数理统计入门教辅

T.-Y. Li (kellty@pku.edu.cn)

2020 年 9 月 19 日

代序

在文档成立初期, 我也曾兴致勃勃想要参与这份笔记的撰写工作, 然而发现自己的知识专业性与写作水平与作者都有相当的差距, 且受时间精力所限, 就不敢下笔了. 兴许是想要宽慰我受伤的内心, LTY 邀请我为这本小册子作序, 我也就斗胆给各位读者介绍一下.

本笔记主要结构如下: 第1章介绍统计学的概况, 阐明 “数据–模型–知识” 的主线; 第2章复习课程中需要用到的概率论知识; 第3章介绍参数估计, 包括点估计与区间估计的概念及主要方法、Bayes 统计及与此相关的统计决策理论; 第4章阐述假设检验的相关理论, 包括问题的提法、主要检验方法与优良性标准; 第5章介绍前述几章的理论在线性模型中的具体应用; 第6章与附录A则单列了一些独立的话题, 对前文的内容作进一步的延申与补充.

作者不愧是概率统计基础极为扎实的统计学博士新生, 在完整回顾课程内容的同时, 自然地引入了测度论的相关语言, 并在不同的话题中介绍了进一步的理解与相关资料. 可以说, 这份笔记是 LTY 以众多经典的概率统计教科书作为药材, 熬出的一碗适合课程学习者饮用的 “大补汤”. 对于想要复习或自学数理统计的读者来说, 这份笔记已基本满足甚至有时超出本科数理统计课程的要求. 为便于初学者使用, 建议先阅读 §1、§2两章与附录 §A.5一节, 对自己的先修知识进行评估与补充, 从而顺利适应其余章节的叙述; 自信的读者可以直接从 §3这一章开始统计学的旅程.

在表意准确完整的前提下, LTY 力求内容与排版的紧凑, 这也是他写东西的一大特色. 因此, 笔记中仅保留了最重要的例子与证明思路, 某些部分也略去特例而直接进入较为一般情况的阐述. 以我对 LTY 其他笔记的阅读情况来看, 这样的风格一定会给初学者带来一定的阅读困难, 请确实做好 “三两行琢磨半小时” 的觉悟. 不过, 若能够仔细完成阅读并对

相关内容有足够的掌握, 相信你一定会和我一样佩服作者流畅简洁的行文与深刻的理解, 并将收获更多新的体悟.

至于对学习数理统计的建议, 我就从这个课程名称出发聊一聊. 首先, "数理" 说明这门课程更看重的是数学的缜密推导而非直觉的想象, 因此需要对各个核心命题的证明给予足够的重视, 并注意总结常用的处理技术. 其次, "统计" 作为一个完整的学科, 有自己的发展历史、内在思想与具体应用, 因此不妨带着历史的视角和应用的视角来进行这门课程的学习. 一方面, 读者可以阅读一些介绍统计学历史的材料, 去思考 "为什么数理统计会出现" 的问题, 具体可参看 §6.4. 另一方面, 读者也可以尝试在学习工作生活日常中使用统计学的思想方法, 带着自己希望解决的问题学知识, 会拥有独属于自己的感触.

LTY 担任数理统计助教之时正值新冠疫情, 身在各地的同学们通过两周一次的线上习题课, 得以窥探更高观点下的数理统计, 堪称幸运. 如今作者将其整理成册, 想必能造福许多正在或准备学习数理统计的读者. 衷心祝愿各位学有所成, 也希望 LTY 和我两个博士小萌新未来能够顺利毕业! xD

Y.-F. Chen (yufan_chen@pku.edu.cn)

2020 年 9 月

目录

1 导论: 从数据中学习

1.1 数理统计学旨在利用数据对未知物进行推断

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

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

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

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

对应的**变量** (variable)

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

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 $\mathbb{N} = \{0, 1, 2, 3, \cdots\}$.

数据集可以用矩阵 $X = (x_{ij})_{1\le j\le p}^{1\le i\le n}$ 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 $(\Omega, \mathscr{F}, \mathbb{P})$ 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).[i)]

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)?

[i)] 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)
$$M = \operatorname{med}(x) = \begin{cases} x_{(k)}, & n = 2k-1 \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}), & n = 2k \end{cases}$$
其中顺序统计量 $x_{(1)} \le \cdots \le x_{(n)}$ 由 $x_1, \dots, x_n$ 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** (interquartile range)
$$IQR = Q_3 - Q_1.$$
- **极差** (range)
$$x_{(n)} - x_{(1)} = \max(x) - \min(x) = \max_{1 \le i, i' \le n}\{x_i - x_{i'}\}.$$

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 $x = (x_i)_{1 \le i \le n}$ 和 $y = (y_i)_{1 \le i \le n}$, 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 $\log : (0, \infty) \to \mathbb{R}$.

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)[ii)]

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

其中

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

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

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

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)

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

与**残差** (residual)

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

绘制 $(\hat{y}_i, \hat{\varepsilon}_i)_{1\le i\le n}$ 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[iii)]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. ⚒

[ii)] 稍加推广将得到 §5 线性模型

[iii)] 推荐https://cosx.org/2012/03/causality2-rcm和统计之都的其他文章

2 重温概率论

2.1 概率空间

为了刻画随机世界, 给出可能性的量化表述, 我们需要一个合适的数学系统——概率空间.

我们关心的随机现象被抽象为集合, 逻辑运算 (且, 或, 非, etc.) 对应成集合论运算 (交, 并, 补, etc.), 由此得到**事件域** $\mathscr{F}$, 这是一个σ 代数, 满足 (思考: 下述公理能够推出什么性质?)

- $\emptyset \in \mathscr{F}$, 表示 "无事发生";
- $A \in \mathscr{F} \implies A^{\complement} \in \mathscr{F}$, 即 $\mathscr{F}$ 对补集运算 (逻辑上的非) 封闭;
- $A_1, \dots, A_n, \dots \in \mathscr{F} \implies \bigcap_{n=1}^{\infty} A_n \in \mathscr{F}$, 即 $\mathscr{F}$ 对可数交运算 (逻辑上的可数多个且) 封闭.

为了补集能够良好定义, 我们指明**全集**为 Ω, 从而 $\mathscr{F}$ 是 Ω 的子集族 (collection of subsets). 可数性是为了在数学上能够恰当地处理无穷的概念, 术语中的 σ 指的就是可数并.

在事件域上定义的**概率**为 $\mathbb{P} : \mathscr{F} \to [0, \infty)$, 适合 (思考: 下述公理能够推出什么性质?)

- $\mathbb{P}(\Omega) = 1$; (如果没有这条就是一般的有限测度)
- 若 $A_1, \dots, A_n, \dots \in \mathscr{F}$ 两两不交, 即 $A_i \cap A_j = \emptyset$, $\forall i \neq j$, 则 $\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$.

我们称 $(\Omega, \mathscr{F}, \mathbb{P})$ 为一个**概率空间** (probability space).

2.2 条件概率与独立性

考虑正概率的事件 $A \in \mathscr{F}$, 称

$$\mathbb{P}(\bullet | A) = \mathbb{P}(\bullet \cap A)/\mathbb{P}(A), \quad \bullet \in \mathscr{F}$$

为**条件于 A 的概率** (probability conditional on A), 这仍然是一个概率测度.

如果 $A, B \in \mathscr{F}$ 满足

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B),$$

则称 A 与 B **独立** (independent), 记为 $A \perp\!\!\!\perp B$. 易见独立性是对称的, 即 $A \perp\!\!\!\perp B \iff B \perp\!\!\!\perp A$.

当 $\mathbb{P}(A) > 0$ 时, 我们有

$$\mathbb{P}(B|A) = \mathbb{P}(B) \iff B \perp\!\!\!\perp A,$$

由此可得到 "B 独立于 A" 的直观理解.

若 $\mathscr{G} \subset \mathscr{F}$ 与 $\mathscr{H} \subset \mathscr{F}$ 满足

$$A \perp\!\!\!\perp B, \quad \forall A \in \mathscr{G},\ B \in \mathscr{H},$$

则称 $\mathscr{G}$ 与 $\mathscr{H}$ 独立, 记为 $\mathscr{G} \perp\!\!\!\perp \mathscr{H}$.

测度论告诉我们一个重要结果: 如果 $\mathscr{G}$ 对交集运算封闭, 那么成立 $\mathscr{G} \perp\!\!\!\perp \mathscr{H} \implies \sigma(\mathscr{G}) \perp\!\!\!\perp \mathscr{H}$, 其中 $\sigma(\mathscr{G})$ 是 $\mathscr{G}$ 扩张而成的一个合适的最小的 σ 代数, 定义为所有包含 $\mathscr{G}$ 的 σ 代数的交集.

扩张, 或者称为延拓, 是数学中很重要的一个概念, 大抵是将某映射的定义域适当扩大, 不改变在初始定义域上的映射取值 (注意值域可能是比较抽象的集合, 配备了某些操作之后被称为空间), 同时在扩大后的定义域上仍然保持某些优良的性质. 与此相对的概念是限制, 即关心局部上可能更加漂亮的性质, 把初始的定义域适当缩小.

2.3 随机变量

为了表示因随机性而变动的量, 称可测映射 (measurable mapping)

$$X : (\Omega, \mathscr{F}, \mathbb{P}) \to (\mathcal{X}, \mathscr{F}_{\mathcal{X}}), \quad \omega \in \Omega \mapsto X(\omega) \in \mathcal{X}$$

为**随机元** (random element), 下面解释可测性. 我们或许关心 X 取值于 $\mathcal{X}$ 的某个子集 B 的可能性, 亦即希望得到 $\mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\})$. 在集合论中我们称 $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$ 为 B 在映射 X 下的原像 (pre-image), 概率论不关心具体的样本点 $\omega \in \Omega$, 将其记为 $\{X \in B\} = X^{-1}(B)$. 由于 $\mathbb{P}$ 定义在 $\mathscr{F}$ 上, 我们需要 $X^{-1}(B) \in \mathscr{F}$. 设所有值得关心的 $B \subset \mathcal{X}$ 组成 $\mathscr{F}_{\mathcal{X}}$, 且 $\forall B \in \mathscr{F}_{\mathcal{X}}$ 都满足 $\{X \in B\} \in \mathscr{F}$, 则称 X 为 $\mathscr{F}/\mathscr{F}_{\mathcal{X}}$ 可测的. 当 $\mathscr{F}_{\mathcal{X}}$ 不引起混淆时, 简记为关于 $\mathscr{F}$ 可测, 写成 $X \in \mathscr{F}$.

注: 由于原像保持交、并、补等集合运算, 且 $\mathscr{F}$ 是 σ 代数, 我们可以将 $\mathscr{F}_{\mathcal{X}}$ 扩张为合适的最小的 σ 代数, 即 $\sigma(\mathscr{F}_{\mathcal{X}})$, 因此可测映射的定义不妨只考虑 $\mathscr{F}_{\mathcal{X}}$ 是 σ 代数 的情况. 特别地, 实数集 $\mathbb{R}$ 的子集族 $\{(-\infty, x] : x \in \mathbb{R}\}$ 生成的 σ 代数 $\mathscr{B}_{\mathbb{R}}$ 称为 $\mathbb{R}$ 上的 *Borel* 代数.

随机元 X 的**分布** (distribution/law) 是 X 诱导的概率测度 $\mathbb{P}\{X \in \bullet\}$, $\bullet \in \mathscr{F}_{\mathcal{X}}$.

⋆ 离散型: 当 $\mathcal{X}$ 是 (至多可数的) 离散点集, 设 $\mathscr{F}_{\mathcal{X}}$ 由 $\mathcal{X}$ 的所有子集组成, 此时 X 的分布由**概率质量函数** (probability mass function, p.m.f.)

$$p_X(x) = \mathbb{P}\{X = x\} = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\}), \quad x \in \mathcal{X}$$

唯一刻画. 思考: 这是概率的哪个性质保证的?

⋆ 实值向量: 当 $\mathcal{X} = \mathbb{R}^n$, 考虑 $\mathscr{F}_{\mathcal{X}}$ 为 $\{\prod_{i=1}^n (-\infty, x_i] : x_1, \dots, x_n \in \mathbb{R}\}$ 生成的 Borel 代数 (最小的 σ 代数), 此时 $X = (X_1, \dots, X_n)^\top$ 的分布由 (累积) **分布函数** (cumulative distribution function, c.d.f.)

$$F_X(x_1, \dots, x_n) = \mathbb{P}\{X_1 \le x_1, \dots, X_n \le x_n\}, \quad x_1, \dots, x_n \in \mathbb{R}.$$

唯一刻画. 若 $F_X : \mathbb{R}^n \to [0,1]$ 可微 (或者更一般地, 绝对连续), 称其导函数

$$f_X := \nabla F_X = \left(\frac{\partial F_X}{\partial x_1}, \dots, \frac{\partial F_X}{\partial x_n} \right)$$

为 X 的**概率密度函数** (probability density function, p.d.f.), 此时称 X 为**连续型随机向量**.

记 $\sigma(X) := X^{-1}(\mathscr{F}_{\mathcal{X}}) := \{X^{-1}(B) : B \in \mathscr{F}_{\mathcal{X}}\}$. 根据前面的定义, X 可测当且仅当 $\sigma(X) \subset \mathscr{F}$. 若随机元 $X_1 : (\Omega, \mathscr{F}) \to (\mathcal{X}_1, \mathscr{F}_{\mathcal{X}_1})$ 与随机元 $X_2 : (\Omega, \mathscr{F}) \to (\mathcal{X}_2, \mathscr{F}_{\mathcal{X}_2})$ 满足 $\sigma(X_1) \perp\!\!\!\perp \sigma(X_2)$, 则称 X_1 与 X_2 **独立**, 记为 $X_1 \perp\!\!\!\perp X_2$.

注: 对任意可测映射 g_1, g_2 成立 $X_1 \perp\!\!\!\perp X_2 \implies g_1(X_1) \perp\!\!\!\perp g_2(X_2)$. 事实上, $\sigma(g(X)) \subset \sigma(X)$.

当 X_1 和 X_2 都是实值随机向量时, $X_1 \perp\!\!\!\perp X_2$ 当且仅当 (思考: $\sigma(X^{-1}(\bullet)) = X^{-1}(\sigma(\bullet))$, $\bullet \subset \mathscr{F}_{\mathcal{X}}$?)

$$F_{X_1,X_2}(x_1, x_2) = F_{X_1}(x_1) F_{X_2}(x_2), \ \forall x_1, x_2.$$

对于连续型随机向量, 刻画独立性只需要

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2), \ \forall x_1, x_2.$$

设随机向量 X 和 Y 有联合 (joint) 概率密度函数 $f_{X,Y}$, 可以证明 X 有概率密度函数

$$f_X(\cdot) = \int f_{X,Y}(\cdot, y) \, \mathrm{d}y,$$

称为 (X, Y) 中 X 的边际 (margin) 概率密度函数. Y **条件于** $X = x$ **的概率密度函数** $f_{Y|X}(\cdot|x)$ 满足

$$f_{X,Y}(x, \cdot) = f_{Y|X}(\cdot|x) f_X(x).$$

2.4 Copula

设连续型实值随机变量 X 有分布函数 F, 易见 F 在 $\overline{\mathbb{R}} = [-\infty, +\infty]$ 上从 0 递增到 1. 定义相应的**分位数函数** (quantile function) 为

$$F^{-1}(p) := \inf\{x \in \mathbb{R} : F(x) \geq p\}, \quad p \in [0,1].$$

当 F 严格递增时, 这与一般的反函数定义相同. 一个简单的性质是 $F(X) \sim \text{Uniform}([0,1])$, 因为

$$\mathbb{P}\{F(X) \leq p\} = \mathbb{P}\{X \leq F^{-1}(p)\} = p, \quad \forall p \in [0,1].$$

反过来, 我们有 $U \sim \text{Uniform}([0,1]) \implies F^{-1}(U) \overset{d}{=} X$, 其中 $\overset{d}{=}$ 表示分布相同 (equal in distribution).

考虑多个连续型实值随机变量 $X_1, \dots, X_k$, 记 X_i 的分布函数为 F_i. 我们称 $(F_1(X_1), \dots, F_k(X_k))$ 的分布函数 $C : [0,1]^k \to [0,1]$ 为相应的 **Copula**, 适合

$$C(F_1(x_1), \dots, F_k(x_k)) = \mathbb{P}\{X_1 \leq x_1, \dots, X_k \leq x_k\}, \quad \forall x_1, \dots, x_k.$$

这个结果称为 **Sklar 定理**, 在金融统计中有颇多应用. 稍作诠释的话, Copula 提取了变量间的相关性, 通过粘合边际能够恰好地表示总体. 人们可以构造各种各样的 Copula, 对真实世界进行建模.

2.5 期望和方差

对于实值随机变量 $X : (\Omega, \mathscr{F}, \mathbb{P}) \to (\mathbb{R}, \mathscr{B}_{\mathbb{R}})$ 和 (可测) 函数 $g : \mathbb{R} \to \mathbb{R}$, 称

$$\mathbb{E}[g(X)] = \int_\Omega g(X(\omega)) \, \mathrm{d}\mathbb{P}(\omega) = \int_{\mathbb{R}} g(x) \, \mathrm{d}F_X(x)$$

为 $g(X)$ 的**均值** (mean) 或**期望** (expectation). 期望算子 $\mathbb{E}$ 是一个线性泛函, 仅适用于可积的随机变量, 严格的定义可以参看基于测度论的概率论, 比如新手友好的 *Williams*《概率和鞅》[iv].

一个重要结果是, 若 $g(X) \geq 0$, 则 $\mathbb{E}[g(X)] = 0 \implies g(X) \overset{\text{a.s.}}{=} 0$, 即 $\mathbb{P}\{g(X) = 0\} = 1$. 其证明可通过 **Markov 不等式**

$$\mathbb{P}\{g(X) \geq \varepsilon\} \leq \mathbb{E}[g(X)]/\varepsilon, \quad \forall \varepsilon > 0$$

完成, 其中需要用到概率的连续性, 即 $\lim\limits_{n\to\infty} A_n = A \implies \lim\limits_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$.

刻画 X 的变动程度的量是其**方差** (variance)

$$\text{Var}(X) = \mathbb{E}[|X - \mathbb{E}X|^2] = \mathbb{E}[X^2] - (\mathbb{E}X)^2.$$

考虑**均方误差** (mean squared error)

$$\text{MSE}(X;\theta) = \mathbb{E}[|X - \theta|^2], \quad \theta \in \mathbb{R},$$

通过**方差偏差分解** (variance-bias decomposition)

$$\text{MSE}(X;\theta) = \text{Var}(X) + |\mathbb{E}X - \theta|^2$$

可以说明 $\theta \mapsto \text{MSE}(X;\theta)$ 在 $\mathbb{E}X$ 处取到最小值 $\text{Var}(X)$. 投影 (projection) 和正交分解的思想在各种内积空间中应用广泛, 这里是 $\mathbb{E} = \text{proj}_{\mathbb{R}}$, 概率论中关于子事件域 $\mathscr{G}$ (随机元 X, resp.) 的条件期望几何直观是 $\text{proj}_{\mathscr{G}}$ ($\text{proj}_{\sigma(X)}$, resp.), 线性模型 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ 中拟合值为 $\hat{\mathbf{y}} = \text{proj}_{\text{Col}(\mathbf{X})} \mathbf{y}$.

预处理随机变量有两个常用变换:

- **中心化** (centralization) $X \mapsto X - \mathbb{E}X$;
- **标准化** (standardization) $X \mapsto \dfrac{X - \mathbb{E}X}{\sqrt{\text{Var}(X)}}$.

[iv] 国内有影印版和中译本, 作为教材适用于大学三年级的课程如 https://courses.maths.ox.ac.uk/node/42189

2.6 连续型随机变量

设连续型实值随机变量 X 有概率密度函数 $f:\mathbb{R}\to\mathbb{R}_+=[0,\infty)$, 其图像

$$\Gamma(f)=\{(x,f(x)):x\in\mathbb{R}\}$$

称为**密度曲线** (density curve), 满足 $\int_{\mathbb{R}}f(x)\,\mathrm{d}x=1$.

- 均值为 $\mathbb{E}X=\int_{\mathbb{R}}xf(x)\,\mathrm{d}x$.
- 中位数为 M s.t. $\int_{-\infty}^{M}f(x)\,\mathrm{d}x=\int_{M}^{+\infty}f(x)\,\mathrm{d}x=1/2$.
- 众数为 $\arg\max_x f(x)$.
- 分布关于 $a\in\mathbb{R}$ 对称 (symmetric) 当且仅当 $f(a+x)=f(a-x)$, $\forall x\in\mathbb{R}$.
- 分布函数为 $x\in\mathbb{R}\mapsto\int_{-\infty}^{x}f(u)\,\mathrm{d}u\in[0,1]$.

2.7 相关性

对于方差有限 (i.e., 二阶矩存在) 的实值随机变量 X 和 Y 可以定义内积 (☕ 试试验证?)

$$\langle X,Y\rangle_{\mathbb{P}}=\mathbb{E}[XY]=\int_{\Omega}X(\omega)Y(\omega)\,\mathrm{d}\mathbb{P}(\omega).$$

注意到中心化可以消除均值, 我们可以用中心化的随机变量的内积来刻画 (相对于均值的) 变动的线性相关性. 定义 X 和 Y 的**协方差** (covariance) 为

$$\mathrm{Cov}(X,Y)=\langle X-\mathbb{E}X,Y-\mathbb{E}Y\rangle_{\mathbb{P}}=\mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)]=\mathbb{E}[XY]-\mathbb{E}X\mathbb{E}Y.$$

于是有 **Cauchy 不等式** (思考: 取等条件是什么?)

$$\mathrm{Cov}(X,Y)^2\le\mathrm{Var}(X)\,\mathrm{Var}(Y).$$

定义 X 和 Y 的**相关系数** (correlation coefficient) 为

$$\mathrm{Corr}(X,Y)=\left\langle\frac{X-\mathbb{E}X}{\sqrt{\mathrm{Var}(X)}},\frac{Y-\mathbb{E}Y}{\sqrt{\mathrm{Var}(Y)}}\right\rangle_{\mathbb{P}}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}\in[-1,1].$$

对于随机向量 $X=(X_1,\dots,X_m)^\top$ 和 $Y=(Y_1,\dots,Y_n)^\top$ 可以自然地定义

$$\mathbb{E}X=(\mathbb{E}X_1,\dots,\mathbb{E}X_m)^\top$$

和

$$\mathrm{Cov}(X,Y)=\mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)^\top]=(\mathrm{Cov}(X_i,Y_j))_{1\le j\le n}^{1\le i\le m}\in\mathbb{R}^{m\times n}.$$

特别地, $\mathrm{Var}(X)=\mathrm{Cov}(X,X)$ 称为随机向量 X 的**方差-协方差矩阵**.

对于容许的常数值矩阵 A,B 和常数值向量 u,v 有

$$\mathrm{Cov}(AX+u,BY+v)=A\,\mathrm{Cov}(X,Y)B^\top.$$

特别地,

$$\mathrm{Var}(AX+u)=A\,\mathrm{Var}(X)A^\top.$$

设随机元 X_1 与 X_2 独立, 则

$$\mathrm{Cov}(\mathbb{1}_{\{X_1\in B_1\}},\mathbb{1}_{\{X_2\in B_2\}})=\mathbb{E}[\mathbb{1}_{\{X_1\in B_1,X_2\in B_2\}}]-\mathbb{E}[\mathbb{1}_{\{X_1\in B_1\}}]\mathbb{E}[\mathbb{1}_{\{X_2\in B_2\}}]=0,\quad \forall B_1,B_2.$$

利用线性组合的逼近, 如

$$g(X)=\lim_{n\to\infty}\sum_{|k|\le n\cdot 2^n} g(k/2^n)\mathbb{1}_{\{X\in[k/2^n,(k+1)/2^n)\}},$$

可得

$$\mathrm{Cov}(g_1(X_1),g_2(X_2))=0,\quad \forall g_1,g_2.$$

上述结论简记为: 独立蕴涵不相关. 反之不成立, 即使边际为正态分布.

2.8 条件期望

给定概率空间 $(\Omega,\mathscr{F},\mathbb{P})$ 上的实数值随机变量 X, 我们会关心部分信息下 X 的表现. 所谓 "部分信息" 指的是 $\mathscr{G}\subset\mathscr{F}$, 不失一般性可以要求 $\mathscr{G}$ 仍然是 σ 代数 (测度常常可以从 $\mathscr{G}$ 扩张到 $\sigma(\mathscr{G})$ 上); 这样的 $\mathscr{G}$ 也可以是某个随机元 ξ 带来的, 亦即 $\mathscr{G}=\sigma(\xi)$. 作为例子, 对于 $A\in\mathscr{F}$, 指示函数 $\mathbb{1}_A$ 反映的信息是 $\sigma(\mathbb{1}_A)=\sigma(\{A\})=\{A,A^{\complement},\emptyset,\Omega\}$, 当 $A=\Omega$ 时常值随机变量 $1=\mathbb{1}_\Omega$ 反映信息 $\sigma(1)=\{\emptyset,\Omega\}$.

利用正交投影的几何直观, 我们给出**条件期望** (conditional expectation) 的一些等价刻画: 我们称关于 $\mathscr{G}$ 可测的随机变量 $\hat{X}$ 是随机变量 X 在给定 $\mathscr{G}$ 时的条件期望, 记为 $\hat{X}=\mathbb{E}[X|\mathscr{G}]$, 当且仅当成立下述条件之一: (试试证明等价性?)

1. 对任意关于 $\mathscr{G}$ 可测的随机变量 Z 满足 $\mathbb{E}\left[\left|X-\hat{X}\right|^2\right]\le\mathbb{E}\left[|X-Z|^2\right]$.
2. 对任意关于 $\mathscr{G}$ 可测的随机变量 Z 满足 $\mathbb{E}\big[(X-\hat{X})Z\big]=0$, 亦即 $\mathbb{E}[XZ]=\mathbb{E}\big[\hat{X}Z\big]$.
3. 对任意事件 $A\in\mathscr{G}$ 满足 $\mathbb{E}[X\mathbb{1}_A]=\mathbb{E}\big[\hat{X}\mathbb{1}_A\big]$.

注意前两条一般要求 X 是二阶矩有限的, 此时与二阶矩有限的 Z 才能良好地运算; 最后一条只需要 X 是一阶矩有限的, 所以更为一般, 是普适的条件期望的定义. 实分析告诉我们这样的 $\hat{X}=\mathbb{E}[X|\mathscr{G}]$ 存在且 a.s. 唯一; 所以更严格一点可以说上面的 $\hat{X}$ 是 $\mathbb{E}[X|\mathscr{G}]$ 的一个版本 (version), 而任意两个 $\mathbb{E}[X|\mathscr{G}]$ 的版本都是 a.s. 相等的.

