

1 概率论

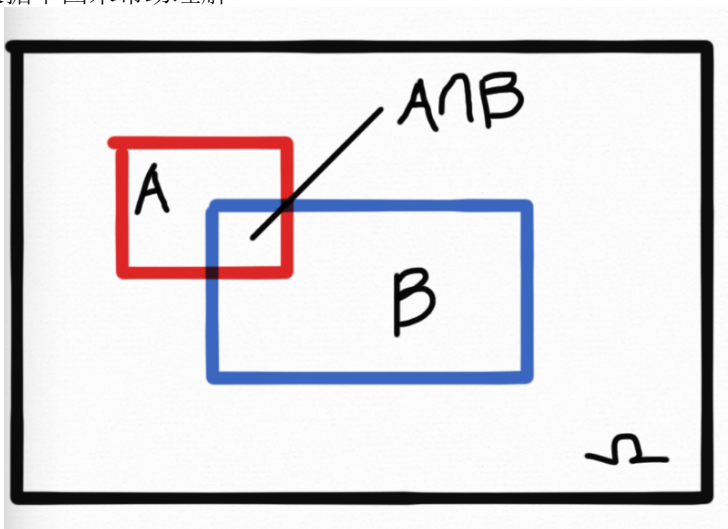
1.1 全概率公式以及贝叶斯公式

条件概率

条件概率的定义是，设有两个事件 A, B ，而 $P(B) \neq 0$ ，则在给定 B 发生的条件下 A 的条件概率记为 $P(A|B)$ ，定义为：

$$P(A|B) = P(AB)/P(B) \quad P(B) \neq 0$$

对于公式的理解，条件概率是在 B 发生的条件下，也就是 B 发生的时候， A 也发生，简言之就是在 B 发生的条件下 A 和 B 同时发生的概率。可以根据下图来帮助理解



全概率公式

设 B_1, B_2, \dots 为有限或无限个事件，它们两两互斥且在每次实验中至少发生一个，即：

$$B_i B_j = \emptyset, i \neq j$$

$$B_1 + B_2 + \dots = \omega$$

把具有这些性质的一组事件成为一个完备事件组。任一事件 B 以及其对立事件组成一个完备事件组。

现考虑任一事件 A 。因 ω 是必然事件，有 $A = A\omega = AB_1 + AB_2 + \cdots$ ，因 B_1, B_2, \dots 两两互斥，显然 AB_1, AB_2, \dots 也两两互斥。

由加法定理有

$$P(A) = P(AB_1) + P(AB_2) + \cdots$$

再由条件概率的定义有：

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \cdots$$

上式成为全概率公式。“全部”概率 $P(A)$ 被分解成了许多部分之和，在较复杂的情况下直接算 $P(A)$ 不容易，但 A 总是伴随着某个 B_i 伴出。

另一角度理解，把 B_i 看作是导致事件 A 发生的一种可能途径，对不同的途径， A 发生的概率即条件概率 $P(A|B)$ 各个不同，而采取哪个途径却是随机的。

贝叶斯公式

根据条件概率的定义以及在全概率公式的假定下

$$P(B_i|A) = P(AB_i)/P(A) = P(B_i)P(A|B_i) / \sum_j P(B_j)P(A|B_j)$$

如果把事件 A 看成是“结果”，把诸事件 B_1, B_2, \dots 看成是导致这结果的可能的“原因”，则可以把全概率公式看作成为“由原因推导结果”；而贝叶斯公式看作是“由结果推导原因”：现在有一个结果 A 已经发生，在众多可能的原因中，是哪一个导致了这结果？

1.2 随即变量及概率分布

一维随机变量

随机变量：顾名思义，就是其值随机而定。可以说随机变量就是实验结果的函数。在实验前，我们不能预知它将取何值，这要凭机会，一旦试验后，取值就确定了。

随机变量按其可能取值的全体的性质，分为两大类：

一类叫离散型随机变量。其特征是只能取有限个值，或随在理论上能取无限个值，但这些值可以毫无遗漏地一个接一个排列出来。

另一类叫连续型随机变量，取值不仅是无穷多，还不能无遗漏地逐一排列，而是充满一个区间。

离散型随机变量分布

定义：设 X 为离散型随机变量，其全部可能值为 $\{a_1, a_2, \dots\}$ 则

$$p_i = P(X = a_i), i = 1, 2, \dots$$

称为 X 的概率函数，也成为概率分布

定义：设 X 为一随机变量，则函数

$$P(X \leq x) = F(x), -\infty < x < \infty$$

称为 X 的分布函数。它对任何随机变量都有定义。

若知道离散型的概率函数，则

$$F(x) = P(X \leq x) = \sum_{\{i: a_i \leq x\}} p_i$$

连续型随机变量分布

定义：设连续性随机变量 X 有概率分布函数 $F(x)$ ，则 $F(x)$ 的倒数 $f(x) = F'(x)$ ，称为 X 的概率密度函数

密度函数解释：取一个点 x ，则按分布函数的定义，事件 $\{x < X \leq x+h\}$ 的概率 ($h > 0$ 为常数)，应为 $F(x+h)-F(x)$ ，所以比值 $[F(x+h)-F(x)]/h$ 可以解释为在 x 点附近 h 这么长的区间 $(x, x+h)$ 内，单位长所占有的概率。令 $h \rightarrow 0$ ，则这个比的极限，即 $F' = f(x)$ ，也就是在 x 点处单位长的概率，或者说它反映了概率在 x 点处的密集程度。

