)f_2(x)$ 是一个带缩放因子的高斯分密度函数
- 多维高斯分布的边缘分布、条件分布也是高斯分布

¹¹⁶ 高斯分布是最重要的分布，它的性质也是最多的，关于高斯分布请一定要参考前面给出的参考资料

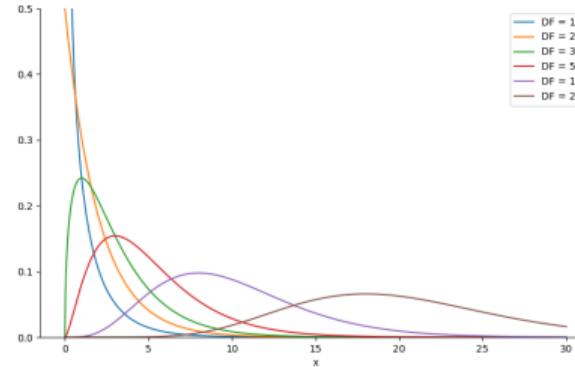


χ^2 分布、t 分布、F 分布是统计中的“三大分布”。设 X_1, X_2, \dots, X_n 是来自总体 $N(0, 1)$ 的样本，则称统计量 χ^2 服从自由度为 n 的 χ^2 分布（自由度指包含独立变量的个数），记作 $\chi^2 \sim \chi^2(n)$ ¹¹⁷

$$\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$

设 $y \sim \chi^2(n)$ 密度函数为

$$p(y) = \frac{1}{2^{n/2}\Gamma(n/2)} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}, \quad y > 0$$



¹¹⁷也称“三大”抽样分布，关于什么是抽样分布会在下一节介绍；注意这里是变量平方的和，前面高斯卷积是变量的和

 χ^2 分布的重要性质

- 可加性: 若 $\chi_1^2 \sim \chi^2(n_1), \chi_2^2 \sim \chi^2(n_2)$ 并且 χ_1^2, χ_2^2 相互独立¹¹⁸, 那么 $\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2)$
- χ^2 分布自由度越大越接近正态分布, 这一点由中心极限定理保证
- $E[\chi^2] = n, D(\chi^2) = 2n, M_X(t) = (\frac{1}{1-2t})^{\frac{n}{2}}, t < \frac{1}{2}$

已知构成 χ^2 的任意 $X_i \sim N(0, 1)$, 所以 $E[X_i^2] = D(X_i) + E^2[X_i] = 1$, 另外 $E[X_i^4]$ 其实就是 X_i 的峰度, 我们在前面已推导过 $E[X_i^4] = 3$ (Jump to)

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

根据方差定义有 $D(X_i^2) = E[X_i^4] - E^2[X_i^2] = 3 - 1 = 2$, 所以

$$D(\chi^2) = D\left(\sum_{i=1}^n X_i^2\right) = \sum_{i=1}^n D(X_i^2) = 2n$$

¹¹⁸ χ_1^2 和 χ_2^2 中没有重复的 X_i



t-分布也称学生氏分布 (Student分布)，设 $X \sim N(0, 1)$, $Y \sim \chi^2(n)$ 且 X, Y 相互独立，则称随机变量 t 服从自由度为 n 的t-分布，记为 $t \sim t(n)$

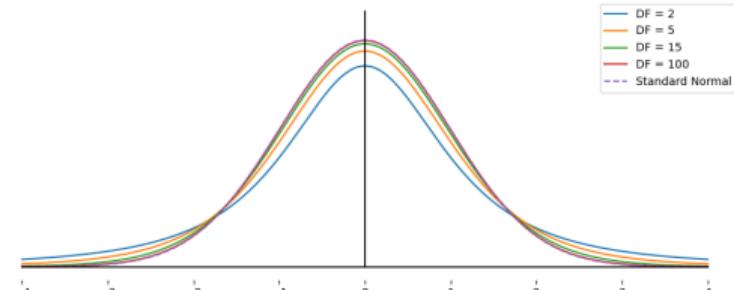
$$t = \frac{X}{\sqrt{Y/n}}$$

当 $n \rightarrow \infty$ 时t-分布将趋向于正态分布。t-分布具有比正态分布更强的鲁棒性，因此在有些时候利用它可以提高对离群点、奇异点的抵抗力

$$\lim_{n \rightarrow \infty} p(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

设 $y \sim t(n)$ 密度函数为

$$p(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{\pi n} \Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$



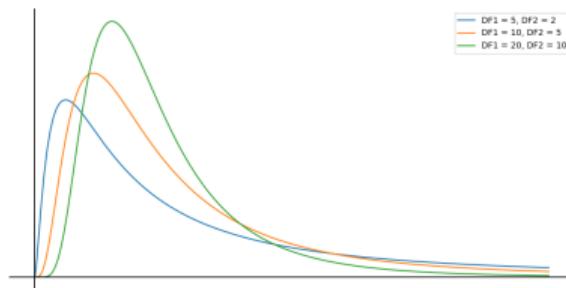


设 $U \sim \chi^2(n_1)$, $V \sim \chi^2(n_2)$ 且 U, V 相互独立, 则称随机变量 F 服从自由度为 (n_1, n_2) 的 F 分布, 记为 $F \sim F(n_1, n_2)$

$$F = \frac{U/n_1}{V/n_2}$$

设 $t \sim F(n_1, n_2)$ 密度函数为

$$p(t) = \frac{\Gamma[(n_1 + n_2)/2](n_1/n_2)^{n_1/2} t^{(n_1/2)-1}}{\Gamma(n_1/2)\Gamma(n_2/2)[1 + (n_1 t/n_2)]^{(n_1+n_2)/2}}, \quad t > 0$$