作为简写, 约定 $\mathbb{E}[X|\xi]:=\mathbb{E}[X|\sigma(\xi)]$. 易见 $\mathbb{E}[X|\mathbb{1}_\Omega]=\mathbb{E}X$. 对于连续型随机向量 (X,Y), 有 $\mathbb{E}[g(Y)|X]=\int g(y)f_{Y|X}(y|X)\,\mathrm{d}y$, 其中 $f_{Y|X}(\cdot|x)$ 是 Y 条件于 $X=x$ 的概率密度函数.

称 $\mathbb{P}(A|\mathscr{G}):=\mathbb{E}[\mathbb{1}_A|\mathscr{G}]$ 为事件 A 在给定 $\mathscr{G}$ 时的**条件概率**; 对于任意 (可测) 分划 $\Omega=A_1\sqcup\cdots\sqcup A_k$, 不难证明 $\mathbb{P}\left(A|\sigma(\{A_1,\dots,A_k\})\right)=\sum_{i=1}^k\mathbb{P}(A|A_i)\mathbb{1}_{A_i}$.

基于反映信息和投影的直观, 我们不加证明地指出下述性质: (其实证明很简单, 不妨试试? ☺)

下面的符号中, $\mathscr{G}$ 和 $\mathscr{H}$ 是子 σ 代数, X 和 Y 是随机变量, a 和 b 是实数.

- 线性: $\mathbb{E}[aX+bY|\mathscr{G}]=a\mathbb{E}[X|\mathscr{G}]+b\mathbb{E}[Y|\mathscr{G}]$.
- 保序: $X\ge Y$ (a.s.) $\implies\mathbb{E}[X|\mathscr{G}]\ge\mathbb{E}[Y|\mathscr{G}]$ (a.s.), 特别地可以取 $Y\equiv 0$.
- 取出已知量: $\sigma(X)\subset\mathscr{G}\implies\mathbb{E}[XY|\mathscr{G}]=X\mathbb{E}[Y|\mathscr{G}]$, 特别地可以取 $Y\equiv 1$.
- *tower property*: $\mathscr{H}\subset\mathscr{G}\implies\mathbb{E}[\mathbb{E}(X|\mathscr{G})|\mathscr{H}]=\mathbb{E}[X|\mathscr{H}]$, 特别地可以取 $\mathscr{H}=\{\emptyset,\Omega\}$.
  (思考: $\mathscr{H}\subset\mathscr{G}\implies\mathbb{E}[\mathbb{E}(X|\mathscr{H})|\mathscr{G}]=?$)

定义**条件方差**为

$$\mathrm{Var}(X|\mathscr{G}) = \mathbb{E}\left[(X - \mathbb{E}[X|\mathscr{G}])^2|\mathscr{G}\right] = \mathbb{E}[X^2|\mathscr{G}] - (\mathbb{E}[X|\mathscr{G}])^2.$$

利用 $\mathrm{proj}_{\mathscr{G}} \bullet := \mathbb{E}[\bullet|\mathscr{G}]$ 可以给出正交分解和条件期望版勾股定理[v]:

$$\mathbb{E}[|X - \mathrm{proj}_{\{\emptyset,\Omega\}} X|^2] = \mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X|\mathscr{G})] + \mathrm{Var}(\mathbb{E}[X|\mathscr{G}]) = \mathbb{E}[|X - \mathrm{proj}_{\mathscr{G}} X|^2] + \mathbb{E}[|\mathrm{proj}_{\mathscr{G}} X - \mathrm{proj}_{\{\emptyset,\Omega\}} X|^2]$$

2.9 矩母函数与特征函数

一些变换为我们研究随机变量 X 的分布 $\mathbb{P}\{X \in \bullet\}$ 提供了有效的工具.

当 X 取值于 $\mathbb{N} = \{0, 1, 2, \cdots\}$ 时, 我们可以考虑**生成函数** (generating function)

$$G_X(s) = \mathbb{E}[s^X] = \sum_{n=0}^{\infty} \mathbb{P}\{X = n\} s^n, \quad s \in [0, 1].$$

对于一般的实数值随机变量 (随机向量类似), 我们可以定义**矩母函数** (moment generating function)

$$M_X(t) = \mathbb{E}[\mathrm{e}^{tX}], \quad t \in \mathbb{R},$$

和**特征函数** (characteristic function)/分布的 Fourier 变换

$$\phi_X(t) = \mathbb{E}[\mathrm{e}^{\sqrt{-1}tX}], \quad t \in \mathbb{R}.$$

一般来说 M_X 未必总是有限, 但是 $\phi_X(\bullet) = M_X(\sqrt{-1}\bullet)$ 足以完全决定 X 的分布.

利用幂级数展式 $\mathrm{e}^x = \sum_{n=0}^{\infty} x^n/n!$ 可以得到

$$M_X(t) = \sum_{k=0}^{n} \frac{\mathbb{E}[X^k] t^k}{k!} + o(t^n),$$

于是 $M_X^{(n)}(0) = \mathbb{E}[X^n]$. 注意: 这些操作在数学上稍欠严谨, 不过应用时趁手即可.

2.10 正态分布及其导出分布

中心极限定理 (见 §2.11) 保证了正态分布是统计推断中最常见的概率分布.

我们直接考虑服从 p 维正态分布 $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的随机向量 $\boldsymbol{X}$, 其矩母函数为

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}[\mathrm{e}^{\boldsymbol{t}'\boldsymbol{X}}] = \mathrm{e}^{\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}}, \quad \boldsymbol{t} \in \mathbb{R}^p.$$

由此可得, $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 具有均值 $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}$ 和协方差矩阵 $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. 利用矩母函数的唯一性, 对任意 $\boldsymbol{A} \in \mathbb{R}^{q \times p}$ 和 $\boldsymbol{b} \in \mathbb{R}^q$, 有 $\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b} \sim \mathcal{N}_q(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}')$. 当 $\boldsymbol{\Sigma}$ 非退化时, $\boldsymbol{X}$ 有概率密度函数

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right\}, \quad \boldsymbol{x} \in \mathbb{R}^p.$$

对于 $\begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right)$, 我们有[vi]: (回忆一下 $\mathrm{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \boldsymbol{\Sigma}_{12}$)

$$\boldsymbol{\Sigma}_{12} = \boldsymbol{0} \implies M_{\boldsymbol{X}_1, \boldsymbol{X}_2}(\boldsymbol{t}_1, \boldsymbol{t}_2) = M_{\boldsymbol{X}_1}(\boldsymbol{t}_1) M_{\boldsymbol{X}_2}(\boldsymbol{t}_2) \implies \boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{X}_2.$$

[v] 更一般地, 可参看https://www.zhihu.com/question/38726155/answer/885319771

[vi] 第一个 "$\implies$" 直接计算可得, 第二个 "$\implies$" 可以参看https://math.stackexchange.com/a/287321

一维的正态分布 $\mathcal{N}(\mu,\sigma^2)$ 有概率密度函数 $f(x)=\frac{1}{\sigma}\varphi\big(\frac{x-\mu}{\sigma}\big)$, $\forall x\in\mathbb{R}$, 其中

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}\exp(-x^2/2),\ \forall x\in\mathbb{R}$$

的图像是著名的钟形曲线 (bell curve).

正态分布的导出分布也很常见[vii], 主要有三种:

- 若 $X_1,\dots,X_n\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,1)$, 则称 $X_1^2+\cdots+X_n^2$ 服从自由度为 n 的**卡方分布**, 即 χ_n^2.
- 若 $X\sim\mathcal{N}(0,1)$ 与 $Q\sim\chi_n^2$ 独立, 则称 $\frac{X}{\sqrt{Q/n}}$ 服从自由度为 n 的 ***t* 分布**, 即 t_n.
- 若 $Q_1\sim\chi_m^2$ 与 $Q_2\sim\chi_n^2$ 独立, 则称 $\frac{Q_1/m}{Q_2/n}$ 服从自由度为 m 和 n 的 ***F* 分布**, 即 $F_{m,n}$.

定理. 设 $X_1,\dots,X_n\overset{\text{i.i.d.}}{\sim}\mathcal{N}(\mu,\sigma^2)$, 则样本均值 $\bar X$ 与样本方差 S^2 独立, 且 $\sqrt{n}(\bar X-\mu)/\sqrt{S^2}\sim t_{n-1}$.

注. 这个不依赖 σ 的随机变量可用来对 μ 进行推断, 参看 §3.5. 注意统计学中重要的参数往往未知.

证明. 记 $\boldsymbol{X}=(X_1,\dots,X_n)'$, 则 $\boldsymbol{X}\sim\mathcal{N}_n(\mu\mathbf{1}_n,\sigma^2\boldsymbol{I}_n)$. 我们有 $\bar X=\frac{1}{n}\mathbf{1}_n'\boldsymbol{X}$ 和

$$\boldsymbol{Z}:=(X_i-\bar X)_{1\le i\le n}=\boldsymbol{X}-\bar X\mathbf{1}_n=(\boldsymbol{I}_n-\tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n')\boldsymbol{X}.$$

可见 $\begin{pmatrix}\bar X\\ \boldsymbol{Z}\end{pmatrix}$ 由 $\boldsymbol{X}$ 线性变换得到, 所以服从联合正态分布, 其均值为

$$\begin{pmatrix}\frac{1}{n}\mathbf{1}_n'\\ \boldsymbol{I}_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\end{pmatrix}\mu\mathbf{1}_n=\begin{pmatrix}\mu\\ \mathbf{0}_n\end{pmatrix},$$

方差-协方差矩阵为

$$\begin{pmatrix}\frac{1}{n}\mathbf{1}_n'\\ \boldsymbol{I}_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\end{pmatrix}(\sigma^2\boldsymbol{I}_n)\begin{pmatrix}\frac{1}{n}\mathbf{1}_n'\\ \boldsymbol{I}_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\end{pmatrix}'=\sigma^2\begin{pmatrix}\frac{1}{n} & \mathbf{0}_n'\\ \mathbf{0}_n & \boldsymbol{I}_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\end{pmatrix}.$$

于是 $\bar X\perp\!\!\!\perp\boldsymbol{Z}$, 从而 $S^2=\frac{1}{n-1}\boldsymbol{Z}'\boldsymbol{Z}$ 作为 $\boldsymbol{Z}$ 的函数也独立于 $\bar X$. 注意到 $\boldsymbol{I}_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$ 是对称幂等矩阵, 取正交矩阵 $\boldsymbol{U}\in\mathbb{R}^{n\times n}$ 适合

$$\boldsymbol{I}_n-\tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n'=\boldsymbol{U}\,\mathrm{diag}(\boldsymbol{I}_{n-1},0)\,\boldsymbol{U}',$$

其中 $\boldsymbol{U}$ 的最后一列是 $\boldsymbol{I}_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$ 的从属于 0 的特征向量 $\frac{1}{\sqrt{n}}\mathbf{1}_n$. 易见 $\boldsymbol{U}'\boldsymbol{X}\sim\mathcal{N}_n(\binom{\mathbf{0}_{n-1}}{\sqrt{n}\mu},\sigma^2\boldsymbol{I}_n)$, 故

$$(n-1)S^2=(\boldsymbol{U}'\boldsymbol{X})'\,\mathrm{diag}(\boldsymbol{I}_{n-1},0)\,\boldsymbol{U}'\boldsymbol{X}\sim\sigma^2\chi_{n-1}^2.$$

结合 $\bar X\sim\mathcal{N}(\mu,\frac{1}{n}\sigma^2)$ 立得 $\sqrt{n}(\bar X-\mu)/\sqrt{S^2}\sim t_{n-1}$. □

2.11 渐近性质

本节试图概览随机世界中的极限. 设 $X_\infty,X_1,X_2,\dots,X_n,\dots$ 是 $(\Omega,\mathscr{F},\mathbb{P})$ 上的一列随机变量.

我们称 X_n **几乎必然收敛** (converges almost surely) 于 X_∞, 记为 $X_n\xrightarrow{\text{a.s.}}X_\infty$, 若

$$\mathbb{P}\{\lim_{n\to\infty}X_n\ne X_\infty\}=0\iff\mathbb{P}\{|X_n-X_\infty|>\varepsilon\text{ i.o.}\}=0,\ \forall\varepsilon>0$$

$$\text{(利用概率 }\mathbb{P}\text{ 的连续性)}\iff\lim_{N\to\infty}\mathbb{P}(\cup_{n=N}^\infty\{|X_n-X_\infty|>\varepsilon\})=0,\ \forall\varepsilon>0$$

注意 $\{A_n\text{ i.o.}\}=\{\sum_{n=1}^\infty\mathbb{1}_{A_n}=\infty\}=\bigcap_{N=1}^\infty\bigcup_{n=N}^\infty A_n$ 表示 A_n 发生无穷多次 (infinitely often).

[vii] 参看https://cosx.org/2013/01/story-of-normal-distribution-2/

称 X_n **依概率收敛** (converges in probability) 于 X_∞, 记为 $X_n \xrightarrow{\mathbb{P}} X_\infty$, 若

$$\lim_{n\to\infty} \mathbb{P}\{|X_n - X_\infty| > \varepsilon\} = 0, \ \forall \varepsilon > 0.$$

易见 $X_n \xrightarrow{\text{a.s.}} X_\infty \implies X_n \xrightarrow{\mathbb{P}} X_\infty$, 但是反之不成立, 例如概率空间 $((0,1], \mathscr{B}_{(0,1]}, \text{Uniform}((0,1]))$ 上的打字机序列

$$\xi_n = \mathbb{1}_{(n/2^k-1,(n+1)/2^k-1]}, \quad 2^k \le n < 2^{k+1}$$

适合 $\mathbb{P}\{|\xi_n| > \varepsilon\} \le 1/2^k \to 0$, 而 $\overline{\lim}\, \xi_n \equiv 1$ 且 $\underline{\lim}\, \xi_n \equiv 0$, 所以 $\lim \xi_n(\omega)$ 对 $\forall \omega \in (0,1]$ 都不存在.

统计学中估计量的相合性/一致性 (consistency) 就是用依概率收敛和几乎必然收敛来刻画的. 为此, 我们常常应用**大数定律** (Law of Large Numbers, 简记为 LLN): 设 $X_1, \dots, X_n, \dots$ 独立同分布, 且 $\mathbb{E}X_1$ 存在, 则 Kolmogorov 证明了*强大数定律* (strong LLN):

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[(n\to\infty)]{\text{a.s.}} \mathbb{E}X_1,$$

这个结果强于 Khinchine 证明的*弱大数定律* (weak LLN):

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[(n\to\infty)]{\mathbb{P}} \mathbb{E}X_1.$$

有的时候我们更关心分布的性质[viii], 希望用 (简洁易操作的) 渐近分布 (asymptotic distribution) 代替 (复杂的) 恰当分布 (exact distribution), 比如用 $\mathcal{N}(0,1)$ 代替 t_{n-1}.

称 X_n **依分布收敛** (converges in distribution) 于 X_∞, 记为 $X_n \xrightarrow{d} X_\infty$, 若分布 $P_n := \mathbb{P}\{X_n \in \bullet\}$ **弱收敛** (converges weakly) 于 $P_\infty := \mathbb{P}\{X_\infty \in \bullet\}$, 记为 $P_n \rightsquigarrow P_\infty$, 即对任意*有界的连续函数* $g : \mathbb{R} \to \mathbb{R}$ 成立

$$\mathbb{E}[g(X_n)] = \int g \,\mathrm{d}P_n \xrightarrow{(n\to\infty)} \int g \,\mathrm{d}P_\infty = \mathbb{E}[g(X_\infty)].$$

事实上, 分布/概率测度的弱收敛在泛函分析中称为*弱 $*$ 收敛* (weak-star convergence) 更为合适.

记 $X \sim P$ 的分布函数为 $F_X(x) = \mathbb{P}\{X \le x\} = P((-\infty, x]), \quad x \in \mathbb{R}$;

特征函数为 $\phi_X(t) = \mathbb{E}\exp(\sqrt{-1}tX) = \int \exp(\sqrt{-1}t\bullet)\,\mathrm{d}P, \quad t \in \mathbb{R}$.

定理. *以下任意一条陈述都是 $P_n \rightsquigarrow P_\infty$ 的等价刻画:*

1. *工具箱/一揽子 (portmanteau) 中的一条: 在 F_{X_∞} 的任意连续点 x 处都成立 $F_{X_n}(x) \to F_{X_\infty}(x)$.*
2. *Lévy 连续性定理: 在 $\forall t \in \mathbb{R}$ 处成立 $\phi_{X_n}(t) \to \phi_{X_\infty}(t)$.*
3. *Skorokhod 表示: 存在 (新) 概率空间 $(\Omega', \mathscr{F}', \mathbb{P}')$ 上一列随机变量 $X'_\bullet \sim P_\bullet$ 适合 $X'_n \xrightarrow{\mathbb{P}'\text{-a.s.}} X'_\infty$.*

可以证明 $X_n \xrightarrow{\mathbb{P}} X_\infty \implies X_n \xrightarrow{d} X_\infty$, 反之, 若 $X_\infty \equiv c \in \mathbb{R}$, 则 $X_n \xrightarrow{d} c \implies X_n \xrightarrow{\mathbb{P}} c$.

前述收敛性 (a.s., in P, in d) 都可以推广到多维向量的情形. 下面罗列一些常用性质.

定理 (连续映射定理 (continuous mapping theorem)**).** *设 (可测) 函数 g 的间断点构成集合 D_g.*

若 $\mathbb{P}\{X_\infty \in D_g\} = 0$, 即 g 在 X_∞ 处 a.s. 连续, 则

$$\begin{cases} X_n \xrightarrow{\text{a.s.}} X_\infty \implies g(X_n) \xrightarrow{\text{a.s.}} g(X_\infty) \\ X_n \xrightarrow{\mathbb{P}} X_\infty \implies g(X_n) \xrightarrow{\mathbb{P}} g(X_\infty) \\ X_n \xrightarrow{d} X_\infty \implies g(X_n) \xrightarrow{d} g(X_\infty) \end{cases}$$

[viii] 光华管理学院的陈松蹊教授会在春季学期开大样本统计理论课程, 用A. W. van der Vaart 所著 Asymptotic Statistics (有影印版)

引理 (联合收敛性). $\begin{cases} X_n \xrightarrow{\text{a.s.}} X_\infty,\ Y_n \xrightarrow{\text{a.s.}} Y_\infty \implies (X_n, Y_n) \xrightarrow{\text{a.s.}} (X_\infty, Y_\infty) \\ X_n \xrightarrow{\mathbb{P}} X_\infty,\ Y_n \xrightarrow{\mathbb{P}} Y_\infty \implies (X_n, Y_n) \xrightarrow{\mathbb{P}} (X_\infty, Y_\infty) \\ X_n \xrightarrow{d} X_\infty,\ Y_n \xrightarrow{d} c \in \mathbb{R} \implies (X_n, Y_n) \xrightarrow{d} (X_\infty, c) \end{cases}$

定理 (基本运算). 利用加减乘除作为二元函数的连续性, 有

- $X_n \xrightarrow{\text{a.s.}} X_\infty,\ Y_n \xrightarrow{\text{a.s.}} Y_\infty \implies X_n \pm Y_n \xrightarrow{\text{a.s.}} X_\infty \pm Y_\infty,\ X_nY_n \xrightarrow{\text{a.s.}} X_\infty Y_\infty,\ X_n/Y_n \xrightarrow{\text{a.s.}} X_\infty/Y_\infty.$
- $X_n \xrightarrow{\mathbb{P}} X_\infty,\ Y_n \xrightarrow{\mathbb{P}} Y_\infty \implies X_n \pm Y_n \xrightarrow{\mathbb{P}} X_\infty \pm Y_\infty,\ X_nY_n \xrightarrow{\mathbb{P}} X_\infty Y_\infty,\ X_n/Y_n \xrightarrow{\mathbb{P}} X_\infty/Y_\infty.$
- *(**Slutsky**)* $X_n \xrightarrow{d} X_\infty,\ Y_n \xrightarrow{d} c \in \mathbb{R} \implies X_n \pm Y_n \xrightarrow{d} X_\infty \pm c,\ X_nY_n \xrightarrow{d} cX_\infty,\ X_n/Y_n \xrightarrow{d} X_\infty/c.$

定理 (**delta 方法**). 作为连续映射定理的一个重要应用, 利用函数的一阶 *Taylor* 展开 *(局部线性近似)* 可得: 设函数 g 在点 $\boldsymbol{c}$ 可微, 任给一列常数 $0 < a_n \to \infty$, 我们有

$$a_n(\boldsymbol{X}_n - \boldsymbol{c}) \xrightarrow{d} \boldsymbol{Y} \implies a_n[g(\boldsymbol{X}_n) - g(\boldsymbol{c})] \xrightarrow{d} [\nabla g(\boldsymbol{c})]^\top \boldsymbol{Y}.$$

注. 当 $\nabla g(\boldsymbol{c}) = \boldsymbol{0}$ 时, 可以考虑更高阶的 *Taylor* 展开 *(局部多项式近似)*. 类似的思想体现于矩的近似式: $\mathbb{E}[g(\boldsymbol{X})] \approx g(\mathbb{E}\boldsymbol{X}) + \frac{1}{2}\mathbb{E}[(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})^\top \nabla^2 g(\mathbb{E}\boldsymbol{X})(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})]$, $\operatorname{Var}(g(\boldsymbol{X})) \approx \operatorname{Var}(\nabla g(\mathbb{E}\boldsymbol{X})^\top (\boldsymbol{X} - \mathbb{E}\boldsymbol{X}))$.

定理 (Lindeberg-Feller **中心极限定理** (central limit theorem, 简记为 CLT)). 对任意 $n = 1, 2, \cdots$, 设 $X_{n,m}$ $(1 \le m \le k_n)$ 是独立的随机向量, 若

- $\sum_{m=1}^{k_n} \mathbb{E}\left[\|X_{n,m}\|^2 \mathbb{1}_{\{\|X_{n,m}\| > \varepsilon\}}\right] \to 0$ $(n \to \infty)$, $\forall \varepsilon > 0$, 且
- $\sum_{m=1}^{k_n} \operatorname{Var}(X_{n,m}) \to \Sigma$ $(n \to \infty)$,

则

$$\sum_{m=1}^{k_n} (X_{n,m} - \mathbb{E}X_{n,m}) \xrightarrow{d} Z \sim \mathcal{N}(0, \Sigma) \quad (n \to \infty).$$

注. (Lévy CLT) 若 ξ_n 独立同分布, 均值为 μ, 方差-协方差矩阵为 Σ, 考虑 $\frac{\xi_m}{\sqrt{n}}$ $(1 \le m \le n)$ 可得

$$\sqrt{n}\left(\frac{\xi_1 + \cdots + \xi_n}{n} - \mu\right) \xrightarrow{d} Z \sim \mathcal{N}(0, \Sigma) \quad (n \to \infty).$$

2.12 简单随机抽样

考虑一批实例, 不妨表示为 $\mathcal{I} = \{1, 2, \cdots, N\}$, 我们关心某特征 $x : \mathcal{I} \to \mathbb{R}^p$ 的统计信息, 为方便起见记 $x_i := x(i)$, $\forall i \in \mathcal{I}$. 最理想的状况是各实例地位等同, 设 $\mathcal{I}$ 的 $N!$ 个排列构成集合 S_N, 我们的观测顺序是随机元 $\kappa \sim \text{Uniform}(S_N)$ 的实现, 观测结果依次记为 $X_j := x_{\kappa(j)}$, $1 \le j \le N$. 受限于成本等原因, 我们往往只能观测 $n \le N$ 次, 称 $(X_1, X_2, \ldots, X_n)$ 为样本容量是 n 的**不放回简单随机抽样** (simple random sample without replacement), 常见于工厂的次品抽检等应用场景.

易见 $\kappa(j) \sim \text{Uniform}(\mathcal{I})$, $\forall j$, 于是

$$\mathbb{E}[X_j] = \mu := \frac{1}{N}\sum_{i \in \mathcal{I}} x_i, \quad \& \quad \operatorname{Var}(X_j) = \Sigma := \frac{1}{N}\sum_{i \in \mathcal{I}} (x_i - \mu)(x_i - \mu)^\top.$$

令 $T_n := \sum_{j=1}^n X_j$, 则 $\mathbb{E}[T_n] = n\mu$, 利用对称性可得

$$\operatorname{Var}(T_n) = n\Sigma + (n^2 - n)\operatorname{Cov}(X_1, X_2).$$

注意到 $T_N = N\mu$ 是常值, 我们有

$$\operatorname{Var}(T_N) = 0 \implies \boxed{\operatorname{Cov}(X_1, X_2) = -\Sigma/(N-1)} \implies \operatorname{Var}(T_n) = \frac{N-n}{N-1} \cdot n\Sigma.$$

读者不难得到 $(X_1, X_2, \ldots, X_n)$ 的样本均值的期望和协方差矩阵, 以及样本协方差矩阵的期望. ☕

3 参数估计

让我们回到统计学. 在 §1.1中说到要从数据 X 推断总体, 现将所有备选总体构成的统计模型记为

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

其中某一 $\theta \in \Theta$ 适合 $P_\theta = \mathbb{P}\{X \in \bullet\}$, 然而是未知信息. 分布族 $\mathcal{P}$ 的指标集 Θ 称为**参数空间** (parameter space), 一般会要求模型 $\mathcal{P}$ **可识别** (identifiable), 即对 $\forall \theta_1, \theta_2 \in \Theta$ 满足

$$\theta_1 \neq \theta_2 \iff P_{\theta_1} \neq P_{\theta_2}.$$

为了推断的便利, 我们考虑 $\Theta \subset \mathbb{R}^d$, 其中 $d < \infty$ 是固定的正整数, 此时模型 $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$ 称为**参数族** (parametric family). 我们往往只对参数 θ 蕴含的部分信息感兴趣, 记为 $\gamma = g(\theta)$, 剩下的碍事无用的部分称为冗余参数 (nuisance parameter).

例. 对于正态分布族 $\{\mathcal{N}(\mu, \sigma^2) : (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\}$, 我们可能仅关心均值 μ, 而认为方差 σ^2 冗余. ♠

通过样本 $X \sim P_\theta$ 对 $\gamma = g(\theta)$ 进行推断, 是统计学的核心任务. 我们的主要手段是构造**统计量** (statistic), 即选取不依赖总体的 (可测) 映射 T 将 X 变换为 $T(X)$, 在不引起混淆时也写成 $T = T(X)$. 注: 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.

