基础数理统计

第六章 模型、统计推断与学习

1. 6.1 引言
2. 6.2 参数和非参数模型
3. 6.3 统计推断的基本概念
   - 6.3.1 点估计
   - 6.3.2 置信集
   - 6.3.3 假设检验

1 6.1 引言
2 6.2 参数和非参数模型
3 6.3 统计推断的基本概念

统计推断

统计推断，在计算机科学中称为“学习”，指利用数据推断产生这些数据分布的过程。例如：
统计推断：

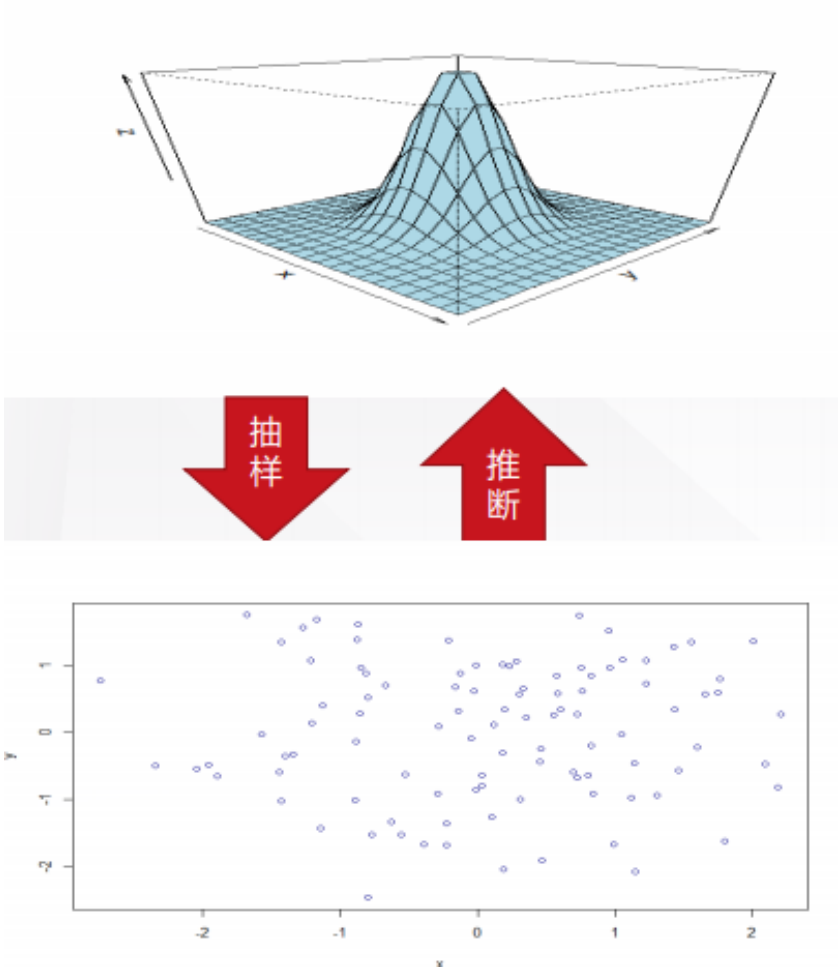
- 给定样本 $X_1, X_2, \ldots, X_n$，怎样去推断 F 及其函数?

统计学习：

- 无监督学习：给定样本 $X_1, \cdots, X_n$, 如何刻画其分布函数 $F(x)$ 或者密度函数 $f(x)$?
- 监督学习：给定样本 $(X_1, Y_1), \cdots, (X_n, Y_n)$, 如何构造一个合适的函数 $f(\cdot)$ 使得

$$Y_i \approx f(X_i), \ i = 1, \ldots, n.$$

抽样与推断



6.2 参数和非参数模型

1 6.1 引言

2 6.2 参数和非参数模型

3 6.3 统计推断的基本概念

统计模型

- 总体：研究对象的全体称为总体；
- 样本：从总体中抽取出来的部分观察值称为样本。通常表示为 $X_1, \cdots, X_n$, 其中 n 为样本大小或样本容量。

定义 1 (统计模型)

