数理统计笔记

作者：肖程哲

时间：November 11, 2022

苟日新，日日新，又日新

目录

第 1 章 导论: 从数据中学习 **1**
1.1 统计模型 1
1.2 统计量 2
第 1 章 练习 2

第 2 章 抽样分布 **3**
2.1 统计量的极限分布 3
2.2 指数族 3
2.3 充分统计量 3
2.4 完全统计量 4

第 3 章 点估计 **5**
3.1 点估计的评价方法 5
3.1.1 无偏性 5
3.1.2 有效性 5
3.1.3 相合性 5
3.1.4 渐进效率 6
3.2 点估计的方法 6
3.2.1 矩估计 6
3.2.2 极大然似估计 7
3.2.3 贝叶斯估计 7
3.2.4 最小二乘估计 7
3.3 最小方差无偏估计 7
3.3.1 均方误差 7
3.3.2 一致最小方差无偏估计 7
3.3.3 充分性原则 7
3.3.4 Cramer-Rao 不等式 7
第 3 章 练习 7

第 4 章 区间估计 **9**
4.1 区间估计的概念 9
4.2 枢轴变量法 9
4.2.1 正态总体参数 9
4.2.2 非正态总体参数 9
4.3 Fisher 的信仰推断法 9
4.4 容忍区间于容忍限 9

第 5 章 假设检验 **10**
5.1 假设检验的概念 10
5.1.1 基本思想 10
5.1.2 Neyman-Pearson 范式 10
5.2 正态总体参数假设检验 11
5.3 然似比检验 11
5.4 一致最优检验 11
5.5 无偏检验 11
5.6 假设检验与区间估计 11
第 5 章 练习 11

第 6 章 线性模型 **12**
6.1 回归分析 12
6.2 最小二乘法 12
6.3 方差分析 12

第 1 章 导论: 从数据中学习

统计学的主要研究内容是如何收集和处理随机数据。数据的随机性一方面来源于自然物理机制中不可避免的随机性；一方面来源于不能被控制（或者不关心）的其他因素。收集数据时，应尽量使其具有代表性（常通过分层抽样实现）。从数据中获取信息，并借此解释和预测数据的过程称为推断 (inference)。统计推断基于概率模型，对两个基本问题：估计 (estimation) 与检验 (testing) 进行回答。为了有效地使用数据进行统计推断，需要给定某些准则去评判不同统计推断方法的优劣。

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 X 的实现(realization)。概率测度 $\mathbb{P}$ 称为总体，观测值或其对应的随机元称为**样本** (sample)，这组观测值的数目 n 称为**样本容量** (sample size)。 ♣

定义 1.2 (总体)

假设样本来源于各分量 X_i 独立同分布的随机向量，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 $x=\{x_1,...,x_n\}$, 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)

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

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

四分位数 (quartile)

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

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. ♣

错题记录

1. (茆 5.5.10) 设 $x_1,...,x_n$ 是来自正态分布 $N(\mu,\sigma^2)$ 的样本，在 μ 已知时给出 σ^2 的一个充分统计量

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 $X_1,\ldots,X_n\sim Binomial(p)$ 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 的充分统计量的充要条件为：

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

♡

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 点估计

考试重点

- 矩估计和极大似然估计
- 点估计的其他评价方法
- UMVUE
- (弱) 相合和强相合
- Fisher 信息量和有效估计
- 贝叶斯估计

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。

3.1 点估计的评价方法

3.1.1 无偏性

定义 3.2 (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:

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

若其为零, 则称此估计为参数 θ 的**无偏估计**

定义 3.3 (渐进无偏估计)

若样本量趋于无穷时，估计式的偏差为零，即:

$$\lim_{n\to\infty} E(\hat{\theta}_n) = \theta$$

则称此估计为参数 θ 的**渐近无偏估计**

3.1.2 有效性

定义 3.4 (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:

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

3.1.3 相合性

定义 3.5 (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**.

注 由于参数是常数，所以依概率收敛于依分布收敛等价。

3.1.4 渐进效率

3.2 点估计的方法

3.2.1 矩估计

定义 3.6 (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 3.7 (矩法估计量)

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

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

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

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

♣

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 $Y \sim \text{Exponential}(1)$. 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

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

的解

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

即为 (λ, a) 的一种矩估计.

命题 3.2 (矩估计的一致性)

3.2.2 极大然似估计

极大然似估计的基本思想：对于参数空间中的每一个参数，计算在此参数下，观测数据的发生概率，选取最大概率对应的参数。

定义 3.8 (然似函数)

设随机变量 $X_1, \cdots, X_n$ 的联合密度函数（或质量函数）为 $f(x_1, \cdots, x_n|\theta)$。对于某一组观测数据 $x_1^*, \cdots, x_n^*$，其**然似函数** (likelihood function) 为：

$$\mathfrak{L}(\theta) = f(x_1^*, \cdots, x_n^*|\theta)$$

对数然似函数 (log likelihood function) 为 $l(\theta) = \log \mathfrak{L}(\theta)$

注 联合密度质量函数代表概率，但联合密度函数不是，代表概率所占比例。然似函数是关于参数 θ 的函数，不是概率，对整个参数空间的积分未必等于一。

定义 3.9 (极大然似估计)