3.1 点估计的概念

如果选取统计量 $\hat{\gamma} = T(X)$ 作为 $\gamma = g(\theta)$ 的猜测, 则称 $\hat{\gamma}$ 是 γ 的**估计量** (estimator), 其具体实现称为**估计值** (estimate). 一些简单的优良性判则如下所示:

- 称 $\hat{\gamma}$ 是**无偏的** (unbiased), 若
$$\mathbb{E}_\theta[\hat{\gamma}] = \gamma, \quad \forall \theta \in \Theta,$$
其中 $\mathbb{E}_\theta[T(X)] = \int_{\mathcal{X}^n} T \,\mathrm{d}P_\theta$. 估计量 $\hat{\gamma}$ 的**偏差** (bias) 定义为 $\mathbb{E}_\theta[\hat{\gamma}] - \gamma$.
- 称 $\hat{\gamma}$ 是**相合的/一致的** (consistent), 若当样本容量 $n \to \infty$ 时, 有
$$\hat{\gamma} \xrightarrow{\mathbb{P}} \gamma, \quad \forall \theta \in \Theta.$$
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.

例. 设 $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$, 其中 $\lambda > 0$ 未知. (思考: 样本 $X = (X_i)_{1\leq i\leq n}$ 相应的统计模型是什么?) 样本均值 $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ 是 λ 的无偏且相合的估计量. (☺ 读者自证不难) ♠

3.2 区域估计的概念

设 $\gamma = g(\theta) \in \Gamma$, 即 $g : \Theta \to \Gamma$. 我们有时候考虑猜测 γ 的一个范围, 而非其确切值. 考虑统计量 $C = C(X)$ 取值为 Γ 的子集, 使得 $\gamma \in C$ 很可能成立, 此时称 C 为 γ 的**置信集** (confidence set). 若

$$\mathbb{P}_\theta\{C \ni \gamma\} \geq 1 - \alpha, \quad \forall \theta \in \Theta,$$

则称 $C = C(X)$ 具有**置信水平** (confidence level) $1 - \alpha$, 其中 $\alpha \in (0, 1)$ 是固定常数, 我们当然希望 α 越小越好. 称

$$\inf_{\theta\in\Theta} \mathbb{P}_\theta\{\gamma \in C(X)\}$$

为 $C = C(X)$ 的**精确置信水平** (confidence size). 显然有置信水平不超过精确置信水平.

例. 设 $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} \text{Uniform}([0,\theta])$, 其中 $\theta\in(0,2)$ 未知. (思考: 样本 $X=(X_i)_{1\le i\le n}$ 相应的统计模型是什么?) 我们用样本极值 $X_{(1)}=\min\limits_{1\le i\le n}X_i$ 和 $X_{(n)}=\max\limits_{1\le i\le n}X_i$ 构建 $C=[X_{(n)},\min\{X_{(1)}+1,2\}]$, 则

$$\mathbb{P}_\theta\{\theta\in C\}=\mathbb{P}_\theta\{\theta\le X_{(1)}+1\}=\mathbb{P}_\theta\{X_{(1)}\ge\theta-1\}=\min\{1,\theta^{-n}\}.$$

进而可得 C 对于 θ 有精确置信水平 2^{-n}. ♠

✍ 下面先介绍一些常用的 (点) 估计方法.

3.3 矩方法

考虑样本 $X=(X_1,\dots,X_n)^\top\sim P_\theta$, 其中 X_i 是 i.i.d. 随机变量. 记总体 k 阶矩为

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

样本 k 阶矩为

$$\hat\mu_k=\frac1n\sum_{i=1}^n X_i^k,\quad k\in\mathbb{N}.$$

我们感兴趣的参数 $\gamma=g(\theta)$ 的**矩(方法)估计量** (method of moments estimator) 定义为

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

其中 $\hat\theta^{\text{MoM}}$ 对某些 (人为) 选定的 k 满足

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

例. 设 $X_1,\dots,X_n$ 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. ♠

值得一提的是, 如果我们用一般的 $\mathcal{X}\times\Theta$ 上的 (向量) 函数 h 代替那些 $(x,\theta)\mapsto x^k-\mu_k(\theta)$, 使得

$$\mathbb{E}_\theta[h(X_i,\theta)]=0,\quad \forall\theta\in\Theta,$$

记 $\hat m(\theta)=\frac1n\sum_{i=1}^n h(X_i,\theta)$, 那么

$$\hat\theta^{\text{GMM}}\in\arg\min_{\theta\in\Theta}\left[\hat m(\theta)^\top W_n(X)\hat m(\theta)\right]$$

就是 θ 的**广义矩**(generalized method of moments)**估计**, 其中 (权重) 统计量 $W_n(X)$ 是正定对称矩阵.

Last but not least: 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.

3.4 极大似然

矩方法仅用到了分布的部分信息, 如果我们对总体有更多了解, 那么就能有更加精细的估计方法.

首先引入似然的概念. 对于 $X = (X_1, \dots, X_n)^\top \sim P_\theta$, 设

$$\mathrm{d}P_\theta = f_\theta \mathrm{d}\nu, \quad \forall \theta \in \Theta,$$

其中 ν 是 $\mathcal{X}^n$ 上控制所有 P_θ 的 (σ 有限) 测度 (比如计数测度或者 $\mathbb{R}^n$ 上的 Lebesgue 测度), 从而 f_θ 是 X 的概率密度函数. 称

$$L_n(\theta; x) := f_\theta(x), \quad \theta \in \Theta$$

为 (容量为 n 的) 样本 x 对应的**似然函数** (likelihood function), 无歧义时可以省略 x. 有时候我们会考虑 (数学上) 更容易处理的**对数似然函数** (log-likelihood function)

$$\ell_n(\theta) := \log L_n(\theta), \quad \theta \in \Theta.$$

注: 独立的数据有乘性 (multiplicative) 似然函数, 通过对数变换成为加性 (additive), 使计算上更便利.

假定一些正则性条件 (regularity conditions), 如光滑性、可积性以及极限 (e.g., 微分和积分) 的可交换性. 我们先来探究一下似然函数的性质. 对数似然函数的一阶导

$$s_n(\theta) := \dot{\ell}_n(\theta) = \frac{\partial \ell_n}{\partial \theta}(\theta) = \frac{1}{L_n(\theta)} \frac{\partial L_n(\theta)}{\partial \theta}, \quad \theta \in \Theta$$

称为**得分函数** (score function). 将前述函数视为依赖 X 的随机变量, 我们有**第一 Bartlett 恒等式** (first Bartlett's identity)

$$\mathbb{E}_\theta[s_n(\theta)] = \int_{\mathcal{X}^n} \frac{1}{\cancel{f_\theta(x)}} \frac{\partial f_\theta(x)}{\partial \theta} \cancel{f_\theta(x)} \, \mathrm{d}\nu(x) = \frac{\partial}{\partial \theta} \underbrace{\int_{\mathcal{X}^n} f_\theta(x) \, \mathrm{d}\nu(x)}_{=1} = 0.$$

继续微分, 可得**第二 Bartlett 恒等式** (second Bartlett's identity)

$$\begin{aligned} 0 = \frac{\partial}{\partial \theta^\top} \int_{\mathcal{X}^n} s_n(\theta) \underbrace{\mathrm{e}^{\ell_n(\theta)}}_{=f_\theta(x)} \mathrm{d}\nu(x) &= \int_{\mathcal{X}^n} \left\{ \left[\frac{\partial^2}{\partial \theta \partial \theta^\top} \ell_n(\theta) \right] f_\theta(x) + s_n(\theta) s_n(\theta)^\top f_\theta(x) \right\} \mathrm{d}\nu(x) \\ &= \mathbb{E}_\theta \left[\ddot{\ell}_n(\theta) \right] + \mathbb{E}_\theta \left[s_n(\theta) s_n(\theta)^\top \right]. \end{aligned}$$

我们称

$$\mathcal{I}_n(\theta) = \mathrm{Var}_\theta(s_n(\theta)) = \mathbb{E}_\theta \left[s_n(\theta) s_n(\theta)^\top \right] = -\mathbb{E}_\theta \left[\ddot{\ell}_n(\theta) \right]$$

为 **Fisher 信息矩阵** (Fisher information matrix). 特别地, 对于 i.i.d. 样本, 我们有 $\underline{\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta)}$.

言归正传. 设

$$\hat{\theta}^{\mathrm{ML}} \in \arg\max_{\vartheta \in \Theta} L_n(\vartheta) = \arg\max_{\vartheta \in \Theta} \ell_n(\vartheta)$$

存在, 则

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

称为 $\gamma = g(\theta)$ 的**极大似然估计量** (maximum likelihood estimator, 简记为 MLE).

- 极大似然估计的**同变性质** (equivariance property): The MLE is equivariant.
  任给 (不依赖总体的) 函数 ψ, 则 $\psi(\hat{\gamma}^{\mathrm{ML}})$ 是 $\psi(\gamma)$ 的一个极大似然估计量.
- 极大似然估计的估计方程 (estimating equation): 在若干正则性条件下, 我们有 $\underline{s_n(\hat{\theta}^{\mathrm{ML}}) = 0}$.

例. 设 $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$, 其中 $p\in(0,1)$ 未知. 似然函数为

$$L_n(p)=\prod_{i=1}^n\left[p^{X_i}(1-p)^{1-X_i}\right]=p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i},$$

记样本均值 $\overline{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$, 则对数似然函数为

$$\ell_n(p)=\log L_n(p)=n\left[\overline{X}_n\log(p)+(1-\overline{X}_n)\log(1-p)\right],$$

进而得分函数为

$$s_n(p)=\dot{\ell}_n(p)=n\left(\frac{\overline{X}_n}{p}-\frac{1-\overline{X}_n}{1-p}\right).$$

注意到

$$\ddot{\ell}_n(p)=-n\left(\frac{\overline{X}_n}{p^2}+\frac{1-\overline{X}_n}{(1-p)^2}\right)<0,\quad \forall p,$$

可得

$$\max\ \ell_n(p)\quad \text{w.r.t.}\quad p\in(0,1)$$

有唯一解

$$\hat{p}^{\text{ML}}=\overline{X}_n\quad \text{s.t. } s_n(\hat{p}^{\text{ML}})=0.$$

易见 $\text{Var}_p(\hat{p}^{\text{ML}})=p(1-p)/n$, 我们称

$$\text{se}(\hat{p}^{\text{ML}})=\sqrt{\widehat{\text{Var}}(\hat{p}^{\text{ML}})}=\sqrt{\hat{p}^{\text{ML}}(1-\hat{p}^{\text{ML}})/n}$$

为 $\hat{p}^{\text{ML}}$ 的标准误 (standard error), 它是 $\sqrt{p(1-p)/n}$ 的极大似然估计. ♠

例. 设 $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} \text{Uniform}([0,\theta])$, 其中 $\theta>0$ 未知. 似然函数为

$$L_n(\theta)=\theta^{-n}\mathbb{1}_{[0\le X_{(1)}\le X_{(n)}\le\theta]},$$

其中 $X_{(1)}=\min X_i$ 和 $X_{(n)}=\max X_i$ 是样本极值. 可见极大似然估计 $\hat{\theta}^{\text{ML}}=X_{(n)}$ 不同于矩估计 $\hat{\theta}^{\text{MoM}}=2\overline{X}_n$, 其中 $\overline{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$ 是样本均值. (☕ 请读者补足细节. 思考: 无偏性? 相合性?) ♠

下面讨论极大似然估计的*渐近性质*. 设真实参数为 θ_0, 并且假定 *i.i.d. 样本*. 记

$$M_n(\vartheta):=\frac{1}{n}\ell_n(\vartheta;X_1,\dots,X_n)=\frac{1}{n}\sum_{i=1}^n\ell_1(\vartheta;X_i),\quad \vartheta\in\Theta.$$

此时根据 LLN 可知 $M_n(\vartheta)\to M_0(\vartheta)$ in probability as $n\to\infty$, 其中

$$M_0(\vartheta):=\mathbb{E}_{\theta_0}[\ell_1(\vartheta;X_i)],\quad \vartheta\in\Theta.$$

为了给出 $\hat{\theta}_n^{\text{ML}}\in\arg\max M_n(\vartheta)$ 逼近 θ_0 的直观, 我们断言 $\{\theta_0\}=\arg\max M_0(\vartheta)$.

引理 (Shannon 信息[ix]不等式)**.** 对 $\forall\vartheta\in\Theta$ 有 $\mathbb{E}_{\theta_0}[\log f_\vartheta(X)]\le\mathbb{E}_{\theta_0}[\log f_{\theta_0}(X)]$.

注. 可识别性将保证等号成立当且仅当 $\vartheta=\theta_0$. 此外, 当然需要假定可积性.

[ix] 如果随机元 ξ 的分布为 $\mathrm{d}P=f\mathrm{d}\nu$, 那么 $H(f):=-\mathbb{E}_{\xi\sim f}[\log(f(\xi))]=-\int\log(f(x))f(x)\,\mathrm{d}\nu(x)$ 称为 $\xi\sim f\mathrm{d}\nu$ 的**熵** (entropy). 熵的符号 H 据传是 η (希腊字母 eta) 的大写, 见https://math.stackexchange.com/q/84719.

证明. 只需考虑 $\vartheta \neq \theta_0$ 的情形, 此时可识别性表明 $f_\vartheta \neq f_{\theta_0}$, ν-a.e. 记

$$D_{\mathrm{KL}}(P_{\theta_0}||P_\vartheta) := \int \log\left(\frac{f_{\theta_0}(x)}{f_\vartheta(x)}\right) f_{\theta_0}(x)\,\mathrm{d}\nu(x) = \mathbb{E}_{\theta_0}[\log f_{\theta_0}(X)] - \mathbb{E}_{\theta_0}[\log f_\vartheta(X)],$$

这称为 **KL 散度** (Kullback–Leibler divergence), 注意其中 θ_0 和 ϑ 地位是不对称的. 为了避免 Jensen 不等式, 我们应用

$$\tfrac{1}{2}\log(t) \leq \sqrt{t} - 1, \quad \forall t \geq 0$$

来得到

$$\begin{aligned} -D_{\mathrm{KL}}(P_{\theta_0}||P_\vartheta) &= \int \log\left(\frac{f_\vartheta(x)}{f_{\theta_0}(x)}\right) f_{\theta_0}(x)\,\mathrm{d}\nu(x) \\ &\leq 2\int \left(\sqrt{\frac{f_\vartheta(x)}{f_{\theta_0}(x)}} - 1\right) f_{\theta_0}(x)\,\mathrm{d}\nu(x) \\ &= 2\int \sqrt{f_\vartheta(x) f_{\theta_0}(x)}\,\mathrm{d}\nu(x) - 2 \quad = -2\,d_{\mathrm{H}}^2(P_{\theta_0}, P_\vartheta), \end{aligned}$$

其中

$$d_{\mathrm{H}}^2(P_{\theta_0}, P_\vartheta) := \frac{1}{2}\int \left[\sqrt{f_{\theta_0}(x)} - \sqrt{f_\vartheta(x)}\right]^2 \mathrm{d}\nu(x) > 0$$

是 **Hellinger 距离**的平方. 由此即证 Shannon 不等式. 另可参看 Pinsker 不等式. □

下述结果的条件在附录 §A.1中有所论证.

定理 (相合性)**.** 设 Θ 上定义了确定函数 $M_0(\cdot)$ 和一列随机函数 $M_n(\cdot)$, 适合

- $\sup_{\vartheta\in\Theta} |M_n(\vartheta) - M_0(\vartheta)| \xrightarrow{\mathbb{P}} 0$ (依概率一致收敛), 且
- $\sup_{\vartheta:|\vartheta-\theta_0|\geq\varepsilon} M_0(\vartheta) < M_0(\theta_0)$, $\forall \varepsilon > 0$.

若一列随机变量 $\hat{\theta}_n$ 满足 $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_{\mathbb{P}}(1)$, 则 $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$.

注. 我们用 $o_{\mathbb{P}}(1)$ 表示一列依概率收敛于 0 的随机变量, 这是概率论中的高阶无穷小.

证明. 既然有依概率一致收敛, 对于单独的 $\theta_0 \in \Theta$, 当然成立 $M_n(\theta_0) = M_0(\theta_0) + o_{\mathbb{P}}(1)$. 于是

$$M_0(\theta_0) - M_0(\hat{\theta}_n) = \underbrace{\Big(M_n(\theta_0) - o_{\mathbb{P}}(1) - M_n(\hat{\theta}_n)\Big)}_{\leq o_{\mathbb{P}}(1) - o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)} + \underbrace{\Big(M_n(\hat{\theta}_n) - M_0(\hat{\theta}_n)\Big)}_{\leq \sup_{\vartheta\in\Theta}|M_n(\vartheta) - M_0(\vartheta)| = o_{\mathbb{P}}(1)} \leq o_{\mathbb{P}}(1).$$

任取 $\varepsilon > 0$, 记 $\eta = M_0(\theta_0) - \sup_{\vartheta:|\vartheta-\theta_0|\geq\varepsilon} M_0(\vartheta) > 0$, 则

$$\Big\{|\hat{\theta}_n - \theta_0| \geq \varepsilon\Big\} \subset \Big\{M_0(\theta_0) - M_0(\hat{\theta}_n) \geq \eta\Big\}$$

概率收敛于零. 注意: 证明中各种 sup 都是考虑到 $\hat{\theta}_n$ 在 Θ 中随机取值. □

关于极大似然估计 $\hat{\theta}_n^{\mathrm{ML}} \in \arg\max \ell_n(\vartheta)$ 的极限分布, 我们先给出启发式 (不严格的) 推导. 利用估计方程可得

$$0 = \dot{\ell}_n(\hat{\theta}_n^{\mathrm{ML}}) \approx \dot{\ell}_n(\theta_0) + \ddot{\ell}_n(\theta_0)\left(\hat{\theta}_n^{\mathrm{ML}} - \theta_0\right) \implies \sqrt{n}\left(\hat{\theta}_n^{\mathrm{ML}} - \theta_0\right) \approx \left(-\tfrac{1}{n}\ddot{\ell}_n(\theta_0)\right)^{-1}\left(\tfrac{1}{\sqrt{n}}\dot{\ell}_n(\theta_0)\right).$$

根据 LLN 和 CLT 分别可得

$$-\tfrac{1}{n}\ddot{\ell}_n(\theta_0) \xrightarrow{\mathbb{P}} -\mathbb{E}_{\theta_0}\Big[\ddot{\ell}_1(\theta_0)\Big] = \mathcal{I}_1(\theta_0)$$

和

$$\frac{1}{\sqrt{n}}\left(\dot{\ell}_n(\theta_0)-n\mathbb{E}_{\theta_0}\left[\dot{\ell}_1(\theta_0)\right]\right)\xrightarrow{d}\mathcal{N}\Big(0,\mathrm{Var}_{\theta_0}\left(\dot{\ell}_1(\theta_0)\right)\Big)=\mathcal{N}(0,\mathcal{I}_1(\theta_0)).$$

注意 $\mathbb{E}_{\theta_0}\left[\dot{\ell}_1(\theta_0)\right]=0$. 由 Slutsky 定理即得

$$\sqrt{n}\left(\hat{\theta}_n^{\mathrm{ML}}-\theta_0\right)\xrightarrow[\approx]{d}\mathcal{I}_1(\theta_0)^{-1}\mathcal{N}(0,\mathcal{I}_1(\theta_0))=\mathcal{N}(0,\mathcal{I}_1(\theta_0)^{-1}),$$

或者写成渐近分布 (用 a 表示 asymptotic)

$$\hat{\theta}_n^{\mathrm{ML}}\overset{a}{\sim}\theta_0+\frac{1}{\sqrt{n}}\mathcal{N}(0,\mathcal{I}_1(\theta_0)^{-1})=\mathcal{N}(\theta_0,\mathcal{I}_n(\theta_0)^{-1}).$$

定理 (MLE 的渐近正态性)**.** *设 $\hat{\theta}_n\in\arg\max\ell_n(\vartheta)$ 良好定义, 适合 $\hat{\theta}_n\xrightarrow{\mathbb{P}}\theta_0$. 如果存在 $\varepsilon>0$ 使得*

- $B_\varepsilon=B(\theta_0;\varepsilon)=\{\vartheta:|\vartheta-\theta_0|<\varepsilon\}\subset\Theta$, *且*
- $\mathbb{E}_{\theta_0}\left[\sup_{\vartheta\in B_\varepsilon}\left\|\dddot{\ell}_1(\vartheta)\right\|\right]<\infty$ (Cramér 三阶导条件),

则 $\sqrt{n}\left(\hat{\theta}_n-\theta_0\right)\xrightarrow{d}\mathcal{N}(0,\mathcal{I}_1(\theta_0)^{-1})$.

注. *当然, 一些自然的正则性条件和 i.i.d. 样本的假定不可或缺.*

注. *作为推论, 利用 delta 方法可得* $\sqrt{n}\left(g(\hat{\theta}_n)-g(\theta_0)\right)\xrightarrow{d}\mathcal{N}\left(0,\dot{g}(\theta_0)^\top\mathcal{I}_1(\theta_0)^{-1}\dot{g}(\theta_0)\right)$.

证明. 根据向量值函数的中值定理, 存在 $\hat{\theta}_n$ 和 θ_0 所连线段上的 $\widetilde{\theta}_n$, 使得

$$\left|\dot{\ell}_n(\hat{\theta}_n)-\dot{\ell}_n(\theta_0)-\ddot{\ell}_n(\theta_0)(\hat{\theta}_n-\theta_0)\right|\le\frac{1}{2}\left\|\dddot{\ell}_n(\widetilde{\theta}_n)\right\|\cdot\left|\hat{\theta}_n-\theta_0\right|^2.$$

利用三阶导条件可得 (笔者偷懒了, 或许有错 ☣)

$$\sqrt{n}\left(\hat{\theta}_n-\theta_0\right)-\left(-\frac{1}{n}\ddot{\ell}_n(\theta_0)\right)^{-1}\left(\frac{1}{\sqrt{n}}\dot{\ell}_n(\theta_0)\right)\xrightarrow{\mathbb{P}}0,$$

然后用 Slutsky 定理结合前述启发式推导即可. □

故此, 极大似然估计是**渐近有效的/最优的** (asymptotically efficient/optimal):

- 分布渐近正态;
- (渐近无偏) 偏差 $\mathbb{E}[\hat{\theta}_n]-\theta_0$ 渐近于 0;
- 渐近协方差矩阵 $\mathcal{I}_n(\theta_0)^{-1}$ 达到了 Cramér-Rao 下界[x)].

3.5 枢轴量方法求置信集

考虑样本 $X\sim P_\theta$, 参数 $\theta\in\Theta$ 未知. 设 $T:\mathcal{X}^n\times\Theta\to\mathcal{T}$ 是确定的 (可测) 函数, 如果 $T(X,\theta)$ 的分布不依赖 θ, 即 $\mathbb{P}_\theta\{T(X,\theta)\in\bullet\}$ 对 $\forall\theta\in\Theta$ 都一样, 则称 $T(X,\theta)$ 为**枢轴量** (pivotal quantity).

如果 $B\subset\mathcal{T}$ 满足 $\mathbb{P}_\theta\{T(X,\theta)\in B\}\ge1-\alpha$, 那么 $g(\theta)$ 的一个 $(1-\alpha)$ 置信集为

$$C(X):=\{g(\theta):T(X,\theta)\in B\}.$$

注意 B 的选取不唯一, 需要凭感觉来确定一个合适的.

[x)] 参看后续 §3.8的阐述

例. 设 $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu,\sigma^2)$, 则 $\sqrt{n}(\bar X-\mu)/\sqrt{S^2}\sim t_{n-1}$ 是枢轴量, 其中 $\bar X$ 和 S^2 分别是样本均值和样本方差. 取分位数 $t_{\frac{\alpha}{2},n-1}$ 适合

$$\mathbb{P}\left\{-t_{\frac{\alpha}{2},n-1}\le \sqrt{n}(\bar X-\mu)/\sqrt{S^2}\le t_{\frac{\alpha}{2},n-1}\right\}=1-\alpha,$$

那么 μ 的一个 $(1-\alpha)$ 置信区间为 $\bar X\pm t_{\frac{\alpha}{2},n-1}\sqrt{S^2/n}$. ♠

设 $T_n:\mathcal{X}^n\times\Theta\to\mathcal{T}$ 是一列确定的可测函数, 如果 $T_n(X_1,\dots,X_n,\theta)$ 的极限分布不依赖 θ, 那么 $T_n(X_1,\dots,X_n,\theta)$ 称为**渐近枢轴量** (asymptotically pivotal quantity).

设 $\gamma=g(\theta)$ 的一列置信集 $C_n=C_n(X_1,\dots,X_n)$ 满足

$$\liminf_{n\to\infty}\mathbb{P}_\theta\{\gamma\in C_n\}\ge 1-\alpha,\quad \forall\theta\in\Theta,$$

则称置信集 C_n 具有**渐近置信水平** (asymptotic confidence level) $1-\alpha$.

例. 设估计量 $\hat\theta_n$ 渐近正态 (比如 MLE): $V_n^{-1/2}(\hat\theta_n-\theta)\xrightarrow{d}\mathcal{N}_p(0_p,I_p)$, 其渐近协方差矩阵 $V_n=V_n(\theta)$ 具有相合的估计量 $\hat V_n$, 即 $V_n^{-1/2}\hat V_nV_n^{-1/2}\xrightarrow{\mathbb{P}}I_p$. 此时

$$\left\|\hat V_n^{-1/2}(\hat\theta_n-\theta)\right\|^2\overset{a}{\sim}\chi_p^2$$

是渐近枢轴量, 进而可以构造 θ 的渐近 $(1-\alpha)$ 置信集

$$\left\{\theta\in\Theta:\left\|\hat V_n^{-1/2}(\hat\theta_n-\theta)\right\|^2\le\chi^2_{\alpha,p}\right\},$$

其中 $\chi^2_{\alpha,p}$ 是分布 χ^2_p 的 $(1-\alpha)$ 分位数. ♠

3.6 Bayes 估计

对于样本 $X\sim \mathrm{d}P_{x|\theta}=f(\cdot|\theta)\mathrm{d}\nu$, 未知参数 $\theta\in\Theta$ 可以视为随机元 ϑ 的实现. 随机元 ϑ 蕴含着我们在观测数据 X 之前已知的信息, 其分布 $\mathrm{d}\Pi_\theta=\pi(\cdot)\mathrm{d}\lambda$ 是**先验的** (prior). 通过条件概率密度公式可得 $\vartheta|X=x$ 的分布 $\mathrm{d}\Pi_{\theta|x}=\pi(\cdot|x)\mathrm{d}\lambda$, 其中

$$\boxed{\text{posterior}\propto\text{likelihood}\times\text{prior}}$$

$$\pi(\theta|x)=f(x|\theta)\pi(\theta)/f(x)\propto_\theta f(x|\theta)\pi(\theta),\quad \theta\in\Theta.$$

由于加入了观测结果 X 的信息, 条件分布 $\Pi_{\theta|x}$ 称为**后验的** (posterior).