统计模型 $\mathcal{F}$ 指一系列分布 (或密度或回归函数)。

参数模型

定义 2 (参数模型)

指一系列可用有限个参数表示的 $\mathcal{F}$，一般具有形式

$$\mathcal{F} = \{f(x, \theta) : \theta \in \Theta\}.$$

其中 θ 表示在参数空间 Θ 中取值的未知参数（或参数向量）。如果 θ 是向量，但仅关心其中一个元素的时候，称其他参数为冗余参数。

例 1

正态分布模型

$$\mathcal{F} = \{f(x : \mu, \sigma) : \quad \mu \in \mathbb{R},\ \sigma > 0\},$$

其中

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

非参数模型

定义 3 (非参数模型)

指一些不能用有限个参数表示的 $\mathcal{F}$.

例 2

$$\mathcal{F} = \{\text{所有 } CDF\}.$$

例 3

$$\mathcal{F} = \left\{ f : f \text{为 } pdf \text{且 } \int (f^{(2)}(x))^2 dx < \infty \right\}.$$

例 4 (函数的非参数估计)

通常情况下，任何 F 的函数称为统计泛函。例如均值，方差和中位数。

例 5 (回归，预测与分类)

给定一组数据 (X_i, Y_i)，$i = 1, \cdots, n$，X 可能会影响 Y。

- X: 预测变量，回归变量，特征变量，自变量
- Y: 输出变量，相应变量，响应变量
- $r(x) = E(Y|X = x)$：回归函数
- Y 的估计：预测（如果是离散的，也叫分类）
- r 的估计：回归估计或曲线估计

1 6.1 引言

2 6.2 参数和非参数模型

3 6.3 统计推断的基本概念

- 6.3.1 点估计
- 6.3.2 置信集
- 6.3.3 假设检验

简单样本

对于简单样本, $X_1, \cdots, X_n$ 独立同分布，来自总体分布 F(或密度函数 f), 通常记为:

$$X_1, \cdots, X_n \ iid \ \sim F \text{ or } X_1, \cdots, X_n \ iid \ \sim f.$$

如果看成多元随机变量，那么 $(X_1, \cdots, X_n)$ 的联合分布函数为:

$$F(x_1, \cdots, x_n) = P(X_1 \leq x_1, \cdots, X_n \leq x_n) = \prod_{i=1}^{n} F(x_i),$$

对应密度函数为:

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

统计量

定义 4 (统计量)

完全基于样本 $X_1, \cdots, X_n$ 所得的量称为统计量。统计量是样本的函数。

常见统计量:

- 样本 m 阶（原点）矩 $\frac{1}{n}\sum_{i=1}^{n} X_i^m$ 以及样本 m 阶中心矩 $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^m$.
- 次序统计量，例如最大值、最小值、中位数等。

统计推断问题

- 点估计
- 置信集
- 假设检验

6.3.1 点估计

1 6.1 引言

2 6.2 参数和非参数模型

3 6.3 统计推断的基本概念
- 6.3.1 点估计
- 6.3.2 置信集
- 6.3.3 假设检验

点估计

点估计是对感兴趣的某一单点 (可以是参数模型中的参数，分布函数或概率密度函数，回归函数在某点的取值等) 的 “最优估计”。

定义 5 (点估计)

令 $X_1, \cdots, X_n$ 为来自某分布的简单随机样本，参数 θ 的点估计记为 $\widehat{\theta}_n = g(X_1, \cdots, X_n)$。

- 估计量的偏差定义为：$\text{bias}(\widehat{\theta}_n) = \mathsf{E}_\theta(\widehat{\theta}_n) - \theta$;
- 如果 $\text{bias}(\widehat{\theta}_n) = 0,\ \theta \in \Theta$，称 $\widehat{\theta}_n$ 是无偏的;
- 如果 $\widehat{\theta}_n \xrightarrow{P} \theta$，称 $\widehat{\theta}_n$ 是相合的;
- $\widehat{\theta}_n$ 的分布称为抽样分布，$\widehat{\theta}_n$ 的标准差称为标准误差，记为 $\text{se} = \text{se}(\widehat{\theta}_n) = \sqrt{\mathsf{V}(\widehat{\theta}_n)}$，通常需要估计;
- 点估计的均方误差定义为 $\text{MSE} = \mathsf{E}_\theta(\widehat{\theta}_n - \theta)^2 = \text{bias}^2(\widehat{\theta}_n) + \mathsf{V}_\theta(\widehat{\theta}_n)$。