数理统计



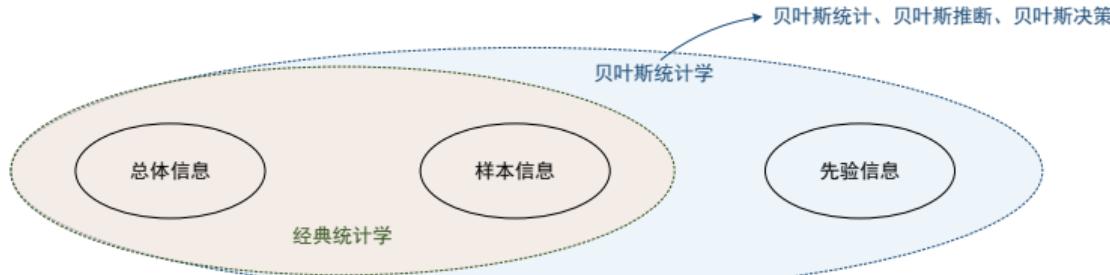
数理统计：使用概率论和其它数学方法，研究如何有效的收集带有随机误差的样本，并对收集的数据进行分析、挖掘，形成有用的结论以为实际应用提供决策依据。





统计推断主要利用三种信息

- 总体信息：将研究对象看作随机变量，变量的总体分布的相关信息，如分布类型；主要通过试验、经验、假设获取，往往代价巨大
- 样本信息：从总体中抽取样本，由样本所提供的信息；是最重要的信息之一
- 先验信息：在抽样之前就已知或预设的关于参数的信息；一般源自经验或某些特殊选择¹¹⁹



¹¹⁹例如前面介绍的无信息先验、共轭先验



研究有关对象的某一项数量指标，对该指标进行试验或观察，所有可能的观察值称为**总体**；每一个可能的观察值称为**个体**；总体中包含的个体数量称为总体的容量，包括有限总体和无限总体。

- 某校有3000名学生，每位学生的身高是一个观察值，所有学生的身高是总体；这是有限总体
- 全中国有14亿人，所有人的身高是总体；理论上是有限总体，但实际上可认为是无限总体

可将总体看作随机变量 X ，在相同条件下对总体 X 进行 n 次重复独立的观察，将结果记为 X_1, X_2, \dots, X_n ，可认为 X_1, X_2, \dots, X_n 是独立同分布的¹²⁰，这样的样本也称为**简单随机样本**。观察完成后得到 X_1, X_2, \dots, X_n 的具体值 x_1, x_2, \dots, x_n ，称为**样本值或样本观测值**。

¹²⁰ 教材上要求随机样本具有两个特性：1.相互独立性，也就是这里所说的独立性；2.代表性，就是要求所有样本服从同一个分布；尽管大多数情况下我们都希望或要求样本是独立同分布的，但也有很多情况样本不是独立同分布的，如序列化样本。序列化样本的一个典型应用就是马尔可夫链，如Diffusion模型中的前后向过程。PRML的第13章主要研究这类问题



设 X_1, X_2, \dots, X_n 是来自总体 X 的一个样本， $g(X_1, X_2, \dots, X_n)$ 是关于 X_1, X_2, \dots, X_n 的函数。若 g 中不含任何未知参数，则称 $g(X_1, X_2, \dots, X_n)$ 是一个统计量。

常用的统计量

- 样本均值： $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
- 样本方差： $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ (样本标准差： $S = \sqrt{S^2}$)
- 样本 k 阶 (原点) 矩： $A_k = \frac{1}{n} \sum_{i=1}^n X_i^k, k = 1, 2, \dots$
- 样本 k 阶中心矩： $B_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k, k = 2, 3, \dots$

X_1, X_2, \dots, X_n 可看作独立同分布的随机变量，所以统计量也可看作随机变量，它（统计量）服从的分布称为**抽样分布**。利用统计量做推断时，常需要知道它的分布。当总体分布已知时抽样分布是确定的，但要求出统计量的精确分布一般仍然是困难的。数理统计中的三大分布： χ^2 分布、 t -分布、 F 分布，都是来自**正态总体**的常用统计量分布。



X_1, X_2, \dots, X_n 是取自正态总体 $N(\mu, \sigma^2)$ 的样本，样本均值 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 和样本方差 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 有

- $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, 即 $\frac{\sqrt{n}}{\sqrt{\sigma^2}}(\bar{X} - \mu) \sim N(0, 1)$
- $\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2(n-1)$, 即 $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
- \bar{X} 与 S^2 相互独立
- $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$

性质1证明：根据前面介绍，独立高斯变量之和仍服从高斯分布，均值、方差为各变量均值方差之和。所以 \bar{X} 服从高斯分布，并且

$$E[\bar{X}] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \mu, \quad D(\bar{X}) = D\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n D(X_i) = \frac{\sigma^2}{n}$$



性质2证明：令 $Z_i = \frac{X_i - \mu}{\sigma}$ ，可见 Z_1, Z_2, \dots, Z_n 相互独立且都服从 $N(0, 1)$ 分布， $\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i = \frac{\bar{X} - \mu}{\sigma}$

$$\begin{aligned}\frac{n-1}{\sigma^2} S^2 &= \sum_{i=1}^n \frac{1}{\sigma^2} (X_i - \bar{X})^2 = \sum_{i=1}^n \left[\frac{(X_i - \mu) - (\bar{X} - \mu)}{\sigma} \right]^2 \\ &= \sum_{i=1}^n (Z_i - \bar{Z})^2 = \sum_{i=1}^n Z_i^2 - 2\bar{Z} \sum_{i=1}^n Z_i + n\bar{Z}^2 = \sum_{i=1}^n Z_i^2 - n\bar{Z}^2\end{aligned}$$