✍ 注: $\pi(\cdot|x)$ 的形状完全取决于 $f(x|\cdot)\pi(\cdot)$, 由此就能确定后验分布, 而 $f(x)=\int_\Theta f(x|\theta)\pi(\theta)\,\mathrm{d}\lambda(\theta)$ 可以视为归一化系数, 即配分函数 (partition function). ↯ 在 Bayes 框架下, 所有推断都基于后验.

例. 设 $X_1,\dots,X_n|\vartheta=\theta\overset{\text{i.i.d.}}{\sim}\mathcal{N}(\theta,\sigma^2)$, 其中 $\sigma^2>0$ 已知. 记 $X=(X_1,\dots,X_n)^\top$ 和 $\bar X=\frac1n\sum_{i=1}^nX_i$. 若 $\vartheta\sim\mathcal{N}(\theta_0,\sigma^2/\kappa_0)$, 其中 $\theta_0\in\mathbb{R}$ 和 $\kappa_0>0$ 已知, 则

$$\begin{aligned}\pi(\theta|x)\propto_\theta f(x|\theta)\pi(\theta)&=(2\pi\sigma^2)^{-\frac n2}\exp\left\{-\sum_{i=1}^n\frac{(x_i-\theta)^2}{2\sigma^2}\right\}(2\pi\sigma^2/\kappa_0)^{-\frac12}\exp\left\{-\frac{(\theta-\theta_0)^2}{2\sigma^2/\kappa_0}\right\}\\&\propto_\theta\exp\left\{-\frac{\sum_{i=1}^n(x_i^2-2x_i\theta+\theta^2)}{2\sigma^2}-\frac{\theta^2-2\theta_0\theta+\theta_0^2}{2\sigma^2/\kappa_0}\right\}\\&\propto_\theta\exp\left\{-\frac{-2n\bar x\theta+n\theta^2}{2\sigma^2}-\frac{\kappa_0\theta^2-2\kappa_0\theta_0\theta}{2\sigma^2}\right\}\\&=\exp\left\{-\frac{(n+\kappa_0)\theta^2-2(n\bar x+\kappa_0\theta_0)\theta}{2\sigma^2}\right\}\quad=\exp\left\{-\frac{\theta^2-2\frac{n\bar x+\kappa_0\theta_0}{n+\kappa_0}\theta}{2\frac{\sigma^2}{n+\kappa_0}}\right\},\end{aligned}$$

此即 $\vartheta|X\sim\mathcal{N}\left(\frac{\kappa_0\theta_0+n\bar X}{\kappa_0+n},\frac{\sigma^2}{\kappa_0+n}\right)$, 其中后验均值是 θ_0 和 $\bar X$ 的加权混合, 且后验方差小于先验方差. ♠

自然地, 我们定义 $\gamma = g(\theta)$ 的 **Bayes 估计量**为

$$\hat{\gamma}^{\text{Bayes}} := \mathbb{E}[g(\vartheta)|X] = \int_{\Theta} g(\cdot)\pi(\cdot|X)\,\mathrm{d}\lambda.$$

为了论证其优良性, 我们引入估计量 $T = T(X)$ 关于 $\mathrm{d}\Pi_\theta = \pi(\cdot)\mathrm{d}\lambda$ 的 **Bayes 风险** (Bayes risk)

$$r_\pi(T) := \int_{\Theta} R(\theta, T)\pi(\theta)\,\mathrm{d}\lambda(\theta),$$

其中风险函数 (risk function)

$$R(\theta, T) := \mathbb{E}[L(\theta, T(X))|\vartheta = \theta]$$

采用了损失函数 (loss function)

$$L(\theta, \tau) := |g(\theta) - \tau|^2.$$

定理. *Bayes 估计量使 Bayes 风险最小化, 即* $\hat{\gamma}^{\text{Bayes}} \in \arg\min r_\pi(\bullet)$.

证明. 由于 $r_\pi(T) = \mathbb{E}[\mathbb{E}[L(\vartheta, T)|\vartheta]] = \mathbb{E}[L(\vartheta, T)] = \mathbb{E}[\mathbb{E}[L(\vartheta, T)|X]]$, 只需说明 $\hat{\gamma}^{\text{Bayes}} = \mathbb{E}[g(\vartheta)|X]$ 使**后验风险** $\mathbb{E}[L(\vartheta, T)|X]$ 最小化. 仿照一维均方误差 (MSE) 的偏差方差分解, 可以得到

$$\mathbb{E}[L(\vartheta, T)|X] = \mathbb{E}\left[|g(\vartheta) - T|^2 \,\middle|\, X\right] = |\mathbb{E}[g(\vartheta)|X] - T|^2 + \operatorname{tr}\left(\operatorname{Var}[g(\vartheta)|X]\right),$$

从而 $T = \mathbb{E}[g(\vartheta)|X]$ 是最优解. □

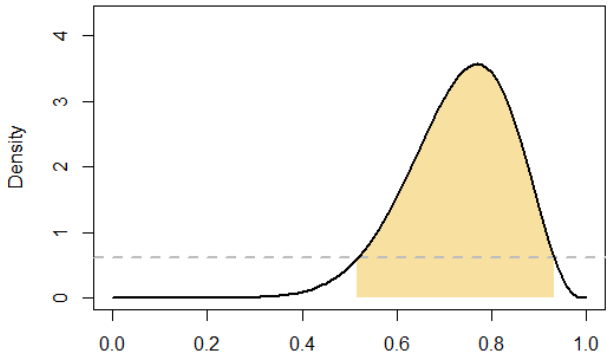
在 Bayesian 分析中, 区域估计的概念有别于频率学派的置信集, 避开了求统计量抽样分布的难题. 我们称取值为集合的统计量 $C = C(X)$ 为参数 $\gamma = g(\theta)$ 的水平 $(1-\alpha)$ 的**可信集** (credible set), 若对 (固定的) 观测结果 x 成立

$$\mathbb{P}(g(\vartheta) \in C(x)|X = x) = \int_{g^{-1}(C(x))} \pi(\theta|x)\,\mathrm{d}\lambda(\theta) \geq 1 - \alpha.$$

如果存在常数 $c = c_{\alpha,x}$, 使得可信集形如

$$C(x) = \{g(\theta) : \pi(\theta|x) \geq c\},$$

那么此 $C = C(x)$ 称为**最大后验密度** (highest posterior density, 简记为 HPD) 可信集[xi].



定理. *设 $\Theta = \mathbb{R}$, 后验密度函数 $\pi(\cdot|x)$ 在 $(-\infty, \theta_0]$ 上单调递增, 在 $[\theta_0, \infty)$ 上单调递减. 在形如 $[a, b]$ 的 θ 的 $(1-\alpha)$ 可信区间中, 具有最短区间长度的一定是 HPD 可信区间.* (证明: 略. ❤)

[xi] 实例可参看 https://stats.stackexchange.com/a/160035

既然 Bayes 统计方法极度依赖于先验分布 Π_θ, 在实践中我们就需要关心如何选取先验. 下面罗列一些原则性的考量. (✍ 补充: de Finetti 定理——引入先验 (在数学上) 是合理的!

- *客观视角.* 利用数据本身来构造先验分布, 这称为**经验贝叶斯方法** (empirical Bayes method, 简称 EB), 可以看作贝叶斯学派与频率学派的结合. 例如: 先选定先验分布的类型, 再用数据估计其 (超) 参数. 不足之处在于, 如果缺少历史数据, 那么现有数据 (可能) 两次被使用.
- *主观视角.* 对 θ 的过往认知可能因人而异, 一个广泛采用的准则是 "**同等无知**", 即先验分布尽可能平滑 (无信息). 例如: 取 $\Pi_\theta = \mathrm{Uniform}(\Theta)$, 即 $\pi(\cdot) \propto 1$; 常用的还有 **Jeffreys 先验** $\pi(\theta) \propto \sqrt{\det \mathcal{I}_n(\theta)}$, 其中 $\mathcal{I}_n(\theta)$ 是 Fisher 信息矩阵. 值得注意的是, 有时候得到的 $\pi(\cdot)$ 未必是良定的 p.d.f. (但是仍然可以作为一个权重函数来使用), 即 $\int_\Theta \pi(\theta)\,\mathrm{d}\lambda(\theta) \neq 1$, 此时 $\mathrm{d}\Pi_\theta = \pi(\cdot)\mathrm{d}\lambda$ 称为**非正常先验** (improper prior). 例如 $\Theta = \mathbb{R}_+$ 时, $\mathrm{Uniform}(\Theta)$ 不存在, 因为 $\int_0^\infty 1\,\mathrm{d}\theta = \infty$.
- *共轭分布族.* 设 $\mathcal{F}$ 是 $\theta \in \Theta$ 的一个分布族, 我们称 $\mathcal{F}$ 关于总体的集合 $\mathcal{P}$ **共轭** (conjugate), 若对任意样本 x, 只要先验分布 $\Pi_\theta \in \mathcal{F}$, 就有后验分布 $\Pi_{\theta|x} \in \mathcal{F}$. 举例如下:

$\theta \sim$	$\mathrm{Beta}(\alpha,\beta)$	$\mathrm{Gamma}(k,\lambda)$	$\mathcal{N}\left(\theta_0, \frac{\sigma^2}{\kappa_0}\right)$
$\pi(\theta) \propto$	$\theta^{\alpha-1}(1-\theta)^{\beta-1}\mathbb{1}_{[0,1]}(\theta)$	$\theta^{k-1}\mathrm{e}^{-\lambda\theta}\mathbb{1}_{(0,\infty)}(\theta)$	$\exp\left\{-\frac{(\theta-\theta_0)^2}{2\sigma^2/\kappa_0}\right\}$
$x_1,\dots,x_n\|\theta \overset{\text{i.i.d.}}{\sim}$	$\mathrm{Bernoulli}(\theta)$	$\mathrm{Poisson}(\theta)$	$\mathcal{N}(\theta,\sigma^2)$
$f(x\|\theta) \propto_\theta$	$\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$	$\theta^{\sum_{i=1}^n x_i}\mathrm{e}^{-n\theta}$	$\exp\left\{-\sum_{i=1}^n \frac{(x_i-\theta)^2}{2\sigma^2}\right\}$
$\theta\|x \sim$	$\mathrm{Beta}(\alpha+n\bar{x},\beta+n-n\bar{x})$	$\mathrm{Gamma}(k+n\bar{x},\lambda+n)$	$\mathcal{N}\left(\frac{\kappa_0\theta_0+n\bar{x}}{\kappa_0+n}, \frac{\sigma^2}{\kappa_0+n}\right)$

在一定的正则性条件下, 我们尝试推导后验分布的大样本性质. 记对数似然函数为 $\ell_n(\theta) = \log f(x|\theta)$, 极大似然估计量 $\hat{\theta}_n$ 适合 $\dot{\ell}_n(\hat{\theta}_n) = 0$. 通过 Taylor 展开得到

$$\ell_n(\theta) \approx \ell_n(\hat{\theta}_n) + \cancel{\dot{\ell}_n(\hat{\theta}_n)^\top(\theta-\hat{\theta}_n)} + \tfrac{1}{2}(\theta-\hat{\theta}_n)^\top \ddot{\ell}_n(\hat{\theta}_n)(\theta-\hat{\theta}_n),$$

其中 $\ell_n(\hat{\theta}_n)$ 仅依赖 x, 从而 (🍵🍵🍵 扩展材料: 最大似然法和贝叶斯的区别? - 子元的回答 - 知乎

$$\begin{aligned}\pi(\theta|x) \propto_\theta \pi(\theta)\exp\{\ell_n(\theta)\} &\approx \pi(\theta)\exp\left\{\ell_n(\hat{\theta}_n) + \tfrac{1}{2}(\theta-\hat{\theta}_n)^\top\ddot{\ell}_n(\hat{\theta}_n)(\theta-\hat{\theta}_n)\right\}\\ &\propto_\theta \pi(\theta)\exp\left\{-\tfrac{1}{2}(\theta-\hat{\theta}_n)^\top\Big(-\ddot{\ell}_n(\hat{\theta}_n)\Big)(\theta-\hat{\theta}_n)\right\}.\end{aligned}$$

可见似然函数 $\exp\{\ell_n(\theta)\}$ 差不多相当于 $\mathcal{N}\left(\hat{\theta}_n, -\ddot{\ell}_n(\hat{\theta}_n)^{-1}\right)$ 带来的权重. 注意到 i.i.d. 样本会导致 $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$, 其中 θ_0 是真实参数, 且 $-\ddot{\ell}_n(\theta)/n \xrightarrow{\mathbb{P}} -\mathbb{E}\ddot{\ell}_1(\theta) = \mathcal{I}_1(\theta)$, 所以 $\theta|x$ 后验分布会渐近 $\mathcal{N}(\theta_0, \mathcal{I}_n(\theta_0)^{-1})$. 这也给出了 Jeffreys 先验 $\pi(\theta) \propto \sqrt{\det \mathcal{I}_n(\theta)}$ 的一种解释.

3.7 基础性理论

给定样本 $X \sim P_\theta$, 其中参数 $\theta \in \Theta$ 未知, 我们希望找出 $\gamma = g(\theta)$ 的最优的估计量, 其中 $g: \Theta \to \Gamma$ 是确定的已知函数. 一般而言, 我们对估计量 $\hat{\gamma} = T(X)$ 定义**风险** (risk)

$$R(\theta, T) := \mathbb{E}_\theta[L(\theta, T(X))] = \int_{\mathcal{X}^n} L(\theta, T(\cdot))\,\mathrm{d}P_\theta,$$

其中 $L: \Theta \times \Gamma \to \mathbb{R}_+$ 是人为选定的**损失函数** (loss function). 例如: 取二次损失 $L(\theta,\gamma) = |g(\theta)-\gamma|^2$ 会得到**均方误差** (mean squared error) $\mathrm{MSE}(\theta;T) = \mathbb{E}_\theta[|g(\theta)-T(X)|^2]$. 在统计决策理论[xii] (statistical decision theory) 的框架下, 最优的估计量应当使风险最小化.

[xii] 参看 https://zhuanlan.zhihu.com/p/102390438

对于统计模型 $\mathcal{P}=\{P_\theta:\theta\in\Theta\}$, 称统计量 $S=S(X)$ 是**充分的** (sufficient), 若对 S 的任一实现值 s 有条件分布 $P_\theta(\bullet|S=s)=\mathbb{P}_\theta(X\in\bullet|S(X)=s)$ 不依赖 θ. 统计学中有一个充分性原则 (sufficiency principle), 说的是统计推断应当仅依赖充分统计量. 例如, 将后面的因子分解定理应用于极大似然.

定理 (Rao-Blackwell)**.** *设 Γ 是凸集, 损失函数 L 对 $\forall\theta\in\Theta$ 满足 $\gamma\in\Gamma\mapsto L(\theta,\gamma)$ 是凸函数. 若 $S=S(X)$ 是充分统计量, 则对 γ 的任一估计量 $\hat\gamma$, 我们有 $\hat\gamma_S:=\mathbb{E}[\hat\gamma|S]$ 适合 $R(\theta,\hat\gamma_S)\le R(\theta,\hat\gamma)$.*

证明. 利用条件期望的 Jensen 不等式, 可得 $\mathbb{E}[L(\theta,\hat\gamma)|S]\ge L(\theta,\hat\gamma_S)$. □

为了寻找充分统计量, 一个简便方法是应用如下 Fisher-Neyman **因子分解** (factorization) 定理:

定理. *设 P_θ 有密度函数 f_θ, 则统计量 $S(X)$ 充分, 当且仅当存在分解 $f_\theta(x)=g_\theta(S(x))h(x)$.*

证明. 如果样本 X 是离散型随机元, 那么容易得到条件概率密度函数 $f_{\theta|S=s}\propto h$ 不依赖 θ. 一般情形需要用到测度论. 参看https://zhuanlan.zhihu.com/p/102499608. □

称统计量 $V=V(X)$ 是**辅助的** (ancillary), 若分布 $\mathbb{P}_\theta\{V\in\bullet\}$ 不依赖 θ.

辅助统计量不携带总体的信息, 而充分统计量包含了总体的全部信息. 如果能够去除多余的信息, 那么似乎应当得到两个无关的统计量.

称统计量 $T=T(X)$ 是**完全的** (complete), 若对任一 (确定的) 函数 ψ, 只要 $\mathbb{E}_\theta[\psi(T)]=0$, $\forall\theta\in\Theta$, 就有 $\psi(T)\overset{\mathbb{P}_\theta\text{-a.s.}}{=\!=}0$, $\forall\theta\in\Theta$.

定理 (Basu)**.** *若 $T=T(X)$ 是完全的充分统计量, $V=V(X)$ 是辅助统计量, 则 T 与 V 独立.*

注. *关于模型 $\{P_\theta:\theta\in\Theta\}$ 的统计概念推出了每一个总体 P_θ 下的概率性质!*

证明. 任取函数 ϕ, 易见 $\mathbb{E}[\phi(V)|T]-\mathbb{E}[\phi(V)]$ 不依赖 θ, 由 T 的完全性立得 $\mathbb{E}[\phi(V)|T]\overset{\text{a.s.}}{=\!=}\mathbb{E}[\phi(V)]$, 所以 V 独立于 T. □

统计学中很多常用的参数族可以归结为**指数族**[xiii] (exponential family). 称 $\mathrm{d}P_\theta=f_\theta\mathrm{d}\nu$ 是指数族分布, 若密度函数形如

$$f_\theta(x)=\exp\{\xi(\theta)^\top T(x)-B(\theta)\}h(x).$$

设 $X\sim f_\theta$, 应用因子分解定理可知, $T=T(X)$ 是充分统计量. 若 $\{\xi(\theta):\theta\in\Theta\}$ 包含一个开集, 则 $T=T(X)$ 是完全的, 其证明手法类似于矩母函数唯一决定分布. 事实上, 为了直观地构造指数族分布, 可以借助矩母函数来进行指数倾斜 (exponential tilting).

3.8 无偏估计

为了找出最优的估计量, 我们对估计量做出一些限制, 从而缩小考察的范围. 特别地, 本节关注于无偏的估计量. 若 $\gamma=g(\theta)$ 存在无偏估计, 则称 γ 是**可估的** (estimable).

考虑二次损失, 相应的风险为均方误差. 对于 $\gamma\in\mathbb{R}$ 的无偏估计量 $T=T(X)$, 我们有

$$\mathrm{MSE}(\theta;T)=\mathrm{Var}_\theta(T(X)).$$

在统计决策理论的框架下, 最优的无偏估计量是**一致最小方差无偏的** (uniformly minimum-variance unbiased, 简记为 UMVU); 若 $T=T(X)$ 是 UMVU 估计量, 则其他无偏估计量 $\widetilde{T}=\widetilde{T}(X)$ 都满足

$$\mathrm{Var}_\theta(T)\le\mathrm{Var}_\theta(\widetilde{T}),\ \forall\theta\in\Theta.$$

[xiii] 参看https://zhuanlan.zhihu.com/p/103110033

定理 (Lehmann-Scheffé). 设统计量 $S = S(X)$ 充分且完全, 参数 $\gamma = g(\theta) \in \mathbb{R}$ 具有无偏估计量 $T = T(X)$, 则对 γ 而言, 存在唯一的 *UMVU* 估计量 $T_* := \mathbb{E}[T|S]$.

证明. 易见 T_* 无偏. 任取其他无偏估计量 $\widetilde{T} = \widetilde{T}(X)$, 我们有 $\mathbb{E}_\theta[\mathbb{E}[\widetilde{T}|S] - T_*] = 0$, $\forall \theta \in \Theta$, 由 S 完全可得 $\mathbb{E}[\widetilde{T}|S] \overset{\text{a.s.}}{=\!=} T_*$. 因为二次损失 $L(\theta, \gamma) = |g(\theta) - \gamma|^2$ 关于 $\gamma \in \mathbb{R}$ 是严格凸的, Rao-Blackwell 定理 (充分性原则) 保证了估计量 T_* 是最优的. □

如果 $\gamma = g(\theta) \in \mathbb{R}$ 的无偏估计量的方差有下界, 那么达到此下界的即为 UMVU 估计量. 为此, 有如下信息不等式:

定理 (Cramér-Rao 下界 (Cramér-Rao lower bound, 简记为 CRLB)). 设 $\Theta \subseteq \mathbb{R}^m$, 在若干正则性条件下, 若 $T = T(X)$ 是 $g(\theta)$ 的无偏估计量, 则 $\operatorname{Var}_\theta(T) \geq \dot{g}(\theta)^\top \mathcal{I}_n(\theta)^{-1} \dot{g}(\theta)$, 其中 $\mathcal{I}_n(\theta)$ 是 *Fisher* 信息.

注. 可以用 $\operatorname{eff}_\theta(T) := \dot{g}(\theta)^\top \mathcal{I}_n(\theta)^{-1} \dot{g}(\theta) / \operatorname{Var}_\theta(T)$ 衡量无偏估计 T 的**效率** (efficiency).

证明. 记得分函数为 $s_n(\theta)$, 可以得到 $\dot{g}(\theta) = \operatorname{Cov}_\theta(s_n(\theta), T)$, 利用 $\mathcal{I}_n(\theta) = \operatorname{Var}_\theta(s_n(\theta))$ 即得所求. □

关于细节和推广, 可参看https://zhuanlan.zhihu.com/p/103274162.

值得一提的是, 无偏性并没有什么特别的好处 (虽然我们总是追求渐近无偏性), 请看下例.

例. 设 $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Uniform}([0, \theta])$, 其中 $\theta \in \mathbb{R}_+$ 是未知参数. 记样本最大值为 $X_{(n)} = \max X_i$, 这是一个充分且完全的统计量 (☕ 读者自证不难). 为了得到 $g(\theta) \in \mathbb{R}$ 的 UMVU 估计量, 我们只需寻找无偏估计量形如 $\psi(X_{(n)})$. 直接计算可得 (☺ 留给读者)

$$\theta^n g(\theta) = \theta^n \mathbb{E}_\theta[\psi(X_{(n)})] = n \int_0^\theta \psi(x) x^{n-1} \,\mathrm{d}x, \quad \forall \theta \in \mathbb{R}_+.$$

对 θ 求导可得

$$n\theta^{n-1} g(\theta) + \theta^n \dot{g}(\theta) = n\psi(\theta)\theta^{n-1}, \quad \forall \theta.$$

由此立得 $g(\theta)$ 的 UMVU 估计量

$$\psi(X_{(n)}) = g(X_{(n)}) + \frac{1}{n} X_{(n)} \dot{g}(X_{(n)}).$$

特别地, θ 的 UMVU 估计量是 $\frac{n+1}{n} X_{(n)}$. 接下来, 我们尝试在 $\{cX_{(n)} : c \in \mathbb{R}_+\}$ 中找出 θ 的最优估计. 可以证明 (☺ 留给读者)

$$\operatorname{MSE}(\theta; cX_{(n)}) = \left(\frac{1}{n+2} c^2 - \frac{2}{n+1} c + \frac{1}{n} \right) n\theta^2,$$

所以最优的系数为 $c^* = \frac{n+2}{n+1}$. 虽然 $\frac{n+2}{n+1} X_{(n)}$ 有偏, 但是比起无偏估计量 $\frac{n+1}{n} X_{(n)}$ 具有更小的 MSE. ♠

思考: https://math.stackexchange.com/questions/2518975/what-is-the-fisher-information-for-a-uniform-distribution? 解答: Fisher information does not exist for distributions with parameter-dependent supports. Using different formulae for the information function, you arrive at different answers. — StubbornAtom Mar 21 '19 at 8:30 ♠

作为补充, 罗列一些稍微深入的话题, 有兴趣的读者可自行查找相关资料. ⌕

- 渐近比较: Hodges-LeCam 估计量超有效 (super efficient).
- 最小最大 (minimax) 估计: 用 Bayes 方法论证容许性 (admissibility).
- 收缩 (shrinkage) 估计: 维数超过 2 时, James-Stein 估计量总有更优的表现.
- 稳健性 (robustness): M-估计/Z-估计中的 Huber 估计量.
- 自助法 (bootstrap): 机器学习中常用布袋法 (bagging) 和提升算法 (boosting), 起源于抽样调查.

4 假设检验

在统计学中, 除了估计之外, 另一类极其重要的问题是**假设检验** (hypothesis testing). 任给一个关于总体的看法, 我们想判断它在多大程度上正确, 是否应该拒绝. 当然, 我们会利用收集到的数据, 以此作为判断依据. 设样本 $X=(X_1,\dots,X_n)^\top$ 的备选总体构成统计模型 $\mathcal{P}=\{P_\theta:\theta\in\Theta\}$, 关于总体的看法往往可以化为对 θ 的描述.

4.1 问题提法

形式地, 考察 $H_0:\theta\in\Theta_0$, 这称为**零假设** (null hypothesis), 约定俗成是一个 "保守的" 断言, 其中 $\Theta_0\subsetneq\Theta$ 是确定的. 作为对立面, $H_1:\theta\in\Theta_1$ 称为**备择假设** (alternative hypothesis), 其中 $\Theta_1:=\Theta\backslash\Theta_0$.

- 若 $\Theta_i=\{\theta_i\}$ 是单点集, 则称假设 H_i 是**简单的** (simple); 否则称为**复合的** (composite).
- 当 $\Theta\subset\mathbb{R}^1$ 时, 设 $\theta_L\le\theta_R$,
  - 若 Θ_i 形如 $[\theta_0,\infty)$ 或 (θ_0,∞) 或 $(-\infty,\theta_0)$ 或 $(-\infty,\theta_0]$, 则称假设 H_i 为**单边的** (one-sided).
  - 若 Θ_i 形如 $\{\theta\in\Theta:\theta<(\le)\,\theta_L\}\sqcup\{\theta\in\Theta:\theta>(\ge)\,\theta_R\}$, 则称假设 H_i 为**双边的** (two-sided).

