第九章 概率模型

第 24 讲 模型、统计推断与学习

黄定江

DaSE @ ECNU

djhuang@dase.ecnu.edu.cn

24.1.1 回顾：从概率到统计

- 从第 7 章内容，我们知道随机变量及其所伴随的概率分布全面描述了随机现象的统计规律性，因此要研究一个随机现象首先要知道它的概率分布。
- 在概率论中，概率分布通常是已知的，或假设为已知的，而一切概率计算和推理，比如求出它的数字特征，讨论随机变量函数的分布，介绍各种常用的分布，就在这已知的基础上得出来。
- 但在实际中，一个随机现象所服从的分布是什么类型可能完全不知道，比如电视机的寿命服从什么分布是不知道的；或者由于现象的某些事实而知道其类型，但不知其分布函数中所含的参数，比如一件产品是合格品还是不合格品服从一个二项分布，但分布中参数 p（不合格品率）却不知道。为了对这些问题展开研究，必须知道它们的分布或分布所含的参数。
- 那么怎样才能知道一个随机现象的分布或参数呢？这是统计学所要解决的一个首要问题。

统计的基本问题：统计推断

- 在统计中我们总是从所要研究的对象全体中抽取一部分进行观测或试验以取得数据或信息，从而对整体作出推断。
- 由于观测或试验是随机现象，依据有限个观测或试验对整体所作出的推论不可能绝对准确，含有一定程度的不确定性，而这种不确定性用概率的大小来表示比较恰当。概率大，推断就比较可靠，概率小，推断就比较不可靠。
- 在统计中，一个基本问题就是依据观测或试验所取得的有限信息对整体如何推断的问题。每个推断必须伴随一定的概率以表明推断的可靠程度。这种伴随有一定概率的推断称为统计推断。

总体、个体和样本

- 在统计中我们把研究的对象全体所构成的集合称为总体，而把组成总体的每一个单元成员称为个体。例如变压器的总体就组成一个总体，其中每一个变压器就是一个个体。
- 在实际中我们所研究的往往是总体中个体的各种数值指标 X，例如变压器的寿命指标，它是一个随机变量。假设 X 的分布函数是 $F(x)$，有时简记为 F。如果我们主要关心的只是这个数值指标 X，为了方便起见，我们可以把这个数值指标 X 的可能取值的全体看作总体，并且称这一总体为具有分布函数 $F(x)$ 的总体，这样就把总体和随机变量联系起来了。这种联系也可以推广到 k 维，这样就和随机向量联系起来。

随机抽样

- 在统计中，我们总是通过观测或试验以取得信息。如果我们按照机会均等的原则随机地选取一些个体进行观测或测试某一指标 X 的数值，我们把这一过程称为随机抽样。假如我们抽取了 n 个个体，且这 n 个个体的某一指标为 $(X_1, X_2, \cdots, X_n)$，我们称这 n 个个体的指标 $(X_1, X_2, \cdots, X_n)$ 为一个样本，n 称作这个样本的容量。在重复取样中每个 X_i 是一个随机变量，从而我们把容量为 n 的样本 $(X_1, X_2, \cdots, X_n)$ 看成一个 n 维随机向量。
- 在一次抽样以后，观测到 $(X_1, X_2, \cdots, X_n)$ 的一组确定的值 $(x_1, x_2, \cdots, x_n)$ 称作容量为 n 的样本的观测值或数据。容量为 n 的样本的观测值 $(x_1, x_2, \cdots, x_n)$ 可以看作一个随机试验的一个结果，它的一切可能的结果的全体构成一个样本空间。它可以是 n 维空间，也可以是其中的一个子集。而样本的一组观测值 $(x_1, x_2, \cdots, x_n)$ 是样本空间的一个点。

