第一章　基 本 概 念

> 一项构思良好的数据分析计划将会赋予组织新的能量，放弃用过往的决定来支配判断，而是转向观察经验里的事实，并从中获得行动力。组织越早转向，就会越快享受到由真正的数据驱动决策所带来的福利。　　——物理学家尼尔斯·玻尔(Niels Bohr,1885年10月7日－1962 年11月18 日)

§1.1　非参数统计的概念与产生

1. 非参数统计研究什么

人天生就有多种不同的思维能力，但并不都转化为决策力。幸运的是，数据是决策的孵化器，可用于规划未来。比如在决定接下来的小长假去哪里旅行的几种备选方案中，就2020年而言，可参考的数据有目的地新冠疫情等级、景区人口密度、空气质量指数、天气预报信息以及景区周边交通路况信息等。在旅行前根据这些数据做功课就可以看作是一项朴素的数据驱动的决策。不过，单就去哪儿这项决策，上述数据是否是影响到决策中最关键的信息却是因人而异。或许在启动上述决策之前，考虑和谁同行以及目的地为什么不是选在图书馆，比如与翻阅非参数统计文献相比，选择走出校园还是留在图书馆才是决策的重点。

个体需求多样化和影响因素繁多是现代复杂决策中的两个难点。良好的决策里离不开清晰的目标、有力的模型和对数据的广泛动员。概率论通过量化目标和计算技术为决策提供了基本框架。随着数学工具的不断发展以及可搜集数据的不断增加，人们已经不满足于仅仅在基本框架上进行决策，而是更希望理论研究能够面对现实问题结合复杂的数据特点给予决策者以有效的指导。为增强模型性能，需要引入巨大的参数量和计算量。

根据数据对背后的分布做出推断是传统统计推断里的一项核心任务。在传统的参数推断框架里，参数通常是作为随机变量（向量）分布族中的显性特征而为人所知的。比如, 研究某类商品的市场占有率,假定在平均的意义下, 某类消费者对待研究商品的喜好或厌倦的表态就是一个随机变量。这个随机变量来自两点分布$B(1,p),(0<p<1)$,p 是两点分布族的待估计的参数；在研究保险公司里某个险种的索赔请求数时,假定索赔请求数来自泊松分布$p(\lambda),(0<\lambda<\infty)$；再比如，在研究气温对农作物产量的影响效果时, 假定平均的意义下, 每测量单元(可能是单产亩产量)产量$y|x$ 服从正态分布$N(\mu+x\beta,\sigma^2)$, 其中x 是试验环境变量，比如日照充足量，μ,β 和σ^2 是待估参数。一般一个推断过程常包括以下几个步骤：假定分布族$\mathcal{F}(\theta)$, 确定要推断的参数和范围，选择用来推断参数的统计量，确定抽样分布, 运用抽样分布估计参数，进行可靠性分析，检验分布特征等。样本被视为从分布族的某个参数族抽取出来的用于估计总体分布的数据代表, 未知的仅仅是总体分布中具

体的参数值。这样，一个实际问题就转化为对分布族中若干个设定好的未知参数通过样本进行推断的过程。通过样本对参数做出估计或进行检验, 从而获知数据背后的分布, 这类推断方法称为**参数方法**。

在许多实际问题中, 我们所关心的数据中对分析问题有帮助的分布可能是多样化的、高维度的且带噪声的，这就为统计推断带来了第二项使命：关注数据的使用。首先在问题表示方面，非参数统计突破传统对参数的认识，扩展了分析问题的范围，能够分析和解释更多与建模相关的理论。比如，当真实信号相对于噪声而言是稀疏微弱分散时，我们收集到的数据可能包含信号也可能不包含信号，参数空间里可能存在着不易推断的区域？我们既不清楚手中的数据里能够提供多少种对分布可能性的表达，也不清楚这些分布差异的大小，用参数分布去表示真实数据的复杂分布也许并不可行。这就需要在统计推断的任务中纳入参数多样性的理解和分析程序的控制等必要的内容。一方面，需要纳入多样化的参数来引导对数据产生丰富分析结构的理解。比如用q 分位数$F^{-1}(q)$ 表示分布的边界，其中F 是数据的分布函数。这个参数特征是分布函数依赖的，而不必借由均值和标准差来定义。如果q取0.9 就可能对应一组特别的人群，比如，经济学家艾略特·哈里斯在一份美国橄榄球联盟球员的健康报告中曾指出：高级别参赛队伍球员的身高和体重较之在低级别参赛队伍而言会呈现出显著优势，但研究显示因高强度的训练和频繁不断的赛事安排，高级别队员的大脑损伤几乎无一例外，这表明职业联赛中均值所指向的群组之外的群体需要获得健康方面的额外关注，也就是说“明者毋掩其弱”，这就需要用于探索异质性结构的参数来发现这些群体。再比如C. 比绍曾讲过一个多项式回归中理想阶数选择的例子，一个阶数较高的模型实现了对训练数据高精度的拟合，却会遭遇系数膨胀问题，这是一个典型的计算效率高但统计效率低下的例子，如何平衡统计效率和计算效率也是非参数推断里需要面对的问题(C.Bishop,2007)。

第二、参数推断和非参数推断虽然在方法论和分析框架上有一些区别，非参数模型与参数模型的结合会带来更细致和更可靠的量化结果。比如，在金融资产管理的可靠性和有效性来自于对风险度量的准确性。传统风险计量方法局限于金融资产波动率为一个恒定的常数体系下，但是研究发现：资产波动率的方差是随时间变化的，这就使得传统风险计量模型关于独立同方差的假设不再适于描述金融资产的运动规律。近几年随着微观数据不断被深入挖掘，稳健的非参模型引起金融资产管理广泛关注，它的优点是：一方面无需设定模型的具体形式极大地提高了模型的自由性和灵活性；另一方面这些稳健的方法允许大范围数据存在相依性，这对于相互关联的金融时间序列数据来说，增强了传统量化模型的功能(解其昌,2015)。

第三、非参数统计推断与更多算法议题产生交集，丰富了跨学科技术决策的传导机制。现有很多数据驱动技术比较偏爱算法，却普遍缺乏统计分析的保障，两者之间的交互借鉴对模型与计算技术的协调发展具有十分重要的意义。比如，重症

疾病再入院研究指出，国际上关于再入院率研究与监管已十分成熟(蒋重阳,2017)。建立在随访数据基础上的健康管理医院的病人不仅再入院率会有明显降低，而该系统也为综合疾病的监测与预防提供了广泛的时空观测视角。比如传染病流行期，一些地理位置相邻的小区的发病率相邻时间段之间会有关联性的升高或下降，传统的通过移动平均结合偏离均值的全局异常检测方法，对局部尺度下的异常响应迟钝，无法满足应急精准施策的灵敏要求而一种通过局部时空区域内外发病密度变化所生成的复合性得分则为预警检测提供了快速响应的统计监测方法。这样一类局部算法框架不仅在快速决策层面优势凸显，而且理论上有良好的统计相合性，实际中还增强了推断的解释性，保证了较低的假阳率(Saligrama,2012)。

