数理统计笔记

作者：肖程哲

时间：August 15, 2022

苟日新，日日新，又日新

目录

第 1 章 导论: 从数据中学习

统计学的主要研究内容是如何处理随机数据。研究者希望设计试验收集数据，从中获取信息，并借此解释和预测 数据。统计推断 (inference) 基于概率模型，对两个基本问题：估计(estimation) 与检验(testing) 进行回答。

1.1 统计模型

定义 1.1 (样本)

在实验中获取的观测值 $\mathbf{x} = (x_1, \cdots, x_n)$ 视为某个概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$ 上的随机元 $\mathbf{X} = (X_1, \cdots, X_n)$ 的实现(realization)。一组观测值构成一个**样本** (sample)，数据的数目 n 称为**样本容量** (sample size)。♣

定义 1.2 (总体)

假设 X_i 独立同分布，其分布 $P(\bullet) := \mathbb{P}\{X_i \in \bullet\}$ 称为**总体** (population)。♣

在统计学中可以认为总体 P 包含我们想知道的一切信息，然而（至少部分）是未知的。我们希望用样本 $\mathbf{x}$ 推断总体 P 的性质。

定义 1.3 (参数)

在总体中固定但未知的常数称为**参数**, 记为 θ. 参数所有可能的取值构成**参数空间**, 记为 Θ, 可以是有限维或者无限维。♣

注 参数的函数同样固定且未知，故也视为参数.

定义 1.4 (统计模型)

统计模型 (statistical model) 是样本 $\mathbf{X}$ 对应的所有可能的总体 P 的集合。在同一统计模型中，不同的总体通过参数区分，所以将模型记为：

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

♣

统计模型代表关于数据产生机制的先验 (prior) 知识。

定义 1.5 (可识别)

若模型 $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ 满足：

$$P_{\theta_1} \neq P_{\theta_2}, \forall \theta_1 \neq \theta_2$$

则称模型 $\mathcal{P}$ **可识别** (identifiable) ♣

定义 1.6

若 Θ 有限维的，则称模型 $\mathcal{P}$ 为**参数族** (parametric family)；若参数空间 Θ 是无限维的，则称为**非参数族**。♣

我们常常只关心参数 Θ 的某些分量的函数 $\gamma = g(\theta)$，剩下的碍事且无用的部分称为**冗余参数** (nuisance parameter)。

1.2 统计量

定义 1.7 (统计量)

给定样本 $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} P$，其中 $P \in \mathcal{P}$ 是未知的总体. 若 $T: (\mathcal{X}_n, \mathscr{F}_{\mathcal{X}}^n) \to (\mathcal{T}, \mathscr{F}_{\mathcal{T}})$ 是已知的 (不依赖 P 的) 可测函数，则称 $T(\mathbf{X}) = T(X_1, \cdots, X_n)$ 为**统计量** (statistic)。 ♣

定义 1.8

统计量的分布称为它的**抽样分布** (sampling distribution), 含有总体 (样本分布) 的一部分信息. ♣

统计量函数, 不依赖未知总体, 给出了一种数据约简 (reduction). 对于实数值样本 $x = \{x_1, \ldots, x_n\}$, 常见的统计量有:

样本均值 (sample mean)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

样本方差 (sample variance)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

由此可得**样本标准差** (sample standard deviation) $s = \sqrt{s^2}$.

样本中位数 (sample median)

$$M = \operatorname{med}(x) = \begin{cases} x_{(k)}, & n = 2k-1 \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}), & n = 2k \end{cases}$$

其中顺序统计量 $x_{(1)} \le \cdots \le x_{(n)}$ 由 $x_1, \ldots, x_n$ 排列得到.

四分位数 (quartile)

$$Q_1 = \operatorname{med}(x \cap (-\infty, M)), \qquad Q_3 = \operatorname{med}(x \cap (M, +\infty)).$$

四分位距 (inter quartile range)

$$IQR = Q_3 - Q_1.$$

极差 (range)

$$x_{(n)} - x_{(1)} = \max(x) - \min(x) = \max_{1 \le i, i' \le n}\{x_i - x_{i'}\}.$$

定义 1.9

五数概括法 (five-number summary): 以 $x_{(1)}, Q_1, M, Q_3, x_{(n)}$ 总结 x, 可用**箱形图** (boxplot) 表示. ♣