1.3 多维随机变量

定义：以 $\{a_{i1}, a_{i2}, \dots\}$ 记 X_i 的全部可能值， $i=1, 2, \dots$ 则事件 $\{X_1 = a_{1j_1}, X_2 = a_{2j_2}, \dots, X_n = a_{nj_n}\}$ 的概率

$$p(j_1, j_2, \dots, j_n) = P(X_1 = a_{1j_1}, X_2 = a_{2j_2}, \dots, X_n = a_{nj_n})$$

$$j_1 = 1, 2, \dots, j_2 = 1, 2, \dots, j_n = 1, 2, \dots$$

称为随即向量 X_1, \dots, X_n 的概率函数或概率分布

定义: 若 $f(x_1, x_2, \dots, x_n)$ 是定义在 R^n 上的非负函数, 使对 R^n 中的任何集合 A , 有

$$P(X \in A) = \int_A \cdots \int f(x_1, \dots, x_n) dx_1 \cdots dx_n$$

则称 f 是 X 的密度函数

边缘分布

设 $X = (X_1, \dots, X_n)$ 是一 n 维随机变量, X 有一定的分布 F , 这是一个 n 维分布。因为 X 的每个分量 X_i 都是因为随机变量, 故它们有各自的分布 $F_i, i = 1, 2, \dots, n$, 这些都是一维分布, 成为随即向量 X 或其分布 F 的边缘分布

虽然一个随即向量 $X = (X_1, \dots, X_n)$ 的分布 F 足以决定其任一分量 X_i 的边缘分布 F_i , 但反过来不对, 即使知道了所有 X_i 的边缘分布, 也不足以决定 X 的分布 F 。边缘分布只考虑了单个变量的情况, 而未涉及它们之间的关系, 而这个信息是包含在 (X_1, \dots, X_n) 的分布之内的。边缘分布就是通常的分布, 无特殊含义。

1.4 条件概率分布

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

离散型随机变量的条件概率分布

设 (X_1, X_2) 为一个二维离散随机向量, X_1 的全部可能取值为 a_1, a_2, \dots ; X_2 的全部可能取值为 b_1, b_2, \dots , 而 (X_1, X_2) 的联合概率分布为

$$p_{ij} = P(X_1 = a_i, X_2 = b_j), i, j = 1, 2, \dots$$

考虑 X_1 在给定 $X_2 = b_j$ 的条件下的条件分布, 依条件概率的定义, 有

$$P(X_1 = a_i | X_2 = b_j) = P(X_1 = a_i, X_2 = b_j) / P(X_2 = b_j) = p_{ij} / P(X_2 = b_j)$$

而根据边缘分布 $P(X_2 = b_j) = \sum_k p_{kj}$, 于是

$$P(X_1 = a_i | X_2 = b_j) = p_{ij} / \sum_k p_{kj}, i = 1, 2, \dots$$

连续型随机变量的条件概率分布

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

$$P(X_1 \leq x_1 | a \leq X_2 \leq b) = \frac{P(X_1 \leq x_1, a \leq X_2 \leq b)}{P(a \leq X_2 \leq b)}$$

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

X_2 的边缘分布的密度函数 $f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1$

$$P(a \leq X_2 \leq b) = \int_a^b f_2(t_2) dt_2$$

由此可得

$$P(X_1 \leq x_1 | a \leq X_2 \leq b) = \frac{\int_{-\infty}^{x_1} dt_1 \int_a^b f(t_1, t_2) dt_2}{\int_a^b f_2(t_2) dt_2}$$

这是 X_1 的条件分布函数, 对 x_1 求导, 得到条件密度函数

$$f_1(x_1 | a \leq X_2 \leq b) = \int_a^b f(x_1, t_2) dt_2 / \int_a^b f_2(t_2) dt_2$$

若在给定 $a=b$ 的情况下, 即在 X_2 给定等于一个值之下, X_1 的条件密度函数。

$$\begin{aligned}
f_1(x_1|x_2) &= f_1(x_1|X_2 = x_2) \\
&= \lim_{h \rightarrow 0} f_1(x_1|x_2 \leq X_2 \leq x_2 + h) \\
&= \lim_{h \rightarrow 0} \frac{1}{h} \int_{x_2}^{x_2+h} f(x_1, t_2) dt_2 / \lim_{h \rightarrow 0} \frac{1}{h} \int_{x_2}^{x_2+h} f_2(t_2) dt_2 \\
&= f(x_1, x_2) / f_2(x_2)
\end{aligned}$$

式子改写成

$$f(x_1, x_2) = f_2(x_2) f_1(x_1|x_2)$$

就是说两个随机变量的联合概率密度，等于其中之一的概率密度乘以在给定这一个之下另一个的概率密度。

式子可以推广到多个变量的场合：设有 n 维随机向量 X_1, \dots, X_n ，其概率密度函数 $f(x_1, \dots, x_n)$ 则

$$f(x_1, \dots, x_n) = g(x_1, \dots, x_k) h(x_{k+1}, \dots, x_n | x_1, \dots, x_k)$$

其中 g 是 X_1, \dots, X_k 的概率密度，而 h 则是在给定 $X_1 = x_1, \dots, X_k = x_k$ 的条件下， X_{k+1}, \dots, X_n 的条件概率密度。