在以上这些例子中，安全的空气质量指数，金融资产动态波动下的稳健风险模型以及在病人再入院巡检数据上的流行病预警指标的设计等，数据变形为参数分不同时按序、分层结构性地参与建模并影响着决策。研究人员需要综合应用场景、模型性能和数据所反馈的信号强弱，找到联结各个参变量之间的影响进而设计和建构出的模型。这样就衍生出两种建模理论：一是静态建模理论，重点关注一次性参数任务中的推断质量，如图1（左）。它的缺陷是明显的，如果模型是现成的，它对有价值的信息的表示可能是不完整的，因为它可能仅仅涵盖了系统中诸多分布中的某一类，而忽略了异质性群体有被表达的平等需要。二是长期持续更新的建模理论，这套理论指的是有能力完成一项综合性任务的参数流程的设计，它综合考虑统计模型的推断模式和训练的研究模式，遵循多种参数一体的系统推断理论。这种建模思想如图1（右）所示，以数据为中心，逐步提炼出数据中的参数信息，探索出适用于分析目标的参数并辅以数据加以巩固。模型在参数调试中持续获得更新,设计的模型也需要满足多解性以及模型动态更新的现实需要。

对于系统性建模，通常会涉及到三种参数：

1.统计类参数：数据中可靠性信息的解读，以增进理解数据并帮助尝试不同的模型空间，其中包括分布中的"显性"参数μ 和σ^2 等，也包括分布"非显性"参数，比如$F^{-1}(q), F$ 等。

2.训练类参数：在训练模型时，需要在模型的弹性空间和训练数据（或测试数据）上对参数进行估计，弹性参数也是模型调节参数。比如绘制直方图和进行密度估计计算中的带宽参数等。

3.平台类参数：为保证分析过程顺利进行而搭设的工作平台参数，例如为程序停止而设置的初始、运行和停止等参数，这些参数是程序依赖的，他们的作用是保证运算的暂时正常输出。另外，在模型训练之后部署应用前的审核参数，这些参数将综合考虑模型损失和数据偏差所带来的模型在使用中的风险，然后再考虑是否将模型部署到生产环境中。

将预设的模型用于信息提炼并使之有助于复杂的决策任务并不容易，许多参

数真实又未知，只能借助已有的数据提炼出来。传统的参数推断方法主要聚焦在统计类参数，多为特定的问题而定制，当模型适用的条件具备时通常有较好的效果。当模型设定有误或需要针对复杂情况进行调整时，仅仅考虑参数模型其估计的效率和精度则呈现明显不足。如此一来，对参数推断过程提出了两项新的范式要求：一方面放松对分布族形态的假定，增加具有稳健特征的推断理论，增强模型对数据的适应性；另一方面，尽量以数据(或个体)为中心来更新所需要的特征, 关注数据和参数特征之间的转换过程，强调递进式更新建模中数据的作用,这类“分布自由”（distribution-free）的方法称为**非参数方法**.事实上，非参数推断比“分布自由”更广泛，Bradley(1968)曾这样说，“分布自由”仅仅关注分布的表示是不是足够准确，而非参数推断则关注一个有统计分布参与的系统中哪些参数应该被作为一个科学问题来获得足够的正视。非参数方法既包括统计类参数也包括训练类参数和平台类参数。

有哪些复杂的问题需要非参数统计的辅助建模呢，我们先看以下几个问题：

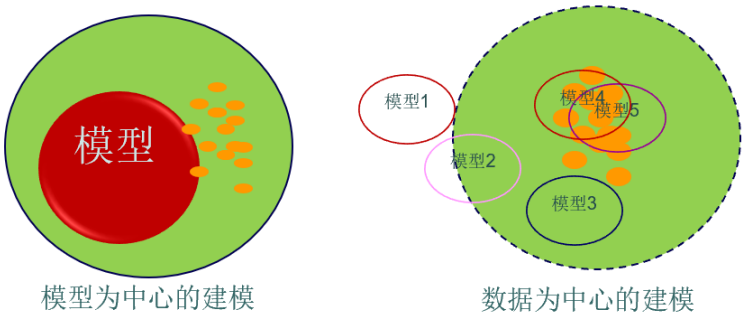

图 1.1 两种建模思想

问题1.1(见chap2\数据:HSK.txt) 在一项关于两种教学方法对留学生学习汉语水平（HSK四级Grade 4）影响是否不同的实验研究中，采用“结构法”（Type I）施教的班级的汉语水平考试平均分数为239 分，采用“功能法”（Type II）施教的班级的汉语水平考试平均分数为240 分，传统的t 检验可以帮助我们分析这个问题.但是应用t 检验的一个基本前提是两组学生的成绩服从正态分布,绘制两组数据的直方图分布, 如图1.2 所示,很难相信数据的分布是单峰对称的. 这样, 应用t 检验会有怎样的问题? 我们将在第3章里进行讨论。

问题1.2(见chap3\数据Gordon.txt) 表1.1 来自美国哈佛医学院高登(G.J. Gordon)2002 年发表在《Cancer Research》上的肺癌微阵列数据，整个数据包括181个组织样本，其中31 个恶性胸膜间皮瘤样本（MPM,Malignant Pleual Mesothelioma),和150个

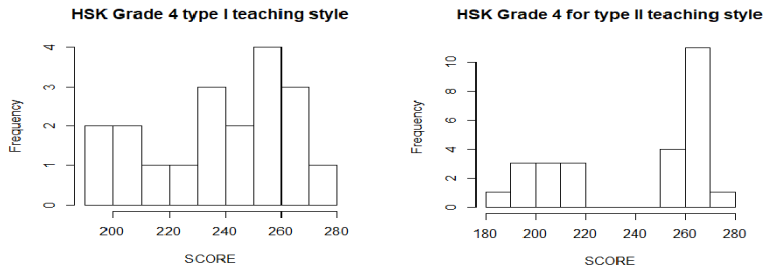

图 1.2 两种教学方法下汉语言水平考试学生成绩的直方图

肺癌样本（ADCA,Adenocarcinoma),每个样本包括12,533 条基因，表1.1 显示了其中的两组基因表现

表1.1 MPM(简称M)和ADCA(简称A)病人基因水平

病人	M1	M2	M3	M4	···	A1	A2	A3	A4	···
基因1	199.1	188.5	284.1	3.8	···	493.8	275.2	189.4	126.8	···
基因2	38.7	82.0	35.6	28.8	···	50.6	51.8	59.2	47.2	···