简单随机样本

- 在实际中，从总体中抽取样本可以有各种不同的方法。为了使抽到的样本能够对总体作出比较可靠的推断，就希望它能很好地代表总体，这就需要对抽样方法提出一些要求。比如：（1）总体中每一个个体有同等机会选入样本；（2）样本的分量 $X_1, X_2, \cdots, X_n$ 是相互独立的随机变量，即样本的每一分量有什么观测结果并不影响其它分量有什么观测结果。这样取得的样本称为简单随机样本。例如返回抽样所得的样本就是简单随机样本。
- 设总体 X 具有分布函数 $F(x)$，$(X_1, X_2, \cdots, X_n)$ 为取自这一总体的容量为 n 的样本，则 $(X_1, X_2, \cdots, X_n)$ 的联合分布函数

$$F^*(x_1, x_2, \cdots, x_n) = \Pi_{i=1}^n F(x_i).$$

又若 X 具有概率密度 f，则 $(X_1, X_2, \cdots, X_n)$ 的概率密度为

$$f^*(x_1, x_2, \cdots, x_n) = \Pi_{i=1}^n f(x_i).$$

直方图、箱线图

- 为了研究总体分布的性质，人们通过试验得到许多观测值，一般来说，这些数据是杂乱无章的，为了利用它们进行统计分析，要将这些数据加以整理，还常借助于表格或图形对它们加以描述。例如，对于连续型随机变量 X 引入频率直方图或数据的箱线图，它们可以使人们对总体 X 的分布有一个粗略的了解。
- 另一方面，我们知道，样本是总体的反映，但是样本所含的信息不能直接用于解决我们所要研究的问题，而需要把样本所含的信息进行数学上的加工，使其浓缩起来，从而解决我们的问题。这在统计学当中，往往通过构造一个合适的依赖于样本的函数——统计量来达到。

统计量

定义 1

设 $(X_1, X_2, \cdots, X_n)$ 是来自总体 X 的一个样本，$g(X_1, X_2, \cdots, X_n)$ 是 $X_1, X_2, \cdots, X_n$ 的函数，若 g 中不含未知参数，则称 $g(X_1, X_2, \cdots, X_n)$ 是一统计量。

显然统计量也是一个随机变量。设 $(x_1, x_2, \cdots, x_n)$ 是相应于样本 $(X_1, X_2, \cdots, X_n)$ 的样本值，则称 $g(x_1, x_2, \cdots, x_n)$ 是 $g(X_1, X_2, \cdots, X_n)$ 的观测值。常用的统计量包括：

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

统计量

它们的观察值分别为：

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$
- $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- $\alpha_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \quad k = 1, 2, \cdots$
- $b_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k, \quad k = 2, 3, \cdots$

这些观察值仍分别称为样本均值，样本方差，样本标准差，样本 k 阶（原点）矩以及样本 k 阶中心矩。

样本矩收敛到总体矩

定理 1

若总体 X 的 k 阶矩 $E\left(X^k\right) \overset{记成}{=} \mu_k$ 存在，则当 $n \to \infty$ 时，$A_k \xrightarrow{P} \mu_k, k = 1, 2, \cdots$。

证明.

因为 $X_1, X_2, \cdots, X_n$ 独立且与 X 同分布，所以 $X_1^k\ X_2^k\ \cdots\ X_n^k$ 独立且与 X^k 同分布，故有

$$E\left(X_1^k\right) = E\left(X_2^k\right) = \cdots = E\left(X_n^k\right) = \mu_k.$$

从而由辛钦大数定理知，$A_k = \frac{1}{n}\sum_{i=1}^n X_i^k \xrightarrow{P} \mu_k, \quad k = 1, 2, \ldots$

进而由关于依概率收敛的序列的性质知道

$$g(A_1, A_2, \cdots, A_k) \xrightarrow{P} g(\mu_1, \mu_2, \cdots, \mu_k)$$

其中 g 为连续函数。 □

这就是下一讲所要介绍的矩估计法的理论根据。

经验分布函数——与总体分布函数 $F(x)$ 相应的统计量

设 $(x_1, x_2, \cdots, x_n)$ 是取自分布为 $F(x)$ 的总体中一个简单随机样本的观测值。若把样本观测值由小到大进行排列，得到 $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$，这里 $x_{(1)}$ 是样本观测值 $(x_1, \cdots, x_n)$ 中最小一个，$x_{(i)}$ 是样本观测值中第 i 个小的数等，则

$$F_n(x) = \begin{cases} 0 & \text{当} x \leqslant x_{(1)} \\ \frac{k}{n} & \text{当} x_{(k)} < x \leqslant x_{(k+1)}, k = 1, 2, \cdots, n-1 \\ 1 & \text{当} x > x_{(n)} \end{cases}$$