点估计–续

定理 1

如果 $bias \to 0$ 且当 $n \to +\infty$ 时 $se \to 0$，则 $\widehat{\theta}_n$ 是相合的。

定义 6 (渐近正态性)

如果

$$\frac{\widehat{\theta}_n - \theta}{se} \rightsquigarrow N(0,1),$$

则称估计量 $\widehat{\theta}_n$ 是渐近正态的。

点估计–续

例 6

$X_1, X_2, \ldots, X_n$ 为来自总体均值为 μ 方差为 σ^2 的简单随机样本，则

(1) $\bar{X}$ 是 μ 的无偏估计；样本方差 S_n^2 为 σ^2 的无偏估计（第三章定理 3)；

(2) $\bar{X}$ 是 μ 的相合估计（大数定律)；

(3) $\bar{X}$ 是渐近正态的（中心极限定理)。

6.3.2 置信集

1 6.1 引言

2 6.2 参数和非参数模型

3 6.3 统计推断的基本概念
- 6.3.1 点估计
- 6.3.2 置信集
- 6.3.3 假设检验

置信集

定义 7 (置信集)

参数 θ 的 $1-\alpha$ 置信区间 $C_n = (a, b)$，其中 $a = a(X_1, X_2, \ldots, X_n)$，$b = b(X_1, X_2, \ldots, X_n)$ 为数据的函数，满足

$$P_\theta(\theta \in C_n) \geq 1-\alpha, \quad \theta \in \Theta. \tag{1}$$

$1-\alpha$ 称为置信区间的覆盖。如果 θ 是向量则用置信集代替置信区间。

评估置信区间的准则:

- 区间大小 $b-a$,
- 置信区间的精度 $1-\alpha$。

置信集–续

注意：

1. θ 是未知参数，为固定值；
2. 式子 (1) 的解释是：对不同的参数值 θ，利用收集到的数据建立置信区间，这些置信区间会有 95% 的概率覆盖真实的参数值。

例子

例 7

$X_1, X_2, \ldots, X_n$ 独立同分布于 $Bernoulli(p)$ 分布，则样本均值 $\bar{X}_n$ 满足

$$P(|\bar{X}_n - p| \geq \epsilon) \leq 2\exp(-2n\epsilon^2) \text{ (第三章定理 4)}$$

取

$$C_n = (\bar{X}_n - \sqrt{\frac{\log(2/\alpha)}{2n}},\ \bar{X}_n + \sqrt{\frac{\log(2/\alpha)}{2n}}), \quad (2)$$

则 $P(p \in C_n) \geq 1 - \alpha$，$C_n$ 为参数 p 的 $1 - \alpha$ 置信区间。

基于正态的置信区间

定理 2 (基于正态的置信区间)

假设 $\widehat{\theta}_n \approx N(\theta, \widehat{se}^2)$，令 Φ 为标准正态分布的 *CDF*，$z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$，令

$$C_n = \left(\widehat{\theta}_n - z_{\alpha/2}\widehat{se},\ \widehat{\theta}_n + z_{\alpha/2}\widehat{se}\right),$$

则 $P_\theta(\theta \in C_n) \to 1-\alpha$。

例 8

令 $X_1, X_2, \ldots, X_n \sim Bernoulli(p)$，则 p 的近似 $1-\alpha$ 置信区间为

$$C_n = (\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}). \tag{3}$$

思考：置信区间 (2) 和 (3) 哪一个更好?