令 $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T$, $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)^T$, 其中 $\mathbf{Y} = \mathbf{A}\mathbf{Z}$, 要求 \mathbf{A} 是正交方阵且第一行元素均为 $\frac{1}{\sqrt{n}}$

$$\mathbf{A} = \begin{pmatrix} 1/\sqrt{n} & 1/\sqrt{n} & \cdots & 1/\sqrt{n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}, \quad \mathbf{A}^T \mathbf{A} = \mathbf{A} \mathbf{A}^T = \mathbf{I}, \quad \sum_{k=1}^n a_{ik} a_{jk} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$



很明显 $Y_i = \sum_{k=1}^n a_{ik} Z_k$ 也服从高斯分布，并且

$$E[Y_i] = E\left(\sum_{k=1}^n a_{ik} Z_k\right) = \sum_{j=1}^n a_{ik} E[Z_k] = 0$$

$$\text{Cov}(Y_i, Y_j) = \text{Cov}\left(\sum_{k=1}^n a_{ik} Z_k, \sum_{l=1}^n a_{jl} Z_l\right) = \sum_{k=1}^n \sum_{l=1}^n a_{ik} a_{jl} \text{Cov}(Z_k, Z_l) = \sum_{t=1}^n a_{it} a_{jt} = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

可见 $\mathbf{Y} \sim N(\mathbf{0}, \mathbf{I})$ ，所以任意 $Y_i \sim N(0, 1)$ ，并且对于 Y_1 和 Y_i 有

$$Y_1 = \sum_{k=1}^n a_{1k} Z_k = \sum_{k=1}^n \frac{1}{\sqrt{n}} Z_k = \sqrt{n} \bar{Z}$$

$$\sum_{i=1}^n Y_i^2 = \mathbf{Y}^T \mathbf{Y} = (\mathbf{A} \mathbf{Z})^T (\mathbf{A} \mathbf{Z}) = \mathbf{Z}^T (\mathbf{A}^T \mathbf{A}) \mathbf{Z} = \mathbf{Z}^T \mathbf{Z} = \sum_{i=1}^n Z_i^2$$



所以有

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n Z_i^2 - n\bar{Z}^2 = \sum_{i=1}^n Y_i^2 - Y_1^2 = \sum_{i=2}^n Y_i^2$$

由于 Y_2, \dots, Y_n 相互独立且 $Y_i \sim N(0, 1)$ ¹²¹, 所以 $\sum_{i=2}^n Y_i^2$ 服从 $\chi^2(n-1)$, 所以

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

性质3证明：因为 $\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i = \frac{\bar{X}-\mu}{\sigma}$, 所以 $\bar{X} = \sigma\bar{Z} + \mu = \frac{\sigma}{\sqrt{n}} Y_1 + \mu$, 所以很明显 \bar{X} 只跟 Y_1 有关, 而前面推导后 S^2 只跟 Y_2, Y_3, \dots, Y_n 有关, 而 Y_i 之间是独立的, 所以 \bar{X} 和 S^2 是独立的

¹²¹ 前面已推导 \mathbf{Y} 的协方差阵为 I , 所以 Y_i 之间相互独立



性质4证明：根据前面三条性质

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1), \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \quad (\text{二者相互独立})$$

根据t-分布定义可知

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



参数估计



参数估计是机器学习中最重要的问题之一，常用两种不同的估计方式

- 点估计：求参数 θ 真实值的一个近似值 $\hat{\theta}$
- 区间估计：求 θ 的一个可能的取值区间以及 θ 真实值属于该区间的概率

如何评价一个点估计结果的优劣？

- 无偏性
- 有效性
- 相合性¹²²

¹²²相合性也称一致性



设有随机变量 X , θ 是其分布的待估参数, X_1, X_2, \dots, X_n 是 X 的一个样本, x_1, x_2, \dots, x_n 是一个具体的样本值。点估计就是要构造一个适当的统计量 $\hat{\theta}(X_1, X_2, \dots, X_n)$, 用它的观察值 $\hat{\theta}(x_1, x_2, \dots, x_n)$ 作为未知参数 θ 的近似值。通常, 称 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 为 θ 的估计量, 称 $\hat{\theta}(x_1, x_2, \dots, x_n)$ 为 θ 的估计值。

例如: 用样本均值去估计随机变量的均值

$$\hat{\lambda} = \hat{E}[X] = \frac{1}{n} \sum_{k=1}^n X_k \text{ (估计量)}, \quad \hat{\lambda} = \hat{E}[X] = \frac{1}{n} \sum_{k=1}^n x_k \text{ (估计值)}$$

点估计法的重点就在于: 如何构造出一个好的估计函数。最常用的估计方法有矩估计和极大似然估计。¹²³

¹²³ 极大后验估计基于极大似然估计, 此处归于极大似然估计



设有随机变量 x , 它有多种不同矩的定义¹²⁴

- n 阶原点矩: $E[x^n]$
- n 阶中心矩: $E[(x - E[x])^n]$
- n 阶绝对矩: $E[|x|^n]$
- n 阶绝对中心矩: $E[|x - E[x]|^n]$

矩估计的基本思想: 用样本原点矩替换总体原点矩

$$A_k = \frac{1}{n} \sum_{j=1}^n x_j^k \quad \xrightarrow{\text{替换}} \quad E[x^k] = \int x^k p(x; \theta) dx$$

¹²⁴通常简称 n 阶矩, 如无特殊说明, n 阶矩也常指 n 阶原点矩; 注意, 要求矩应先证明矩存在



设总体的密度函数为 $f(x; \theta_1, \theta_2, \dots, \theta_m)$, 求其 m 阶矩并写出参数关于矩的函数

$$E[x^k] = \int x^k f(x; \theta_1, \theta_2, \dots, \theta_m) \rightarrow \theta_k = h^{-1}(E[x^k])$$