要检验12,533组基因的中位数水平有什么不同，应该用什么方法？如果其中只有少数几个检验的$p-$值$< \alpha = 0.05$，用什么方法可以找到这几个检验所对应的基因呢？我们将在第4章讨论该问题。

问题1.3 该数据来自Rousseeuw(1987),是有关CYG OB1星团的天文观测数据(具体见chap6\ 数据CYGOB1)，响应变量为对数光强（log light intensity 用logli表示），解释变量为对数温度（log surface temperature 用logst 表示），如图1.3制作了这两个变量的散点图，最小二乘(LS) 回归呈现出令人匪夷所思的走向，那么应该怎样估计才可以将数据中的主要模式比较准确地刻画出来，我们将在第6 章给出详细讨论。

以上这些问题, 并不总是能够在传统的参数框架结构中找到对应的答案,因为在传统的参数框架中参数与特定的分布绑定，而且估计是一次性的。而数据驱动的方法将会打破这两个限制，在数据中寻找稳定的信息特征。总而言之,非参数统计是统计学的一个分支. 相对于参数统计而言,非参数统计有以下几个突出的特点.

（1）.非参数统计方法对总体的假定相对较少, 效率高, 结果一般有较好的稳健性, 即不会因为对总体假设错误导致结论出现重大偏差. 在经典的统计框架中, 正态分布一直是最引人瞩目的, 可以刻画许多相对确定的好问题. 而数据的不确定性是复杂的，可以分为三个方面：系统内在的随机性、可见数据集的有限性和不完备

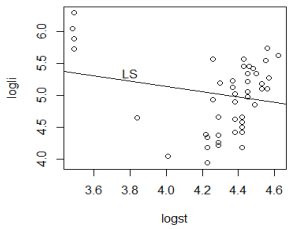

图 1.3 恒星表面对数温度和发光强度散点图与LS拟合线

的建模。正态分布并不能涵盖所有的推断兴趣, 在对探索性问题建模时，对总体做服从正态分布的假设并不总是合适的。在宽松总体假设下的推断反而可获得更为可信的结论，以“找形”来获得对数据真实的理解是非参数推断区别于参数推断的一个特征。

（2）.非参数统计可以处理多种类型的数据，许多统计量由秩序型、计数型和评分型数据合成，追求在总体宽松的假设下获得稳健的估计. 我们知道,统计数据按照数据类型可以分为两大类：分类数据(包括类别数据和顺序数据),连续数据(包括等距数据和比例数据). 拿检验来说, 一般而言,参数统计中常用的是正态型数据,其在理论上容易得到较好的结果, 然而在实践中,分布不符合正态假定的数据非常多. 比如：在满意度调查中数据只有顺序, 没有大小, 这时很多流行的参数模型无能为力。而连续数据如果过度测量则会产生大量累积测量误差，测量误差会导致对分布的判断失真, 统计推断就会失效。这个时候将连续数据转化为顺序数据或定性数据, 表面上看损失了一些信息，但丢掉噪声干扰的丢弃是有价值的：不仅可以消除测量误差的影响，而且降低了量纲的影响，增强了数据的可比性，推断结果更稳健.

（3）.非参数统计思想容易理解, 易于计算。作为统计学的分支, 非参数统计统计思想非常深刻, 其方法继承了参数统计推断的理论, 容易发展成算法. 特别是伴随着计算机技术的发展, 近代非参数统计更强调运用大量计算求解问题, 这些问题很容易通过编写程序求解, 计算结果也更容易解释. 非参数统计方法在处理小样本问题时, 可能涉及到一些不常见的统计表, 过去会对一些非专业的使用者造成不便。现如今很多统计软件, 如R 软件中已存储了现成的统计表供人们计算和使用, 一些统计量的精确分布或近似分布都可以从软件中轻松地获取, 取代了以往编制粗糙且不精确的表.而事实上，在现代的许多参数统计推断中，在诊断过程、图形展现和拟合优度检验方面使用了大量的非参数统计量。

当然, 非参数方法也有一些弱点, 比如当人们对总体有充分的了解且足以确定其分布类型时, 非参数方法就不如参数方法更具有针对性, 且有效性可能也会差一

些.这样来看，非参数统计并非要取代参数统计, 而是继承和发展了参数统计。

2. 非参数统计的历史

在早期形成阶段，非参数统计强调与分布无关（distribution free）的思想。最早有关非参数统计推断的历史记载是1701 年苏格兰数理统计学家约翰·阿巴斯诺特(John Arbuthnot) 提出了单样本符号检验。1900年卡尔·皮尔逊(Karl Pearson)提出了列联表和拟合优度检验等适用于计数类型数据分布的检验方法。同一时期的1904 年，心理学家查尔斯·斯皮尔曼(Charles Spearman) 提出了Spearman等级相关系数检验方法，1937 年弗里德曼(Friedman) 提出了可用于区组实验设计的Q检验法，随后肯德尔(Maurice Kendall) 提出了τ 相关系数检验方法，1939 年斯米尔诺夫(Smirnov) 还提出来了著名的K-S正态性检验，同时期费歇尔·欧文(Fisher Erwin) 提出Fisher 精确性检验作为χ^2 独立性检验的补充。一般认为, 非参数统计概念形成于20 世纪40∼50 年代之间，其中化学家威尔科克森(F.Wilcoxon) 做出了突出贡献，1945 年威尔科克森提出两样本秩和检验,1947 年亨利·B·曼(Henry B. Mann) 和D·兰塞姆·惠特尼(D. Ransom Whitney) 将结果推广到两组样本量不等的一般情况,而1975 年班贝尔(Bamber) 发现ROC 曲线面积与Mann-Whitney 统计量之间的天然等价关系。

继威尔科克森之后的50∼60 年代,多元位置参数的估计和检验理论相继建立起来,这些理论极大地丰富了实验设计不同情况下的数据分析方法。1950 年，科克伦(Cochran) 补充了分类数据的Q检验法。1951 年，布朗和沐德(Brown & Mood)提出中位数检验法。德宾(Durbin) 提出均衡不完全区组实验设计方法。1952 年，克鲁斯卡尔(Kruskal) 和沃利斯(Wallis) 提出KW 秩检验, 麦金太尔(Mclntyre)提出排序集抽样方法用以提取更有效的总体信息。1958 年，布洛斯(Bross)提出了非参数Ridit 检验法。1960 年，科恩(Cohen) 提出Kappa 一致性检验。这些方法在小样本检验和异常数据诊断方面获得了成功的应用。1948年，皮特曼(Pitman)回答了非参数统计方法相对于参数方法来说的效率问题。1956 年,霍吉斯(J.L.Hodges)和莱曼(E.L.Lehmann) 则发现了一个令人吃惊的结果, 与正态模型中的t 检验相比,秩检验能经受住有效性的较小损失.而对于厚尾分布所产生的数据, 秩检验统计量可能更为有效.第一本论述非参数统计的著作《非参数统计》于1956 年由西格尔(S.Siegel)出版。汤姆斯海特曼斯波格(Thomas P.Hettmansperger) 所著的《基于秩的统计推断》(Statistical Inference Based on Ranks)一书在1956∼1972年被引用了1824 次.