我们要在

$$H_0:\theta\in\Theta_0\overset{\text{vs.}}{\longleftrightarrow}H_1:\theta\in\Theta_1\quad(\Theta_0\cap\Theta_1=\emptyset)$$

之间做一个抉择. 为此, 引入随机化决策准则, 我们定义一个取值于 $[0,1]$ 的统计量 $\varphi=\varphi(X)$, 称为**检验函数** (test function), 代表拒绝 H_0 的概率. 也就是说, 观测到样本 $X=x$ 之后, 生成一个随机数 $u\sim\text{Uniform}([0,1])$, 当且仅当 $u\le\varphi(x)$ 时我们选择拒绝 H_0. 事实上, 这包含了确定的决策准则:

- 检验函数可以形如 $\varphi(x)=\mathbb{1}_K(x)$, 其中 K 是样本取值空间的子集, 称为**拒绝域** (rejection region).
  - 拒绝域一般形如 $K=\{x:T(x)>c\}$, 其中 T 是确定的 (可测) 实值函数, c 是确定的实数. 此时, $T=T(X)$ 称为**检验统计量** (test statistic), 而 $c\in\mathbb{R}$ 称为**临界值** (critical value).
    * 称两个检验统计量是**等价的** (equivalent), 若它们能导出相同的检验函数;
      检验统计量等价的一个充分条件是——二者之间存在单调的双射.

为了衡量拒绝零假设 H_0 的力度, 定义检验函数 $\varphi=\varphi(X)$ 的**功效函数** (power function) 为

$$\pi(\theta;\varphi):=\mathbb{E}_\theta[\varphi(X)]=\int\varphi(x)\,\mathrm{d}P_\theta(x),\quad\forall\theta\in\Theta=\Theta_0\sqcup\Theta_1.$$

在检验中我们可能会犯两种错误:

- **第一类错误** (type I error)/ 假阳性 (false positive): 当 H_0 为真时, 拒绝 H_0.
- **第二类错误** (type II error)/ 假阴性 (false negative): 当 H_1 为真时, 没有拒绝 H_0.

对于检验函数 $\varphi=\varphi(X)$, 定义检验的**精确水平** (size) 为

$$\alpha(\varphi):=\sup_{\theta\in\Theta_0}\pi(\theta;\varphi).$$

设 $\alpha\in(0,1)$, 若 $\alpha(\varphi)\le\alpha$, 则称检验具有**水平** (level) α. 通过控制检验水平 α, 可以消减第一类错误. 为了消减第二类错误, 我们需要提高检验的**功效** (power) $\inf_{\theta\in\Theta_1}\pi(\theta;\varphi)$. 记

$$\beta(\varphi):=1-\inf_{\theta\in\Theta_1}\pi(\theta;\varphi).$$

如果对于 $\forall\theta\in\Theta_1$, 当样本容量 $n\to\infty$ 时, 均有 $\pi(\theta;\varphi)\to1$, 那么检验称为**相合的** (consistent).

我们将检验中的错误总结成表格:

	Retain H_0	Reject H_0
H_0 true	✓	Type I Error (α)
H_1 true	Type II Error (β)	✓ (power)

4.2 统计显著性和 p 值

一般来说, 我们希望拒绝零假设 H_0. 检验水平 α 也称为**显著性水平** (significance level). 例如, 考虑 $\theta \in \mathbb{R}$ 衡量了某个作用的强弱, 零假设 $H_0 : \theta = 0$ 说的是该作用不存在, 如果我们通过一个水平 α 的检验拒绝了 H_0, 那么可以称 θ 在水平 α 下是显著的. 当然, 这里的显著性完全是统计学概念, 并未涉及实际效果的大小. 在实践中, 人们常常取显著性水平为 $\alpha = 0.05$, 然后再设计检验.

大名鼎鼎的 **p 值** (p-value, 是 probability value 的简写) 与显著性密切联系而又稍有区别. 考虑如下非随机的检验, 拒绝域 $K = K_{T,c}$ 依赖于检验统计量 T 和临界值 c, 此时功效函数为

$$\pi(\theta; T, c) = \mathbb{E}_\theta[\mathbb{1}_K(X)] = \mathbb{P}_\theta\{X \in K_{T,c}\}, \quad \theta \in \Theta = \Theta_0 \sqcup \Theta_1.$$

观测到样本 $X = x$ 之后, 记 $t := T(x)$ 为检验统计量的实现值, 我们定义 ✍ 注意: p-value 是统计量!

$$p\text{-value} := \sup_{\theta \in \Theta_0} \pi(\theta; T, t).$$

当 p-value < 0.05 时, 可以认为存在*较强的证据以反对* (strong evidence against) 零假设 $H_0 : \theta \in \Theta_0$. 作为比较, 显著性水平

$$\alpha(T, c) = \sup_{\theta \in \Theta_0} \pi(\theta; T, c)$$

中的临界值 $c \in \mathbb{R}$ 是人为选取并且*预先确定*的数. 特别地, 若拒绝域形如

$$K = \{x : T(x) \geq c\},$$

则功效函数 $\pi(\theta; T, c)$ 关于 c 单调递减, 进而可见显著性水平 $\alpha(T, c)$ 关于 c 单调递减, 此时成立

$$x \in K \iff t \geq c \iff p\text{-value} \leq \alpha(T, c).$$

不难发现, 当拒绝域形如 $K = \{x : T(x) \leq c\}$ 时, 也有类似结论.

一般地, 任给一族非随机检验 $\{\varphi_\alpha\}_{\alpha \in (0,1)}$, 其中检验函数 φ_α 具有水平 α, 则相应的 p 值定义为

$$p\text{-value} := \widehat{\alpha}(X) = \inf\{\alpha \in (0,1) : \varphi_\alpha(X) = 1\}.$$

延伸阅读: 如何看待 Nature 上 800 名科学家联名反对统计学意义, 放弃 “p 值决定论”?
斯坦福教授: 没有 p 值, 期刊将充斥 “无可辩驳的废话”

4.3 假设检验与置信集的联系

如果用拒绝域 K 来检验零假设 H_0, 那么 $A := K^\complement$ 称为**接受域** (acceptance region). 在一定程度上, 参数 $\gamma = g(\theta)$ 的置信集 (请回顾 §3.2) 与假设检验的接受域可以互相转化.

定理. *对于 $\forall \theta_0 \in \Theta$, 用 $A(\theta_0)$ 作为接受域对零假设 $H_0 : \theta = \theta_0$ 的检验具有水平 α, 当且仅当 $\gamma = g(\theta)$ 的置信集 $C(X) := \{g(\vartheta) : X \in A(\vartheta)\}$ 具有置信水平 $1 - \alpha$.*

证明. 二者都划归为 $\mathbb{P}_\theta\{X \notin A(\theta)\} \leq \alpha$, $\forall \theta \in \Theta$. □

因此, 在进行区域估计时, 可以通过假设检验的方法来构造置信集. 反之亦然. 如下所示:

$$X \in A(\theta) \iff g(\theta) \in C(X).$$

4.4 似然比检验

对于假设检验问题 $H_0 \leftrightarrow H_1$, 我们首先从 *Bayesian* 视角考察一下. 基于样本 X 更新信念之后, 我们自然地选择拒绝 H_0 当且仅当后验满足 $\mathbb{P}(H_0|X) < \mathbb{P}(H_1|X)$. 为此, 形式地计算得到

$$\frac{\mathbb{P}(H_1|X)}{\mathbb{P}(H_0|X)} = \frac{\mathbb{P}(X|H_1)\mathbb{P}(H_1)}{\mathbb{P}(X|H_0)\mathbb{P}(H_0)} > 1 \iff \frac{\mathbb{P}(X|H_1)}{\mathbb{P}(X|H_0)} > \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)}.$$

基于过往的认知, 先验 $\mathbb{P}(H_i)$ 一般是确定的, 所以我们拒绝 H_0 当且仅当 $\mathbb{P}(X|H_1)/\mathbb{P}(X|H_0)$ 比较大.

更严格一点, 设样本 X 的总体 P_θ 具有密度函数 f_θ, 记似然函数为 $L_n(\theta) := f_\theta(X)$, $\theta \in \Theta$. 对于 $H_0 : \theta \in \Theta_0 \leftrightarrow H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$, 定义**广义似然比统计量** (generalized likelihood ratio statistic) 为

$$\Lambda = \Lambda_n(X) := \frac{\sup_{\theta\in\Theta} L_n(\theta; X)}{\sup_{\theta_0\in\Theta_0} L_n(\theta_0; X)}.$$

人们常常借此构造检验 $\varphi = \mathbb{1}_K(X)$, 取拒绝域形如 $K = \{x : \Lambda_n(x) > c\}$, 临界值 $c \in \mathbb{R}_+$ 经过适当的选取可以让检验的水平合乎所需. 在一定的正则性条件下, 有大样本性质:

定理 (Wilks)**.** *当零假设 H_0 为真时, 上述似然比统计量满足 $2\log(\Lambda_n) \xrightarrow[(n\to\infty)]{d} \chi^2_{\dim(\Theta)-\dim(\Theta_0)}$.*

证明. 主要是利用极大似然估计量的渐近正态性, 通过 delta 方法进行论证. 读者有兴趣的话可参看 https://www.statlect.com/fundamentals-of-statistics/likelihood-ratio-test □

4.5 Neyman-Pearson 范式

对于检验函数 $\varphi = \varphi(X)$, 记功效函数为 $\pi(\theta; \varphi) = \mathbb{E}_\theta[\varphi(X)]$. 经典的 Neyman-Pearson 假设检验思想是: 控制住显著性水平 $\alpha(\varphi) = \sup_{\theta\in\Theta_0} \pi(\theta; \varphi)$, 并且对每个 $\theta \in \Theta_1$ 尽可能最大化 $\pi(\theta; \varphi)$.

✍ 检验完全取决于检验函数, 所以我们将不区分这两个概念.

称检验函数 $\varphi^* = \varphi^*(X)$ 是一**致最大功效的** (uniformly most powerful, 简记为 UMP) 水平 α 的检验, 若 $\alpha(\varphi^*) \le \alpha$, 且对任一水平不超过 α 的检验函数 $\varphi = \varphi(X)$, 均成立

$$\pi(\theta; \varphi^*) \ge \pi(\theta; \varphi), \quad \forall \theta \in \Theta_1.$$

简单起见, 先考察 $\Theta = \{\theta_0, \theta_1\}$ 是两点集的情形. 设总体 P_θ 具有密度函数 f_θ, 记似然函数为 $L_n(\theta) := f_\theta(X)$, $\theta \in \Theta$. 定义似然比统计量 $\Lambda := L_n(\theta_1)/L_n(\theta_0)$.

定理 (Neyman-Pearson lemma, **NP 引理**)**.** *对于 $H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta = \theta_1$, 任给 $\alpha \in (0, 1)$, 取 $\lambda \in \mathbb{R}_+$ 和 $\delta \in [0, 1]$ 适合 $\mathbb{P}_{\theta_0}\{\Lambda > \lambda\} + \delta\,\mathbb{P}_{\theta_0}\{\Lambda = \lambda\} = \alpha$, 则似然比检验*

$$\varphi^* := \mathbb{1}_{\{\Lambda>\lambda\}} + \delta\mathbb{1}_{\{\Lambda=\lambda\}}$$

是 UMP 水平 α 检验.

证明. 由定义,

$$\alpha(\varphi^*) = \pi(\theta_0; \varphi^*) = \mathbb{E}_{\theta_0}[\varphi^*] = \mathbb{P}_{\theta_0}\{\Lambda > \lambda\} + \delta\,\mathbb{P}_{\theta_0}\{\Lambda = \lambda\} = \alpha.$$

任取检验函数 φ, 由于 $0 \le \varphi \le 1$, 我们有

$$\big(\varphi^*(x) - \varphi(x)\big)\big(f_{\theta_1}(x) - \lambda f_{\theta_0}(x)\big) \ge 0, \quad \forall x.$$

将上式变形为

$$\big(\varphi^*(x) - \varphi(x)\big) f_{\theta_1}(x) \ge \lambda\big(\varphi^*(x) - \varphi(x)\big) f_{\theta_0}(x),$$

然后对 x 积分, 立得

$$\pi(\theta_1;\varphi^*)-\pi(\theta_1;\varphi)\geq\lambda\big(\pi(\theta_0;\varphi^*)-\pi(\theta_0;\varphi)\big).$$

由此可见, 只要 $\pi(\theta_0;\varphi)\leq\alpha$, 就有 $\pi(\theta_1;\varphi^*)\geq\pi(\theta_1;\varphi)$. □

不得不说, 仅适用于简单假设的 NP 引理实在过于局限, 我们对此稍加推广.

设 $\Theta\subseteq\mathbb{R}^1$, 总体 P_θ 具有密度函数 f_θ, 似然函数记为 $L_n(\theta):=f_\theta(X)$, $\theta\in\Theta$. 称统计模型 $\{P_\theta:\theta\in\Theta\}$ 具有关于 (实数值) 统计量 $V=V(X)$ 的**单调似然比** (monotone likelihood ratio), 若对 $\forall\theta_0,\theta_1\in\Theta:\theta_0\leq\theta_1$, 有

$$\frac{L_n(\theta_1;x)}{L_n(\theta_0;x)}=g_{\theta_0,\theta_1}(V(x)),\quad\forall x,$$

其中 $g_{\theta_0,\theta_1}(\cdot)$ 是单调递增的函数.

定理 (Karlin-Rubin)**.** *设统计模型具有关于充分统计量的单调似然比, 对于 $H_0:\theta\leq\theta_*\leftrightarrow H_1:\theta>\theta_*$, 任给 $\alpha\in(0,1)$, 存在似然比检验是 UMP 水平 α 检验.*

证明. 划归为简单假设, 应用 NP 引理即可, 略. 或许需要用到后面无偏检验的概念. □

例. 一维的指数族具有关于充分统计量的单调似然比. 设 $L_n(\theta;x)=f_\theta(x)=\exp\{Q(\theta)V(x)-B(\theta)\}h(x)$, 我们有 $L_n(\theta_1;x)/L_n(\theta_0;x)=\exp\{\big(Q(\theta_1)-Q(\theta_0)\big)V(x)-\big(B(\theta_1)-B(\theta_0)\big)\}$, 只需 $Q(\cdot)$ 单调递增. ♠

4.6 无偏检验

有时候 UMP 检验未必存在, 我们考虑对检验做出一些限制, 从而缩小考察的范围. 特别地, 本节关注于无偏的检验.

称 $H_0:\theta\in\Theta_0\leftrightarrow H_1:\theta\in\Theta_1$ 的检验函数 $\varphi=\varphi(X)$ 是**无偏的** (unbiased), 若功效函数 $\pi(\theta;\varphi)=\mathbb{E}_\theta[\varphi(X)]$ 满足

$$\pi(\theta;\varphi)\geq\alpha(\varphi),\quad\forall\theta\in\Theta_1,$$

其中 $\alpha(\varphi)=\sup_{\theta\in\Theta_0}\pi(\theta;\varphi)$ 是显著性水平.

例. *对于简单假设, 在 Neyman-Pearson 引理中定义的似然比检验是无偏的.*

证明. 注意 $\Lambda(x)=f_{\theta_1}(x)/f_{\theta_0}(x)$. 利用

$$\begin{aligned}\varphi^*-\alpha&=(1-\alpha)\mathbb{1}_{\{\Lambda>\lambda\}}+\delta\mathbb{1}_{\{\Lambda=\lambda\}}-\alpha\mathbb{1}_{\{\Lambda\leq\lambda\}}\\&\geq\lambda\big((1-\alpha)\mathbb{1}_{\{\Lambda>\lambda\}}+\delta\mathbb{1}_{\{\Lambda=\lambda\}}-\alpha\mathbb{1}_{\{\Lambda\leq\lambda\}}\big)/\Lambda\end{aligned}$$

可以得到

$$\begin{aligned}\mathbb{E}_{\theta_1}[\varphi^*]-\alpha&\geq\lambda\,\mathbb{E}_{\theta_1}\big[\big((1-\alpha)\mathbb{1}_{\{\Lambda>\lambda\}}+\delta\mathbb{1}_{\{\Lambda=\lambda\}}-\alpha\mathbb{1}_{\{\Lambda\leq\lambda\}}\big)/\Lambda\big]\\&=\lambda\,\mathbb{E}_{\theta_0}\big[(1-\alpha)\mathbb{1}_{\{\Lambda>\lambda\}}+\delta\mathbb{1}_{\{\Lambda=\lambda\}}-\alpha\mathbb{1}_{\{\Lambda\leq\lambda\}}\big]\\&=\lambda\,\mathbb{E}_{\theta_0}[\varphi^*-\alpha]\quad=0.\end{aligned}$$

□

定理. *若 φ^* 是 UMP 水平 α 检验, 则 φ^* 无偏.*

证明. 与检验函数 $\varphi(x)\equiv\alpha$ 比较. □

称无偏的检验函数 $\varphi^*=\varphi^*(X)$ 是**一致最大功效无偏的** (uniformly most powerful unbiased, 简记为 UMPU) 水平 α 的检验, 若 $\alpha(\varphi^*)\leq\alpha$, 且对任一水平不超过 α 的无偏检验函数 $\varphi=\varphi(X)$, 均成立

$$\pi(\theta;\varphi^*)\geq\pi(\theta;\varphi),\quad\forall\theta\in\Theta_1.$$

4.7 卡方检验

考虑一列独立重复试验, 每次试验的结果有 $m \geq 2$ 种可能, 第 $j \in [m] = \{1, 2, \cdots, m\}$ 种结果出现的概率为 $p_j \in (0, 1)$. 我们关心 $\mathbf{p} = (p_1, p_2, \ldots, p_m)^\top$, 参数空间为单纯形 (simplex)

$$\Delta^{m-1} = \left\{ \mathbf{p} \in \mathbb{R}^m : p_j > 0,\ \forall j \in [m];\ \&\ \sum_{j=1}^m p_j = 1 \right\}.$$

设 n 次观测得到 $\xi_1, \ldots, \xi_n \overset{\text{i.i.d.}}{\sim} ([m], \mathbf{p})$, 适合 $\mathbb{P}\{\xi_1 = e_j\} = p_j$, 其中 $e_j = (\mathbb{1}_{[k=j]})_{1 \leq k \leq m} \in \mathbb{R}^m$. 将第 j 种试验结果出现的次数记为 $O_j = \sum_{i=1}^n \mathbb{1}_{\{\xi_i = e_j\}}$, 频率记为 $\hat{p}_j = O_j/n$. 显然有 $\sum_{j=1}^m O_j = n$, 即 $\sum_{j=1}^m \hat{p}_j = 1$. 事实上, 统计量 $(O_1, \ldots, O_m)^\top \sim \text{Multinomial}(n, \mathbf{p})$ 具有概率质量函数

$$f(o_1, \ldots, o_m; \mathbf{p}) = \mathbb{P}\{O_j = o_j,\ j \in [m]\} = \binom{n}{o_1, \ldots, o_m} p_1^{o_1} \cdots p_m^{o_m} \mathbb{1}_{[o_1 + \cdots + o_m = n]},$$

其中 $\binom{n}{o_1, \ldots, o_m} = \frac{n!}{o_1! \cdots o_m!}$ 是多项式系数. 对数似然函数为

$$\ell_n(\mathbf{p}) = \log f(O_1, \ldots, O_m; \mathbf{p}) = \log \binom{n}{O_1, \ldots, O_m} + n \sum_{j=1}^m \hat{p}_j \log p_j,$$

从而 $\mathbf{p}$ 的极大似然估计恰为 $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_m)^\top = \frac{1}{n} \sum_{i=1}^n \xi_i$, 这是因为 Shannon 信息不等式保证了

$$\sum_{j=1}^m \hat{p}_j \log p_j < \sum_{j=1}^m \hat{p}_j \log \hat{p}_j, \quad \forall \mathbf{p} \in \Delta^{m-1} \setminus \{\hat{\mathbf{p}}\}.$$

易见 $\mathbb{E}\xi_1 = \sum_{j=1}^m p_j e_j = \mathbf{p}$ 和 $\mathbb{E}\xi_1\xi_1^\top = \sum_{j=1}^m p_j e_j e_j^\top$, 进而

$$\text{Var}(\xi_1) = \mathbb{E}\xi_1\xi_1^\top - \mathbf{p}\mathbf{p}^\top = \left[\sum_{j=1}^m p_j (1 - p_j) e_j e_j^\top \right] - \sum_{j \neq k} p_j p_k e_j e_k^\top.$$

根据 CLT, 我们有 $\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}) \xrightarrow{d} \mathcal{N}_m(0_m, \text{Var}(\xi_1))$.

任给 m_0 维的子参数空间 $\Delta_0 \subset \Delta^{m-1}$, 考察假设检验问题

$$H_0 : \mathbf{p} \in \Delta_0 \leftrightarrow H_1 : \mathbf{p} \notin \Delta_0.$$

设 $\mathbf{p}$ 受 H_0 限制时具有极大似然估计 $\mathbf{p}^* = (p_1^*, \ldots, p_m^*)^\top$, 则似然比统计量

$$\Lambda = f(O_1, \ldots, O_m; \hat{\mathbf{p}}) / f(O_1, \ldots, O_m; \mathbf{p}^*)$$

满足

$$2 \log \Lambda = 2\big(\ell_n(\hat{\mathbf{p}}) - \ell_n(\mathbf{p}^*)\big) = 2n \sum_{j=1}^m \hat{p}_j \log \frac{\hat{p}_j}{p_j^*}.$$

利用二阶 delta 方法可以证明 Wilks 定理的结果, 即 $2 \log \Lambda \xrightarrow[H_0]{d} \chi^2_{m - m_0 - 1}$. 为了便于计算, 我们对上述检验统计量进行一些近似处理. 利用在 $z = z_0$ 附近的 Taylor 展开

$$z \log \frac{z}{z_0} = (z - z_0) + \frac{1}{2z_0}(z - z_0)^2 + \cdots,$$

以及 $\sum_{j=1}^m \hat{p}_j = 1 = \sum_{j=1}^m p_j^*$, 可以得到

$$2 \log \Lambda \approx 2n \sum_{j=1}^m \left\{ (\hat{p}_j - p_j^*) + \frac{1}{2p_j^*}(\hat{p}_j - p_j^*)^2 \right\} = \sum_{j=1}^m \frac{(n\hat{p}_j - np_j^*)^2}{np_j^*}.$$

记期许观测数为 $E_j = np_j^*,\ j \in [m]$, 则

$$\chi^2 = \sum_{j=1}^{m} \frac{(n\hat{p}_j - np_j^*)^2}{np_j^*} = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j}$$

就是著名的 **Pearson 卡方统计量**[xiv]. 可以证明, 当 $n \to \infty$ 时, 有 $\chi^2 \xrightarrow[H_0]{d} \chi^2_{m-m_0-1}$, 进而可以构造一个检验水平渐近 α 的拒绝域

$$\left\{(\xi_1, \dots, \xi_n)^\top : \chi^2 > \chi^2_{\alpha, m-m_0-1}\right\}.$$

下面给出两个常见的应用场景.

例 (拟合优度 (goodness of fit) 检验)**.** 设 $Y_1, \dots, Y_n$ 独立同分布且具有分布函数 F. 任给分布函数 F_0, 考虑假设检验问题

$$H_0 : F = F_0 \leftrightarrow H_1 : F \neq F_0.$$

一种方法是把 Y_1 的值域分划成 $\mathcal{Y}_1 \sqcup \cdots \sqcup \mathcal{Y}_m$, 然后处理 $p_j = \mathbb{P}\{Y_1 \in \mathcal{Y}_j\},\ j \in [m]$. ♠

例 (独立性检验)**.** 考虑样本空间的两个分划 $\Omega = \bigsqcup_{i=1}^r A_i = \bigsqcup_{j=1}^c B_j$, 在 n 次独立重复试验中 $A_i \cap B_j$ 发生的次数记为 n_{ij}, 则 A_i 和 B_j 分别发生 $n_{i\cdot} = \sum_{j=1}^c n_{ij}$ 和 $n_{\cdot j} = \sum_{i=1}^r n_{ij}$ 次. 由此立得列联表 (contingency table):

	$\dots$	B_j	$\dots$	总和
$\vdots$		$\vdots$		$\vdots$
A_i	$\cdots$	n_{ij}	$\cdots$	$n_{i\cdot}$
$\vdots$		$\vdots$		$\vdots$
总和	$\dots$	$n_{\cdot j}$	$\dots$	n

为了探究 $\{A_i\}$ 是否独立于 $\{B_j\}$, 值得关心的是 $p_{ij} = \mathbb{P}(A_i \cap B_j)$. 记 $p_{i\cdot} = \sum_{j=1}^c p_{ij}$ 和 $p_{\cdot j} = \sum_{i=1}^r p_{ij}$, 则需要检验

$$H_0 : p_{ij} = p_{i\cdot} p_{\cdot j},\ \forall i, j.$$

可以证明 (☕ 留给读者), 导出的 Pearson 卡方统计量为

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - n\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n})^2}{n\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n}},$$

且当 $n \to \infty$ 时, 有 $\chi^2 \xrightarrow[H_0]{d} \chi^2_{(r-1)(c-1)}$. ♠

作为补充, 附录 §A.4介绍了两种基于极大似然和二次型的检验.

4.8 非参数检验

一般来说, 我们无法确知总体的形式, 不存在一个有限维的参数完全刻画分布, 此时统计模型称为**非参数的** (nonparametric); 当然, 可以施加一些最低限度的限制, 比如光滑性和可积性. 本节考虑连续型随机变量 $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P = \mathbb{P}\{X_1 \in \bullet\}$, 记分布函数为

$$F(x) := \mathbb{P}\{X_1 \leq x\} = P((-\infty, x]) = \int \mathbb{1}_{(-\infty, x]}\, \mathrm{d}P, \quad x \in \mathbb{R}.$$

任给一个分布函数 F_0 或者一个分布族 $\mathcal{P}_0$, 我们想探究 $H_0 : F = F_0$ 或者 $H_0 : P \in \mathcal{P}_0$.