随机变量的独立性

定义： 设 n 维随机向量 $\{X_1, \dots, X_n\}$ 的联合密度函数为 $f(x_1, \dots, x_n)$ ，而 X_i 的边缘密度函数为 $f_i(x_i), i = 1, 2, \dots, n$ 如果：

$$f(x_1, \dots, x_n) = f_1(x_1) \cdot \dots \cdot f_n(x_n)$$

就称随机变量 $\{X_1, \dots, X_n\}$ 相互独立

定义： 设 $\{X_1, \dots, X_n\}$ 都是离散型随机变量，若对任何常数 a_1, \dots, a_n 都有：

$$P(X_1 = a_1, \dots, X_n = a_n) = P(X_1 = a_1) \cdot \dots \cdot P(X_n = a_n)$$

就称随机变量 $\{X_1, \dots, X_n\}$ 相互独立

1.5 随机变量的函数的概率分布

已知某个或某些随机变量 X_1, \dots, X_n 的分布, 另有一些随机变量 Y_1, \dots, Y_m , 它们都是 X_1, \dots, X_n 的函数:

$$Y_i = g_i(X_1, \dots, X_n), i = 1, \dots, m$$

比如在数理统计中, X_1, \dots, X_n 是原始的观察或试验数据, Y_1, \dots, Y_m 则是为某种目的将这些数据加工而得到的量, 成为统计量。比如 X_1, \dots, X_n 的算术平均值 $\bar{X} = (X_1, \dots, X_n)/n$. \bar{X} 就是 X_1, \dots, X_n 的函数。

两个重要的特殊函数:

Γ 函数 $\Gamma(x)$: 通过积分

$$\Gamma(x) = \int_0^{\infty} \exp^{-t} t^{x-1} dt, x > 0$$

$$\Gamma(1) = 1, \Gamma(1/2) = \sqrt{2}, \Gamma(x+1) = x\Gamma(x)$$

β 函数 $\beta(x, y)$: 通过积分

$$\beta(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt, x > 0, y > 0$$

$$\beta(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$$

2 数理统计

2.1 基本概念

当我们用试验或观察的方法研究一个问题时, 首先要通过适当的观察或试验以取得必要的数据, 然后对数据进行分析, 以对所提出的问题作出尽可能正确的结论。

之所以说尽可能正确, 是因为数据一般带有随机性误差, 不只是通常意义下得因测量不准造成的误差。由于数据带有随机误差, 所以作出的结论, 也有可能出错。

统计的两大问题: 参数估计和假设检验

参数估计：比如模型为指数函数，估计参数 λ 为多少，从而求得平均值 $1/\lambda$ ， λ 一般是未知的。所以可以从选取的样本计算样本的平均值来估计 $1/\lambda$

假设检验：假设不符合但被接受，假设符合但可能被拒绝

总体：是指与所研究的问题有关的对象的全体构成的集合

总体的概率分布：是指数分布还是正态分布等等，总体就是一个概率分布，只要服从同一概率分布，就可以视为同类总体。

样本：是按一定的规定从总体中抽出的一部分个体。总体中的每一个个体有同等的被抽出的机会。

统计量：完全由样本所决定的量。统计量只依赖于样本，而不能依赖于其它位置的量。特别是，它不依赖于总体分布中所包含的未知参数。例如，设 $X_1 + \dots + X_n$ 是从正态总体 $N(\mu, \sigma^2)$ 中取出的样本，则 $\bar{X} = (X_1 + \dots + X_n)/n$ 是统计量，因为它完全由样本决定，但 $\bar{X} - \mu$ 不是统计量，因为 μ 未知， $\bar{X} - \mu$ 不是完全由样本所决定。

统计量可以看作是对样本的一种加工，它把样本中所含的（某一方面）的信息集中起来。例如 \bar{X} 可以用于估计未知的 μ 。可以这样看：原始数据 X_1, \dots, X_N 中的每一个，都包含有 μ 的若干信息，但这些事杂乱无章的，一经集中到 \bar{X} 就有了明确的概念。

统计量有：样本均值 $\bar{X} = (X_1 + \dots + X_n)/n$

样本方差 $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$

样本矩，分为样本原点矩和样本中心矩。

$$a_k = (X_1^k + \dots + X_n^k) / n$$

称为 k 阶样本原点矩

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

称为 k 阶样本中心矩

2.2 矩估计，极大似然估计，贝叶斯估计

设有一个统计总体，以 $f(x, \theta_1, \dots, \theta_k)$ 及其概率密度函数（若总体分布为连续型），或其概率函数（若其总体分布为离散型）。这分布包含 k 个未知参数 $\theta_1, \dots, \theta_k$

例如对于正态总体 $N(\mu, \sigma^2)$, 有 $\theta_1 = \mu, \theta_2 = \sigma^2$ 而

$$f(x, \theta_1, \theta_2) = (\sqrt{2\pi\theta_2})^{-1} \exp(-\frac{1}{2\theta_2}(x - \theta_1)^2), -\infty < x < \infty$$

若总体有二项分布 $B(n, p)$, 则 $\theta_1 = p$, 而

$$f(x, \theta_1) = \binom{n}{x} \theta_1^x (1 - \theta_1)^{n-x}, x = 0, 1, \dots, n$$