20世纪60年代, 霍吉斯(J.L.Hodges)和莱曼(E.L.Lehmann)从秩检验统计量出发，推导出了若干估计量和置信区间，以HL 估计量和Theilsen估计量为代表，由检验统计量推导出估计量引发了推断理论的一次新的变迁(参见文献莱曼(E.L.Lehmann,1975), 沃尔夫(Wolfe,1999))。约翰·图基(John Tukey,1960) 较早地注意到传统估计量的不稳健性和效率低下的问题，打开了数据分析的大门.之后, 非参数统计的应用和研究

获得巨大的发展,其中较有代表性的是20 世纪60 年代中后期，考克斯(D.R.Cox)和弗古森(Ferguson)最早将非参数方法应用于生存分析. 1930∼1970 年非参数统计体系大厦得以建立。约翰·渥施(John Walsh)分别于1962，1965和1968年相继出版了一部三卷有关非参数方法的指南。萨维奇(I.R.Savage) 于1962 年编纂了一部有关非参数统计的文献志。20 世纪六七十年代比较流行的两本教材有詹姆斯·布拉德利(James V.Bradley) 于1958年出版的《不依赖于分布的统计检验》(Distribution-free Statistical Tests) 以及吉本斯(Jean D. Gibbons)于1970 年编纂的《非参数统计推断》(Non-parametric Statistical Inference). 而这段时期恰恰是数据科学的萌芽期，由于不同学科之间还没有形成很厚的壁垒，很多统计学家实际上一生都在从事着对其他学科的研究，他们对于其他领域的眼界十分开阔。这段时间是传统统计通往机器学习的过渡期，也是整个非参数话语体系正式形成期。他们在解决化学、生物、心理等快速发展领域中现实问题的过程中发展出一种全新的数据分析理论，这些方法一边借着参数推断已形成的渐进工具阐释优良性理论，同时通过成熟的分布表技术推广方法的应用，发展存在于数据本身的“秩序”、“稳健性”、“有效性”和“局部表示”等潜在特征。这些统计方法在当时的推断文化中看似不具有核心话语权。但是随着信息技术的发展，却以见微知著的独特力量，连接着数据分析的传统与未来。

进入20世纪70∼80年代, 继埃夫龙(Efron)1979年提出自助法(Bootstrap)方法之后,非参数方法借助计算机技术和大量计算提出大量的稳健估计和预测方法,比如，置换检验(Permutation Tests) 和多重检验等在包括生物医学等诸多领域取得长足发展. 以休伯(P.J.Huber) 和汉佩尔(F.Hampel) 为代表的统计学家从实际数据出发,为衡量估计量的稳健性提出了新准则. 20世纪90 年代非参数统计将其早期的检验和估计优势扩展到非参数回归领域,典型的方法有核方法(kernel)、样条(spline)及小波(wavelet), 相关文献有欧班克(Eubank,1988), 哈特(Hart,1997), 瓦博(Wahba,1990), 格林和斯沃曼(Green&Silverman,1994),范建青和吉贝尔(Fan&Gijbels, 1995), 哈尔德(Härdle,1998), 大卫·多诺霍和约翰斯通(I. M. Johnstone&D.L.Donoho,1994)。20世纪90年代以后, 算法建模思想发展飞快, 成为非参数统计的新宠儿。非参数统计借助其独有的化繁为简的灵活性能，在半参数模型、模型（变量）选择和降维方法中显出巨大优势并成为大尺度统计推断中的领跑者，代表性的成果有切夫·黑斯帖和罗伯特·提布施拉尼(Trevor Hastie&Robert Tibshirani,1990)，丹尼斯·库克和李冰((R.Dennis Cook&Bing Li,2002)，范建青和李润泽(Fan&Li,2001) 以及大卫·多诺霍和金加顺(D.Donoho&J.Jin2004,2015)。弗拉基米尔·万普尼克(Vladimir Naumovich Vapnik,1974) 等从结构风险的角度规范了面向预测的模型选择框架. 里奥·布莱曼Leo Brieman(1984;2001) 等为数据驱动文化敞开了大门。大规模计算和自动化技术的飞速发展，非参数统计不仅为机器学习输送了大量的新方法,其中的

统计推断与机器学习彼此渗透，相互叠加，共同推动数据科学的进展.

§1.2 假设检验回顾

假设检验问题是统计推断和决策问题的基本形式之一，其核心内容是利用样本所提供的信息对关于总体的某个假设进行检验。相对于探索型数据分析，假设检验是典型的推断型数据分析。基本的假设检验是从两个相互对立的命题即假设开始的：原假设和备择假设. 对这两个相互对立的假设而言,一般还要假设分布族和数据, 比如假设分布族是正态的，那么对总体的选择就可以简化为对位置参数或形状参数的选择. 假设一般都以参数的形式出现,记作θ. 原假设记作$H_0 : \theta \leqslant \theta_0$；备择假设记作$H_1 : \theta > \theta_0$. 当然, 这里给出的是一个常规的单边检验问题. 类似地,如果猜测是另一个方向的或无倾向性的,则有单边检验问题($H_1 : \theta < \theta_0$) 或双边检验问题($H_1 : \theta \neq \theta_0$). 假设检验的基本原理是小概率事件在一次试验中是不会发生的。如果在H_0成立的时候，一次试验中某个小概率事件发生了，则表明原假设H_0 不成立。从某种意义上说，假设检验的过程很类似于数学中的反证法。