显然，$F_n(x)$ 是一非减左连续函数，且满足

$$F_n(-\infty) = 0 \text{ 和} F_n(+\infty) = 1$$

由此可见，$F_n(x)$ 是一个分布函数, 称作经验分布函数 (或子样分布函数)。

经验分布函数——与总体分布函数 $F(x)$ 相应的统计量

对于经验分布函数 $F_n(x)$，格里汶科 (Glivenko) 在 1933 年证明了以下的结果：对于任一实数 x，当 $n \to \infty$ 时 $F_n(x)$ 以概率 1 一致收敛于分布函数 $F(x)$，即

$$P\left\{\lim_{n\to\infty} \sup_{-\infty<x<\infty} |F_n(x) - F(x)| = 0\right\} = 1$$

因此，对于任一实数 x 当 n 充分大时，经验分布函数的任一个观察值 $F_n(x)$ 与总体分布函数 $F(x)$ 只有微小的差别，从而在实际上可当作 $F(x)$ 来使用。

抽样分布

定义 2

统计量的分布称为抽样分布。

在使用统计量进行统计推断时，常需要知道它的分布。当总体的分布函数已知时，抽样分布是确定的，然而要求出统计量的精确分布，一般来说是困难的。
来自正态总体的几个常用的统计量的抽样分布：

- χ^2 分布
- t 分布，也称为学生分布
- F 分布

上述三个分布称为统计学的三大分布，它们在数理统计中有着广泛的应用。

总结

统计学的基本问题之一是统计推断，在数据科学或机器学习领域，称之为学习，是指利用数据（样本）去推断产生这些数据（样本）分布的过程。因此这里又有几个问题：

- 如何根据样本去推断？
- 推断结果的评价？

统计推断问题

统计学的基本问题之一是统计推断，在数据科学或机器学习领域，称之为学习，是指利用数据（样本）去推断产生这些数据（样本）的总体分布的过程。一个典型的统计推断问题是：

- 给定样本 $X_1, ..., X_n \sim F$，怎样去推断总体分布 F?
- 某些情况下，只需推断分布 F 的某种性质，如数字特征，包括均值方差等。

通常把数据服从的一系列分布称为概率模型或统计模型。

定义 3

（概率模型）概率模型 $\mathfrak{F}$ 是指一系列分布（或密度或回归函数），通常也称为统计模型。

统计推断问题

本课程我们只讨论总体分布是连续型和离散型两种情形。为了简便起见，我们引入一个对两种情形通用的概念——概率函数的概念。我们称随机变量（总体）X 的概率函数为 $f(x)$ 的意思是指：

- 在连续情形时，$f(x)$ 是 $X = x$ 的密度函数值；
- 在离散情形时，$f(x)$ 是 $X = x$ 的概率。

一般地，在实际推断中，我们对样本总体分布情况的了解有两种可能性：一种是其形式已知，并且可以用有限个参数来表示（虽然这些参数可能是未知的）；另一种是其形式未知，或者其形式已知但不能用有限个参数来表示。由此我们引出分布的表示：参数和非参数模型。

24.2.1 分布的表示：参数与非参数模型

定义 4

（参数模型）参数模型是指一系列可用有限个参数表示的概率模型 $\mathfrak{F}$。

一般地，参数模型可以用一族带参数 θ 的概率函数来表示，具有如下形式:

$$\mathfrak{F} = \{f(x;\theta) : \theta \in \Theta\},$$

其中 $f(x;\theta)$ 是总体（也就是随机变量）X 的概率函数，参数 θ（可能是标量或向量）除了只知道它的可能取值范围为 Θ 外，其它一无所知。今后我们称 Θ 为参数空间。如果 θ 是向量，但仅关心其中的一个元素的时候，则称其它参数为冗余参数。例如 $\{N(\mu;1) : \mu \in R\}$ 是 μ 取实数值的一族正态分布。