得到 m 个方程并用样本矩替代矩, 解方程组得到参数估计值

$$\begin{cases} \theta_1 = h^{-1}(E[x^1]) \\ \theta_2 = h^{-1}(E[x^2]) \\ \dots \\ \theta_m = h^{-1}(E[x^m]) \end{cases} \xrightarrow{\text{替换}} \begin{cases} \theta_1 = h^{-1}(A_1) \\ \theta_2 = h^{-1}(A_2) \\ \dots \\ \theta_m = h^{-1}(A_m) \end{cases}$$

设有随机变量 X , (x_1, x_2, \dots, x_n) 是 X 的一组样本, 利用矩估计法求 X 所属分布参数

- X 服从伯努利分布 $B(1, p)$: 已知伯努利分布期望为 $E[X] = p$, 所以有 $p = \frac{1}{n} \sum_{i=1}^n x_i$
- X 服从指数分布 $E(\lambda)$: 已知指数分布期望为 $E[X] = \frac{1}{\lambda}$, 所以有 $\frac{1}{\lambda} = \frac{1}{n} \sum_{i=1}^n x_i$



矩估计法是一种经典的估计方法，它的优点是简单直观，但它也存在一些缺点

- 并非所有情况矩都存在
- 求矩并写出参数关于矩的函数很多时候是困难的
- 矩估计的结果可能是不唯一的

设 X 服从泊松分布 $P(\lambda)$ ， (x_1, x_2, \dots, x_n) 是 X 的一组样本。已知泊松分布有 $E[X] = \lambda$ ， $D(X) = \lambda = E[X^2] - E^2[X]$ ， λ 的矩估计有两种不同的结果¹²⁵

- 根据 $E[X] = \lambda$ ，有 $\hat{\lambda} = \frac{1}{n} \sum_{j=1}^n x_j$
- 根据 $D(X) = \lambda = E[X^2] - E^2[X]$ ，有 $\hat{\lambda} = \frac{1}{n} \sum_{j=1}^n x_j^2 - \left(\frac{1}{n} \sum_{j=1}^n x_j \right)^2$

¹²⁵ 矩估计结果不唯一时，通常优先选取低阶估计结果



设随机变量 X 有分布列 $P(X = x; \theta)$ 或概率密度函数 $f(x; \theta)$, 其中 $\theta = (\theta_1, \theta_2, \dots, \theta_n)$ 。设 (x_1, x_2, \dots, x_n) 是 X 的一组样本, 将样本的联合分布概率或联合概率密度看成是 θ 的函数用 $L(\theta)$ 表示, 称 $L(\theta)$ 为似然函数

$$L(\theta) = \prod_{i=1}^n P(X_i = x_i; \theta) \quad (\text{或}) \quad L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$

若 $\hat{\theta}$ 使得 $L(\theta)$ 取得最大值 $\max L(\theta)$, 则称 $\hat{\theta}$ 为 θ 的极大似然估计量

在具体计算时, 通常进一步求 $L(\theta)$ 的对数 $\ln L(\theta)$, 并称其为对数似然函数

- $\ln z$ 是严格递增函数, 当 $\ln z$ 取到最大值时 z 也取到最大值, 因此 $\ln L(\theta)$ 取到最大值时 $L(\theta)$ 也取到最大值
- $\ln z$ 是关于 z 的凸函数, 若 \hat{z} 使得 $(\ln z)' = 0$ 则 $\ln z$ 一定在 \hat{z} 处取得极大值
- 对数操作可将 $L(\theta)$ 中的连乘转成求和, 提高计算的精准度



极大似然估计的具体计算：写出样本集的对数似然函数 $\ln L(\theta)$ ，令 $\frac{d}{d\theta} \ln L(\theta) = 0$ 解方程得到 $\hat{\theta}$ 作为对 θ 的估计值。若 $\theta = (\theta_1, \theta_2, \dots, \theta_m)$ 则解方程组

$$\frac{\partial}{\partial \theta_1} \ln L(\theta) = 0, \quad \dots, \quad \frac{\partial}{\partial \theta_m} \ln L(\theta) = 0$$

设总体 X 服从正态分布 $N(\mu, \sigma^2)$ ，其中 μ, σ^2 未知。 x_1, x_2, \dots, x_n 是取自该总体的一组样本，利用极大似然估计法求 μ, σ^2 的估计量。已知高斯密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

因此似然及对数似然函数为

$$L(\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \rightarrow \ln L(\theta) = \sum_{i=1}^n \ln \frac{1}{\sqrt{2\pi}\sigma} - \sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2} = -\frac{n}{2} \ln 2\pi - n \ln \sigma - \sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}$$



先求 μ 的估计量

$$\frac{\partial}{\partial \mu} \ln L = \frac{\partial}{\partial \mu} \left(-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2} \right) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) \rightarrow \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0 \rightarrow \mu = \frac{1}{n} \sum_{i=1}^n x_i$$

再求 σ^2 的估计量

$$\frac{\partial}{\partial \sigma^2} \ln L = \frac{\partial}{\partial \sigma^2} \left(-n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2$$

令偏导数等于0

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = 0 \rightarrow -n\sigma^2 + \sum_{i=1}^n (x_i - \mu)^2 = 0 \rightarrow \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

所以对于高斯分布，其极大似然估计有

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = S_n^2$$



极大似然估计比矩估计更加通用，但极大似然估计也有固有缺点

- 有较严重的小样本偏差，即常说的过拟合
- 对奇异值较为敏感
- 似然函数可能不可导或形式复杂无闭式解
- 若似然函数关于参数是非凸的，则有可能得到局部最优解

常见解决办法

- 引入先验做极大后验估计或增加正则化项
- 数据预处理或调整似然函数
- 通过再参数化（reparameter）或数值计算等方法实现计算
- 接受局部最优或基于不同初始值多次求解后选择最优



同一个参数有多种不同的估计方法，如何评价估计结果的优劣？常用标准有三种

- 无偏性：多次估计时，估计值的均值与参数真实值无偏差
- 有效性：无偏估计的方差，表示一次估计值与真实值间的波动
- 相合性：随着样本数增加，估计值与真实值的接近程度