参数估计问题的提法一般是: 设有了从总体中抽出的样本 X_1, \dots, X_n (独立同分布), 依据这些样本对参数 $\theta_1, \dots, \theta_k$ 的未知值做出估计。为了要估计 θ_1 , 需要构造出适当的统计量 $\hat{\theta}_1 = \hat{\theta}_1(X_1, \dots, X_n)$ 。当有了样本 X_1, \dots, X_n 就代入函数 $\hat{\theta}_1 = \hat{\theta}_1(X_1, \dots, X_n)$ 算出一个值, 用来作为 θ_1 的估计值。由于未知参数 θ_1 是数轴上的一个点, 用 $\hat{\theta}_1$ 去估计 θ_1 , 等于用一个点去估计另一个点, 所以这就叫做点估计。

矩估计法

设总体分布为 $f(x, \theta_1, \dots, \theta_k)$, 则它的矩 (原点矩和中心距都可以, 以原点矩为例),

$$\alpha_m = \int_{-\infty}^{\infty} x^m f(x, \theta_1, \dots, \theta_k) dx$$

或

$$\sum_i x_i^m f(x_i, \theta_1, \dots, \theta_k)$$

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

$$\alpha_m = \alpha_m(\theta_1, \dots, \theta_k) \approx a_m = \sum_{i=1}^n X_i^m / n$$

并让上面的近似式改为等式, 得到

$$\alpha_m(\theta_1, \dots, \theta_k) = a_m, m = 1, \dots, k$$

解此方程组，得根 $\hat{\theta}_i = \hat{\theta}_i(X_1, \dots, X_n), i = 1, \dots, k$ 。就以 $\hat{\theta}_i$ 作为 θ_i 的估计

关于矩估计量有如下结论：

定理：设总体 X 的均值 $E(X) = \mu$ ，方差 $D(X) = \sigma^2$ ， (X_1, \dots, X_n) 为取自该总体的样本，则 \bar{X} 是 μ 的矩估计量， S_N^2 是 σ^2 的矩估计量， S_n 是 σ 的矩估计量

极大似然估计

设总体有分布 $f(X; \theta_1, \dots, \theta_k)$ ， X_1, \dots, X_n 为这总体中抽出的样本，则样本 X_1, \dots, X_n 的分布为

$$f(X_1; \theta_1, \dots, \theta_k) f(X_2; \theta_1, \dots, \theta_k) \dots f(X_n; \theta_1, \dots, \theta_k)$$

记为 $L(X_1, \dots, X_n; \theta_1, \dots, \theta_k)$

固定 $\theta_1, \dots, \theta_k$ 而看作是 X_1, \dots, X_n 的函数时， L 是一个概率密度函数或概率函数。

当 X_1, \dots, X_n 固定，而把 L 看作是 $\theta_1, \dots, \theta_k$ 的函数时，它成为似然函数用似然程度最大的点 $\theta_1^*, \dots, \theta_k^*$ ，满足条件

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

这个估计 $\theta_1^*, \dots, \theta_k^*$ 就叫做定 $\theta_1, \dots, \theta_k$ 的极大似然估计

之后可以对上式两边取对数，再求偏导

与据估计法不同，极大似然估计法要求分布有参数的形式，比如总体分布毫无所知而要估计其均值方差，极大似然法就无能为力

贝叶斯法

对点估计问题，矩估计和极大似然估计，未知参数 θ 就是简单的是一个未知数，在抽取样本之前，我们对 θ 没有任何了解，所有的信息全部来自样本

贝叶斯学派认为，在进行抽样之前，我们对 θ 有一定的知识，叫先验知识，表示这种知识在实验之前就有了。这种先验知识必须用 θ 的某种概率分布表达出来，这概率分布叫做 θ 的先验分布，这个分布总结了我们在实验之前对未知参数 θ 的知识。

贝叶斯统计的一个基本要求是：你必须设法去定出一个 θ 的先验密度 $h(\theta)$ ，甚至出于你自己的主观认识

如果已定下先验密度之后，怎么去得出参数 θ 的估计？

设总体有概率密度 $f(X, \theta)$ ，从这总体抽样本 X_1, \dots, X_n ，则样本的密度为 $f(X_1, \theta) \cdots f(X_n, \theta)$ 。它可视为在给定 θ 值时 X_1, \dots, X_n 的密度， $(\theta, X_1, \dots, X_n)$ 的联合密度为

$$h(\theta)f(X_1, \theta) \cdots f(X_n, \theta)$$

由此算出 X_1, \dots, X_n 的边缘密度为：

$$p(X_1, \dots, X_n) = \int h(\theta)f(X_1, \theta) \cdots f(X_n, \theta)d\theta$$

积分的范围，要看参数 θ 的范围而定

在给定 X_1, \dots, X_n 的条件下， θ 的条件密度为

$$h(\theta|X_1, \dots, X_n) = h(\theta)f(X_1, \theta) \cdots f(X_n, \theta)/p(X_1, \dots, X_n)$$

按照贝叶斯学派的观点，这个条件密度代表了我们现在（在取得样本 X_1, \dots, X_n 之后）对 θ 的知识，它综合了 θ 的先验信息（以 $h(\theta)$ 反映）与由样本带来的信息，把上式成为 θ 的后验密度，因为他是在做了试验之后得到的。

	问 题	先验知识	当前知识	后验(现在)知识
贝叶斯公式	事件 B_1, \dots, B_n 中那一个发生了？	$P(B_1),$ $\dots, P(B_n)$	事件 A 发生了	$P(B_1 A), \dots,$ $P(B_n A)$
此处的问题	$\theta = ?$	$h(\theta)$	样本 X_1, \dots, X_n	后验密度(2.11)