在假设检验的理论框架形成的过程中，有一个著名的假设检验故事—女士品茶试验。穆里尔 • 布里斯托(Muriel Bristol) 博士是著名统计学家费歇尔.R.A.(Ronald Fisher) 在试验站的同事，她声称自己能够通过奶茶的口味判断奶茶中里是先加的奶还是先加的茶。显然，一般人很难察觉出这种细微的口感差别。为了验证该女士的正确性，费希尔设计了一个试验，他预备了8 杯奶茶，其中4 杯先加茶后加奶，另外4杯是先加奶后加茶。将8 杯奶茶的顺序随机打乱，布里斯托博士对哪些奶茶是先加的茶哪些奶茶是先加的奶并不知道。原假设H_0 是布里斯托博士不能通过奶茶口味成功分辨出奶和茶加入的先后顺序。如果原假设成立，而布里斯托博士猜对了全部奶茶简爱的顺序，这就等价于布里斯托博士完全靠猜的方式分辨出全部奶茶，可以计算得到她全部猜对的可能性是$\frac{1}{70} = 0.014$,（为什么是这个结果，可以通过思考题来回答）。这是一个非常小的概率，表示如果加奶的先后顺序对于判断没有影响，随机猜全部答对的可能性几乎是0，而从史料记载来看，布里斯托全部答对，那么她“没有任何分辨能力”这个假设就与数据的客观结果不相容，于是可以拒绝这个假设，在显著性水平为5% 的情况下,拒绝原假设，统计上呈现显著结果。

在假设检验的基本原理中，有个限定是“一次试验”，而并非重复试验。如果试验重复很多遍，即便是小概率事件，无论它的概率多么小总会发生，这是著名的墨菲定律。

假设检验的基本原理是, 先假定原假设成立, 样本被视为通过合理设计所获得的总体的代表. 一旦总体分布确定, 那么统计量的抽样分布也就确定了, 从而理论上样本应该体现总体的特点,统计量的值应该位于其抽样分布的中心位置附近, 不

会距离中心位置太远. 这显然是原假设成立的一个几乎必然的结果,就像在理想环境下投一枚均匀硬币100 次, 正面和负面出现的次数应近乎相等,因为这是在硬币均匀假设前提下几乎必然的抽样结果. 然而, 假设检验里硬币的正负面胜算的真值是未知的, 在一次固定投币次数的实验中,当发现硬币正面和负面呈现的次数之间有较大差异, 一种直觉是硬币不均匀. 用逆否命题进行推断是假设检验的本质. 当然, 正负差大到多少才可以认为硬币是不均匀的，需要测量样本远离中心位置程度大小的一个工具,如果样本量的值偏离抽样分布的中心位置过远,则从小概率原理很难发生的统计观点出发,认为有很大的把握认为这个试验是从假定总体中取得的,几乎必然地认为这些样本与备择命题更匹配,从而拒绝数据对原假设的支持, 接受数据对备择假设的支持.“过远”是一个统计的概念, 在假设检验中用显著性衡量. 几乎必然的含义是,虽然拒绝原假设的依据是样本偏离了原假设的分布, 然而在零假设下产生特殊样本的可能性和随机性却是存在的, 承认差距存在并不表示判断是绝对准确的, 随机性的发生不可避免, 但是如果样本超出了假设理论分布可以允许的边界,则可以认为样本呈现出的差异性已经超出了随机性可以解释的范围,这种差异是由于数据与假设分布的不同而导致的必然结果. 所以，假设检验的实质是对数据来源的分布做比较，当某一种分布相对于另一种分布而言产生数据的可能性更大，就可以生成一种检验的标准，这就是Neymman-Pearson 引理的核心思想。

一般而言，对假设检验问题而言, 讨论以下4组基本概念：

1. 如何选择原假设和备择假设：在数学上,原假设和备择假设没有实际含义,形式对称,采取接受或拒绝结论也是对称的.但在实践中,检验的目的是试图将样本中表现出来的特点升华为更一般的分布或分布的特点,是部分数据特征能否推广至整体分布的过程. 因而,如果所建立的猜想与样本的表现相背离, 则这个推断的过程基本上是“空想”,也就是说与数据的支持不相符.这样的假设检验问题是没有意义的, 当然也不可能期待有拒绝假设检验的结果, 参见习题1.1. 假设不应该是随意设定的, 而是应该根据数据的表现来设定的.如果数据背离理想的抽样分布, 从小概率原理来看,提出了可能拒绝原假设的证据, 接受备择假设,认为是分布上的差异导致了样本对原假设分布的偏移. 因此, 通常将样本显示出的特点作为对总体的猜想, 并优先被选作备择假设.与备择假设相比, 原假设的设定则较为简单,它是相对于备择假设而出现的. 如此建立在实践经验基础上的假设才是有意义的假设.

2. 检验的$p-$值和显著性水平的作用：从假设检验的整个过程来看,起关键作用的是和检验目的相关的检验统计量$T = T(X_1, X_2, \cdots, X_n)$ 和在原假设之下检验统计量的分布情况. 原假设下统计量的分布是已知的,这样才能通过统计量判断数据是否远离了原假设所支持的参数分布。以单一总体正态分布均值μ是否等于μ_0 的检验来看，选择检验统计量$T = \frac{\bar{X}-\mu_0}{\sqrt{\sigma^2/\sqrt{n}}}$, 如果$|T|$ 大意味着备择假设$\mu \neq \mu_0$ 的可能性

更大,那么就要计算概率$P_{H_0}(|T| > t_0), t_0 = \frac{\bar{X}-\mu_0}{\sqrt{\sigma^2/\sqrt{n}}}$; 这个概率称为检验的$p-$值.如果$p-$ 值很小,这说明统计量反映出样本在原假设下是小概率事件, 这时如果拒绝原假设, 则决策错误的可能性是非常小的, 等于$p-$值,这个错误称为第 I 类错误. 通常情况下, 统计计算软件都输出$p-$ 值. 传统意义上, 一般先给出第 I 类错误的概率α,称它为检验的显著性水平, 如果检验的显著性水平$\alpha > p$, 那么拒绝原假设, $p-$值可以认为是拒绝原假设的最小的显著性水平. 对于双边检验, $p-$ 值是双边尾概率之和, 是单边检验p 值的2 倍. p值的概念如图1.4和图1.5所示. 如果是一笔样本，同一个显著性水平，双边检验是更不易拒绝的,如果能够拒绝双边检验, 则更能拒绝单边检验, 但反之不对.p-值真正的作用是用来测量数据和原假设不相容的可能性，是原假设为真时获得极端结果的可能性。

3. 两类错误：只要通过样本决策,就不可避免真实情况和数据推断不一致的情况发生, 此时会犯决策错误. 在假设检验中, 有可能犯两类错误.