[xiv] 关于 Pearson 残差, 参看https://stats.stackexchange.com/q/45479

在 §2.4中我们知道 $F(X_1),\dots,F(X_n) \overset{\text{i.i.d.}}{\sim} \text{Uniform}([0,1])$, 那么自然地, 一个检验 $H_0: F = F_0$ 的思路是将 $F_0(X_1),\dots,F_0(X_n)$ 与 $U_1,\dots,U_n \overset{\text{i.i.d.}}{\sim} \text{Uniform}([0,1])$ 进行比较. 回忆一下, 由 $X_1,\dots,X_n$ 从小到大排序得到的 $X_{(1)},\dots,X_{(n)}$ 称为**顺序统计量** (order statistics), 给出了样本的一些描述 (见 §1.2); 利用分布函数的单调性, 可见 $\{F(X_{(i)})\}_{1\le i\le n}$ 是 $\{F(X_i)\}_{1\le i\le n}$ 的顺序统计量. 我们来探究一下 $(F(X_1),\dots,F(X_n)) \overset{d}{=} (U_{(1)},\dots,U_{(n)})$ 的性质: 其联合概率密度函数为

$$f_{1:n}(u_{(1)},\dots,u_{(n)}) = n!\mathbb{1}_{[0\le u_{(1)}\le\cdots\le u_{(n)}\le 1]}.$$

对于任一 $i\in\{1,\cdots,n\}$, 可得 $U_{(i)}$ 的概率密度函数

$$\begin{aligned} f_{i,n}(u_i) &= \int_{\mathbb{R}^{n-1}} f_{1:n}(u_1,\dots,u_i,\dots,u_n)\,\mathrm{d}u_1\dots\mathrm{d}u_{i-1}\mathrm{d}u_{i+1}\dots\mathrm{d}u_n \\ &= n!\left(\int_0^{u_i}\mathrm{d}u_{i-1}\int_0^{u_{i-1}}\mathrm{d}u_{i-2}\cdots\int_0^{u_2}\mathrm{d}u_1\right)\mathbb{1}_{[0,1]}(u_i)\left(\int_{u_i}^1\mathrm{d}u_{i+1}\int_{u_{i+1}}^1\mathrm{d}u_{i+2}\cdots\int_{u_{n-1}}^1\mathrm{d}u_n\right) \\ &= n!\frac{u_i^{i-1}}{(i-1)!}\mathbb{1}_{[0,1]}(u_i)\frac{(1-u_i)^{n-i}}{(n-i)!} \quad = \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i+1)}u_i^{i-1}(1-u_i)^{n-i}\mathbb{1}_{[0,1]}(u_i), \end{aligned}$$

即

$$U_{(i)} \sim \text{Beta}(i, n-i+1),$$

从而 $\mathbb{E}U_{(i)} = i/(n+1)$. 我们可以绘制如下两种散点图.

- **PP 图** (P 表示概率 probability): 对 $i=1,2,\cdots,n$ 画出 $(\frac{i}{n+1}, F_0(X_{(i)}))$.
- **QQ 图** (Q 表示分位数 quantile): 对 $i=1,2,\cdots,n$ 画出 $(F_0^{-1}(\frac{i}{n+1}), X_{(i)})$.

通过与直线 $x\mapsto x$ 对比, 可以得到 $H_0: F = F_0$ 的直观判断. 注意到 $\sigma X_1+\mu$ 的分布函数为 $F(\frac{\bullet-\mu}{\sigma})$, 其中 $\mu\in\mathbb{R}$ 称作位置参数, $\sigma\in(0,\infty)$ 称作尺度参数. 判断 F 是否属于**位置尺度族** (location-scale family) $\mathcal{F}_0 := \{F_0(\frac{\bullet-\mu}{\sigma}) : \mu\in\mathbb{R}, \sigma\in(0,\infty)\}$ 时, 可以考察上述 QQ 图是否接近线性.

顺序统计量可以视为经验分布 P_n 的分位数, 其中随机测度 P_n 是 X_i $(i=1,2,\cdots,n)$ 处 Dirac 点测度的算术平均, 对应**经验分布函数** (empirical distribution function)

$$F_n(x) := P_n((-\infty,x]) = \int \mathbb{1}_{(-\infty,x]}\,\mathrm{d}P_n = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(-\infty,x]}(X_i) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i\le x\}}, \quad x\in\mathbb{R}.$$

Kolmogorov-Smirnov 检验采用了统计量

$$D_n := \sup_{x\in\mathbb{R}}|F_n(x)-F_0(x)| = \max_{1\le i\le n}\left\{\max\left(\tfrac{i}{n}-F_0(X_{(i)}), F_0(X_{(i)})-\tfrac{i-1}{n}\right)\right\}.$$

定理 (Glivenko-Cantelli)**.** *当 $H_0: F = F_0$ 为真时, 成立 $D_n \xrightarrow{\text{a.s.}} 0$.*

证明. 读者有兴趣的话, 可参看https://math.stackexchange.com/q/1672177. □

正态性检验 (normality test)[xv]是统计学中很重要的一个话题, 不过有人吐槽[xvi]说: "检验总体的分布在我看来几乎没有什么用处,... 一、若拒绝零假设, 即数据不服从某种分布, 那么往往会使得下面要做的工作的前提假设不成立——这显然会很惨; 二、若不拒绝零假设——这几乎是无用的结论, 因为不拒绝这个零假设, 不代表能拒绝其它零假设, 因此你仍然不知道数据是什么分布——这显然更惨".

在经典的非参数统计中, **秩统计量** (rank statistics) $R_i = \min\{k : X_{(k)} = X_i\} = 1+\sum_{j=1}^n \mathbb{1}_{\{X_j<X_i\}}$ 也很常见, 其分布不依赖 F, 可用来检验 F 的对称性之类问题. ♈

[xv] 参看 B. W. Yap, & C. H. Sim (2011). "Comparisons of various types of normality tests". *Journal of Statistical Computation and Simulation*. **81**(12): 2141-2155. https://doi.org/10.1080/00949655.2010.520163.

[xvi] 原文可见于https://yihui.org/cn/2009/02/test-statistical-distributions/

5 线性模型

为了刻画变量之间的关系, 最常见也最基础的是考虑**线性模型** (linear models). 将第 $i \in \{1, \cdots, n\}$ 个观测结果记为 $(Y_i, \boldsymbol{X}_i)$, 其中随机变量 Y_i 是响应变量 (response), 随机向量 $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$ 是解释变量 (explanatory variable, 也称为协变量 (covariate)). 设

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \varepsilon_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, 2, \cdots, n,$$

其中 $\boldsymbol{\beta} = (\beta_j)_{1 \le j \le p} \in \mathbb{R}^p$ 是未知参数, ε_i 是误差 (error)/噪声 (noise); 对分布给出一些假定, 就能指明一个统计模型. 我们往往将 i 看作独立的个体, 于是观测结果应该两两不相关. 通过一些矩条件, 可以论证估计的优良性质; 进而考虑正态总体的话, 我们将得以做出一些更深入的推断.

我们可以把线性模型写成更紧凑的形式:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

其中

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

线性代数让许多推导变得更加简洁, 也便于我们抓住一些本质.

5.1 回归分析

回归 (regression) 在统计学中已经是一个通用的术语, 稍微背离了其本意[xvii]. 回归分析的主要目标是研究变量之间的相关性, 希望明确形如 $\boldsymbol{Y} \approx f(\boldsymbol{X})$ 的函数关系. 当然, 线性关系是最简单的一种, 却也有着成熟优美的理论和相当强大的实用性. 通过一些变换, 很多模型都能体现出线性关系.

例. 在 Cobb-Douglas 生产函数模型中, 产量 Q 与资本 K 和劳动量 L 之间满足 $Q = AK^\gamma L^\delta$, 其中 A 表示技术水平, γ 和 δ 表示弹性. 通过对数变换, 有 $\log Q = \log A + \gamma \log K + \delta \log L$. ♠

我们常常可以要求 $\boldsymbol{X}_i$ 的第一项为 $X_{i1} \equiv 1$, 由此可以通过参数 β_1 纳入常数的影响, 反映整体的平均值. 实践中, $\boldsymbol{X}$ 常常可以人为控制或设计, 或者至少比较容易观测到实现值, 所以不失一般性地, 我们可以认为 $\boldsymbol{X}$ 取常值; 为了不放弃 $\boldsymbol{X}$ 的随机性, 后面的推导也可以看作条件于 $\boldsymbol{X}$ 之后所进行的.

5.2 估计: 最小二乘法

记欧氏范数为 $\|\bullet\| := \sqrt{\bullet'\bullet}$. 回归系数 $\boldsymbol{\beta}$ 的**普通最小二乘** (ordinary least squares, 简记为 OLS) **估计量**为 $\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$, 适合**正规方程** (normal equation)

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y},$$

这通过一阶条件 $\partial\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2/\partial\boldsymbol{\beta}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \boldsymbol{0}_p$ 或者投影性质 (见附录 §A.5) 可以看出. 事实上,

$$\hat{\boldsymbol{Y}} := \boldsymbol{X}\hat{\boldsymbol{\beta}} = \mathrm{proj}_{\mathrm{Col}(\boldsymbol{X})}\boldsymbol{Y}$$

存在且唯一, 称为 $\boldsymbol{Y}$ 的**拟合值** (fitted values); 由此来确定 $\hat{\boldsymbol{\beta}}$ 则需要额外的条件. 为得到良定的 $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$, 其中 $\boldsymbol{c} \in \mathbb{R}^p$, 应有 $\boldsymbol{c} = \boldsymbol{X}'\boldsymbol{l} \in \mathrm{Col}(\boldsymbol{X}')$ 对某 $\boldsymbol{l} \in \mathbb{R}^n$ 成立, 此时称 $\boldsymbol{c}'\boldsymbol{\beta}$ 是线性可估的 (linearly estimable).

[xvii] 参看https://stats.stackexchange.com/a/11089

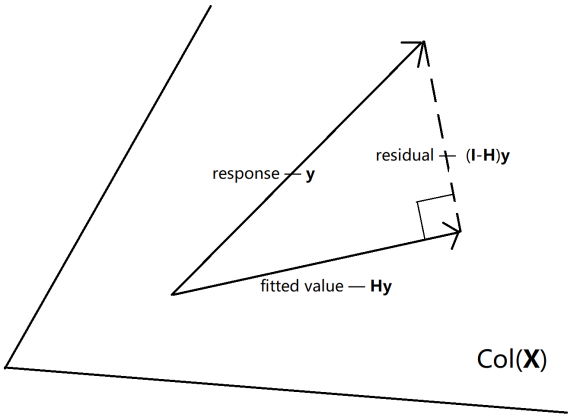
假定①$\boldsymbol{X}$ 列满秩, 即 $\operatorname{rank}(\boldsymbol{X})=p$, 从而 $\boldsymbol{b}\in\mathbb{R}^p\mapsto\boldsymbol{X}\boldsymbol{b}\in\mathbb{R}^n$ 是单射. 此时 $\boldsymbol{X}'\boldsymbol{X}\in\mathbb{R}^{p\times p}$ 可逆, 我们能够唯一地得到

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

记投影矩阵 (对称幂等矩阵, 见 §A.5)

$$\boldsymbol{P}_{\boldsymbol{X}}=\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}',$$

则 $\hat{\boldsymbol{Y}}=\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y}$; 有时我们也将 $\boldsymbol{P}_{\boldsymbol{X}}$ 记为 $\boldsymbol{H}$, 称作**帽子矩阵** (hat matrix), 因为 $\boldsymbol{H}$ "给 $\boldsymbol{Y}$ 戴上了帽子".



残差 (residual) 定义为

$$\hat{\boldsymbol{\varepsilon}}:=\boldsymbol{Y}-\hat{\boldsymbol{Y}}=(\boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y},$$

它落入 $\operatorname{Col}(\boldsymbol{X})^\perp$ 中, 即 $\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}}=\boldsymbol{0}_p$. 于是, 我们有

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{Y}}.$$

假定②$\mathbb{E}[\boldsymbol{\varepsilon}]=\boldsymbol{0}_n$ 和③$\operatorname{Var}(\boldsymbol{\varepsilon})=\sigma^2\boldsymbol{I}_n$, 其中 $\sigma^2>0$ 是未知参数. 由 $\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, 易知 $\mathbb{E}[\boldsymbol{Y}]=\boldsymbol{X}\boldsymbol{\beta}$ 和 $\operatorname{Var}(\boldsymbol{Y})=\sigma^2\boldsymbol{I}_n$. 直接计算可得

$$\mathbb{E}[\hat{\boldsymbol{\beta}}]=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}[\boldsymbol{Y}]=\boldsymbol{\beta},$$

以及

$$\operatorname{Var}(\hat{\boldsymbol{\beta}})=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\operatorname{Var}(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

对于拟合值 $\hat{\boldsymbol{Y}}$, 有

$$\mathbb{E}[\hat{\boldsymbol{Y}}]=\boldsymbol{P}_{\boldsymbol{X}}\mathbb{E}[\boldsymbol{Y}]=\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X}\boldsymbol{\beta}=\boldsymbol{X}\boldsymbol{\beta},$$

和

$$\operatorname{Var}(\hat{\boldsymbol{Y}})=\boldsymbol{P}_{\boldsymbol{X}}\operatorname{Var}(\boldsymbol{Y})\boldsymbol{P}_{\boldsymbol{X}}'=\boldsymbol{P}_{\boldsymbol{X}}(\sigma^2\boldsymbol{I}_n)\boldsymbol{P}_{\boldsymbol{X}}=\sigma^2\boldsymbol{P}_{\boldsymbol{X}}.$$

定理 (Gauss-Markov)**.** *任取 $\boldsymbol{c}\in\mathbb{R}^p$, 在 $\boldsymbol{c}'\boldsymbol{\beta}$ 的所有形如 $\boldsymbol{l}'\boldsymbol{Y}$ ($\boldsymbol{l}\in\mathbb{R}^n$) 的无偏估计量中, $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ 具有最小的方差. 换言之, OLS 估计量是*__最优线性无偏估计量__ *(best linear unbiased estimator, 简记为 BLUE).*

证明. 设 $\boldsymbol{l}'\boldsymbol{Y}$ 是 $\boldsymbol{c}'\boldsymbol{\beta}$ 的无偏估计量, 其中 $\boldsymbol{l}\in\mathbb{R}^n$. 由无偏性定义可得 $\boldsymbol{l}'\boldsymbol{X}=\boldsymbol{c}'$, 从而 $\boldsymbol{c}'\hat{\boldsymbol{\beta}}=\boldsymbol{l}'\hat{\boldsymbol{Y}}$. 于是,

$$\operatorname{Var}(\boldsymbol{l}'\boldsymbol{Y})-\operatorname{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}})=\boldsymbol{l}'[\operatorname{Var}(\boldsymbol{Y})-\operatorname{Var}(\hat{\boldsymbol{Y}})]\boldsymbol{l}=\sigma^2\boldsymbol{l}'(\boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{l}\geq 0,$$

其中 $\boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{X}}=\operatorname{proj}_{\operatorname{Col}(\boldsymbol{X})^\perp}$ 作为投影矩阵是半正定的. □

对于残差 $\hat{\varepsilon}$, 有

$$\mathbb{E}[\hat{\varepsilon}] = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}_n, \quad \mathrm{Var}(\hat{\varepsilon}) = \mathbb{E}[\hat{\varepsilon}\hat{\varepsilon}'] = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})(\sigma^2\boldsymbol{I}_n)(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})' = \sigma^2(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}),$$

以及

$$\mathrm{Cov}(\hat{\varepsilon}, \hat{\boldsymbol{Y}}) = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})(\sigma^2\boldsymbol{I}_n)\boldsymbol{P}_{\boldsymbol{X}}' = \sigma^2(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{0}_{n\times n}.$$

残差平方和 (residual sum of squares)

$$\mathrm{RSS} := \|\hat{\varepsilon}\|^2 = \hat{\varepsilon}'\hat{\varepsilon} = \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}$$

满足

$$\mathbb{E}[\mathrm{RSS}] = \mathbb{E}\,\mathrm{tr}\,(\hat{\varepsilon}\hat{\varepsilon}') = \mathrm{tr}\,(\mathbb{E}[\hat{\varepsilon}\hat{\varepsilon}']) = \mathrm{tr}\,(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\,\sigma^2 = (n-p)\sigma^2.$$

由此, 我们得到了 σ^2 的一个无偏估计量

$$s^2 := \mathrm{RSS}/(n-p).$$

5.3 推断: F 检验

事实上, 本节可以视为三大经典假设检验 (§A.4) 的特例.

为了推断, 进一步假定[4]误差服从正态分布, 即 $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}_n, \sigma^2\boldsymbol{I}_n)$. 此时 $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$, 不难看出 OLS 估计 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ 恰为 $\boldsymbol{\beta} \in \mathbb{R}^p$ 的极大似然估计, 这不依赖 σ^2; 通过一些计算, 可得 σ^2 的极大似然估计为 $\hat{\sigma}^2 = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2/n = (n-p)s^2/n$, 不具有无偏性. 考虑一般的线性假设

$$H_0 : \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r},$$

其中 $\boldsymbol{r} \in \mathbb{R}^q$, 且 $\boldsymbol{R} \in \mathbb{R}^{q\times p}$ 行满秩, 即 $\mathrm{rank}\,(\boldsymbol{R}) = q$.

- 似然比检验. 令 $V_0 := \{\boldsymbol{X}\boldsymbol{b} : \boldsymbol{b} \in \mathbb{R}^p \text{ s.t. } \boldsymbol{R}\boldsymbol{b} = \boldsymbol{r}\}$, 易见 V_0 是 p 维空间 $\mathrm{Col}(\boldsymbol{X}) \subset \mathbb{R}^n$ 的 $(p-q)$ 维子空间. 通过简单的推导, 可得

  $$(\boldsymbol{\mu}, \sigma^2) \in V_0 \times (0,\infty) \mapsto \log f(\boldsymbol{Y}|\boldsymbol{\mu}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{\|\boldsymbol{Y} - \boldsymbol{\mu}\|^2}{2\sigma^2}$$

  在 ✍ 注: 关于 $\boldsymbol{\mu}$ 的估计其实就是最小二乘法.

  $$(\tilde{\boldsymbol{\mu}}, \tilde{\sigma}^2) := \left(\mathrm{proj}_{V_0}\boldsymbol{Y}, \|\boldsymbol{Y} - \tilde{\boldsymbol{\mu}}\|^2/n\right)$$

  处取到最大值. 于是 Wilks 检验统计量为

  $$2\left(\log f(\boldsymbol{Y}|\hat{\boldsymbol{Y}}, \hat{\sigma}^2) - \log f(\boldsymbol{Y}|\tilde{\boldsymbol{\mu}}, \tilde{\sigma}^2)\right) = -n\log\frac{\hat{\sigma}^2}{\tilde{\sigma}^2} = n\log\frac{\|\boldsymbol{Y} - \tilde{\boldsymbol{\mu}}\|^2}{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2}.$$

  利用正交分解

  $$\begin{matrix}\boldsymbol{I}_n - \mathrm{proj}_{V_0} \\ \text{rank: } n-(p-q)\end{matrix} = \underset{n-p}{(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})} + \underset{q}{(\boldsymbol{P}_{\boldsymbol{X}} - \mathrm{proj}_{V_0})},$$

  可以得到

  $$\|\boldsymbol{Y} - \tilde{\boldsymbol{\mu}}\|^2 = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 + \|\hat{\boldsymbol{Y}} - \tilde{\boldsymbol{\mu}}\|^2.$$

  于是, 一个等价的检验统计量为

  $$F := \frac{\|\hat{\boldsymbol{Y}} - \tilde{\boldsymbol{\mu}}\|^2/q}{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2/(n-p)},$$

  其中分子和分母用自由度 (degree of freedom) 进行了修正.

引理. 设 $\boldsymbol{u}_1,\dots,\boldsymbol{u}_r\in\mathbb{R}^n$ 构成规范正交系, 若 $\boldsymbol{\xi}\sim\mathcal{N}_n(\boldsymbol{0}_n,\boldsymbol{I}_n)$, 则 $\boldsymbol{u}_1'\boldsymbol{\xi},\dots,\boldsymbol{u}_r'\boldsymbol{\xi}\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,1)$. 进一步地, 有 $\|\sum_{j=1}^r\boldsymbol{u}_j\boldsymbol{u}_j'\boldsymbol{\xi}\|^2=\sum_{j=1}^r(\boldsymbol{u}_j'\boldsymbol{\xi})^2\sim\chi_r^2$. □

当 H_0 为真时, 有 $\boldsymbol{X}\boldsymbol{\beta}\in V_0$. 利用 $\boldsymbol{Y}\sim\mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n)$ 可得

$$\begin{pmatrix}\hat{\boldsymbol{Y}}-\tilde{\boldsymbol{\mu}}\\ \boldsymbol{Y}-\hat{\boldsymbol{Y}}\end{pmatrix}=\begin{pmatrix}\boldsymbol{P}_{\boldsymbol{X}}-\mathrm{proj}_{V_0}\\ \boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{X}}\end{pmatrix}\boldsymbol{Y}\overset{H_0}{\sim}\mathcal{N}_{2n}\left(\boldsymbol{0}_{2n},\sigma^2\begin{pmatrix}\boldsymbol{P}_{\boldsymbol{X}}-\mathrm{proj}_{V_0} & \boldsymbol{0}_{n\times n}\\ \boldsymbol{0}_{n\times n} & \boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{X}}\end{pmatrix}\right),$$

从而

$$\|\hat{\boldsymbol{Y}}-\tilde{\boldsymbol{\mu}}\|^2\overset{H_0}{\sim}\sigma^2\chi_q^2\ \perp\!\!\!\perp\ \|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\|^2\sim\sigma^2\chi_{n-p}^2,$$

于是

$$F=\frac{\|\hat{\boldsymbol{Y}}-\tilde{\boldsymbol{\mu}}\|^2/q}{\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\|^2/(n-p)}\overset{H_0}{\sim}F_{q,n-p}.$$

由此可以确定临界值 $c\in\mathbb{R}_+$, 当 $F>c$ 时我们选择拒绝 $H_0:\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{r}$.

- *Wald* 检验. 由于 $\hat{\boldsymbol{Y}}=\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y}$ 和 $\hat{\boldsymbol{\varepsilon}}=(\boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}$ 服从联合正态分布, 利用不相关性可以得到二者的独立性, 进而

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{Y}}\sim\mathcal{N}_p(\boldsymbol{\beta},\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})\ \perp\!\!\!\perp\ (n-p)s^2=\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}\sim\sigma^2\chi_{n-p}^2.$$

当 $H_0:\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{r}$ 为真时, 有 $\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\overset{H_0}{\sim}\mathcal{N}_q(\boldsymbol{0}_q,\sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')$, 进而

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})\overset{H_0}{\sim}\sigma^2\chi_q^2\ \perp\!\!\!\perp\ (n-p)s^2\sim\sigma^2\chi_{n-p}^2,$$

于是

$$W:=\frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})/q}{s^2}\overset{H_0}{\sim}F_{q,n-p}.$$

事实上, $W=F$. 为此, 我们将说明 $\|\hat{\boldsymbol{Y}}-\tilde{\boldsymbol{\mu}}\|^2=(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})$.

- *Lagrange* 乘子法. 在 H_0 限制下, 考虑求解 $\tilde{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2$. 我们引入

$$(\boldsymbol{\beta},\boldsymbol{\lambda})\mapsto\tfrac{1}{2}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})+\boldsymbol{\lambda}'(\boldsymbol{r}-\boldsymbol{R}\boldsymbol{\beta}),$$

分别对 $\boldsymbol{\beta}$ 和 $\boldsymbol{\lambda}$ 求导得到一阶条件

$$\boldsymbol{X}'\boldsymbol{X}\tilde{\boldsymbol{\beta}}-\boldsymbol{X}'\boldsymbol{Y}-\boldsymbol{R}'\tilde{\boldsymbol{\lambda}}=\boldsymbol{0}_p,\quad\&\quad\boldsymbol{r}-\boldsymbol{R}\tilde{\boldsymbol{\beta}}=\boldsymbol{0}_q.$$

直接计算可得 $\tilde{\boldsymbol{\lambda}}=(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{r}-\boldsymbol{R}\hat{\boldsymbol{\beta}})$, 进而

$$\tilde{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}),$$

于是 $\hat{\boldsymbol{Y}}-\tilde{\boldsymbol{\mu}}=\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}\tilde{\boldsymbol{\beta}}=\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})$ 满足

$$(\hat{\boldsymbol{Y}}-\tilde{\boldsymbol{\mu}})'(\hat{\boldsymbol{Y}}-\tilde{\boldsymbol{\mu}})=(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}).$$

♦ t 检验. 特别地, 当 $q=1$ 时, 由

$$\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}\overset{H_0}{\sim}\mathcal{N}(0,\sigma^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')\ \perp\!\!\!\perp\ (n-p)s^2\sim\sigma^2\chi_{n-p}^2$$

构造的

$$T:=(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})/\sqrt{s^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'}\overset{H_0}{\sim}t_{n-p}$$

能够更灵活地处理单边检验.

✍ 借助上述检验, 相应地可以得到 $\boldsymbol{R}\boldsymbol{\beta}$ 的置信集 (见 §4.3).

5.4 诊断: 检查假定

在线性模型的推导中, 我们假定了一些条件. 在实际应用中, 为了说明线性模型的合理性, 需要对其进行检查 (checking) 和诊断 (diagnostics).

⓪ $\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$? 为了让线性关系比较明确, 我们会剔除离群值 (outlier). 有时, 我们会对响应变量做一些变换, 如 **Box-Cox 变换** $y\mapsto\begin{cases}(y^\lambda-1)/\lambda, & \lambda>0\\ \log y, & \lambda=0\end{cases}$, 其中参数 λ 通过一些准则来适当选取; 不过, 这也可能带来其他问题.

① $\operatorname{rank}(\boldsymbol{X})=p$? 如果 $\boldsymbol{X}$ 不是列满秩的, 我们就遇到了**多重共线性** (multicollinearity) 的问题. 一般地, 我们可以去掉一些 $\boldsymbol{X}$ 的列, 或者将一些列组合起来, 这也是一种降维 (dimension reduction).

② $\mathbb{E}[\boldsymbol{\varepsilon}|\boldsymbol{X}]=\boldsymbol{0}_n$? 如果 $\boldsymbol{\varepsilon}$ 与 $\boldsymbol{X}$ 存在较强的相关性, 可以考虑剔除一些解释变量. 在计量经济学中, 一个有趣的想法是借助工具变量 (instrumental variable).

③ $\operatorname{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X})=\sigma^2\boldsymbol{I}_n$? 做变换 $\boldsymbol{Y}\mapsto\operatorname{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X})^{-1/2}\boldsymbol{Y}$ 是一种解决方法, 其中 $\operatorname{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X})$ 需要估计. 检查**异方差** (heteroskedasticity) 时, 绘制 (相对于拟合值的) **残差图** (residual plot) 能给出直观; 有时会考虑标准化的残差, 即 $\hat{\varepsilon}_i/\sqrt{s^2(1-h_{ii})}$, 其中 h_{ii} 是 $\boldsymbol{P}_{\boldsymbol{X}}$ 的第 i 个主对角元.

④ $\boldsymbol{\varepsilon}|\boldsymbol{X}$ 正态? 一般来说, 我们相信中心极限定理成立. 关于正态性检验, 在 §4.8已有介绍.

为了模型的优良性, 还有许多其他操作, 比如常常用 AIC 和 BIC 等信息准则 (information criteria) 进行变量选择[xviii] (variable selection). 实践出真知!