贝叶斯学派的下一个重要观点：在得出后验分布后，对参数 θ 的任何统计推断，都只能基于这个后验分布。

那么如何使用这个后验概率呢？可以结合某种准则去进行，**对点估计问题**，一个常用的方法是取后验分布的均值作为 θ 的估计。

例如，作 n 次独立试验，每次观察某事件 A 是否发生，A 在每次试验中发生的概率为 p ，要依据试验结果去估计 p

以往都是用频率估计概率的方法去处理。这种方法不用 p 的先验知识。以下用贝叶斯统计的观点来处理这个问题。

引进 $X_i = 1, X_i = 0$, 视第 i 次试验时 A 发生与否而定, $i = 1, \dots, n$. $P(X_i = 1) = p, P(X_i = 0) = 1 - p$, 因此 (X_1, \dots, X_n) 的概率函数为 $p^x(1-p)^{n-x}$, x 是 $X_i = 1$ 发生的次数, 取 p 得先验密度 $h(p)$, 则 p 的后验密度为

$$h(p|X_1, \dots, X_n) = \frac{h(p)p^x(1-p)^{n-x}}{\int_0^1 h(p)p^x(1-p)^{n-x}dp}, 0 \leq p \leq 1$$

此分布的均值 (其实就是期望, 注意上式得分布是个对 p 定积分, 已经是个无关 p 的式子了)

$$\begin{aligned}\tilde{p} &= \tilde{p}(X_1, \dots, X_n) = \int_0^1 ph(p|X_1, \dots, X_n)dp \\ &= \frac{\int_0^1 h(p)p^{x+1}(1-p)^{n-x}dp}{\int_0^1 h(p)p^x(1-p)^{n-x}dp}\end{aligned}$$

\tilde{p} 就是 p 在先验分布 $h(p)$ 之下的贝叶斯估计

上式中概率 p 是需要进行估计的, $h(p)$ 是先验知识, 需要我们选择, 那么如何选择呢? 贝叶斯本人提出过“同等无知”的原则, 即实现认为 p 取 $[0,1]$ 内一切值都有可能, 也就是说在 $[0,1]$ 内均匀分布, 这作为 p 的先验分布。这时根据均与分布, 可知概率密度 $h(p)=1$, 当 $0 \leq p \leq 1$, 上式的两个积分都可以用 β 函数表示出, 可得:

$$\tilde{p} = \frac{\beta(X+2, n-x+1)}{\beta(X+1, n-x+1)}$$

最终算得

$$\tilde{p} = \frac{X+1}{n+2}$$

这个估计与频率 x/n , 有些差别, 当 n 很大时并不显著, 而在 n 很小的颇为显著。从一个角度看, 当 n 相当小时, 用贝叶斯估计比用 x/n 合理。因为当 n 很小的时候, 试验结果可能出现 $X=0$ 或 $X=n$ 的极端结果, 这时, 依 X/n 应该把 p 估计为 0 或 1, 这就太极端了。而在这两种情况下, 按照贝叶斯估计, 分别给出估计值为 $1/(n+2)$ 和 $(n+1)/(n+2)$ 就留有一定的余地。

联想: 这跟在自然语言处理里面的平滑处理很像, 一句话虽然没有收录在语料库中, 或者人们根本不会去说, 但不代表这句话不会出现, 即出现的概率为 0, 尤其是某些新词突然出现, 比如“蓝瘦香菇”。

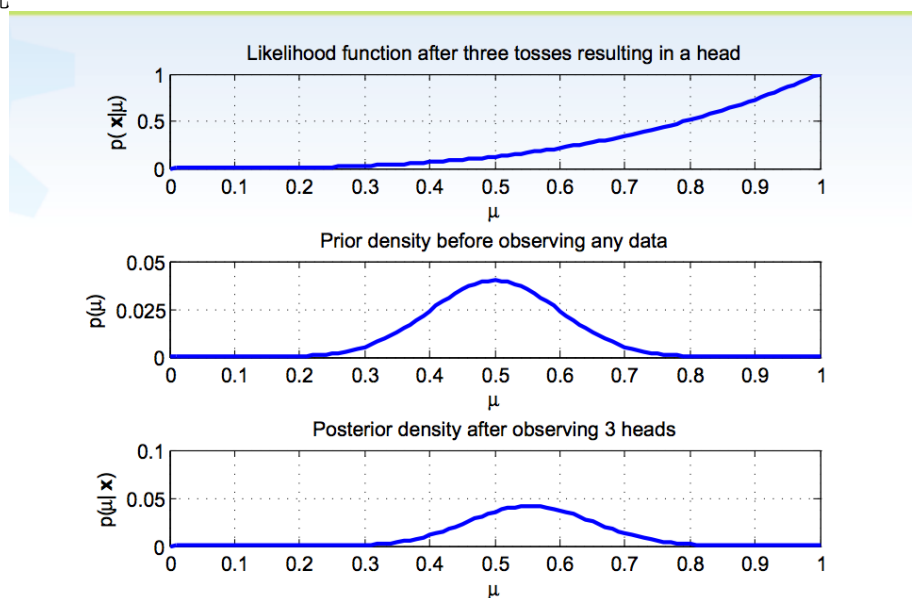
这个同等无知原则也成为贝叶斯原则，被广泛应用到其他情况，不过随着所估计的参数的范围和性质不同，该原则表现的形式也不同。

