



数理统计笔记

作者：肖程哲

时间：August 5, 2022



苟日新，日日新，又日新

目录

第 1 章 导论: 从数据中学习	1	3.2 极大然似法	6
1.1 利用数据推断	1	3.3 评价方法	6
1.2 提炼数据信息的尝试——描述性统计 .	1	3.4 最小方差无偏估计	6
1.3 刻画变量之间的关系	2	3.5 Bayes 估计	6
1.4 通过试验设计和抽样调查得到数据 .	3		
第 2 章 数据简化原理	4	第 4 章 区间估计	7
2.1 充分性原理	4	4.1 区间估计狱的求法	7
2.2 似然原理	5	4.2 区间估计量的评价方法	7
2.3 同变性原理	5		
第 3 章 点估计	6	第 5 章 线性模型	8
3.1 矩法	6	5.1 回归分析	8
		5.2 最小二乘法	8
		5.3 方差分析	8

第1章 导论: 从数据中学习

1.1 利用数据推断

统计是处理带有随机性的数据 (观测结果) 的艺术. 人们设计试验 (experiments) 并收集数据, 然后希望统计学家能通过数据分析来学到一些知识, 从而有助于解释 (explanation) 和预测 (prediction). 统计推断 (inference) 的两个基本问题是估计 (estimation) 和检验 (testing), 数理统计学家致力于构建数学的理论, 基于概率模型提出研究数据的方法, 对这两个问题进行回答.

一个典型的数据集 (dataset)/样本 (sample) 形如

$$\{x_i : 1 \leq i \leq n\} = \{x_1, \dots, x_n\},$$

其中 i 作为标签 (label) 指代实例 (case/instance/subject)

$$x_i = (x_{ij})_{1 \leq j \leq p} = (x_{i1}, \dots, x_{ip}) \in \mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_p,$$

对应的变量 (variable)

$$x_{\cdot j} = (x_{ij})_{1 \leq i \leq n} = (x_{1j}, \dots, x_{nj}) \in \mathcal{X}_j^n, \quad j = 1, \dots, p$$

刻画了实例的属性 (attribution)/特征 (feature). 每个实例的第 j 个变量的取值空间都为 \mathcal{X}_j , 一般分为两种:

- 分类 (categorical) 数据的取值空间——离散点集, 比如名称 (nominal) 和顺序 (ordinal).
- 数值 (numerical)/定量 (quantitative) 数据的取值空间——实数集 \mathbb{R} 的子集, 比如计数 (counting) 常用非负整数集 $\mathbb{N} = \{0, 1, 2, 3, \dots\}$.

数据集可以用矩阵 $X = (x_{ij})_{1 \leq i \leq n}^{1 \leq j \leq p}$ 表示出来, 第 i 行第 j 列的数据 x_{ij} 为实例 i 的第 j 个变量. 我们称实例的数目 n 为样本容量 (sample size), 变量的数目 p 为样本的维数 (dimensionality).

数据处理方法的严格性由概率论来保证. 统计学预设了数据的随机性, 将 x_i 视为某个概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 上的随机元 X_i 的实现 (realization)/观测结果 (observation). 统计分析得到的结论一般是关于分布 $\mathbb{P}\{(X_1, \dots, X_n) \in \bullet\}$ 的推断——在统计学中可以认为分布包含我们想知道的一切信息, 然而 (至少部分) 是未知的, 我们试图用收集到的样本 (已知信息) 来揣度其性质. 这个未知的分布称为总体 (population), 所有备选 (candidate) 总体构成所谓统计模型 (statistical model).¹

1.2 提炼数据信息的尝试——描述性统计

绘制 (plot) 数据的图示 (pattern) 能够提供直观印象. 注意有时需要先对数据进行变换, 比如计算出占总数的比例 (proportion). 常用的统计图示有:

- 条形图 (bar graph): 数值数据 vs 分类数据, 分类数据等宽, 数值数据的长度表示大小.
- 饼状图 (pie chart): 表现出每一类数据占总数的比例.
- 直方图 (histogram): 数据的频次 vs 数值.

描述统计图示可以考虑形状 (shape)、中心 (center) 和延展 (spread), 比如:

- 离群值 (outlier)? 对称 (symmetric)? 单峰 (unimodal)?
- 众数 (mode)?
- 右偏 (skewed to the right)?

¹更数学一点的总结可参看<https://zhuanlan.zhihu.com/p/101355754>

用确定的(不依赖未知总体的)函数作用于样本,即得统计量(statistic),这给出了一种数据约简(reduction).对于实数值样本 $x = \{x_1, \dots, x_n\}$,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)

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

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

四分位数 (quartile)

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

四分位距 (inter quartile range)

$$IQR = Q_3 - Q_1.$$

极差 (range)

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

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

1.3 刻画变量之间的关系

考虑同一批实例的两个变量 $x = (x_i)_{1 \leq i \leq n}$ 和 $y = (y_i)_{1 \leq i \leq n}$,我们或许认为 y 是值得关心的结果(outcome),并猜想 x 对 y 造成了影响——此时称 y 为**响应变量**(response variable),称 x 为**解释变量**(explanatory variable).

常用的图示是**散点图**(scatterplot),对每个实例*i*绘制数据点 (x_i, y_i) . 我们往往期待线性关系,为此,可以考虑对数据进行变换,比如 $\log : (0, \infty) \rightarrow \mathbb{R}$.