设 $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ 是 θ 的一个估计量， θ 取值空间为 Θ ，若对任意 $\theta \in \Theta$ 下式成立，则称 $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ 是 θ 的一个无偏估计，否则称为有偏估计

$$E[\hat{\theta}(X_1, X_2, \dots, X_n)] = \theta$$

(在有限样本集下估计量可能是有偏的) 若样本量趋于无穷时估计值是无偏的，则称 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 是 θ 的一个渐近无偏估计

$$\lim_{n \rightarrow \infty} E[\hat{\theta}(X_1, X_2, \dots, X_n)] = \theta$$



设 (X_1, X_2, \dots, X_n) 是取自总体 X 的一组样本，总体 X 服从区间 $(0, \theta)$ 的均匀分布，其中 θ 未知。分析 θ 的矩估计量 $\hat{\theta}_1$ 和极大似然估计量 $\hat{\theta}_2$ 的无偏性。

解： X 服从 $(0, \theta)$ 的均匀分布，所以其密度函数如下

$$f(x) = \begin{cases} \frac{1}{\theta}, & 0 \leq x \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

X 的期望 $E[X]$ 为

$$E[X] = \int_{-\infty}^{+\infty} xf(x)dx = \int_0^{\theta} x \frac{1}{\theta} dx = \frac{1}{\theta} \frac{x^2}{2} \Big|_0^{\theta} = \frac{\theta}{2}$$

根据矩估计思想用样本矩替代总体矩，因此矩估计值为

$$\frac{\theta}{2} = E[X] \rightarrow \frac{\hat{\theta}_1}{2} = \frac{1}{n} \sum_{i=1}^n X_i \rightarrow \hat{\theta}_1 = \frac{2}{n} \sum_{i=1}^n X_i = 2\bar{X}$$



根据极大似然估计思想，样本的似然为

$$L(\theta) = \prod_{i=1}^n p(x_i) = \prod_{i=1}^n \frac{1}{\theta^n} = \frac{1}{\theta^n} \rightarrow \ln L(\theta) = n \ln \frac{1}{\theta}$$

可见， θ 取值越小 $\frac{1}{\theta}$ 越大， $L(\theta)$ 的取值也就越大。但是很明显，任意 X_i 应介于 $0 \sim \theta$ 之间，所以 θ 的最小取值也应为 X_i 的最大取值，所以 θ 的极大似然估计为

$$\hat{\theta}_2 = \max_{1 \leq i \leq n} \{X_i\}$$

验证这两种估计结果的无偏性。因为 $\hat{\theta}_1$ 满足下式，所以均匀分布的一阶矩估计是无偏的

$$E[\hat{\theta}_1] = E[2\bar{X}] = 2E[\bar{X}] = 2E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{2}{n} \sum_{i=1}^n E[X_i] = \frac{2}{n} \sum_{i=1}^n \frac{\theta}{2} = \theta$$



对于 $\hat{\theta}_2$ 有 $E[\hat{\theta}_2] = E[\max\{X_1, X_2, \dots, X_n\}]$ ，令 $z = \max\{X_1, X_2, \dots, X_n\}$ 可将 z 看作一个随机变量，所以 $E[\hat{\theta}_2] = E[z]$ 。而 z 的累积分布函数为

$$F(z') = F(z \leq z') = F(\max\{X_1, X_2, \dots, X_n\} \leq z') = \prod_{i=1}^n P(X_i \leq z')$$

已知任意 X_i 都服从 $(0, \frac{1}{\theta})$ 的均匀分布，所以有

$$P(X_i \leq z') = \begin{cases} 0, & z' < 0 \\ \frac{z'}{\theta}, & 0 \leq z' \leq \theta \\ 1, & z' > \theta \end{cases} \rightarrow F(z') = \prod_{i=1}^n P(X_i \leq z') = \begin{cases} 0, & z' < 0 \\ \left(\frac{z'}{\theta}\right)^n, & 0 \leq z' \leq \theta \\ 1, & z' > \theta \end{cases}$$

所以 z' 的密度函数为¹²⁶

$$\frac{d}{dz'} F(z') = \begin{cases} n \frac{(z')^{n-1}}{\theta^n}, & 0 \leq z' \leq \theta \\ 0, & \text{otherwise} \end{cases} \rightarrow f(z) = \begin{cases} n \frac{z^{n-1}}{\theta^n}, & 0 \leq z \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

¹²⁶ z' 的密度函数实际就是 z 的密度函数，只是符号不同而已



所以 z 的期望为¹²⁷

$$E[\hat{\theta}_2] = E[z] = \int_0^\theta z n \frac{z^{n-1}}{\theta^n} dz = \frac{n}{\theta^n} \int_0^\theta z^n dz = \frac{n}{\theta^n} \frac{z^{n+1}}{n+1} \Big|_0^\theta = \frac{n\theta}{n+1}$$

很明显 $E[\hat{\theta}] \neq \theta$, 所以此例的极大似然估计结果有偏估计。但是因为有

$$\lim_{n \rightarrow \infty} E[\hat{\theta}_2] = \lim_{n \rightarrow \infty} \frac{n}{n+1} \theta = \lim_{n \rightarrow \infty} \frac{1}{1 + \frac{1}{n}} \theta = \theta$$

所以 $\hat{\theta}_2$ 是渐近无偏估计。或者我们可以将极大似然估计结果 $\hat{\theta}_2$ 修正为 $\hat{\theta}'_2$ 使其成为无偏估计

$$\hat{\theta}'_2 = \frac{n+1}{n} \hat{\theta}_2 \quad \rightarrow \quad E[\hat{\theta}'_2] = E\left[\frac{n+1}{n} \hat{\theta}_2\right] = \frac{n+1}{n} E[\hat{\theta}_2] = \frac{n+1}{n} \frac{n}{n+1} \theta = \theta$$