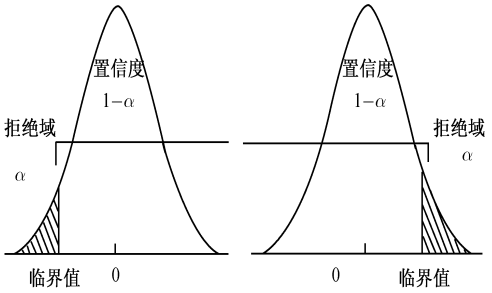

图 1.4 单边检验的p值

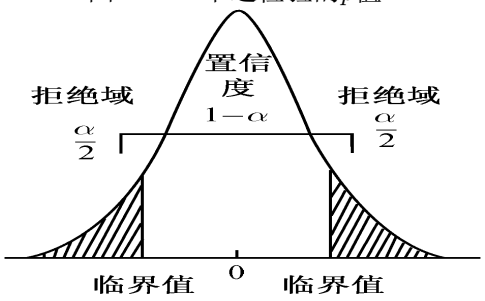

图 1.5 双边检验的p值

当拒绝原假设而实际的情况是原假设为真时,犯第 I 类错误,这个错误一般由事先给出描述数据支持的命题和原假设差异显著性的α控制, 这表示拒绝原假设时出现决策错误的可能性不会超过α, 因此拒绝原假设的决策可靠程度较高；当原假设不能

被拒绝,而实际情况是备择假设为真时, 犯第Ⅱ类错误,此时表现为在原假设下样本统计量的$p-$值较大.当不能拒绝原假设时, 如果选择接受原假设, 则会出现取伪错误.假设检验的目的是给出临界值用于决策, 一个好的决策应该尽量让犯两类错误的概率都小, 然而这在很多情况下是不现实的,因为在理论上, 犯第I, II类错误的概率彼此之间相互制衡, 不可能同时很小.为了度量犯两类错误的概率, 定义势函数如下：

定义1.1(检验的势)　对一般的假设检验问题：$H_0:\theta\in\Theta_0\leftrightarrow H_1:\theta\in\Theta_1$, 其中$\Theta_0\bigcap\Theta_1=\varnothing$, 检验统计量为$T_n$. 拒绝原假设的概率, 也就是样本落入拒绝域$W$的概率为检验的**势**, 记为

$$g_{T_n}(\theta)=P(T_n\in W),\quad \theta\in\Theta=\Theta_0\bigcup\Theta_1.$$

由定义1.1可知, 当$\theta\in\Theta_0$时, 检验的势是犯第Ⅰ类错误的概率, 一般由显著性水平α 控制；当$\theta\in\Theta_1$时, 检验的势是不犯第Ⅱ类错误的概率, $1-g(\theta)$是犯第Ⅱ类错误的概率. 我们用势函数将犯两类错误的概率统一在一个函数中。一个有意义的检验, 势函数理论上应该越大越好, 低势的检验说明检验在区分原假设和备择假设方面的价值不大.

1933年J.奈曼(Neyman Jerzy)和E.S.皮尔逊(Egon Pearson)提出了著名的Neyman-Pearson 引理。考虑两个简单检验问题：$H_0:\theta=\theta_0$; $H_1:\theta=\theta_1$. 记$f_0(x)$ 和$f_1(x)$分别对应着随机变量X 在H_0 和H_1 下的密度函数，$X\in(\mathcal{X},\mathcal{F})$, 有$\int_{\mathcal{X}}f_i(x)\mathrm{d}x=1,(i=0,1)$。Neyman－Pearson 引理要表达的是,如果对水平α, 存在$W=W_\alpha\subset\mathcal{X}$ 和W上的似然比

$$W_\alpha=\left\{x:\frac{f_1(x)}{f_0(x)}\geqslant k_\alpha\right\},k_\alpha\geqslant 0$$
$$W_\alpha^c=\left\{x:\frac{f_1(x)}{f_0(x)}<k_\alpha\right\}$$

那么,似然比检验是简单检验问题水平为α 的一致最优势检验,k_α 满足$P(x\in W(\alpha))=\alpha$。

证明：　令W_α满足$P(x\in W_\alpha)=\alpha$,记W'为另一个水平为α 的检验拒绝域。那么，对任意一个密度函数$f(x)$，

$$\begin{aligned}&\int_{W_\alpha}f(x)\mathrm{d}x-\int_{W'}f(x)\mathrm{d}x\\=&\int_{W_\alpha\cap W'}f(x)\mathrm{d}x+\int_{W_\alpha\cap W'^c}f(x)\mathrm{d}x-\int_{W'\cap W_\alpha}f(x)\mathrm{d}x-\int_{W'\cap W_\alpha^c}f(x)\mathrm{d}x\\=&\int_{W_\alpha\cap W'^c}f(x)\mathrm{d}x-\int_{W'\cap W_\alpha^c}f(x)\mathrm{d}x\end{aligned}\tag{1.1}$$

第一种情况，如果$f = f_0$，上述表达式一定非负，因为$W^{'}$对应的假设检验的水平不会超过α ，而W_α 满足$P(x \in W_\alpha) = \alpha$。也就是说，

$$\int_{W_\alpha} f_0(x)\mathrm{d}x \geqslant \int_{W'} f_0(x)\mathrm{d}x \tag{1.2}$$

这意味着，

$$\int_{W_\alpha \cap W'^c} f_0(x)dx \geqslant \int_{W' \cap W_\alpha^c} f_0(x)\mathrm{d}x \tag{1.3}$$

第二种情况，如果$f = f_1$，当$x \in \mathcal{X}_\alpha$时，有$f_1(x) \geqslant k_\alpha f_0(x)$。因此，

$$\int_{W_\alpha \cap W'^c} f_1(x)\mathrm{d}x \geqslant k_\alpha \int_{W_\alpha \cap W'^c} f_0(x)\mathrm{d}x, \quad k_\alpha \int_{W' \cap W_\alpha^c} f_0(x)\mathrm{d}x \geqslant \int_{W' \cap W_\alpha^c} f_1(x)\mathrm{d}x \tag{1.4}$$

上述两个结果表明，无论$f = f_0$还是$f = f_1$，

$$\int_{W_\alpha} f(x)\mathrm{d}x - \int_{W'} f(x)\mathrm{d}x = \int_{W_\alpha \cap W'^c} f(x)\mathrm{d}x - \int_{W' \cap W_\alpha^c} f(x)\mathrm{d}x \geqslant 0 \tag{1.5}$$

事实上，上述推理证明的是Neyman-Pearson引理的的充分性条件，说明了似然比检验是一致最优势检验，Neyman-Pearson 引理的必要条件是，在简单原假设对简单备择假设的情形，最优势检验一定是似然比检验，此处不再赘述，详细的证明过程可参见文献George Casella, Roger L.Berger(2002)。这表明似然比可以用于构造一致最优势检验。

下面通过一个单边检验的问题观察势函数的特点.