思考：其实从贝叶斯的角度来看，极大似然也算是贝叶斯估计的一种，似然函数 L 其实可以写作 $P(L|\theta)$ ，本质上也是概率密度函数，只是忽略了先验信息罢了。比如扔硬币，对于似然估计，用频率估计概率，一般都是先做了假设：硬币是均匀材质，正反两面情况出现的机会相等，其实已经利用到了先验信息。但是要是试验的次数很少，只有几次，或者硬币的质地并不均匀，则似然估计就不是那么适用。

贝叶斯公式其实也可以表示成：

$$p(\theta|X) = \frac{p(\theta)p(X|\theta)}{p(X)} = \frac{\text{likelihood} \cdot \text{prior}}{p(X)}$$

对扔三次硬币，针对不同的先验信息，极大似然以及贝叶斯估计的不同表现



<http://www.cs.tut.fi/~hehu/SSP/lecture10.pdf>

2.3 点估计的优良性准则

在考虑估计量的优劣时，必须从某种整体性能去衡量它，而不能看它在个别样本下得表现如何。整体性有两种意义：一是指估计量的某种特性，具有这种特性就是好的，否则就是不好的，如无偏性；二是指某种具体的数量性指标，两个估计量，指标小者为优，比如均方误差。

估计量的无偏性

设某统计总体的分布包含未知参数 $\theta_1, \dots, \theta_k$, X_1, \dots, X_n 是从该总体抽出的样本，要估计 $g(\theta_1, \dots, \theta_k)$ 。g 为一已知函数。设 $\hat{g}(X_1, \dots, X_n)$ 是一个估计量，如果对任何可能的 $\theta_1, \dots, \theta_k$ 都有

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

称 \hat{g} 是 $g(\theta_1, \dots, \theta_k)$ 的一个无偏估计量。

估计量的无偏性有两个含义。第一个含义是没有系统性的偏差，不论用什么样的估计量 \hat{g} 去估计 g，总是时而偏低，时而偏高。无偏性表示，把这些正负偏差在概率上平均起来，其值为 0。

另一个含义是估计量由无偏性，则在大量次数使用取平均时，能以接近于 100% 的把握无限逼近被估计量。如果没有无偏性，则无论是用多少次，其平均也会与真值保持一定距离——这距离就是系统误差

最小方差无偏估计

一个参数往往有不只一个无偏估计，从这些众多的无偏估计中，挑选出最优的。这牵涉到两个问题：一是为优良性制定一个准则，二是在已定的准则之下，如何找到最优者。

1. 均方误差，设 X_1, \dots, X_n 是从某一带参数 θ 的总体中抽出的样本，要估计 θ 。若我们采用估计量 $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ ，则其误差为 $\hat{\theta}(X_1, \dots, X_n) - \theta$ 。这误差随样本 X_1, \dots, X_n 的具体值而定，也是随机的，因为其本身无法取为优良性指标。我们把它平方以消除符号，得 $(\hat{\theta}(X_1, \dots, X_n) - \theta)^2$ ，然后取它的均值：

$$M_{\hat{\theta}}(\theta) = E_{\theta} [\hat{\theta}(X_1, \dots, X_n) - \theta]^2$$

2. 最小方差无偏估计 (MVU 估计)。若局限于无偏估计的范围，且采用

均方误差的准则, 则两个无偏估计 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 的比较, 归结为其方差的比较: 方差小者为优

估计量的相合性与渐近正态性

定义: 设总体分布依赖于参数 $\theta_1, \dots, \theta_k$, $g(\theta_1, \dots, \theta_k)$ 是 $\theta_1, \dots, \theta_k$ 之一给定函数。设 X_1, X_2, \dots, X_n 为该总体中抽出的样本, $T(X_1, \dots, X_n)$ 是 $g(\theta_1, \dots, \theta_k)$ 的一个估计量, 如果对任意给定的 $\varepsilon > 0$ 有

$$\lim_{n \rightarrow \infty} P_{\theta_1, \dots, \theta_k}(|T(X_1, \dots, X_n) - g(\theta_1, \dots, \theta_k)| \geq \varepsilon) = 0$$

而且这对 $\theta_1, \dots, \theta_k$ 一切可能取的值都成立, 则称 $T(X_1, \dots, X_n)$ 是 $g(\theta_1, \dots, \theta_k)$ 的一个相合估计。

如果当样本大小无限增加时, 估计量依概率收敛于被估计的值, 则称该估计量是相合估计。

渐近正态性: 当样本大小 $n \rightarrow \infty$ 时, 其分布都渐近于正态分布

总结: 估计量的相合性和渐近正态性成为估计量的大样本性质, 指的是: 这种性质都是对样本大小 $n \rightarrow \infty$ 来谈的, 对一个固定的 n , 相合性与渐近正态性都是无意义。与此相对, 估计量的无偏性概念是对固定的样本大小来谈的, 不需要样本大小趋于无穷。这种性质成为小样本性质。因此大小样本性质之分不在于样本的具体大小如何, 而在于样本大小趋于无穷与否。