定义 5

非参数模型指一些不能用有限个参数表示的概率模型 $\mathfrak{F}$。

例如，$\mathfrak{F}_{\text{所有}} = \{\text{ 所有 CDF}\}$ 就是非参数模型。非参数模型是相对参数模型来说的。

例 1

（一维参数估计） 令 $X_1, \cdots, X_n$ 为相互独立的 *Bernoulli(p)* 观察值，问题是如何估计参数 p。

例 2

（二维参数估计） 假设 $X_1, \cdots, X_n \sim F$ 并假设 *PDF* $f \in \mathfrak{F}$，其中 $\mathfrak{F}$ 满足高斯分布。这种情况下就有两个参数 μ 和 σ，目标是根据数据去估计这两个参数，如果仅关心估计 μ 的值，则 μ 就是感兴趣的参数而 σ 就是冗余参数。

例 3

（*CDF* 的非参数估计） 令 $X_1, \cdots, X_n$ 是来源于 *CDF* 为 F 的独立观察值，问题是在假设 $F \in \mathfrak{F}_{\text{所有}} = \{$ 所有 *CDF*$\}$ 的前提下如何去估计 F.

例 4

（非参数密度估计） 令 $X_1, \cdots, X_n$ 是来源于 *CDF* 为 F 的独立观察值，令 $f = F'$ 为 *PDF*。假设要估计 *PDF* f. 如果仅假设 $F \in \mathfrak{F}_{\text{所有}}$ 是不可能估计 f 的，需要假设 f 的光滑性，例如，假设 $f \in \mathfrak{F} = \mathfrak{F}_{\text{DENS}} \cap \mathfrak{F}_{\text{SOB}}$，其中，$\mathfrak{F}_{\text{DENS}}$ 表示所有密度函数的集合

$$\mathfrak{F}_{SOB} = \{f\colon \int \left(f''(x)\right)^2 dx < \infty\} \tag{1}$$

集合 $\mathfrak{F}_{SOB}$ 称为**索伯列夫空间**（*Sobolev space*），它表示一系列“波动不大”的函数的集合。

例 5

(函数的非参数估计) 令 $X_1, ..., X_n \sim F$。假定要在仅假设 μ 存在的条件下去估计 $\mu = T(F) = \int xdF(x)$，通常情况下，任何 F 的函数称为**统计泛函**，其他一些统计泛函的例子有方差 $T(F) = \int x^2 dF(X) - (\int xdF(x))^2$，中位数 $T(F) = F^{-1}(1/2)$。

例 6

（回归，预测与分类）假设有成对的观察值 $(X_1, Y_1), \cdots, (X_n, Y_n)$，如 X_i 表示第 i 个患者的血压，Y_i 表示该患者能活多久. X 称为**预测变量** 或**回归变量**或**特征变量**或**自变量**，Y 称为**输出变量**或**响应变量**或**相应变量**. 称 $r(x) = \mathbb{E}(Y \mid X = x)$ 为**回归函数**. 如果假设 $r \in \mathfrak{F}$，其中，$\mathfrak{F}$ 是有限维的，如直线集，则称模型为**参数回归模型**，如果假设 $r \in \mathfrak{F}$，其中，$\mathfrak{F}$ 不是有限维的，则称模型为**非参数回归模型**. 对一个新的病人，根据他的 X 值去预测 Y 称为**预测**，如果 Y 是离散的（例如，生或死），则称为**分类**，如果目标是估计函数 r，则称为**回归估计**或**曲线估计**，有时回归模型也记为

$$Y = r(X) + \varepsilon$$

其中，$\mathbb{E}(\varepsilon) = 0$，通常也用这种方式来描述回归模型，为进一步理解，定义 $\varepsilon = Y - r(X)$，则 $Y = Y + r(X) - r(X) = r(X) + \varepsilon$. 此外，$\mathbb{E}(\varepsilon) = \mathbb{E}\mathbb{E}(\varepsilon \mid X) = \mathbb{E}(\mathbb{E}(Y - r(X)) \mid X)$
$= \mathbb{E}(\mathbb{E}(Y \mid X) - r(X)) = \mathbb{E}(r(X) - r(X)) = 0.$