可使然似函数取最大值的参数被称为**极大然似估计**，即

$$\hat{\theta} = \max_{\theta \in \Theta} \mathfrak{L}(\theta)$$

注 由于对数函数单调递增，最大化然似函数等价于最大化对数然似函数。

若样本来源变量独立同分布，即 $f(x_1, \cdots, x_n|\theta) = \prod_{i=1}^n f(x_i|\theta)$，则其然似函数和对数然似函数分布可写为：

$$\mathfrak{l}(\theta) = \prod_{i=1}^n f(x_i^*|\theta), \quad l(\theta) = \sum_{i=1}^n \log f(x_i^*|\theta)$$

命题 3.3 (极大然似估计的不变性)

设 $\hat{\theta}$ 是参数 θ 的极大然似估计，那么对于参数 θ 的任意函数 $\gamma = g(\theta)$，其极大然似估计 $\hat{\gamma} = g(\hat{\theta})$

证明

3.2.3 贝叶斯估计

3.2.4 最小二乘估计

3.3 最小方差无偏估计

3.3.1 均方误差

3.3.2 一致最小方差无偏估计

3.3.3 充分性原则

3.3.4 Cramer-Rao 不等式

错题记录

1. （茆 6.1.2）
2. （茆 6.1.8）
3. （茆 6.1.9）
4. （茆 6.1.11）
5. （茆 6.1.12-14）
6. （茆 6.2.3(2)）
7. （茆 6.2.8）
8. （茆 6.3.9）
9. （茆 6.3.14）
10. （茆 6.4.3）
11. （茆 6.4.4）
12. （茆 6.4.5）
13. （茆 6.4.6）
14. （茆 6.4.9）
15. （茆 6.4.11）
16. （茆 6.4.12）
17. （茆 6.4.13）
18. （茆 6.4.14）
19. （茆 6.4.16-20）
20. （茆 6.5.2）
21. （茆 6.5.12-16）
22. （茆 6.6.11）
23. （茆 6.6.13）
24. （茆 6.6.15-18）
25. （茆 6.6.20-24）

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

5.1 假设检验的概念

5.1.1 基本思想

基本思想：将样本空间 Ω 拆分为两个不相交的集合 Ω_0, Ω_A 即

$$\Omega_0 \cap \Omega_A = \varnothing,\ \Omega_0 \cup \Omega_A = \Omega$$

接下来通过数据选择应该接受两个假设

$$H_0 : \Theta \in \Omega_0,\ H_A : \Theta \in \Omega_A$$

中的哪一个。

问：是否能用点估计代替假设检验？即若估计结果 $\hat{\Theta} \in \Omega_0$，则接受假设 H_0；或反之。

假设某一总体遵循二项分布 $B(n, p)$，做出如下假设：

$$H_0 : p = 0.5,\ H_A : p \neq 0.5$$

若 $n = 10^5$，估计值 $\hat{p} = 0.50001$。按估计结果应认为 H_A 成立，然而一般按情理而言，似乎 H_0 更恰当。这体现出点估计与假设检验的不同：点估计的结果是客观的；而假设检验的结果包含一定的主观性。

若假设可用一个参数的集合表示，该假设检验问题称为**参数假设检验**问题，否则称为**非参数假设检验**问题。上例就是一个参数假设检验问题，而对假设“总体为正态分布”作出检验的问题则是一个非参数假设检验问题。

定义 5.1 (原假设与对立假设)

设有来自某一个参数分布族 $\{F(x, \theta) | \theta \in \Theta\}$ 的样本 $x_1, \cdots, x_n$，其中 Ω 为参数空间。设 $\Theta_0 \in \Theta$, $\Theta_0 \neq \varnothing$，则命题 $H_0 : \Theta \in \Omega_0$ 称为或**原假设**或零假设 (null hypothesis)；并称命题 $H_A : \Theta \in \Omega_A$, $\Theta_A = \Theta - \Theta_0$ 为 H_0 的**对立假设**或备择假设 (alternative hypothesis)。

♣

定义 5.2

如果假设（原假设或对立假设）只含一个点，则称之为**简单假设** (simple hypothesis)，否则称为**复合假设** (composite hypothesis)。

♣

当 H_0 为简单假设时，其形式可写成 $H_0 : \theta\theta_0$，此时的备择假设通常有如下三种可能：

$$H_A' : \theta \neq \theta_0,\ H_A'' : \theta < \theta_0,\ H_A''' : \theta > \theta_0,$$

称 H_0 vs H_A' 为双侧假设或双边假设；H_0 vs H_A'' 以及 H_0 vs H_A''' 为单侧假设或单边假设。

5.1.2 Neyman-Pearson 范式

定义 5.3

对于样本空间 S，若有一种规则将其划分为两个互不相交的区域 $W, \overline{W}$，并且当样本 $\mathbf{x} \in W$ 时，选择假设 H_A，否则接受假设 H_0，则将这种划分称为**检定** (test)，将 W 与 $\overline{W}$ 分别称为**拒绝域** (rejection region, RR) 与**接受域** (acceptance region, AR)。

♣

根据检定结果与实际参数的不同，参见表5.1有 4 种可能的情况：

表 5.1: 检验的 4 种情况

数据情况	总体情况 H_0 为真	H_1 为真
$\mathbf{x} \in W$	一类错误	正确
$\mathbf{x} \in W^c$	正确	二类错误

5.2 正态总体参数假设检验

5.3 然似比检验

5.4 一致最优检验

5.5 无偏检验

5.6 假设检验与区间估计

错题记录

1. （茆 7.1.1）
2.

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