2.4 区间估计

设 X_1, \dots, X_n 是从该总体中抽出的样本。所谓 θ 的区间估计, 就是满足条件 $\hat{\theta}_1(X_1, \dots, X_n) < \hat{\theta}_2(X_1, \dots, X_n)$ 的两个统计量 $\hat{\theta}_1, \hat{\theta}_2$ 为端点的区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 。一旦有了样本 X_1, \dots, X_n , 就把 θ 估计在区间 $[\hat{\theta}_1(X_1, \dots, X_n), \hat{\theta}_2(X_1, \dots, X_n)]$ 之内, 这里有两个要求:

1. θ 要以很大的可能性落在区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 内, 也就是说, 概率 $P_\theta(\hat{\theta}_1(X_1, \dots, X_n) \leq \theta \leq \hat{\theta}_2(X_1, \dots, X_n))$ 要尽可能大

2. 估计的精密密度要尽可能高。区间的长度尽可能小

定义: 给定一个很小的数 $\alpha > 0$ 。如果对参数 θ 的任何值, 概率 $P_\theta(\hat{\theta}_1(X_1, \dots, X_n) \leq \theta \leq \hat{\theta}_2(X_1, \dots, X_n))$ 都等于 $1 - \alpha$, 则称区间估计 $[\hat{\theta}_1, \hat{\theta}_2]$ 的置信系数为 $1 - \alpha$ 。

区间估计也常称为置信区间。意思是对该区间能包含未知参数 θ 可置信到何种程度。

构造合理的区间估计的方法:

枢轴变量法

例如, 设 X_1, \dots, X_n 为抽自正态总体 $N(\mu, \sigma^2)$ 的样本, σ^2 已知, 要求 μ 的区间估计

先找一个 μ 的良好点估计, 在此可以选择样本均值 \bar{X} 。由总体为正态易知

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

以 Φ 记为 $N(0,1)$ 的分布函数。对 $0 < \beta < 1$, 用方程

$$\Phi(\mu_\beta) = 1 - \beta$$

定义记号 μ_β 为分布 $N(0,1)$ 的上 β 分位点。其意义是: $N(0,1)$ 分布中大于 μ_β 的那部分的概率就是 β 。如下图画的是 $N(0,1)$ 的密度函数, 涂黑的部分标出的面积就是 β

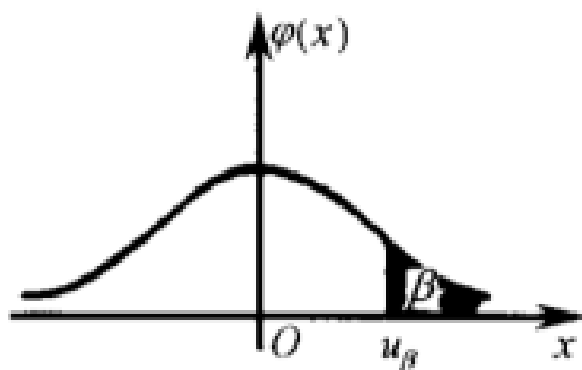


图 4.2

由上面的式子以及 $\Phi(-t) = 1 - \Phi(t)$ 有

$$\begin{aligned} P(-\mu_{\alpha/2} \leq \sqrt{n}(\bar{X} - \mu)/\sigma \leq \mu_{\alpha/2}) &= \Phi(\mu_{\alpha/2}) - \Phi(-\mu_{\alpha/2}) \\ &= (1 - \alpha/2) - \alpha/2 = 1 - \alpha \end{aligned}$$

可以改写为:

$$P(\bar{X} - \sigma\mu_{\alpha/2}/\sqrt{n} \leq \mu \leq \bar{X} + \sigma\mu_{\alpha/2}/\sqrt{n})$$

此式指出：

$$[\hat{\theta}_1, \hat{\theta}_2] = [\bar{X} - \sigma\mu_{\alpha/2}/\sqrt{n}, \bar{X} + \sigma\mu_{\alpha/2}/\sqrt{n}]$$

可作为 μ 的区间估计，置信系数为 $1 - \alpha$

方法可总结为：

1. 找到一个与要估计的参数 $g(\theta)$ 有关的统计量 T ，一般是其良好的点估计（此例 T 为 \bar{X} ）

2. 设法找出 T 和 $g(\theta)$ 的某一函数 $S(T, g(\theta))$ ，其分布 F 要与 θ 无关（此例中， $S(T, g(\theta))$ 为 $\sqrt{n}(\bar{X} - \mu)/\sigma$ ，分布 F 就是 Φ ， S 成为枢轴变量。

3. 对任何常数 $a < b$ ，不等式 $a \leq S(T, g(\theta)) \leq b$ 要能够改写成为等价的形式 $A \leq g(\theta) \leq B$ ， A, B 只与 T, a, b 有关而与 θ 无关

4. 取分布 F 的上 $\alpha/2$ 分位点 $\omega_{\alpha/2}$ 和上 $1 - \alpha/2$ 分位点 $\omega_{1-\alpha/2}$ 。有 $F(\omega_{\alpha/2}) - F(\omega_{1-\alpha/2}) = 1 - \alpha$ ，因此

$$P(\omega_{1-\alpha/2} \leq S(T, g(\theta)) \leq \omega_{\alpha/2}) = 1 - \alpha$$