第 2 章 抽样分布

2.1 统计量的极限分布

2.2 指数族

2.3 充分统计量

若样本数据量大，可能难以解释。试验者希望提取样本值的一些关键特征以概括样本中的信息。这类数据简化（缩减）在计算统计学中通常以样本函数的形式实现，例如，样本均值、样本方差、最大观测值和最小观测值就是四个概括样本关键特征的统计量。

任意一个统计量 $T(X)$ 都定义了一种数据简化方式。如果试验者只观测统计量 $T(x)$ 而非整个样本 x，则他必将满足 $T(x) = T(y)$ 的 x 和 y 视作两个相同的样本，尽管事实可能并非如此。不同的统计量对数据中的信息划分有不同的方法。

依据某统计量简化样本数据可以看成样本空间 $\mathcal{X}$ 上的一个划分。设 $\mathcal{T} = \{t|\exists x \in \mathcal{X}, \text{s.t.} \quad t = T(x)\}$ 为 $\mathcal{X}$ 在 $T(x)$ 下的象。则 $A_t = \{x|T(x) = t, t \in \mathcal{T}\}$ 为 $\mathcal{X}$ 若干划分。

原始数据包含了所有信息，规律或随机部分。若进行转化，则将丢失信息，可能有用，也可能无用。其中界限由假设的统计模型判断。转化后的可能结果：

1. 留下部分有用信息：完备统计量
2. 留下所有有用信息：充分统计量
3. 不留下有用信息：辅助统计量

例题 2.1 设 $X_1, \ldots, X_n \sim Binomial(p) \quad \text{i.i.d.}$，设 $T(X) = \sum X_i$。若 $T = t$ 已知，则实验结果与 p 无关，由于：

$$P(X|T) = \frac{P(X)}{P(T)} = \frac{p^t(1-p)^{n-t}}{\binom{n}{t}p^t(1-p)^{n-t}} = \frac{1}{\binom{n}{t}}$$

注 $T \sim Binomial(n, p)$ 与 p 有关，而 $X|T$ 与参数无关。即原始数据经 T 转化后的 $T(X)$，仍包含所有关于参数的信息；而余下的 $X|T$ 不再包含参数信息。

定义 2.1 (充分统计量)

假设样本 X，满足分布 $P(\theta)$，若统计量 $S(X)$，$P(X|S)$ 与 θ 无关，则称 $S(X)$ 为关于 θ 的充分统计量。♣

注 充分统计量的判断与统计模型有关，模型不当可能导致充分统计量实际不"充分"。

定理 2.1 (分解定理)

$S(X)$ 为关于 θ 的充要条件为：

$$\exists g(), h() \text{s.t.} f(X|\theta) = g(S(X), \theta)h(X)$$

♡

注 直观理解：$P(X) = P(S)P(X|S)$

证明 对于离散情况：

充分：

$$P(S = s) = \sum_{S(x)=s} P(X = x) = g(s, \theta) \sum_{S(x)=s} h(x)$$

$$P(X = x|S = s) = \frac{P(X = x)}{P(S = s)} = \frac{h(x)}{\sum_{S(x)=s} h(x)}$$

与 θ 无关。

必要：
令

$$g(s,\theta)=P(T=s|\theta),h(x)=P(X=x|S=s)$$

即可

定理 2.2

若 $S(X)$ 为关于 θ 的充分统计量，则 θ 的极　然似估计可表示为 S 的函数。

证明 然似函数为 $g(S,\theta)h(X)$。由于 $h(X)$ 为定值，故只需求 $g(S,\theta)$ 的极值情况，故 θ 的取值可由 S 的函数表示。

为使数据尽可能精简，摈弃无用信息，定义极小充分统计量。

定义 2.2 (极小充分统计量)

若统计量 M 满足

$$\forall S,\exists h \quad \text{s.t.} \quad M=h(S)$$

则其为关于 θ 的充分统计量

2.4 完全统计量

第 3 章 参数估计

3.1 点估计的概念

定义 3.1 (点估计)

用于估计参数 $\gamma = g(\theta)$ 的统计量 $T(\mathbf{X})$ 称为**估计量** (estimator), 记为 $\hat{\gamma} = T(X)$. 估计式是随机变量, 若给定统计模型, 则其分布由参数 θ 决定, 将观测数值代入估计式后得到的值称为**估计值**.