24.2.2 统计推断的基本概念

有了总体分布的参数和非参数模型表示，接下来我们需要对总体进行参数和非参数统计推断：

- 对于参数模型：我们的任务是，如何根据已知的信息，在分布族 $\{f(x;\theta) : \theta \in \Theta\}$ 中选定一个分布作为总体的分布。用统计的语言就是根据已知信息估计出未知参数 θ 的值。这样，就能使总体的分布从不明确变成明确的了。
- 对于非参数模型：我们的任务是，在没有关于总体累积分布函数 F 或者概率函数 $f(x)$ 的任何假设或者仅有一般性假设（例如连续分布、对称分布等）的前提下，作出一个累积分布函数 F 或者一个概率函数 $f(x)$ 的一致估计。

24.2.2 统计推断的基本概念

- 参数模型推断属于参数统计问题，非参数模型推断属于非参数统计问题。
- 例如，检验“两个总体有相同分布”这个假设，若假定两总体的分布分别为正态分布 $N(\mu_1, \sigma_2)$ 和 $N(\mu_2, \sigma_2)$，则问题只涉及三个实参数 μ_1, μ_2, σ_2，这是参数统计问题。若只假定两总体的分布为连续，此外一无所知，问题涉及的分布不能用有限个实参数刻画，则这是非参数统计问题。

本课程我们主要讨论参数统计推断；对于非参数统计推断，我们主要限定在对概率密度函数或回归函数的非参估计讨论。

统计推断的基本概念

研究统计推断的的方法有多种，最主要的有两大类方法：

- 古典的频率统计推断
- 贝叶斯推断

许多统计推断问题可以归入以下三类：

- 点估计
- 置信区间
- 假设检验

下面对这三类问题做一个简单的介绍。

1. 点估计

- 点估计：是指对感兴趣的某一单点提供“最优估计”。感兴趣的点可以是参数模型、分布函数 F、概率函数 f 和回归函数 r 等中的某一参数，或者可以是对某些随机变量的未来值 Y 的预测。
- 参数的点估计：假设总体 X 的分布函数的形式已知，但它的一个或多个参数未知，借助于总体 X 的一个样本来估计总体未知参数的值称为参数的点估计。
- 记 θ 的点估计为 $\hat{\theta}$ 或 $\hat{\theta}_n$。注意 θ 是固定且未知的，而估计 $\hat{\theta}$ 依赖于数据，所以它是随机的。

点估计问题的数学描述

估计量

设 $X_1, X_2, \cdots, X_n$ 是取自总体 X 的一个样本。我们构造一个统计量 $\hat{\theta} = u(X_1, X_2, \cdots, X_n)$ 作为参数 θ 的估计，称这个统计量 $\hat{\theta}$ 为参数 θ 的一个估计量。

估计值

若 $(x_1, x_2, \cdots, x_n)$ 是样本 $(X_1, X_2, \cdots, X_n)$ 的一组观测值，则 $\hat{\theta} = u(x_1, x_2, \cdots, x_n)$ 就是 $\hat{\theta}$ 的一个点估计值或简称估计值。

点估计问题

如果分布簇中含有 k 个未知参数，即 $\{f(x; \theta_1, \cdots, \theta_k) : (\theta_1, \cdots, \theta_k) \in \Theta\}$，则需要构造 k 个统计量 $\hat{\theta_1} = u_1(X_1, X_2, \cdots, X_n)$，$\cdots$，$\hat{\theta_k} = u_k(X_1, X_2, \cdots, X_n)$ 分别作为 $\theta_1, \cdots, \theta_k$ 的估计量。这种问题又称为参数的点估计问题。

点估计问题的数学描述

- 由上面看到，要求参数 θ 的估计值，必须先构造一个估计量，然后把样本观测值代入估计量得到一个估计值。
- 寻找估计量是寻找参数 θ 的估计值的一个前提，绝不是针对一组具体的观测值去定一个估计值，因为对于一组观测值所决定的估计值是不可能知道这个估计的好坏的，而必须从总体出发，在大量重复取样的情况下，才能评价估计的好坏。
- 研究估计的好坏，一个很自然的想法是研究参数 θ 的一个估计量与参数 θ 的真值之间的偏差在统计意义下是大？还是小呢？
- 在统计意义下，偏差小的估计量可以认为是较好的估计量。

在下一讲介绍估计量的构造方法之前，下面我们先简要介绍估计量的评价。

估计量的评价

对于同一参数，用不同的估计方法求出的估计量可能不相同，原则上任何统计量都可以作为未知参数的估计量，一个自然的问题是，采用哪一个估计量为好？这涉及用什么样的标准来评价估计量的问题。主要有三个评价标准：