6.3.3 假设检验

1 6.1 引言

2 6.2 参数和非参数模型

3 6.3 统计推断的基本概念
- 6.3.1 点估计
- 6.3.2 置信集
- 6.3.3 假设检验

例 9

某保健品生产企业正常时所生产的保健品中维生素 *D3* 的含量标准为每片 *7.5ug*, 质检部门某天抽检 *36* 片该保健品的维生素 *D3* 的含量数据如下,

$$7.56,\ 7.32,\ 7.23,\ 7.6,\ 7.54,\ 7.34,\ 7.45,\ 7.54,\ 7.43,$$
$$7.33,\ 7.34,\ 7.86,\ 7.29,\ 7.49,\ 7.59,\ 7.48,\ 7.49,\ 7.68,$$
$$7.07,\ 7.47,\ 7.58,\ 7.66,\ 7.59,\ 7.7,\ 7.11,\ 7.62,\ 7.59,\ 7.46,$$
$$7.54,\ 7.41,\ 7.61,\ 7.36,\ 7.37,\ 7.4,\ 7.15,\ 7.67.$$

按照经验，每片保健品中 *D3* 的含量服从正态分布，方差 $\sigma^2 = 0.12^2$, 那么根据该抽检结果该保健品维生素 *D3* 的含量符合每片 *7.5ug* 的标准吗?

用随机变量 X 表示每片保健品中 D3 的含量，则 $X \sim N(\mu, 0.12^2)$, 需要判断 $\mu = 7.5$ 是否成立。可以先假设 $\mu = 7.5$, 称为**原假设**，记为 $H_0 : \mu = 7.5$; 该假设的对立面即 $\mu \neq 7.5$ 称为**备择假设**，记为 $H_1 : \mu \neq 7.5$.

此为假设检验的第一步：根据实际问题提出合理的原假设和备择假设。

原假设与备择假设

(1) 原假设与备择假设往往不能交换，把哪一个作为原假设往往要根据实际问题而定.

(2) 一般把没有充分理由不能轻易否定的命题作为原假设，只有具备充分理由时才拒绝它，而把其他容许的命题做为备择假设.

假设检验的基本法则

基于概率的反证法:

(1) 假设 H_0 成立, 然后基于这个假设构造一个小概率事件, 保证这个小概率事件在 H_0 成立时几乎不会在一次抽样（或试验）中发生。

(2) 如果根据样本数据，发现这个小概率事件发生了，就有理由认为 H_0 不成立，即做出拒绝 H_0 的决策；否则就没有充分理由拒绝 H_0。

基本原理：小概率事件原理，即概率很小的事件在一次实验中几乎不会发生。

置信水平，临界值，拒绝域

1. 若 $H_0 : \mu = 7.5$ 成立，则 $|\bar{X} - 7.5|$ 应该很小，“’$|\bar{X} - 7.5|$ 较大” 即为小概率事件，我们引入 α 来表示这个小概率，并称其为**显著性水平** (α 一般选取为 0.1, 0.05, 0.01 等)
2. 同时引入 C 表示差距，并称其为**临界值**，满足 $P(|\bar{X} - 7.5| > C) = \alpha$.
3. 如果 $|\bar{X} - 7.5|$ 大于某一个值 T，我们就有理由拒绝原假设 $H_0 : \mu = 7.5$. 这个区域称为**拒绝域**，本例中，拒绝域为

$$\mathcal{W} = \{(X_1, \ldots, X_n) : |\bar{X} - 7.5| > C\}.$$

检验统计量

给定显著性水平，确定临界值和拒绝域的方法：构造**检验统计量**。上例中，取

$$U = \frac{\bar{X} - 7.5}{\sigma/\sqrt{36}} \overset{H_0 \text{成立时}}{\sim} N(0,1),$$

由 $P(|\bar{X} - 7.5| > C) = \alpha$ 得，取 $\alpha = 0.05$, 拒绝域为

$$\mathcal{W} = \{(X_1, \ldots, X_n) : |\frac{\bar{X} - 7.5}{\sigma/\sqrt{36}}| > u_{\alpha/2} = 1.96\}.$$

代入样本观测值即得到检验统计量的值为 -1.5, 在拒绝域外面，故不拒绝 H_0.
作业：1，2，3