得到某个具体的区间后， μ 是一个虽然未知，但其值确定的数。区间包含 μ ，或者不包含，二者只居其一。说这区间的置信系数是 0.95，其确切的含义应当是：它是根据所有的数据，用一个置信系数为 0.95 的方法作出的。可见置信系数一词是针对方法：用这方法作出的区间估计，平均 100 此种 95 次包含所要估计的值。一旦算出具体区间，就不能再说它有 95% 的机会包含要估计的值了。比如一个人擅长挑选西瓜：他挑选的西瓜，平均 100 个中有 95 个好的。某天他给你挑一个，结果或好或坏，必居其一，不是 95% 的好。但是考虑到它挑瓜的技术，我对他挑的比较放心，这就是置信系数。

大样本法

主要利用中心极限定理，以建立枢轴变量

贝叶斯法

在有了先验分布密度 $h(\theta)$ 和样本 X_1, \dots, X_n 后，算出后验密度 $h(\theta|X_1, \dots, X_n)$ ，再找两个数 $\hat{\theta}_1, \hat{\theta}_2$ 都与 X_1, \dots, X_n 有关，使得：

$$\int_{\hat{\theta}_1}^{\hat{\theta}_2} h(\theta|X_1, \dots, X_n) d\theta = 1 - \alpha$$

区间 $\hat{\theta}_1, \hat{\theta}_2$ 的意思是：在所得后验分布之下， θ 落在这区间的概率为 $1 - \alpha$

3 假设检验

Statistics	Computer Science	Meaning
estimation	learning	using data to estimate an unknown quantity
classification	supervised learning	predicting a discrete Y from $X \in \mathcal{X}$
clustering	unsupervised learning	putting data into groups
data	training sample	$(X_1, Y_1), \dots, (X_n, Y_n)$
covariates	features	the X_i 's
classifier	hypothesis	a map from covariates to outcomes
hypothesis	—	subset of a parameter space Θ
confidence interval	—	interval that contains unknown quantity with a prescribed frequency
directed acyclic graph	Bayes net	multivariate distribution with specified conditional independence relations
Bayesian inference	Bayesian inference	statistical methods for using data to update subjective beliefs
frequentist inference	—	statistical methods for producing point estimates and confidence intervals with guarantees on frequency behavior
large deviation bounds	PAC learning	uniform bounds on probability of errors

4 不等式

4.1 Markov and Chebychev Inequalities

Markov's Inequality: X 是非负的随机变量，同时假设 $E(X)$ 存在，对任意的 $t > 0$ ，有：

$$P(X > t) \leq \frac{E(X)}{t}$$

PROOF:

$$\begin{aligned}
 E(X) &= \int_0^\infty xf(x)dx = \int_0^t xf(x)dx + \int_t^\infty xf(x)dx \\
 &\geq \int_t^\infty xf(x)dx \\
 &\geq t \int_t^\infty f(x)dx = tP(X > t)
 \end{aligned}$$

Chebyshev's inequality: 让 $\mu = E(X)$, $\sigma^2 = V(X)$, 然后有：

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}, P(|Z| \geq k) \leq \frac{1}{k^2}, Z = (X - \mu)/\sigma$$

PROOF:

use Markov's inequality to conclude that:

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

the second part follows by setting $t = k\sigma$

$$P(|X - \mu| \geq k\sigma) = P(|X - \mu|/\sigma \geq k) = P(|Z| \geq k) \leq \frac{1}{k^2}$$

4.2 Hoeffding's Inequality

Theorem:

let Y_1, \dots, Y_n be independent observations such that $E(Y_i) = 0$, and $a_i \leq Y_i \leq b_i, \epsilon > 0$, for any $t > 0$

$$P\left(\sum_{i=1}^n Y_i \geq \epsilon\right) \leq e^{-t\epsilon} \prod_{i=1}^n e^{t^2(b_i - a_i)^2/8}$$

Theorem:

let $X_1, \dots, X_n \sim \text{Bernoulli}(p)$, then for any $\epsilon > 0$,

$$P(|\bar{X}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2}, \bar{X}_n = \sum_{i=1}^n X_i/n$$

Fix $\alpha > 0$ and let

$$\epsilon_n = \left\{ \frac{1}{2n} \log\left(\frac{2}{\alpha}\right) \right\}^{1/2}$$

By Hoeffding's inequality,

$$P(|\bar{X}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2} = \alpha$$

let $C = (\bar{X}_n - \epsilon, \bar{X}_n + \epsilon)$.

$$P(p \notin C) = P(|\bar{X}_n - p| > \epsilon) \leq \alpha$$

So

$$P(p \in C) \geq 1 - \alpha$$

所以随机变量 p 以 $1 - \alpha$ 的概率落在随机区间 C 中

4.3 Cauchy-Schwarz and Jensen Inequalities

Cauchy-Schwarz inequality: If X and Y have finite variances then

$$E|XY| \leq \sqrt{E(X^2)E(Y^2)}$$

Jensen's Inequality: If g is convex then

$$Eg(X) \geq g(EX)$$

if g is concave then

$$Eg(X) \leq g(EX)$$

PROOF:

假设直线 $L(x) = a + bx$, 与 $g(x)$ 的切点在 $E(X)$, 因为 g 是凸函数, 曲线位于直线 L 之上, 所以

$$Eg(X) \geq E(L) = E(a + bX) = a + bE(X) = L(E(X)) = g(E(X))$$

由上式可知, $EX^2 \geq (EX)^2, E(1/X) \geq 1/E(X)$

参考文献

- [1] 陈希孺: 概率论与数理统计
- [2] 茆诗松: 贝叶斯统计
- [3] all of statistics