对于实值变量,常用的统计量有:

- **样本协方差** (sample covariance)

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

- **样本相关系数** (sample correlation coefficient)

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right),$$

其中 s_x 和 s_y 是相应的样本标准差.

我们常常在散点图中画出**回归直线** (regression line)²

$$y = \hat{\alpha} + \hat{\beta}x,$$

其中

$$(\hat{\alpha}, \hat{\beta}) = \arg \min_{(\alpha, \beta)} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

²稍加推广将得到§??线性模型

是最小二乘法 (method of least squares) 的解, 适合

$$\hat{\beta} = s_{xy}/s_x^2, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

回归直线是一种简单的线性拟合 (fitting), 并且给出了一种还算有道理 (?) 的预测 (prediction) 方法——沿着直线外推 (extrapolation). 在模型的训练集 (training set) 上, 回归直线得到拟合值 (fitted value)

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i, \quad i = 1, \dots, n$$

与残差 (residual)

$$\hat{\varepsilon}_i = y_i - \hat{y}_i, \quad i = 1, \dots, n.$$

绘制 $(\hat{y}_i, \hat{\varepsilon}_i)_{1 \leq i \leq n}$ 得到的残差图 (residual plot) 可以直观地反映拟合效果, 这里 \hat{y}_i 是 x_i 的线性变换 (画图时二者几乎没有区别), 容易推广到多个解释变量的情形.

回归模型能够捕捉变量之间的 (线性) 相关性 (association), 但是未必蕴涵因果关系 (causation). 对响应变量有影响但是难以观测的变量称为潜变量 (latent/lurking variable), 无法甄别的解释变量之间存在混杂 (confoundedness), 这些都让模型显得不那么可靠. 因果推断 (causal inference) 是统计学中方兴未艾的一个领域, 有人认为 2019 年炸药奖应该颁发给开创因果分析研究范式³的 Rubin、Angrist 和 Imbens, 而不是将实验引入贫困研究的 Banerjee、Duflo 和 Kremer. [狗头]

(顺便分享 [xkcd 漫画](#))

1.4 通过试验设计和抽样调查得到数据

数据可能来自于轶闻 (anecdote) 或者可从某些机构获得 (available), 不过在统计学中收集数据的常规方法是试验 (experiment) 和抽样调查 (survey sampling). 这部分内容不宜在入门课程中占据过多学时, 稍作了解即可, 有兴趣的同学可以参看方开泰 *et al.* 《试验设计与建模》以及冯士雍 *et al.* 《抽样调查理论与方法》. 尽管如此, 让数据具有好的概率结构 (比如独立性) 是统计理论中极其重要的部分, 窃以为要诀是让选取的样本具有代表性和利用有限的样本有效解决问题.

设计试验对试验点 (experimental unit) 施加特定的处理 (treatment), 一般是不同因子 (factor) 的不同水平 (level) 的组合, 然后观测输出 (outcome) 来获得数据. 试验设计能保证数据的优良性, 多快好省地提供统计分析的素材, 在业界应用广泛. 好的试验应该满足下述准则: 随机 (randomized)、对照 (comparative) 和重复 (repeated). 识别因果应该需要试验是双盲 (double-blind) 的. 同一区组 (block) 的试验有近似的试验环境, 通过区组设计可以减少系统误差的干扰.

抽样调查意为从总体中抽取样本, 根据方法不同可分为概率抽样 (probability sampling) 和非概率抽样 (non-probability sampling). 非概率抽样不遵循科学的原则, 无法保证样本具有代表性, 比如根据主观经验进行抽样, 或者出于道德考虑仅对志愿者进行调查. 概率抽样是严格地按照给定的概率抽取样本, 包括简单随机抽样 (simple random sampling) 和分层随机抽样 (stratified random sampling).

特别注意, 试验设计和抽样调查在实践中都会遇到各种各样的问题, 需要项目组织者审慎对待.

³ 推荐<https://cosx.org/2012/03/causality2-rcm>和统计之都的其他文章

第2章 数据简化原理

内容提要

- | | |
|---|---|
| <input type="checkbox"/> 充分统计量 (sufficient statistic) | <input type="checkbox"/> 完备统计量 (complete statistic) |
| <input type="checkbox"/> 极小充分统计量 (minimal sufficient statistic) | <input type="checkbox"/> 似然原理 |
| <input type="checkbox"/> 辅助统计量 (ancillary statistic) | <input type="checkbox"/> 同变性原理 |

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 $\mathcal{T} = \{t | \exists x \in \mathcal{X}, \text{s.t. } t = T(x)\}$ 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

2.1 充分性原理

例题 2.1 设 $X_1, \dots, X_n \sim \text{Binomial}(p)$ 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：

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

注 $T \sim \text{Binomial}(n, p)$ 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 似估计可表示为 S 的函数。



证明 似然函数为 $g(S, \theta)h(X)$ 。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

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

则其为关于 θ 的充分统计量



2.2 似然原理

2.3 同变性原理

第3章 点估计

3.1 矩法

3.2 极大然似法

3.3 评价方法

3.4 最小方差无偏估计

3.5 Bayes 估计

第4章 区间估计

4.1 区间估计的求法

4.2 区间估计量的评价方法

第 5 章 线性模型

5.1 回归分析

5.2 最小二乘法

5.3 方差分析