5.5 方差分析

为了比较不同处理 (treatment) 的效果, 我们常常考虑**方差分析** (analysis of variance, ANOVA). 单因素方差分析 (one-way ANOVA) 模型为

$$\underset{\text{(observation)}}{Y_{\ell r}}=\underset{\text{(overall mean)}}{\mu}+\underset{\text{(the }\ell_{\text{th}}\text{ treatment effect)}}{\tau_\ell}+\underset{\text{(error)}}{\varepsilon_{\ell r}}\quad(r=1,\cdots,n_\ell;\ \ell=1,\cdots,g)$$

表示第 ℓ 种处理在第 r 次重复时观测到的效果, 其中 $\sum_{\ell=1}^g n_\ell\tau_\ell=0$, 且 $\varepsilon_{\ell r}\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,\sigma^2)$. 易见

$$\begin{pmatrix}Y_{11}\\ \vdots\\ Y_{1n_1}\\ \vdots\\ Y_{g1}\\ \vdots\\ Y_{gn_g}\end{pmatrix}_{n\times1}=\begin{pmatrix}1_{n_1} & 1_{n_1} & & \\ 1_{n_2} & & 1_{n_2} & \\ \vdots & & & \ddots \\ 1_{n_g} & & & 1_{n_g}\end{pmatrix}_{n\times(g+1)}\begin{pmatrix}\mu\\ \tau_1\\ \vdots\\ \tau_g\end{pmatrix}_{(g+1)\times1}+\begin{pmatrix}\varepsilon_{11}\\ \vdots\\ \varepsilon_{1n_1}\\ \vdots\\ \varepsilon_{g1}\\ \vdots\\ \varepsilon_{gn_g}\end{pmatrix}_{n\times1},$$

其中 $n=\sum_{\ell=1}^g n_\ell$ 是总观测次数. ✍ 注: 我们关心的 $\tau_1-\tau_g$ 之类的参数[xix]常常可估.

[xviii] 参看https://cosx.org/2015/08/some-basic-ideas-and-methods-of-model-selection/

[xix] 参看https://stats.stackexchange.com/a/221861

记

$$Y_{\ell r} = \underbrace{\bar{Y}_{\cdot\cdot}}_{=\hat{\mu}} + (\underbrace{\bar{Y}_{\ell\cdot} - \bar{Y}_{\cdot\cdot}}_{=\hat{\tau}_\ell}) + (\underbrace{Y_{\ell r} - \bar{Y}_{\ell\cdot}}_{=\hat{\varepsilon}_{\ell r}}),$$

其中 $\bar{Y}_{\ell\cdot} = \frac{1}{n_\ell}\sum_{r=1}^{n_\ell} Y_{\ell r}$, 且 $\bar{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{\ell=1}^{g}\sum_{r=1}^{n_\ell} Y_{\ell r}$. 我们有**平方和分解**

$$\sum_{\ell=1}^{g}\sum_{r=1}^{n_\ell}(Y_{\ell r} - \bar{Y}_{\cdot\cdot})^2 = \sum_{\ell=1}^{g} n_\ell(\bar{Y}_{\ell\cdot} - \bar{Y}_{\cdot\cdot})^2 + \sum_{\ell=1}^{g}\sum_{r=1}^{n_\ell}(Y_{\ell r} - \bar{Y}_{\ell\cdot})^2,$$

这是因为交叉项相加得到

$$\sum_{\ell=1}^{g}\underbrace{\sum_{r=1}^{n_\ell}(Y_{\ell r} - \bar{Y}_{\ell\cdot})}_{=0}(\bar{Y}_{\ell\cdot} - \bar{Y}_{\cdot\cdot}) = 0.$$

来源	平方和	自由度
组间 (between-group)	$B = \sum_{\ell=1}^{g} n_\ell(\bar{Y}_{\ell\cdot} - \bar{Y}_{\cdot\cdot})^2$	$g-1$
组内 (within-group)	$W = \sum_{\ell=1}^{g}\sum_{r=1}^{n_\ell}(Y_{\ell r} - \bar{Y}_{\ell\cdot})^2$	$n-g$
总和	$B+W = \sum_{\ell=1}^{g}\sum_{r=1}^{n_\ell}(Y_{\ell r} - \bar{Y}_{\cdot\cdot})^2$	$n-1$

检验 $H_0 : \tau_1 = \cdots = \tau_g = 0$ 的 F 统计量为

$$\frac{B/(g-1)}{W/(n-g)} \overset{H_0}{\sim} F_{g-1,n-g}.$$

多因素 ANOVA 模型可以类似推导. 作为例子, 我们考虑*两因素均衡方差分析* (two-way balanced ANOVA) 模型 (思考: 如何写成矩阵形式?)

$$Y_{ijr} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijr}, \quad i = 1,\cdots,I;\ j = 1,\cdots,J;\ r = 1,\cdots,R,$$

其中 μ 反映整体的*平均*, α_i 表示第 i 种水平因素 A 的*主效应*, β_j 表示第 j 种水平因素 B 的*主效应*, γ_{ij} 表示第 i 种水平因素 A 和第 j 种水平因素 B 的*交互效应*, ε_{ijr} 表示随机误差. 不失一般性, 我们假定 $\sum_{i=1}^{I}\alpha_i = \sum_{j=1}^{J}\beta_j = \sum_{i=1}^{I}\gamma_{ij} = \sum_{j=1}^{J}\gamma_{ij} = 0$, 且 $\varepsilon_{ijr} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2)$. 记

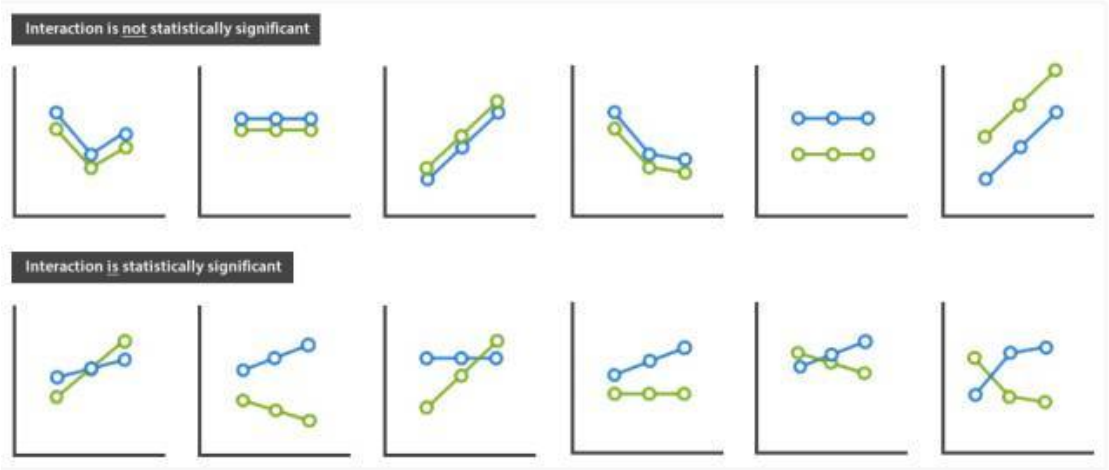
$$Y_{ijr} = \underbrace{\bar{Y}_{\cdots}}_{=\hat{\mu}} + (\underbrace{\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots}}_{=\hat{\alpha}_i}) + (\underbrace{\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdots}}_{=\hat{\beta}_j}) + (\underbrace{\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots}}_{=\hat{\gamma}_{ij}}) + (\underbrace{Y_{ijr} - \bar{Y}_{ij\cdot}}_{=\hat{\varepsilon}_{ijr}}),$$

其中 $\cdot$ 表示对相应的指标求平均. 直接计算可得如下**平方和分解**

$$\begin{aligned}
\sum_{i,j,r}(Y_{ijr} - \hat{\mu})^2 &= \sum_{i,j,r}\left(\hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij} + \hat{\varepsilon}_{ijr}\right)^2 \\
&= \sum_{i,j,r}\left[\hat{\alpha}_i^2 + \hat{\beta}_j^2 + \hat{\gamma}_{ij}^2 + \hat{\varepsilon}_{ijr}^2 + 2\hat{\alpha}_i\hat{\beta}_j + 2\hat{\alpha}_i\hat{\gamma}_{ij} + 2\hat{\beta}_j\hat{\gamma}_{ij} + 2\left(\hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij}\right)\hat{\varepsilon}_{ijr}\right] \\
&= JR\sum_i \hat{\alpha}_i^2 + IR\sum_j \hat{\beta}_j^2 + R\sum_{i,j}\hat{\gamma}_{ij}^2 + \sum_{i,j,r}\hat{\varepsilon}_{ijr}^2 + \\
&\quad + 2R\underbrace{\sum_i \hat{\alpha}_i}_{=0}\underbrace{\sum_j \hat{\beta}_j}_{=0} + 2R\sum_i \hat{\alpha}_i\underbrace{\sum_j \hat{\gamma}_{ij}}_{=0} + 2R\sum_j \hat{\beta}_j\underbrace{\sum_i \hat{\gamma}_{ij}}_{=0} + 2\sum_{i,j}\left(\hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij}\right)\underbrace{\sum_r \hat{\varepsilon}_{ijr}}_{=0}
\end{aligned}$$

来源	平方和	自由度
因素 A	$\mathrm{SS}_{\mathrm{A}} = JR\sum_i(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots})^2$	$I-1$
因素 B	$\mathrm{SS}_{\mathrm{B}} = IR\sum_j(\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdots})^2$	$J-1$
交互	$\mathrm{SS}_{\mathrm{A}\times\mathrm{B}} = R\sum_{i,j}(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots})^2$	$(I-1)(J-1)$
误差	$\mathrm{SS}_{\mathrm{E}} = \sum_{i,j,r}(Y_{ijr} - \bar{Y}_{ij\cdot})^2$	$IJ(R-1)$
总和	$\mathrm{SS}_{\mathrm{T}} = \sum_{i,j,r}(Y_{ijr} - \bar{Y}_{\cdots})^2$	$IJR-1$

如果存在显著的交互效应, 那么因素的主效应将难以清晰地诠释, 这会在轮廓图 (profile plot) 中表现出来[xx].



首先, 考虑检验 $H_0 : \gamma_{ij} = 0,\ \forall i, j$, 相应的检验统计量为

$$\frac{\mathrm{SS}_{\mathrm{A}\times\mathrm{B}}/[(I-1)(J-1)]}{\mathrm{SS}_{\mathrm{E}}/[IJ(R-1)]} \overset{H_0}{\sim} F_{(I-1)(J-1),IJ(R-1)}.$$

如果上述 F 统计量比较大, 我们将拒绝 H_0; 否则继续. 此时, 模型的正交结构[xxi]使我们得以分别检验因素 A 和因素 B 的主效应.

- 检验 $H_{01} : \alpha_1 = \cdots = \alpha_I = 0$ 的统计量为

$$\frac{\mathrm{SS}_{\mathrm{A}}/(I-1)}{\mathrm{SS}_{\mathrm{E}}/[IJ(R-1)]} \overset{H_{01}}{\sim} F_{I-1,IJ(R-1)}.$$

- 检验 $H_{02} : \beta_1 = \cdots = \beta_J = 0$ 的统计量为

$$\frac{\mathrm{SS}_{\mathrm{B}}/(J-1)}{\mathrm{SS}_{\mathrm{E}}/[IJ(R-1)]} \overset{H_{02}}{\sim} F_{J-1,IJ(R-1)}.$$

读者如果对分布之类的细节有困惑, 可参看https://zhuanlan.zhihu.com/p/47181027.

方差分析的理论比较标准, 其应用主要在于试验设计, 对业界来说颇具价值.

[xx] 参看https://www.spss-tutorials.com/spss-two-way-anova-interaction-significant/

[xxi] 参看https://stats.stackexchange.com/q/228797

6 杂集

考虑到时间限制和入门课程的要求, 下面的内容仅作简要描述.

6.1 两样本比较

考虑两个实值样本 $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} P$ 和 $Y_1,\dots,Y_m \overset{\text{i.i.d.}}{\sim} Q$, 其中 P 和 Q 是未知的连续型分布, 我们希望比较 P 和 Q.

设 $P=\mathcal{N}(\mu_X,\sigma_X^2)$, $Q=\mathcal{N}(\mu_Y,\sigma_Y^2)$, 且两样本独立. 我们关心 $\mu_X-\mu_Y$, 它的一个自然的估计量是样本均值之差

$$\bar{X}-\bar{Y}\sim\mathcal{N}(\mu_X-\mu_Y,\tfrac{1}{n}\sigma_X^2+\tfrac{1}{m}\sigma_Y^2).$$

当 $\sigma_X^2=\sigma_Y^2$ 时, 估计 σ^2 可以采用加权混合的样本方差

$$s_{\text{pooled}}^2=\frac{(n-1)s_X^2+(m-1)s_Y^2}{n+m-2},$$

此时有枢轴量 (回顾 §3.5)

$$\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{(\frac{1}{n}+\frac{1}{m})s_{\text{pooled}}^2}}\sim t_{n+m-2}.$$

当 $\sigma_X^2\neq\sigma_Y^2$ 时, 情况比较复杂, 人称 *Behrens-Fisher* 问题 (problem), 从略.

还有一些其他思路, 稍作罗列如下:

- *Mann-Whitney* 检验. 采用非参数统计, 不假定分布的形式, 我们关心 $H_0:P=Q$. 将两样本放在一起, 计算 X_i 的秩 (rank) 统计量 $R_i=\sum_{i'=1}^n\mathbb{1}_{\{X_{i'}\le X_i\}}+\sum_{j=1}^m\mathbb{1}_{\{Y_j\le X_i\}}$, 秩和 $T_X=\sum_{i=1}^n R_i$ 将作为检验统计量. 所谓 "秩", 其实就是排名, 对于连续型随机变量, 取值相等 (平局, tie) 应为零概率事件. 当 H_0 为真时, $R_1,\dots,R_n$ 等同于从 $\{1,2,\cdots,n+m\}$ 中不放回简单随机抽样 (见 §2.12), 有 $\mathbb{E}[T_X]=\frac{n(n+m+1)}{2}$ 和 $\mathrm{Var}(T_X)=\frac{nm(n+m+1)}{12}$; 易见 $T_X=\frac{n(n+1)}{2}+\sum_{i=1}^n\sum_{j=1}^m\mathbb{1}_{\{Y_j\le X_i\}}$, 利用 U 统计量的结论, 可以证明当 m 和 n 成比例时, 有 $(T_X-\mathbb{E}[T_X])/\sqrt{\mathrm{Var}(T_X)}$ 渐近 $\mathcal{N}(0,1)$.
- 配对样本 (paired sample). 设 (X_i,Y_i) 是 i.i.d. 样本, 此时 $m=n$, 我们用 $D_i=X_i-Y_i$ 来做推断. 此时, 我们允许 X 和 Y 存在相关, 最终只需要探究 D 的性质, 如利用单样本 t 检验.
- *Wilcoxon* 符号秩检验 (signed rank test). 考虑配对样本的非参数检验, 零假设: D_i 的分布关于 0 对称. 记 D_i 绝对值的秩 $R_i^+=\sum_{j=1}^n\mathbb{1}_{\{|D_j|\le|D_i|\}}$, 符号秩统计量定义为 $W_+=\sum_{i=1}^n R_i^+\mathbb{1}_{\{D_i>0\}}$. 计算可得 $W_+=\binom{n}{2}U+\sum_{i=1}^n\mathbb{1}_{\{D_i>0\}}$, 其中 $U=\binom{n}{2}^{-1}\sum_{i<j}\mathbb{1}_{\{D_i+D_j>0\}}$. 当零假设为真时, 可以证明: $\mathbb{1}_{\{D_i>0\}}\overset{\text{i.i.d.}}{\sim}\text{Bernoulli}(\frac{1}{2})$ 且与 $(R_i^+)_{1\le i\le n}$ 独立, 由此容易得到 $\mathbb{E}[W_+]=\frac{n(n+1)}{4}$ 和 $\mathrm{Var}(W_+)=\frac{n(n+1)(2n+1)}{24}$; 利用 U 统计量的结论, 有 $(W_+-\mathbb{E}[W_+])/\sqrt{\mathrm{Var}(W_+)}$ 渐近 $\mathcal{N}(0,1)$.

6.2 自助法简介

在实际情况中, 统计量服从的分布可能未知或不易处理, Efron 提出的**自助法** (bootstrap) 很好地解决了这一问题.

具体而言, 设 $X=\{X_1,X_2,\dots,X_n\}$ 为取自某总体的 i.i.d. 样本, $T_n=T_n(X)$ 是我们所关心的统计量. 我们尝试用经验分布近似总体, 通过重抽样 (resampling) 得到自助样本 $X^*=\{X_1^*,X_2^*,\dots,X_n^*\}$, 其中每个 X_i^* 从 X 中有放回地随机抽取, 然后据此计算 $T_n^*=T_n(X^*)$; 重复此过程得到多个 T_n^*, 用它们来得到 T_n 的特征, 如偏差、标准误、置信区间等等. 上述方法是非参数自助法, 并不需要总体具有特殊的参数形式.

在参数自助法中, 我们假定原样本 X 来自于一个指定参数形式的分布 P_θ, 如正态分布. 在此框架下, 首先通过 X 得到参数 θ 的估计 $\hat{\theta}$, 从而得到估计出的总体 $P_{\hat{\theta}}$. 接下来仿照非参数自助法的操作, 从 $P_{\hat{\theta}}$ 中重抽样, 最终得到相应统计量的某种特征结果. 如果我们有信心认定样本服从某种特定形式的分布, 那么参数自助法给出的推断结果将更为精确和可靠. 但是, 如果预设的分布不能很好地拟合样本数据, 那么利用参数自助法得到的结果将会比较糟糕.

6.3 统计建模的框架

统计学界有一句流传甚广的名言: "all models are wrong, but some are useful." 我们希望从数据获取一些信息, 一个套路就是去拟合一个统计模型, 然后做一些自然而然的推断. 模型应当适应数据的特点, 因此前面的独立正态线性模型 (LM) 是远远不够的, 之后将稍微罗列一些常见的模型. 其实, 主要的思路就是放松模型假设, 考虑相关的、非正态的、非线性的数据结构, 从参数模型推广到非参数模型以及半参数模型.

例如, 可以认为不同的个体是独立的, 对不同个体测量某变量得到的是截面 (cross-section) 数据, 可能标准线性模型就能发掘很多信息; 但是, 如果对某一个个体在不同时间多次测量某变量, 得到的时间序列 (time series) 理应具有某种相关性, 那么就应该借助时间序列的理论来进行推断; 而且更一般地, 在纵向 (longitudinal) 研究中, 有多个个体在多个时间被记录某变量, 由此得到的面板 (panel) 数据既体现出个体间的独立性又体现出个体内的相关性, 就需要更复杂的手法进行处理.

考虑指数分散族分布, 引入连结函数, 将得到广义线性模型 (GLM). 为了刻画分组/聚类的数据, 在确定效应之外再加入随机效应, 将得到混合 (效应) 模型, 其中线性混合模型[xxii](LMM) 适用于正态的响应变量, 广义线性混合模型 (GLMM) 适用于较为一般的非正态数据.

除了让随机性更加一般, 我们还可以让解释变量的组合更加灵活. 非参数模型包括核回归 (KR)、加性模型 (AM)、广义加性模型 (GAM) 等等; 半参数模型包括部分线性模型 (PLM)、单指标模型 (SIM)、变系数模型 (VCM) 等等.

这完全是报菜名式的介绍, 有兴趣的读者自可学习深入课程[xxiii]或者查阅相关资料.

6.4 统计学掌故

统计学思想在现代社会中愈发重要, 了解一些轶闻/八卦或许会让学习过程增添一些趣味, 在此笔者诚挚地推荐 *Salsburg*《女士品茶》; 此外, 陈希孺《数理统计学简史》和 *Jaynes*《*Probability Theory: The Logic of Science*》文笔流畅而充满哲思, 闲暇时不妨翻阅, 开卷有益.

贵校已故的许宝Ⓕ先生也是概率统计学界宗师级人物, 可参看 http://www.math.pku.edu.cn/misc/probstat/xbl.htm, 所谓 "道德文章垂范人间".

[xxii] 参看https://cosx.org/2014/04/lmm-and-me/

[xxiii] 参看https://zhuanlan.zhihu.com/p/106896222

A 附录: more about...

A.1 极大似然估计渐近性质的条件

沿用 §3.4中的记号, 不妨用 $W_i(\vartheta)$ 来代替 $\ell_1(\vartheta;X_i)$, 其中 $W_i(\vartheta):=\ell_1(\vartheta;X_i)-\ell_1(\theta_0;X_i)$, $\vartheta\in\Theta$ 是独立同分布的随机函数列. 事实上, 这不会影响估计方法, 而且或许有更佳的可积性. 记其期望为 $\mu(\vartheta):=\mathbb{E}[W_i(\vartheta)]$, $\vartheta\in\Theta$. 回忆一下, Shannon 信息不等式表明 $\mu(\vartheta)<\mu(\theta_0)=0$, $\forall\vartheta\in\Theta\setminus\{\theta_0\}$.

为了刻画一致收敛性, 我们引入 $C(\Theta;\mathbb{R})$ 上的范数

$$\|g\|_\infty:=\sup_{\vartheta\in\Theta}|g(\vartheta)|,\quad g\in C(\Theta;\mathbb{R}).$$

当 Θ 是紧集时, 熟知 $C(\Theta;\mathbb{R})$ 在 $\|\cdot\|_\infty$ 下构成可分[xxiv]的 *Banach* 空间 (☺ 仅仅提一下而已).

定理 (随机 (连续) **函数的 LLN**). *设独立同分布的随机元 $W_1,W_2,\dots$ 取值于 $C(\Theta;\mathbb{R})$, 且 Θ 是紧集. 若 $\mathbb{E}\|W_i\|_\infty<\infty$, 则 $\overline{W}_n:=\frac{1}{n}\sum_{i=1}^n W_i$ 满足 $\|\overline{W}_n-\mu\|_\infty\xrightarrow{\text{a.s.}}0$, 其中 $\mu=\mathbb{E}W_i\in C(\Theta;\mathbb{R})$.*

证明. 记 $B(\vartheta;\delta):=\{\tau\in\Theta:|\tau-\vartheta|<\delta\}$. 引入**连续模** (modulus of continuity)

$$\omega_\delta(W_i,\vartheta):=\sup_{\tau\in B(\vartheta;\delta)}|W_i(\tau)-W_i(\vartheta)|,\quad i\in\mathbb{N},\ \vartheta\in\Theta,\ \delta>0.$$

令 $\lambda_\delta(\vartheta):=\mathbb{E}[\omega_\delta(W_i,\vartheta)]$. 易见 $|\mu(\tau)-\mu(\vartheta)|\le\lambda_\delta(\vartheta)$, $\forall\tau\in B(\vartheta;\delta)$.

引理. *当 $\delta\searrow0$ 时, $\|\lambda_\delta\|_\infty\to0$.*

引理的证明. 任取 $\vartheta\in\Theta$, 由 W_i 连续可知 $\omega_\delta(W_i,\vartheta)\xrightarrow{\delta\searrow0}0$, 而 $\omega_\delta(W_i,\vartheta)\le2\|W_i\|_\infty$, 利用控制收敛定理立得 $\lambda_\delta(\vartheta)\xrightarrow{\delta\searrow0}0$. 为了将逐点收敛性加强为一致收敛性, 我们可借助数学分析中的 **Dini 定理**: *如果紧集上的连续函数列逐点单调收敛于连续函数, 那么收敛是一致的.* 易见 λ_δ 随 $\delta\searrow0$ 单调递减, 只需说明 $\lambda_\delta(\cdot)$ 是紧集 Θ 上的连续函数. 利用 $\omega_\delta(W_i,\cdot)$ 的连续性, 再次控制收敛即可. □

引理的一个显然的推论是 μ 的一致连续性. 任取 $\varepsilon>0$, 存在 $\delta\in(0,\varepsilon)$, 使得

$$|\mu(\tau)-\mu(\vartheta)|\le\varepsilon,\quad\forall\vartheta,\tau\in\Theta:|\vartheta-\tau|\le\delta.$$

由于 Θ 紧, 存在有限子集 $F\subset\Theta$, 使得 $\Theta=\bigcup_{\vartheta\in F}B(\vartheta;\delta)$, 进而

$$\begin{aligned}
\|\overline{W}_n-\mu\|_\infty&=\max_{\vartheta\in F}\sup_{\tau\in B(\vartheta;\delta)}\left|\overline{W}_n(\tau)-\mu(\tau)\right|\\
&\le\max_{\vartheta\in F}\sup_{\tau\in B(\vartheta;\delta)}\Big[\left|\overline{W}_n(\tau)-\overline{W}_n(\vartheta)\right|+\left|\overline{W}_n(\vartheta)-\mu(\vartheta)\right|+\left|\mu(\vartheta)-\mu(\tau)\right|\Big]\\
&\le\max_{\vartheta\in F}\sup_{\tau\in B(\vartheta;\delta)}\left|\overline{W}_n(\tau)-\overline{W}_n(\vartheta)\right|+\max_{\vartheta\in F}\left|\overline{W}_n(\vartheta)-\mu(\vartheta)\right|+\varepsilon\\
&\le\max_{\vartheta\in F}\Big(\frac{1}{n}\sum_{i=1}^n\underbrace{\sup_{\tau\in B(\vartheta;\delta)}|W_i(\tau)-W_i(\vartheta)|}_{=\omega_\delta(W_i,\vartheta)}\Big)+\max_{\vartheta\in F}\left|\overline{W}_n(\vartheta)-\mu(\vartheta)\right|+\varepsilon\\
&\xrightarrow{\text{a.s.}}\max_{\vartheta\in F}\lambda_\delta(\vartheta)+\max_{\vartheta\in F}|\mu(\vartheta)-\mu(\vartheta)|+\varepsilon\quad\le\|\lambda_\delta\|_\infty+\varepsilon,
\end{aligned}$$

这一步用了有限维随机向量的 SLLN, 最后让 $\varepsilon\searrow0$ 即证. □

[xxiv] 这是 Stone-Weierstrass 定理的一个应用.

定理. 设 $\mu \in C(\Theta;\mathbb{R})$, 且 Θ 是紧集. 若 $\mu(\vartheta) < \mu(\theta_0)$, $\forall \vartheta \neq \theta_0$, 则 $\sup_{\vartheta:|\vartheta-\theta_0|\geq\varepsilon} \mu(\vartheta) < \mu(\theta_0)$, $\forall \varepsilon > 0$.