例1.1 假设总体X来自Poisson(泊松)分布$\mathcal{P}(\lambda)$, 简单随机抽样$X_1, X_2, \cdots, X_n$, 假设检验问题$H_0 : \lambda \geqslant 1 \leftrightarrow H_1 : \lambda < 1$.根据假设检验的步骤, 可以选取充分统计量$\sum_{i=1}^{n} X_i$为检验统计量, 检验的目的是选择使犯第 I 类错误的概率较小的检验域, 即使$\alpha(\lambda) = P\left(\sum_{i=1}^{n} X_i < C\right)$ 足够小. 可以看出，$\alpha(\lambda)$是分布的函数. 我们在样本量$n = 10$时, 对$C = 5$和$C = 7$考虑了检验势函数随分布的参数λ_0从0变化到2的情况. 在原假设下, 我们注意到检验

$$\alpha(\lambda) = P(\text{拒绝原假设}|\text{原假设为真}) = P\left(\sum_{i=1}^{n} X_i < C|\lambda \in H_0\right),$$

$$\beta(\lambda) = 1 - P(\text{拒绝原假设}|\text{备择假设为真}) = 1 - P\left(\sum_{i=1}^{n} X_i < C|\lambda \in H_1\right).$$

检验犯各类错误的概率随分布参数的变化曲线如图1.6所示.

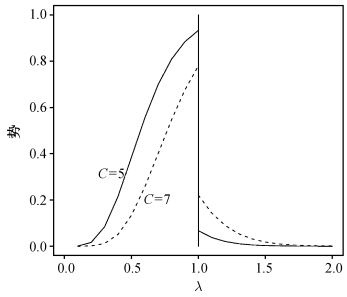

图 1.6 检验势函数随分布参数变化曲线

在图1.6中,右侧的两条曲线分别是$C = 5$(实线)和$C = 7$(虚线)时犯第Ⅰ类错误的概率曲线, 我们观察发现犯第Ⅰ类错误的概率在原假设下随着λ的增加而减小, 犯第Ⅰ类错误的概率在$\lambda = 1$ 处达到最大,这与Neyman-Pearson 引理体现的控制第Ⅰ类错误在边界分布上达到最大的思想是一致的. 其中$C = 5$的检验第Ⅰ类错误概率比$C = 7$的检验第Ⅰ类错误的概率低, 这是因为$C = 5$比$C = 7$的检验更倾向支持备择假设. 两个检验犯第Ⅱ类错误的概率在图像的左侧, 随着λ的减小而减小, 犯第Ⅱ类错误的概率在$\lambda = 1$处达到最大. 在单边检验中, 真实的λ越远离临界分布1, 犯第Ⅱ类错误越小.

上面的两个例子都说明, 即便β是可以计算的, 当α很小时, β 也可能很大. 也就是说, 如果做接受原假设的决策, 则可能存在着很大的潜在决策风险, 比如当参数的真值（比如λ）和要比较的参考值（比如λ_0）比较接近时更应该尽量避免接受原假设. 实际上,不能拒绝原假设的原因很多,可能是证据不足, 比如样本量太少,也可能是模型假设的问题, 也可能是检验效率低, 当然也包括原假设本身就是对的.

结合Neyman-Pearson引理，我们看到，如果将假设检验当作两类分布的分类问题的话，拒绝域通过设置临界值定义了一个决策。这个决策当样本落入拒绝域选择拒绝原假设选择备择假设；当样本没有落入拒绝域，选择原假设。当真实参数在备择空间里，样本落入到接受域时，有一片区域的错误概率是非常高的，这相当于这个决策失效的区域，类别在这个区域的归属相对不确定。如果真值落在这些区域里，避免做出决策是更合适的选择，这个区域被称为弃权区（reject option），表示决策豁免。这个拒绝选项区有多大，受什么因素影响，将在第4章中进行更详细的讨论。和势有关的检验问题, 我们将在1.3 节中详细介绍.

4.置信区间和假设检验的关系

以单变量位置参数为例来说,置信区间和双边检验有密切的联系. 比如，有参数θ的估计量$\hat{\theta}$, 用$\hat{\theta}$构造一个θ的$100(1-\alpha)\%$ 置信区间如下：

$$(\hat{\theta} - C_\alpha,\ \hat{\theta} + C_\alpha). \tag{1.6}$$

这是数据所支持的总体(参数)可能的取值范围, 这个区间的可靠性为$100(1-\alpha)\%$. 如果猜想的θ_0不在该区间内, 则可以拒绝原假设, 认为数据所支持的总体与猜想的总体不一致. 当然, 由于区间端点取值的随机性, 也可能因为一次性试验结果的偶然性而犯错误。犯错误的概率恰好是区间不包含总体参数的可能性α. 反之, 如果θ_0在区间中, 则表示不能拒绝原假设, 但是这并没有表明θ就是θ_0,而仅仅表达了不拒绝θ_0.从这一点来看, 置信区间和假设检验虽然对总体推断的角度不同, 但推断的结果却可能是一致的.

§1.3 经验分布和分布探索

§1.3.1 经验分布

一个随机变量$X \in \mathbb{R}$的分布函数(右连续)定义为：$F(x) = P(X \leqslant x), \forall x \in \mathbb{R}$. 对分布函数最直接的估计是应用经验分布函数. 经验分布函数的定义是：当有独立随机样本$X_1, X_2, \cdots, X_n$时,对$\forall x \in \mathbb{R}$, 定义

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leqslant x). \tag{1.7}$$

这里$I(X \leqslant x)$是示性函数：

$$I(X \leqslant x) = \begin{cases} 1, & X \leqslant x, \\ 0, & X > x. \end{cases}$$

如果对$\forall i = 1, 2, \cdots, n$,定义伴随变量$Y_i = I(X_i \leqslant x), Y_i$ 服从贝努利分布$B(1, p)$,.除此之外，还可以定义一个离散型随机变量Z,是在$\{x_1, x_2, \cdots, x_n\}$ 上均匀分布的随机变量，Z的分布函数就是$\hat{F}_n(x)$.

定理1.1 令$X_1, X_2, \cdots, X_n$的分布函数为F, $\hat{F}_n$为经验分布函数, 于是有以下结论成立：

(1) 对$\forall x, E(\hat{F}_n(x)) = F(x), \mathrm{var}(\hat{F}_n(x)) = \dfrac{F(x)(1-F(x))}{n}$; 于是, MSE $= \dfrac{F(x)(1-F(x))}{n} \to 0$,而且$\hat{F}_n(x) \xrightarrow{P} F(x)$.

(2) (Glivenko-Cantelli 定理) $\sup\limits_x |\hat{F}_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0$.