- 无偏性
- 有效性
- 相合性

设 $X_1, X_2, \cdots, X_n$ 是取自总体的一个样本，$\theta \in \Theta$ 是包含在总体 X 的分布中的待估参数，这里 Θ 是 θ 的取值范围。首先给出无偏性。

无偏性

无偏性（数学期望)

若估计量 $\hat{\theta} = u(X_1, X_2, \cdots, X_n)$ 的数学期望 $E(\hat{\theta})$ 存在，且对任意的 $\theta \in \Theta$ 有

$$E(\hat{\theta}) = \theta,$$

则称 $\hat{\theta}$ 是 θ 的无偏估计量。

估计量的偏差定义为

$$bias(\hat{\theta}) = E(\hat{\theta}) - \theta,$$

称为以 $\hat{\theta}$ 作为 θ 的估计的系统误差。无偏估计的实际意义就是无系统误差。

有效性

为了比较参数 θ 的两个无偏估计量 Θ_1 和 Θ_2 哪个更好，需要引入估计量的有效性。

有效性 (风险小)

设 $\hat{\theta}_1 = u(X_1, X_2, \cdots, X_n)$ 和 $\hat{\theta}_2 = u(X_1, X_2, \cdots, X_n)$ 都是 θ 的无偏估计量，若对于任意的 $\theta \in \Theta$ 有

$$D(\hat{\theta}_1) \leq D(\hat{\theta}_2),$$

且至少对于某一个 $\theta \in \Theta$ 上式中的不等号成立，则称 Θ_1 较 Θ_2 有效，也即方差越小，越有效。

相合性

前面讲的无偏性和有效性都是在样本容量 n 固定的前提下提出的，我们自然希望随着样本容量的增大，也即收集的数据越来越多的时候，一个估计量的值稳定于待估参数的真值。这样需要引入估计量的相合性要求。

相合性（一致性）

设 $\hat{\theta} = u(X_1, X_2, \cdots, X_n)$ 为参数 θ 的估计量，若对任意的 $\theta \in \Theta$，当 $n \to \infty$ 时 $\hat{\theta} = u(X_1, X_2, \cdots, X_n)$ 依概率收敛于 θ，则称 $\hat{\theta}$ 为 θ 的相合估计量。

即对任意的 $\theta \in \Theta$ 都满足：对于任意的 $\varepsilon > 0$，有

$$\lim_{n \to \infty} P\left\{\left|\hat{\theta} - \theta\right| \geq \varepsilon\right\} = 0,$$

则称 $\hat{\theta}$ 为 θ 的相合估计量。

均方误差评价

点估计的质量好坏有时也用均方误差或 MSE 来评价，均方误差定义为

$$\mathrm{MSE} = \mathbb{E}_\theta \left(\hat{\theta} - \theta\right)^2$$

要注意 $E_\theta(\cdot)$ 是关于如下分布的期望而不是关于 θ 分布的平均，该分布由数据得来，具体如下

$$f(x_1, \cdots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

均方误差评价

定理 2

MSE 可写成如下形式:

$$\mathrm{MSE} = \mathrm{bias}^2\left(\hat{\theta}\right) + D\left(\hat{\theta}\right)$$

证明：令 $\bar{\theta} = \mathbb{E}_\theta\left(\hat{\theta}\right)$, 则

$$\begin{aligned}
\mathbb{E}_\theta\left(\hat{\theta} - \theta\right)^2 &= \mathbb{E}_\theta\left(\hat{\theta} - \bar{\theta} + \bar{\theta} - \theta\right)^2 \\
&= \mathbb{E}_\theta\left(\hat{\theta} - \bar{\theta}\right)^2 + 2\left(\bar{\theta} - \theta\right)\mathbb{E}_\theta\left(\hat{\theta} - \bar{\theta}\right) + \mathbb{E}_\theta\left(\bar{\theta} - \theta\right)^2 \\
&= \left(\bar{\theta} - \theta\right)^2 + \mathbb{E}_\theta\left(\hat{\theta} - \bar{\theta}\right)^2 \\
&= \mathrm{bias}^2\left(\hat{\theta}\right) + D\left(\hat{\theta}\right).
\end{aligned}$$