¹²⁷注意 $0^n = 0$ 若 $n > 0$



设 X_1, X_2, \dots, X_n 是总体 X 的一组样本， X 服从正态分布 $N(\mu, \sigma^2)$ 。分析其样本均值、样本方差的无偏性。

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

样本均值是总体均值的无偏估计

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

但是样本方差是总体方差的有偏估计，先对 S_n^2 变形

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



所以

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

所以

$$E[S_n^2] = E\left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2\right] - E[(\bar{X} - \mu)^2] = \frac{1}{n} \sum_{i=1}^n E[(X_i - \mu)^2] - E[(\bar{X} - \mu)^2]$$

因为任意 X_i 都服从 $N(\mu, \sigma^2)$, 而 $E[(z - \mu_z)^2]$ 正好是 z 的方差, 所以

$$E[(X_i - \mu)^2] = D(X_i) = \sigma^2, \quad E[(\bar{X} - \mu)^2] = D(\bar{X})$$

注意到 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, 所以根据方差的计算规则¹²⁸

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

¹²⁸注意这里的 X_i 之间是相互独立的



所以最终有

$$E[S_n^2] = \frac{1}{n} \sum_{i=1}^n \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2$$

所以 S_n^2 是 σ^2 的有偏估计。通常在利用样本估计总体的方差时会对 S_n^2 做修正

$$S'_n{}^2 = \frac{n}{n-1} S_n^2 = \frac{n}{n-1} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

但是同样的， S_n^2 是 σ^2 的渐近无偏估计

$$\lim_{n \rightarrow \infty} E[S_n^2] = \lim_{n \rightarrow \infty} \frac{n-1}{n} \sigma^2 = \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right) \sigma^2 = \sigma^2$$



一个参数可以有多个无偏估计，无偏估计之间可以用估计值与真实值的偏差波动进一步衡量估计结果的优劣，即为有效性：设 $\hat{\theta}_1, \hat{\theta}_2$ 是 θ 的两个无偏估计，若有 $D(\hat{\theta}_1) \leq D(\hat{\theta}_2)$ 则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效。

前面已知 $(0, \theta)$ 均匀分布的矩估计 $\hat{\theta}_1$ 是无偏估计，修正后的 $\hat{\theta}'_2$ 也是无偏估计。分析这两个无偏估计的有效性。

$$\hat{\theta}_1 = 2\bar{X}, \quad \hat{\theta}'_2 = \frac{n+1}{n}\hat{\theta}_2 = \frac{n+1}{n} \max\{X_i\}$$

因为 $(0, \theta)$ 范围均匀分布的方差为 $D(X) = \frac{\theta^2}{12}$ ¹²⁹，所以任意 X_i 的方差为 $\frac{\theta^2}{12}$ ，所以有

$$D(\hat{\theta}_1) = D(2\bar{X}) = D\left(\frac{2}{n} \sum_{i=1}^n X_i\right) = \frac{4}{n^2} \sum_{i=1}^n D(X_i) = \frac{\theta^2}{3n}$$

¹²⁹前面介绍过， (a, b) 范围的均匀分布其方差为 $\frac{(b-a)^2}{12}$



对于极大似然估计，前面已推导了 $\hat{\theta}_2$ 的密度函数，所以有

$$f(z) = \begin{cases} n \frac{z^{n-1}}{\theta^n}, & 0 \leq z \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

已知 $D(z) = E[z^2] - E^2[z]$ ，前面已推导

$$E[\hat{\theta}_2] = E[z] = \frac{n}{n+1}\theta$$

又因为有

$$E[z^2] = \int_0^\theta z^2 f(z) dz = \int_0^\theta \frac{n}{\theta^n} z^{n+1} dz = \frac{n}{n+2} \frac{1}{\theta^n} z^{n+2} \Big|_0^\theta = \frac{n}{n+2} \theta^2$$

所以有

$$D(z) = E[z^2] - E^2[z] = \frac{n}{n+2} \theta^2 - \left(\frac{n}{n+1} \theta \right)^2 = \frac{n(n+1)^2 - n^2(n+2)}{(n+2)(n+1)^2} \theta^2 = \frac{n}{(n+2)(n+1)^2} \theta^2$$



所以对 $\hat{\theta}'_2$ 有

$$D(\hat{\theta}'_2) = D\left(\frac{n+1}{n}\hat{\theta}_2\right) = \frac{(n+1)^2}{n^2}D(\hat{\theta}_2) = \frac{(n+1)^2}{n^2} \frac{n}{(n+2)(n+1)^2}\theta^2 = \frac{1}{n(n+2)}\theta^2$$

若 $n = 1$ 有

$$D(\hat{\theta}_1) = \frac{1}{3n}\theta^2 = \frac{1}{3}\theta^2, \quad D(\hat{\theta}'_2) = \frac{1}{n(n+2)}\theta^2 = \frac{1}{3}\theta^2$$

若 $n \geq 2$ 则有

$$D(\hat{\theta}'_2) = \frac{1}{n(n+2)}\theta^2 < \frac{1}{n(1+2)}\theta^2 = \frac{1}{3n}\theta^2 = D(\hat{\theta}_1)$$

所以总有 $D(\hat{\theta}'_2) \leq D(\hat{\theta}_1)$, 所以对于 $(0, \theta)$ 均匀分布, 修正后的极大似然估计结果比矩估计更有效。



对于某种估计方法，若随着样本容量的增加，估计值稳定趋于待估参数的真实值，则称该估计具有相合性：设 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 为参数 θ 的估计量，若 $n \rightarrow \infty$ 时 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 依概率收敛于 θ ，则称 $\hat{\theta}$ 为 θ 的相合估计量。即对于任意 $\varepsilon > 0$ ，总有

$$\lim_{n \rightarrow \infty} P\{|\hat{\theta} - \theta| < \varepsilon\} = 1 \quad \text{或者} \quad \lim_{n \rightarrow \infty} P\{|\hat{\theta} - \theta| \geq \varepsilon\} = 0$$