将数据代入估计式得到的估计值并不是真实的参数值，若使用新的样本，很可能会得到不同的估计值。

定义 3.2 (标准误差)

估计式的**标准误差** (standard error) 定义为其抽样分布的标准差, 即:

$$\sqrt{\mathrm{Var}_\theta(\hat{\theta})}$$

定义 3.3 (无偏估计)

估计式的**偏差** (standard error) 定义为其抽样分布的均值与实际参数的偏差, 即:

$$E_\theta(\hat{\theta}) - \theta$$

若其为零, 则称次估计为**无偏估计**

定义 3.4 (相合)

若当样本容量 $n \to \infty$ 时, 有

$$\hat{\gamma} \xrightarrow{\mathbb{P}} \gamma, \quad \forall \theta \in \Theta.$$

则称 $\hat{\gamma}$ 是**相合的/一致的** (consistent), 如果上式中 $\xrightarrow{\mathbb{P}}$ 可以增强为 $\xrightarrow{\text{a.s.}}$, 则称 $\hat{\gamma}$ 是**强相合的**.

3.2 矩估计

定义 3.5 (总体矩与样本矩)

若 $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} P_\theta$，记 k 阶**总体矩**为:

$$\mu_k(\theta) = \mathbb{E}_\theta[X_i^k], \quad k \in \mathbb{N}$$

记 k 阶**样本矩**为：

$$\hat{\mu_k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad k \in \mathbb{N}$$

命题 3.1

k 阶样本矩是关于总体分布 k 阶矩的无偏估计。

证明

定义 3.6

参数 $\gamma = g(\theta)$ 的**矩**(方法)**估计量** (method of moments estimator) 定义为

$$\hat{\gamma}^{\mathrm{MoM}} = g(\hat{\theta}^{\mathrm{MoM}}),$$

其中 $\hat{\theta}^{\mathrm{MoM}}$ 对选定的 k 满足

$$\mu_k(\hat{\theta}^{\mathrm{MoM}}) = \hat{\mu}_k.$$

矩方法的步骤：

1. 将低阶矩写为参数的函数，一般阶数与参数个数相同 $(\mu_i)_n = f((\theta)_n)$；
2. 找出上一步骤的反函数，通过矩表达参数 $(\theta)_n = f^{-1}((\mu_i)_n)$；
3. 将样本矩代入，得到参数的估计式 $\hat{\theta} = f^{-1}(\hat{\mu}_i)$

例题 3.1 泊松分布的矩估计 设 $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} P(\lambda)$，则其一阶矩 $\mu_1 = \lambda$，所以 $\lambda = \mu_1$，其矩估计为：

$$\hat{\lambda} = \hat{\mu_1} = \overline{X}$$

例题 3.2 设 $X_1, \ldots, X_n$ 是独立同分布的连续型随机变量, 具有 p.d.f.

$$f_{\lambda,a}(x) = \lambda \mathrm{e}^{-\lambda(x-a)} \mathbb{1}_{[x>a]}, \quad x \in \mathbb{R},$$

其中 $\lambda > 0$ 和 $a \in \mathbb{R}$ 未知 (注: 相应的统计模型是带有位置 (location) 参数和速率 (rate) 参数的指数分布). 易见 $X_i \overset{d}{=} a + Y/\lambda$, 其中 $Y \sim \mathrm{Exponential}(1)$. 利用 $\mathbb{E}[Y] = 1$ 和 $\mathbb{E}[Y^2] = 2$, 可以得到

$$\mu_1(\lambda, a) = a + 1/\lambda, \quad \& \quad \mu_2(\lambda, a) = a^2 + 2a/\lambda + 2/\lambda^2.$$

方程

$$\mu_k(\hat{\lambda}^{\mathrm{MoM}}, \hat{a}^{\mathrm{MoM}}) = \hat{\mu}_k, \quad k = 1, 2$$

的解

$$\hat{\lambda}^{\mathrm{MoM}} = 1\Big/\sqrt{\hat{\mu}_2 - \hat{\mu}_1^2}, \quad \& \quad \hat{a}^{\mathrm{MoM}} = \hat{\mu}_1 - \sqrt{\hat{\mu}_2 - \hat{\mu}_1^2}$$

即为 (λ, a) 的一种矩估计.

注 矩方法得到的估计量往往可以援引大数定律 (LLN) 来说明相合性.

3.3 极大然似估计

3.4 一致最小方差无偏估计

3.5 Cramer-Rao 不等式

3.6 概率密度函数的核估计

第 4 章 区间估计

4.1 区间估计的概念

4.2 枢轴变量法

4.2.1 正态总体参数

4.2.2 非正态总体参数

4.3 Fisher 的信仰推断法

4.4 容忍区间于容忍限

第 5 章 假设检验

第 6 章 线性模型

6.1 回归分析

6.2 最小二乘法

6.3 方差分析