推导过程中用到了如下事实：$\mathbb{E}_\theta\left(\hat{\theta} - \bar{\theta}\right) = \bar{\theta} - \bar{\theta} = 0$

估计量的评价

$\hat{\theta}$ 的分布称为抽样分布，$\hat{\theta}$ 的标准差称为标准误差，记为 se，

$$se = se(\hat{\theta}) = \sqrt{D(\hat{\theta})}.$$

通常标准误差依赖于未知分布 F，在另外一些情况下，se 是未知量，但通常去估计它，估计的标准误差记为 $\hat{se}$。

估计量的评价

例 7

在抛硬币的试验中，令 $X_1, \cdots, X_n \sim \text{Bernoulli}(p), \hat{p}_n = n^{-1}\sum X_i$，则：

（1）$\mathbb{E}(\hat{p}_n) = n^{-1}\sum \mathbb{E}(X_i) = p$，所以 $\hat{p}_n$ 是无偏的。

（2）标准误差为 $se = \sqrt{D(\hat{p}_n)} = \sqrt{p(1-p)/n}$，估计的标准误差为 $\hat{se} = \sqrt{\hat{p}(1-\hat{p})/n}$。

（3）因为 $\mathbb{E}(\hat{p}_n) = p$，所以 $bias = p - p = 0$，$se = \sqrt{p(1-p)/n} \to 0$，因此 $\hat{p}_n \xrightarrow{P} p$，即 $\hat{p}_n$ 是一致估计量，是相合的。

今后将要遇到的许多估计量都近似服从正态分布。

定义 6

如果 $\frac{\hat{\theta}_n - \theta}{se} \rightsquigarrow N(0,1)$，则称估计量 $\hat{\theta}_n$ 是渐进正态的。

点估计的方法

那么怎样构造估计量呢？参数的点估计方法包括：

- 矩估计（频率学派）
- 极大似然估计（频率学派）
- 极大后验估计（贝叶斯学派）
- 贝叶斯估计（贝叶斯学派）

2. 置信区间

对于未知参数 θ，除了求出它的点估计 $\hat{\theta}$ 外，我们还希望估计出一个范围，并希望知道这个范围包含参数 θ 真值的可信程度。这样的范围通常以区间的形式给出，同时还给出此区间包含参数 θ 真值的可信程度。这种形式的估计称为区间估计，这样的区间即所谓的置信区间。

置信区间 设总体 X 的分布函数 $F(x:\theta)$ 含有一个未知参数 $\theta, \theta \in \Theta$（$\Theta$ 是 θ 可能取值的范围），对于给定值 α（$0<\alpha<1$），若由来自 X 的样本 $X_1, X_2, \cdots, X_n$ 确定的两个统计量 $\underline{\theta}=\underline{\theta}(X_1, X_2, \cdots, X_n)$ 和 $\bar{\theta}=\bar{\theta}(X_1, X_2, \cdots, X_n)$（$\underline{\theta}<\bar{\theta}$），对于任意 $\theta<\Theta$ 满足

$$P\left\{\underline{\theta}\left(X_1, X_2, \cdots, X_n\right)<\theta<\bar{\theta}\left(X_1, X_2, \cdots, X_n\right)\right\} \geqslant 1-\alpha \tag{2}$$

则称随机区间 $C_n=(\underline{\theta}, \bar{\theta})$ 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。

置信区间

- 当 X 是连续型随机变量时，对于给定的 α，我们总是按要求 $P\{\underline{\theta} < \theta < \bar{\theta}\} = 1 - \alpha$ 求出置信区间。
- 而当 X 是离散型随机变量时，对于给定的 α, 常常找不到区间 $(\underline{\theta}, \bar{\theta})$ 使得 $P\{\underline{\theta} < \theta < \bar{\theta}\}$ 恰为 $1 - \alpha$。此时我们去找区间 $(\underline{\theta}\ \bar{\theta})$ 使得 $P\{\underline{\theta} < \theta < \bar{\theta}\}$ 至少为 $1 - \alpha$，且尽可能地接近 1—α。
- C_n 是随机的而 θ 是固定的。
- 如果 θ 是向量，则用置信集（例如球面或者椭圆面）代替置信区间。