相合性是指估计结果会随样本容量的增加而稳定提高，直至提高到任意精度内。如果一个估计不具有相合性，说明它的精度不随样本容易的增加而提高，这种估计被认为是非常不好的，一般不会考虑这种估计方法。



$\hat{\theta}(X_1, X_2, \dots, X_n)$ 会随 (X_1, X_2, \dots, X_n) 取值变化而变化，因此可看作是一个随机变量。若 $\hat{\theta}$ 是 θ 的一个无偏估计且 $\lim_{n \rightarrow \infty} D(\hat{\theta}) = 0$ ，则 $\hat{\theta}$ 是 θ 的一个相合估计

证明：因为 $\hat{\theta}$ 是无偏估计，所以 $E[\hat{\theta}] = \theta$ ，根据切比雪夫不等式，对任意 $\varepsilon > 0$ 总有

$$P(|\hat{\theta} - \theta| \geq \varepsilon) = P(|\hat{\theta} - E[\hat{\theta}]| \geq \varepsilon) \leq \frac{D(\hat{\theta})}{\varepsilon^2}$$

根据题意，当 $n \rightarrow \infty$ 时 $\lim_{n \rightarrow \infty} D(\hat{\theta}) = 0$ ，所以有

$$P(|\hat{\theta} - \theta| \geq \varepsilon) = P(|\hat{\theta} - E[\hat{\theta}]| \geq \varepsilon) = 0$$

所以说 $\hat{\theta}$ 是 θ 的一个相合估计量



设 (X_1, X_2, \dots, X_n) 是取自总体 $X \sim N(0, \sigma^2)$ 的样本，令 $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2$ 是 σ^2 的估计值，则 $\hat{\sigma}^2$ 是 σ^2 的相合估计量

证明：因为

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n} \sum_{i=1}^n X_i^2\right] = \frac{1}{n} \sum_{i=1}^n E[X_i^2] = \sigma^2$$

所以 $\hat{\sigma}^2$ 是 σ^2 的无偏估计量。又因为 $X_i \sim N(0, \sigma^2)$ ，所以 $\frac{X_i}{\sigma} \sim N(0, 1)$ ，所以根据 χ^2 分布定义有

$$\frac{1}{\sigma^2} \sum_{i=1}^n X_i^2 = \left(\frac{X_1}{\sigma}\right)^2 + \left(\frac{X_2}{\sigma}\right)^2 + \dots + \left(\frac{X_n}{\sigma}\right)^2 = \chi^2(n)$$

已推导 $D(\chi^2(n)) = 2n$ ，所以 $D\left(\frac{1}{\sigma^2} \sum_{i=1}^n X_i^2\right) = 2n$ ，所以

$$D(\hat{\sigma}^2) = D\left(\frac{1}{n} \sum_{i=1}^n X_i^2\right) = D\left(\frac{\sigma^2}{n} \sum_{i=1}^n \frac{X_i^2}{\sigma^2}\right) = \frac{\sigma^4}{n^2} D\left(\sum_{i=1}^n \frac{X_i^2}{\sigma^2}\right) = \frac{2\sigma^4}{n}$$

很明显，当 $n \rightarrow \infty$ 时 $\lim_{n \rightarrow \infty} D(\hat{\sigma}^2) = 0$ ，所以 $\hat{\sigma}^2$ 是 σ^2 的相合估计



对于点估计有效性有以下几点注意

- 相合性讨论较少，但一般都默认估计值应具有相合性，否则不予考虑¹³⁰
- 机器学习中讨论较多的是无偏性
- 根据大数定律，矩估计一般都具有相合性

¹³⁰但也不绝对，尤其在小样本分析、启发式算法等情况下



点估计是估计 θ 的一个近似值，但很多时候这种估计是不够的。区间估计给出 θ 的一个取值范围，以及这个范围包含 θ 真实值的可信程度

设总体 X 的密度函数为 $f(x; \theta)$, (X_1, X_2, \dots, X_n) 是 X 的一组样本。设 $\underline{\theta} = \underline{\theta}(X_1, X_2, \dots, X_n)$ 和 $\bar{\theta} = \bar{\theta}(X_1, X_2, \dots, X_n)$ ($\underline{\theta} < \bar{\theta}$) 是 θ 的两个统计量¹³¹，若对任意给定的 α ($0 < \alpha < 1$) 总有

$$P\{\underline{\theta} < \theta < \bar{\theta}\} \geq 1 - \alpha$$

则称随机区间 $(\underline{\theta}, \bar{\theta})$ 是 θ 的置信水平为 $1 - \alpha$ 的置信区间， $\underline{\theta}$ 和 $\bar{\theta}$ 分别称为置信水平为 $1 - \alpha$ 的双侧置信区间的置信下限和置信上限， $1 - \alpha$ 称为置信水平

对随机变量 X 反复抽样，每组样本得到一个区间 $(\underline{\theta}, \bar{\theta})$ ，真实值 θ 要么属于这个区间要么不属于这个区间¹³²。在所得区间中，包含 θ 的约占 $(1 - \alpha) * 100\%$ ，不包含的约占 $\alpha * 100\%$

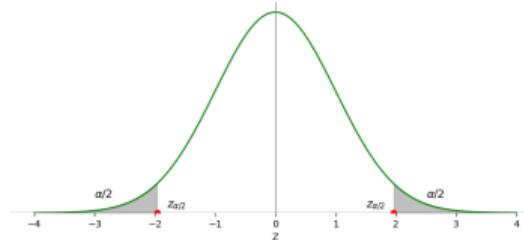
¹³¹注意， $\underline{\theta}$ 和 $\bar{\theta}$ 是用同一组样本得到的