证明. 由于 $\{\vartheta \in \Theta : |\vartheta - \theta_0| \geq \varepsilon\}$ 闭, 作为紧集的子集仍然保持紧, 所以连续函数 μ 在其上能够取到最大值, 从而目标不等式严格成立. □

当 Θ 是紧集时, 我们已经成功建立了所需条件. 对于一般的 Θ, 假设存在紧集 D 适合

$$\mathbb{E}\big[\sup_{\vartheta\in\Theta\setminus D} W_i(\vartheta)\big] < \mathbb{E}[W_i(\theta_0)] = \mu(\theta_0).$$

利用 LLN 可以说明如下事件的概率趋近于一:

$$\sup_{\vartheta\in\Theta\setminus D} \overline{W}_n(\vartheta) \leq \frac{1}{n}\sum_{i=1}^n \sup_{\vartheta\in\Theta\setminus D} W_i(\vartheta) < \frac{1}{n}\sum_{i=1}^n W_i(\theta_0) = \overline{W}_n(\theta_0).$$

于是 $\hat\theta_n \in \arg\max_{\vartheta\in\Theta} \overline{W}_n(\vartheta)$ 落入紧集 D 的概率趋近于一, 从而划归为已解决的问题.

A.2 求解极大似然估计的算法

很多时候 MLE 难以显式计算, 为此本节简要介绍一些实用的数值近似算法.
选定参数初值 $\hat\theta^{(0)}$, 我们关注从 $\hat\theta^{(t)}$ 推导 $\hat\theta^{(t+1)}$ 的迭代步骤, 其中 $t = 0, 1, 2, \cdots$ 表示迭代次数.

- (**Newton–Raphson**) 记对数似然函数为 $\ell_n(\vartheta)$, $\vartheta \in \Theta$. 注意到

$$0 = \dot\ell_n(\hat\theta^{\mathrm{ML}}) \approx \dot\ell_n(\hat\theta^{(t)}) + \ddot\ell_n(\hat\theta^{(t)})\big(\hat\theta^{\mathrm{ML}} - \hat\theta^{(t)}\big),$$

  为了 $\hat\theta^{(t+1)} \approx \hat\theta^{\mathrm{ML}}$, 采用

$$\hat\theta^{(t+1)} = \hat\theta^{(t)} - \ddot\ell_n(\hat\theta^{(t)})^{-1}\dot\ell_n(\hat\theta^{(t)}).$$

- (**Fisher's scoring**) 注意到 $-\ddot\ell_n(\theta)$ 在 $\mathbb{P}_\theta$ 下的期望是 Fisher 信息矩阵 $\mathcal{I}_n(\theta)$, 前述 Newton 法可以稍作修改, 变为

$$\hat\theta^{(t+1)} = \hat\theta^{(t)} + \mathcal{I}_n(\hat\theta^{(t)})^{-1}\dot\ell_n(\hat\theta^{(t)}).$$

- (**EM 算法**, expectation-maximization) 在许多实际应用[xxv]中, 我们会遇到缺失数据 (missing data). 将样本写成 $X = (Y, Z)$, 其中 Y 是可观测变量, 而 Z 是无法观测的隐变量. 此时, 概率密度函数分解为

$$f_{Y,Z|\theta}(y,z) = f_{Z|Y,\theta}(z|y) f_{Y|\theta}(y).$$

  – (E 步, Expectation step) 计算

$$Q(\theta|\hat\theta^{(t)}) = \mathbb{E}_{Z|Y,\hat\theta^{(t)}}[\log f_{Y,Z|\theta}(Y,Z)] = \int_{\mathcal{Z}^n} \big(\log f_{Y,Z|\theta}(Y,z)\big) f_{Z|Y,\hat\theta^{(t)}}(z|Y)\,\mathrm{d}z.$$

  – (M 步, Maximization step) 求解

$$\hat\theta^{(t+1)} \in \arg\max_{\theta\in\Theta} Q(\theta|\hat\theta^{(t)}).$$

关于收敛性, 可以用 (条件版本的) Shannon 信息不等式推出 $f_{Y|\hat\theta^{(t)}}(Y)$ 关于 t 单调递增. 事实上,

$$\log f_{Y|\theta}(Y) - \log f_{Y|\hat\theta^{(t)}}(Y) \geq Q(\theta|\hat\theta^{(t)}) - Q(\hat\theta^{(t)}|\hat\theta^{(t)}), \quad \forall\theta\in\Theta.$$

[xxv] 参看 Jeff Bilmes (1998) *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*. https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.613

A.3 Markov Chain Monte Carlo 方法

应用 Bayes 方法时, 涉及后验分布的显式计算可能极其困难. 实践中, 近年来流行的数值逼近技术是 **Markov Chain Monte Carlo** (MCMC). 所谓 Monte Carlo 方法, 其实就是随机模拟/统计模拟, 这是 von Neumann 用摩纳哥的赌城 Monte Carlo 来命名的; 很早之前的 Buffon 投针实验[xxvi)]被认为是随机模拟方法的起源.

一般地, 我们想求出

$$\int_\Theta g(\theta)\pi(\theta|x)\,\mathrm{d}\lambda(\theta),$$

其中后验密度函数 $\pi(\bullet|x)\propto f(x|\bullet)\pi(\bullet)$ 具有难以计算的配分函数 $f(x)=\int_\Theta f(x|\bullet)\pi(\bullet)$. 注意到

$$\frac{\pi(\theta'|x)}{\pi(\theta|x)}=\frac{f(x|\theta')\pi(\theta')}{f(x|\theta)\pi(\theta)},\quad \theta,\theta'\in\Theta$$

避开了难以处理的配分函数 $f(x)$, 所以我们想充分利用这样的比值. 目标是抽样得到一列 ϑ_t, $t\in\mathbb{N}$, 适合

$$\frac{1}{T}\sum_{t=1}^{T} g(\vartheta_t)\xrightarrow[(T\to\infty)]{\text{a.s.}}\sum_{i\in S} g(\theta_i)\pi_i, \tag{$\star$}$$

其中 S 是至多可数的指标集: 类似 Riemann 和, 我们预先将 Θ 分划成足够小的 Θ_i, 即 $\Theta=\bigsqcup_{i\in S}\Theta_i$, 取定 $\theta_i\in\Theta_i$, 并且记 $\pi_i=\int_{\Theta_i}\pi(\bullet|x)$. 显然有

$$\frac{\pi_j}{\pi_i}=\frac{\int_{\Theta_j} f(x|\bullet)\pi(\bullet)}{\int_{\Theta_i} f(x|\bullet)\pi(\bullet)},\quad \forall i,j\in S$$

是容易计算的. 形如($\star$)的结论称为**遍历定理** (ergodic theorem), 对于非周期 (aperiodic) 且不可约 (irreducible) 的马氏链 (Markov chain), 只要 $(\pi_i)_{i\in S}$ 是平稳分布, 就能使($\star$)成立.

我们先简介一下马氏链的基本概念. 设 $Y_0,Y_1,Y_2,\dots$ 是一列取值于 S 的随机元, 则 $Y=(Y_t)_{t\in\mathbb{N}}$ 称为一个离散时间**随机过程** (stochastic process in discrete time). 若对 $\forall t\in\mathbb{N}$ 都有

$$\mathbb{P}(Y_{t+1}\in\bullet|Y_t,Y_{t-1},\dots,Y_0)=\mathbb{P}(Y_{t+1}\in\bullet|Y_t),$$

即分布满足 "给定现在之后, 未来与过去无关", 则称 Y 是一个 Markov 过程, 具有 **Markov 性质**. 特别地, 若 $\mathbb{P}(Y_{t+1}\in\bullet|Y_t)=\mathbb{P}(Y_1\in\bullet|Y_0)$, $\forall t$, 则称 Markov 过程 Y 是**时齐的** (time-homogeneous). 当 S 至多可数时, 我们称 Markov 过程 Y 为马氏链, 此时 S 称为状态空间, 每个 $i\in S$ 都称作一个状态 (state). 我们主要关心时齐马氏链. 定义**一步转移概率** (transition probabilities)

$$p(i\to j):=\mathbb{P}(Y_{t+1}=j|Y_t=i),\quad i,j\in S.$$

我们称 $\mu^0:=\mathbb{P}\{Y_0\in\bullet\}$ 为马氏链 Y 的**初始分布** (initial distribution). 演化时间 t 之后, 可得分布 $\mu^t:=\mathbb{P}\{Y_t\in\bullet\}=\mu^0p^t$, 其中

$$\mu^0p^t\{j\}=\sum_{i_0,i_1,\dots,i_{t-1}\in S}\mu^0\{i_0\}p(i_0\to i_1)p(i_1\to i_2)\cdots p(i_{t-2}\to i_{t-1})p(i_{t-1}\to j),\quad \forall j\in S.$$

称 S 上的概率分布 μ 为 (以 p 为转移概率的马氏链的) **平稳分布/不变分布** (stationary/invariant distribution), 若 $\mu p=\mu$, 即

$$\mu\{j\}=\mu p\{j\}:=\sum_{i\in S}\mu\{i\}p(i\to j),\quad \forall j\in S.$$

[xxvi)] 参看https://mste.illinois.edu/activity/buffon/

显而易见, 若 $\mu^0=\mu$ 是平稳分布, 则 $\mu^t=\mu,\ \forall t\in\mathbb{N}$. 不难发现, 当**细致平衡** (detailed balance) 条件

$$\mu\{i\}p(i\to j)=\mu\{j\}p(j\to i),\quad \forall i,j\in S$$

成立时, 概率分布 μ 一定是平稳分布.

回到原先的问题, 我们希望构造以 $(\pi_i)_{i\in S}$ 为平稳分布的马氏链 $(\vartheta_t)_{t\in\mathbb{N}}$. 为此, 只需构造转移概率即可, 而初始分布可以任意选定. 利用细致平衡条件, 尝试寻找 p 满足

$$\frac{p(i\to j)}{p(j\to i)}=\frac{\pi_j}{\pi_i},\quad \forall i,j\in S.$$

显然只需考虑 $j\neq i$ 的情形, 之后令 $p(i\to i):=1-\sum_{j\neq i}p(i\to j)$ 即可. 下述转移概率满足要求:

$$p(i\to j)=q(i\to j)a(i\to j),\quad j\neq i,$$

其中 q 是人为选取的**提议** (proposal) 概率, 每个 $q(i\to\bullet)$ 都是 $S\setminus\{i\}$ 上的分布, 例如 Uniform$(S\setminus\{i\})$; 而 a 称作**接受** (acceptance) 概率, 定义为

$$a(i\to j):=\min\left(\frac{\pi_jq(j\to i)}{\pi_iq(i\to j)},1\right),\quad j\neq i.$$

由此立得 **Metropolis-Hastings 算法**:

1. 任取初值 $i_0\in S$, 令 $t=0$;
2. 生成 $i_{t+1}\sim q(i_t\to\bullet)$;
3. 生成 $u\sim\text{Uniform}([0,1])$, 若 $u>\frac{\pi_{i_{t+1}}q(i_{t+1}\to i_t)}{\pi_{i_t}q(i_t\to i_{t+1})}$, 则置 $i_{t+1}\leftarrow i_t$;
4. 置 $t\leftarrow t+1$, 利用 $\vartheta_t=\theta_{i_t}$ 更新($\star$)中的计算结果;
5. 重复第 2–4 步.

具体实现时, 可以忽略 t 较小时产生的 ϑ_t, 此阶段称为预热 (warm-up), 从而可以减少初值的影响.

当 $\theta=(\theta_1,\theta_2,\dots,\theta_m)$ 具有较高的维数 m 时, 我们可以应用下述称为 **Gibbs 采样**的降维技巧. 在迭代过程中, 从旧的样本 θ^{old} 得到新的样本 θ^{new} 的步骤可以逐次进行:

1. 用条件分布 $\theta_1\mid x,\{\theta_s=\theta_s^{\text{old}}\}_{s>1}$ 生成 θ_1^{new};
2. 用条件分布 $\theta_2\mid x,\theta_1=\theta_1^{\text{new}},\{\theta_s=\theta_s^{\text{old}}\}_{s>2}$ 生成 θ_2^{new};

$\vdots$

k. 用条件分布 $\theta_k\mid x,\{\theta_r=\theta_r^{\text{new}}\}_{r<k},\{\theta_s=\theta_s^{\text{old}}\}_{s>k}$ 生成 θ_k^{new};

$\vdots$

m. 用条件分布 $\theta_m\mid x,\{\theta_r=\theta_r^{\text{new}}\}_{r<m}$ 生成 θ_m^{new}.

A.4 经典的检验统计量

对于参数 $\theta\in\Theta\subseteq\mathbb{R}^p$, 考虑假设检验问题

$$H_0:\theta\in\Theta_0=\{\theta:g(\theta)=0\}\overset{\text{vs}}{\longleftrightarrow}H_1:\theta\in\Theta_1=\Theta\setminus\Theta_0,$$

其中 $g:\mathbb{R}^p\to\mathbb{R}^q$ 是已知的可微函数. 我们不加声明地假定一些正则性条件. 记对数似然函数为 $\ell_n(\cdot)$, 设未知参数 θ 的 (不受限) 极大似然估计

$$\hat\theta\in\arg\max_{\theta\in\Theta}\ell_n(\theta)$$

和 θ 在 H_0 下的受限极大似然估计

$$\hat{\theta}_0 \in \arg\max_{\theta\in\Theta_0} \ell_n(\theta)$$

都是良好定义的. 熟知 Wilks 定理给出了似然比检验的大样本性质 (回顾 §4.4):

$$2\big(\ell_n(\hat{\theta}) - \ell_n(\hat{\theta}_0)\big) \xrightarrow[H_0]{d} \chi^2_{\dim(\Theta)-\dim(\Theta_0)}.$$

下面介绍[xxvii]两个与似然比统计量渐近等价的检验统计量——Wald 统计量和得分统计量; 这里我们说一列检验统计量 T_n 与似然比统计量渐近等价, 指的是

$$2\big(\ell_n(\hat{\theta}) - \ell_n(\hat{\theta}_0)\big) - T_n \xrightarrow[H_0]{\mathbb{P}} 0,$$

从而根据 Slutsky 定理有

$$T_n \xrightarrow[H_0]{d} \chi^2_{\dim(\Theta)-\dim(\Theta_0)}.$$

- **Wald 检验**采用的统计量为

  $$W_n = g(\hat{\theta})^\top \widehat{\mathrm{Var}}(g(\hat{\theta}))^{-1} g(\hat{\theta}) = g(\hat{\theta})^\top \big(\dot{g}(\hat{\theta})\mathcal{I}_n(\hat{\theta})^{-1}\dot{g}(\hat{\theta})^\top\big)^{-1} g(\hat{\theta}),$$

  其中 $\mathcal{I}_n(\cdot)$ 是 Fisher 信息矩阵. 注意 $\mathrm{Var}_\theta(\hat{\theta}) \approx \mathcal{I}_n(\theta)^{-1}$, 且 $\mathrm{Var}(g(\hat{\theta})) \approx \mathrm{Var}(\dot{g}(\theta)(\hat{\theta}-\theta))$.

  ♢ 一般来讲, 我们拒绝 $H_0 : g(\theta) = 0$ 时, 一个合理的依据应该是 $\hat{\gamma} := g(\hat{\theta})$ 与 0 的距离比较大, 而且这要考虑到 $\hat{\gamma}$ 的变动能力. 我们举个例子, 设 $\hat{\gamma} \sim \mathcal{N}_q(\gamma, D)$, 其中 γ 是待估参数, D 是正定矩阵. 定义 $\hat{\gamma}$ 到 $\mathcal{N}_q(\gamma_0, D)$ 的 **Mahalanobis 距离**为

  $$\|\hat{\gamma} - \gamma_0\|_{D^{-1}} = \sqrt{(\hat{\gamma}-\gamma_0)^\top D^{-1}(\hat{\gamma}-\gamma_0)}.$$

  由于 $D^{-1/2}(\hat{\gamma}-\gamma_0) \sim \mathcal{N}_q(D^{-1/2}(\gamma-\gamma_0), I_q)$, 当 $\gamma = \gamma_0$ 时就会成立

  $$\|\hat{\gamma} - \gamma_0\|^2_{D^{-1}} = \big\|D^{-1/2}(\hat{\gamma}-\gamma_0)\big\|^2 \sim \chi^2_q.$$

- **得分检验** (由 Rao 引入, 又称 **Lagrange 乘子检验**) 采用的统计量为

  $$R_n = \dot{\ell}_n(\hat{\theta}_0)^\top \mathcal{I}_n(\hat{\theta}_0)^{-1} \dot{\ell}_n(\hat{\theta}_0),$$

  其中 $\dot{\ell}_n(\theta) = \frac{\partial}{\partial\theta}\ell_n(\theta)$ 是得分函数.

  ♢ 引入 Lagrange 函数

  $$(\theta, \lambda) \mapsto \ell_n(\theta) - \lambda^\top g(\theta),$$

  一阶条件为

  $$\dot{\ell}_n(\hat{\theta}_0) = \dot{g}(\hat{\theta}_0)\hat{\lambda}.$$

  直观上, Lagrange 乘子 λ 影响越大, 我们就越倾向于拒绝 $H_0 : g(\theta) = 0$. 注意到 $\dot{\ell}_n(\hat{\theta}_0)$ 是 $\hat{\lambda}$ 通过线性变换得到的, 所以利用 $\dot{\ell}_n(\hat{\theta}_0)$ 构造 Mahalanobis 距离应当是一个合理的选择.

✍ 似然比检验、Wald 检验和得分检验是三大经典假设检验, 它们在大样本下是等价的. Wald 检验相对于其他两个检验的优势在于, 只需要对无约束模型进行估计, 从而降低了计算量. 得分检验只需要在零假设下对受限模型的似然函数进行估计, 因此备择假设的确切性质不如其他两个检验那么明晰.

[xxvii] 证明略. 有兴趣的读者可以参看https://statlect.com/fundamentals-of-statistics/Wald-test 和https://statlect.com/fundamentals-of-statistics/score-test.

A.5 投影矩阵

设 V 是 $\mathbb{R}^n$ 的线性子空间, 则 $\boldsymbol{y}\in\mathbb{R}^n$ 向 V 的**正交投影** (orthogonal projection) 为

$$\operatorname{proj}_V(\boldsymbol{y}):=\underset{\boldsymbol{v}\in V}{\arg\min}\,\|\boldsymbol{y}-\boldsymbol{v}\|,$$

其中 $\|\bullet\|:=\sqrt{\bullet'\bullet}$ 是欧氏范数. 考虑等价刻画

$$\|\boldsymbol{y}-(\operatorname{proj}_V(\boldsymbol{y})+\boldsymbol{u})\|^2-\|\boldsymbol{y}-\operatorname{proj}_V(\boldsymbol{y})\|^2=\|\boldsymbol{u}\|^2-2(\boldsymbol{y}-\operatorname{proj}_V(\boldsymbol{y}))'\boldsymbol{u}\geq 0,\quad \forall\boldsymbol{u}\in V,$$

这当且仅当成立

$$\boldsymbol{y}-\operatorname{proj}_V(\boldsymbol{y})\in V^\perp:=\{\boldsymbol{w}\in\mathbb{R}^n:\boldsymbol{w}'\boldsymbol{v}=0,\forall\boldsymbol{v}\in V\}.$$

因此, 投影算子 proj_V 是由正交分解 $\mathbb{R}^n=V\bigoplus V^\perp$ 唯一确定的线性变换. 对于 $\boldsymbol{v}\in\mathbb{R}^n\setminus\{\boldsymbol{0}_n\}$, 常常记

$$\operatorname{proj}_{\boldsymbol{v}}:=\operatorname{proj}_{\operatorname{span}(\boldsymbol{v})}=\tfrac{v}{\|v\|}\tfrac{v'}{\|v\|}.$$

对于正交系 $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_k\in\mathbb{R}^n\setminus\{\boldsymbol{0}_n\}$, 有 $\operatorname{proj}_{\operatorname{span}(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_k)}=\sum_{i=1}^k\operatorname{proj}_{\boldsymbol{v}_i}$.

称矩阵 $\boldsymbol{P}\in\mathbb{R}^{n\times n}$ 是**幂等的** (idempotent), 若 $\boldsymbol{P}^2=\boldsymbol{P}$. 再设 $\boldsymbol{P}$ 对称, 通过谱分解[xxviii] (spectral decomposition) 可以得到

$$\boldsymbol{P}=(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n)\operatorname{diag}\big(\boldsymbol{I}_r,\boldsymbol{0}_{(n-r)\times(n-r)}\big)(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n)'=\sum_{i=1}^r\boldsymbol{e}_i\boldsymbol{e}_i',$$

其中特征向量 $\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n$ 构成 $\mathbb{R}^n$ 的规范正交基, $r=\operatorname{tr}(\boldsymbol{P})=\operatorname{rank}(\boldsymbol{P})$. 容易发现, 从属于特征值 1 的特征向量 $\boldsymbol{e}_1,\ldots,\boldsymbol{e}_r$ 张成 $\boldsymbol{P}$ 的列空间 (column space), 即

$$\operatorname{span}(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_r)=\operatorname{Col}(\boldsymbol{P}):=\{\boldsymbol{P}\boldsymbol{v}:\boldsymbol{v}\in\mathbb{R}^n\},$$

所以 $\boldsymbol{P}=\operatorname{proj}_{\operatorname{Col}(\boldsymbol{P})}$, 且 $\boldsymbol{I}_n-\boldsymbol{P}=\operatorname{proj}_{\operatorname{Col}(\boldsymbol{P})^\perp}$.

定理. *任给 $\boldsymbol{X}\in\mathbb{R}^{n\times p}$, 设 $(\boldsymbol{X}'\boldsymbol{X})^-\in\mathbb{R}^{p\times p}$ 是 $\boldsymbol{X}'\boldsymbol{X}$ 的广义逆[xxix] (generalized inverse), 即满足 $\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X}=\boldsymbol{X}'\boldsymbol{X}$, 则*

$$\boldsymbol{P}_{\boldsymbol{X}}:=\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'$$

是 $\mathbb{R}^n$ 向 $\operatorname{Col}(\boldsymbol{X}):=\{\boldsymbol{X}\boldsymbol{b}:\boldsymbol{b}\in\mathbb{R}^p\}$ 的正交投影, 这不依赖 $(\boldsymbol{X}'\boldsymbol{X})^-$ 的选取方法.

证明. 任取 $\boldsymbol{v}\in\mathbb{R}^n$, 将其写成 $\boldsymbol{v}=\boldsymbol{x}+\boldsymbol{w}$, 其中 $\boldsymbol{x}\in\operatorname{Col}(\boldsymbol{X})$, $\boldsymbol{w}\in\operatorname{Col}(\boldsymbol{X})^\perp$; 这种表示存在且唯一, 因为 $\mathbb{R}^n=\operatorname{Col}(\boldsymbol{X})\bigoplus\operatorname{Col}(\boldsymbol{X})^\perp$. 由于 $\boldsymbol{X}'\boldsymbol{w}=\boldsymbol{0}_p$, 我们有 $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{w}=\boldsymbol{0}_n$ 和 $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{v}=\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{x}$. 为了证明 $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{x}=\boldsymbol{x}$, 只需

$$\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X}=\boldsymbol{X},$$

或者等价地, $\boldsymbol{u}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X}=\boldsymbol{u}'\boldsymbol{X}$, $\forall\boldsymbol{u}\in\mathbb{R}^n$. 为此, 只需 $\boldsymbol{u}'\boldsymbol{X}=\boldsymbol{z}'\boldsymbol{X}'\boldsymbol{X}$ 对某一 $\boldsymbol{z}\in\mathbb{R}^p$ 成立. 于是, 我们的目标化为

$$\operatorname{Col}(\boldsymbol{X}')=\operatorname{Col}(\boldsymbol{X}'\boldsymbol{X}).$$

显然 $\operatorname{Col}(\boldsymbol{X}'\boldsymbol{X})\subset\operatorname{Col}(\boldsymbol{X}')$, 所以利用

$$\operatorname{rank}(\boldsymbol{X}'\boldsymbol{X})=\operatorname{rank}(\boldsymbol{X})=\operatorname{rank}(\boldsymbol{X}')$$

即可; 至于这个秩的关系式, 注意 $\boldsymbol{X}\boldsymbol{b}=\boldsymbol{0}_n\iff\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}=\boldsymbol{0}_p$, $\forall\boldsymbol{b}\in\mathbb{R}^p$. □

✍ 通过投影矩阵构造正态随机向量的二次型, 将会得到卡方分布, 这很有用;
参看 **Cochran 定理** https://zhuanlan.zhihu.com/p/85314314

[xxviii] 参看https://zhuanlan.zhihu.com/p/75250722
[xxix] 参看https://zhuanlan.zhihu.com/p/75283604

B 后记

笔者的概率论老师说过, 学懂一门课的检验标准是能否完整清晰地向初学者讲授这门课, 大意如此. 由于当初学习数理统计异常痛苦而低效, 即使期末取得了一个还算不错的总评成绩, 笔者仍然深觉自己缺少扎实的统计学基础. 为此, 趁着大四稍有闲暇, 笔者尝试着去担任了国发院数理统计的助教, 希望或许能够查漏补缺, 也尽力帮助选课的同学避免一些无谓的困难.

这份笔记是作为习题课的 tutorial 而准备的, 不过并不完全局限于课程内容及其编排方式, 而是按照笔者的浅薄认识从各个地方吸取了一些经验, 希望简明扼要地呈现尽量多的知识点, 为学弟学妹们留下一份方便的入门参考资料.

感谢 CYF 同学作序, 他有着足够高的水平, 提供的意见必然充满建设性.

一切可能的错误和不足都归于笔者, 惭愧惭愧.

C 资料推荐

笔者才疏学浅[xxx], 谨罗列一些与本课程相关/相当的网络资源, 希望对读者有所帮助.

- Stanford Stats 200: 以 Rice 书为教材.
- Oxford A9: 有自编讲义, 作业包括 R 的练习.
- MIT 14-381: 从 Casella&Berger 书中选取材料.
- UIC Stat 411: 以 Hogg 书为教材, notes 尤其是最后一章 “What else is there to learn?” 非常好.
- Columbia GU4204: 取材于 DeGroot&Schervish 书, 也挺好的.
- CMU 36-705: 由 Wasserman 讲他自己的书《All of Statistics》, 涉及不少进阶内容.
- PSU STAT 553: 稍微强调渐近性质.
- A. W. van der Vaart 的书 An Introduction to Mathematical Statistics以及一个初等渐近统计的讲义 MATHEMATISCHE STATISTIEK: 脑残粉不解释.

应用多元统计分析

[xxx] https://www.zhihu.com/question/34941826/answer/788612729