置信区间的两种解释

1. 式(2)的含义如下：若反复抽样多次（各次得到的样本的容量相等，都是 n）。每个样本值确定一个区间 $(\underline{\theta}, \bar{\theta})$，每个这样的区间要么包含 θ 的真值，要么不包含 θ 的真值。按伯努利大数定理，在这么多的区间中，包含 θ 真值的约占 $100(1-\alpha)\%$, 不包含 θ 真值的约仅占 $100\alpha\%$。例如, 若 $\alpha = 0.05$, 反复抽样 1000 次, 则得到的 1000 个区间中不包含 θ 真值的约仅为 50 个。该解释并没有错误，但用处不大，因为人们很少反复地多次重复相同的试验
2. 第 1 次，对于参数 θ，收集到数据并建立了 95% 的置信区间，第 2 次，对于参数 θ_2，收集到数据并建立了 95% 的置信区间，第 3 次，对于参数 θ_3。收集到数据并建立了 95% 的置信区间，继续这一过程，对一系列不相关参数 $\theta_1, \theta_2, \cdots$ 建立置信区间, 则这些置信区间有 95% 的概率覆盖真实的参数值，这一解释不需要反复地重复同一试验。

置信区间举例

例 8

报纸每天都会报道民意调查地结果。例如，报道称“有 83% 的公众对飞行员随身配备真枪飞行的做法表示赞同”，通常你还会看到诸如这样的陈述“该调查有 95% 的概率在 4 个百分点的范围内变动”。意思就是赞同飞行员随身配备真枪飞行的做法的人数所占的比例 p 的 95% 的置信区间是 $83\% \pm 4\%$，如果以后都按这种方式建立置信区间，则有 95% 的区间将包括真实的参数值，即使每天估计的量不同（不同的民意测验），这一结论也是正确的。

置信区间举例

例 9

在抛硬币的试验中，令 $C_n = (\hat{p}_n - \varepsilon, \hat{p}_n + \varepsilon)$，其中 $\varepsilon^2 = \log(2/\alpha)/(2n)$，由霍夫丁不等式得，对任意 p

$$\mathbb{P}(p \in C_n) \geqslant 1 - \alpha$$

因此，C_n 是 $1-\alpha$ 置信区间。

3. 假设检验

- 统计推断的另一类重要问题是假设检验问题。在总体的分布函数完全未知或只知其形式，但不知其参数的情况。为了推断总体的某些未知特性，提出某些关于总体的假设。
- 例如，提出总体服从泊松分布的假设，又如对于正态总体提出数学期望等于 μ_0 的假设等。我们要根据样本对所提出的假设作出是接受，还是拒绝的决策。假设检验是作出这一决策的过程。

假设检验

在假设检验中，从缺省理论，即原假设开始，通过数据是否提供显著性证据来支持拒绝该假设，如果不能拒绝，则保留原假设。

例 10

（检验硬币是否均匀）令

$$X_1, \cdots, X_n \sim \text{Bernoulli}(p)$$

为 n 次独立的硬币投掷结果，假设要检验硬币是否均匀，令 H_0 表示硬币是均匀的假设，H 表示硬币不是均匀的假设，H_0 称为原假设，H_1 称为备择假设，可以将假设写成

$$H_0 : p = 1/2 \text{ 对比} H_1 : p \neq 1/2.$$

如果 $T = |\hat{p}_n - \frac{1}{2}|$ 的值很大，则有理由拒绝 H_0，当详细讨论假设检验的时候，将会确定出拒绝 H_0 的精确 T 值。

其它推断

参数估计和非参数估计也分别称为参数推断和非参数推断。除了这两种推断，统计推断还包括：

- 独立性推断
- 因果推断

本讲小结

参数与非参数模型

- 参数密度估计
- 函数（CDF）的非参数估计
- 非参数密度估计

统计推断的基本概念

- 点估计
- 置信区间
- 假设检验

本讲只对参数与非参数模型，统计推断的基本概念做了简单的介绍，关于概率函数的估计方法，特别是参数估计和非参数估计的方法，没有涉及，将在下一讲进行详细介绍！