¹³²这可以看作一个伯努利分布，根据伯努利大数定律可得此结论



设总体 $X \sim N(\mu, \sigma^2)$, σ^2 为已知 μ 未知, 设 X_1, X_2, \dots, X_n 是来自总体的样本, 求 μ 的置信水平为 $1 - \alpha$ 的置信区间。已知 \bar{X} 是 μ 的无偏估计, 并且 $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ 。设 $z_{\alpha/2}$ 是标准正态分布的 α 分位点, 则一定有

$$P \left\{ \left| \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right| < z_{\alpha/2} \right\} = 1 - \alpha$$



因此有

$$P \left\{ \bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right\} = 1 - \alpha$$

所以 μ 的一个置信水平为 $1 - \alpha$ 的置信区间为¹³³

$$\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right) \xrightarrow{\text{或写作}} \left(\bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right)$$

¹³³ 实际操作中, 只需要将样本均值、样本数量、总体标准差和标准正态分布的 α 分位点的具体值代入, 便可得到一个具体的置信区间



构造未知参数 θ 的置信区间的计算过程可总结为以下几步

- ① 求出 θ 的一个点估计 $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ （矩估计、极大似然估计或无偏估计）
- ② 构造 $\hat{\theta}$ 和 θ 的一个函数 $G = G(\hat{\theta}, \theta)$, 要求 G 中除 θ 外不含任何其它未知参数且 G 的分布已知或分位数可知
- ③ 确定一组 $\{a, b\}$, 使得 $P\{a \leq G(\hat{\theta}, \theta) \leq b\} = 1 - \alpha$
- ④ 将 $a \leq G(\hat{\theta}, \theta) \leq b$ 等价变形为 $\underline{\theta} \leq \theta \leq \bar{\theta}$, $[\underline{\theta}, \bar{\theta}]$ 就是 θ 的 $1 - \alpha$ 置信区间

在这一过程中有几点需要注意

- G 又称为枢轴变量或主元
- $\underline{\theta}(X_1, X_2, \dots, X_n)$ 和 $\bar{\theta}(X_1, X_2, \dots, X_n)$ 仅是样本的函数
- 通常 $1 - \alpha$ 的置信区间不是唯一的, 最理想的情况是选择 $\underline{\theta} - \bar{\theta}$ 平均长度最短的区间, 如果难以做到则通常直接选择两个尾部各 $\frac{\alpha}{2}$ 的区间



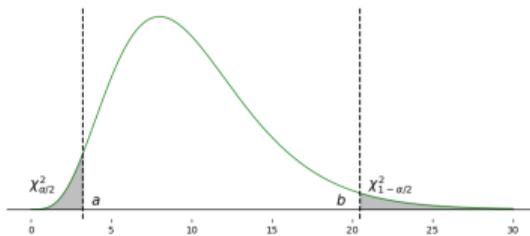
设 X 服从正态分布 $N(\mu, \sigma^2)$, μ 已知 σ^2 未知, (X_1, X_2, \dots, X_n) 为 X 的一组样本, 求 σ^2 的 $1 - \alpha$ 的置信区间。当 μ 已知时, σ^2 的无偏估计是 $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$, 取 $G(\hat{\sigma}^2, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \sim \chi^2(n)$ ¹³⁴, 取 $a < b$ 满足

$$P\left(a \leq \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \leq b\right) = 1 - \alpha$$

因为 χ^2 分布不是对称的，所以寻找等尾置信区间使得 $a = \chi^2_{a/2}(n), b = \chi^2_{1-a/2}(n)$

$$\chi^2_{a/2}(n) \leq \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \leq \chi^2_{1-a/2}(n)$$

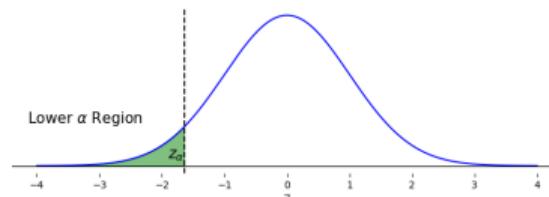
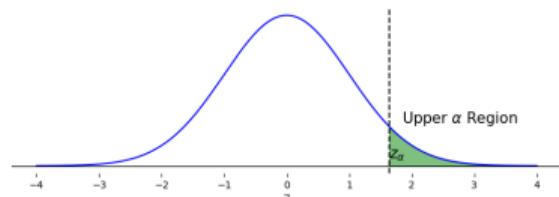
$$\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{1-\alpha/2}(n)} \leq \sigma^2 \leq \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{\alpha/2}(n)}$$



$\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 是 μ 未知时 σ^2 的无偏估计; $\frac{X_i - \mu}{\sigma}$ 服从 $N(0, 1)$, 所以 $\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n)$



很多时候用到单侧置信区间：若有统计量 $\bar{\theta} = \bar{\theta}(X_1, X_2, \dots, X_n)$ ，使得 $P_{\theta}(\theta \leq \bar{\theta}) = 1 - \alpha$ ，则称 $(-\infty, \bar{\theta}(X_1, X_2, \dots, X_n)]$ 为 θ 的单侧 $1 - \alpha$ 置信区间， $\bar{\theta}(X_1, X_2, \dots, X_n)$ 为 θ 的单侧 $1 - \alpha$ 置信区间的置信上限¹³⁵



区间估计有以下几点注意

- 枢轴变量 $G(\hat{\theta}, \theta)$ 多采用常用分布¹³⁶，便于通过查表得到 a 和 b 的值
- 大多数情况下都采用等尾置信区间

¹³⁵ 类似地，还有置信下限，此处从略

¹³⁶ 如正态分布、t 分布、F 分布、 χ^2 分布等