(3) (Dvoretzky-Kiefer-Wolfowitz(DKW) 不等式) 对$\forall \varepsilon > 0$,

$$P(\sup_x |\hat{F}_n(x) - F(x)| > \varepsilon) \leqslant 2\mathrm{e}^{-2n\varepsilon^2}. \tag{1.8}$$

由DKW不等式, 我们可以构造一个置信区间. 令$\varepsilon_n^2 = \ln(2/\alpha)/(2n), L(x) = \max\{\hat{F}_n(x) - \varepsilon_n, 0\}, U(x) = \min\{\hat{F}_n(x) + \varepsilon_n, 1\}$, 根据式(1.3)可以得到

$$P(L(x) \leqslant F(x) \leqslant U(x)) \geqslant 1 - \alpha.$$

也就是说, 可以得到如下推论.

推论1.1 令

$$L(x) = \max\{\hat{F}_n(x) - \varepsilon_n, 0\}, \tag{1.9}$$

$$U(x) = \min\{\hat{F}_n(x) + \varepsilon_n, 1\}, \tag{1.10}$$

其中

$$\varepsilon_n = \sqrt{\frac{1}{2n}\ln\left(\frac{2}{\alpha}\right)},$$

那么

$$P(L(x) \leqslant F(x) \leqslant U(x)) \geqslant 1 - \alpha.$$

例1.2 1966年Cox和Lewis的一篇研究报告给出了神经纤维细胞连续799次激活的等待时间的分布拟合. 求数据的经验分布函数, 可以编写程序, 也可以调用函数ecdf. 我们根据定理1.1编写了函数求解经验分布函数的95%的置信区间, 程序如下:

```
data(nerve)
nerve.sort=sort(nerve)
nerve.rank=rank(nerve.sort)
nerve.cdf=nerve.rank/length(nerve)
plot(nerve.sort,nerve.cdf)
N=length(nerve)
segments(nerve.sort[1:(N-1)], nerve.cdf[1:(N-1)],
+ nerve.sort[2:N], nerve.cdf[1:(N-1)])
alpha=0.05
band=sqrt(1/(2*length(nerve))*log(2/alpha))
Lower.95= nerve.cdf-band
Upper.95= nerve.cdf+band
lines(nerve.sort,Lower.95,lty=2)
lines(nerve.sort,Upper.95,lty=2)
```

如图1.7所示, 分段左连续函数即为经验分布函数, 上下两条虚线分别是95%上下置信限.

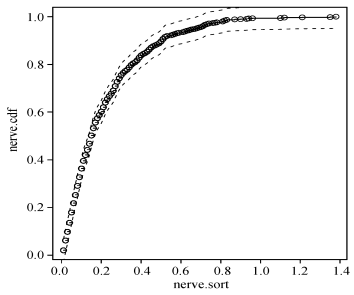

图 1.7 经验分布函数及分布函数的置信区间变化曲线

§1.3.2 生存函数

很多实际问题关心随机事件的寿命, 比如零件损坏的时间、病人生病的生存时间等, 这时需要用生存分析来回答. 生存函数是生存分析中基本的概念, 它是用分布函数来定义的：

$$S(t) = P(T > 0) = 1 - F(t),$$

其中, T是服从分布F的随机变量. 这里, 我们更习惯于用生存函数而不是累积分布, 尽管两者给出同样的信息. 于是, 可以用经验分布函数估计生存函数：

$$S_n(t) = 1 - F_n(t),$$

表示寿命超过t的数据占的比例.

例1.3(数据见chap1\pig.rar) 数据来自受不同程度结核病毒感染的几内亚猪的死亡时间. 其中实验组分为五组, 每组安排72只猪, 组内受同等程度结核病毒感染. 1∼5组感染病毒的程度依次增大, 标记为1, 2, 3, 4, 5. 对照组包含107只猪, 没有受到感染. 对这些试验观察两年以上, 记录猪死亡时间. 这个例子中, 我们用经验分布函数估计生存函数, 研究受不同程度结核病毒感染的几内亚猪的生存情况. 其生存函数如图1.8所示.

粗实线对应于对照组, 其他的线（细实线和虚线）从上到下分别为1∼5的实验组, 图1.8中的经验生存函数直观地描述了受感染猪的生存情况. 超过规定时间的存活比率在图1.8中表现出来, 可以看出：随着病毒剂量的增加, 猪的寿命有很大程度的下降, 第5组猪的寿命和第3组寿命相比几乎差了100天. 该图比列表更有效地展示了数据.

危险函数是生存分析中另一项重要内容,它表示一个生存时间超过给定时间的个体瞬时死亡率. 生存图形可以非正式地表现危险函数. 如果一个个体在时刻t仍

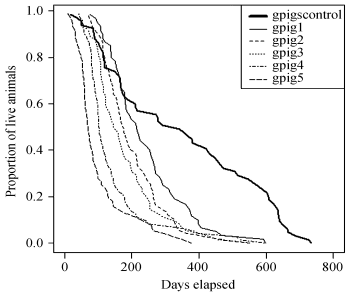

图 1.8 几内亚猪经验生存函数

然存活, 那么个体在时间范围$(t, t+\delta)$死亡的概率为(假设密度函数f在t上是连续的)

$$\begin{aligned} P(t \leqslant T \leqslant t+\delta | T \geqslant t) &= \frac{P(t \leqslant T \leqslant t+\delta)}{P(T \geqslant t)} \\ &= \frac{F(t+\delta) - F(t)}{1 - F(t)} \\ &\approx \frac{\delta f(t)}{1 - F(t)}. \end{aligned}$$

危险函数定义为

$$h(t) = \frac{f(t)}{1 - F(t)}.$$

$h(t)$是一个存活时间超过规定时间的个体瞬时死亡率. 如果T是一个产品零件的寿命, $h(t)$可以解释成零件的瞬时损坏率. 危险函数还可以表示为

$$h(t) = -\frac{\mathrm{d}}{\mathrm{d}t} \ln[1 - F(t)] = -\frac{\mathrm{d}}{\mathrm{d}t} \ln S(t).$$

上式说明危险函数是对数生存函数斜率的负数.

考虑一个指数分布的例子:

$$\begin{aligned} F(t) &= 1 - \mathrm{e}^{-\lambda t}, \\ S(t) &= \mathrm{e}^{-\lambda t}, \\ f(t) &= \lambda \mathrm{e}^{-\lambda t}, \\ h(t) &= \lambda. \end{aligned}$$

如果一个零件的损坏时间服从指数分布, 由于指数分布的“无记忆”特性, 零件损坏的可能性不依赖于它使用的时间, 但这不符合零件损坏的规律. 一个合理的

零件损坏时间分布应该是：它的危险函数是U形曲线. 新零件刚开始损坏的概率(危险函数值)较大, 因为制造过程中一些缺陷在使用之初会很快暴露出来. 然后危险函数值会下降. 当用到一段时间后, 零件老化, 危险函数值会再度上升. 这个过程体现了危险函数的作用.

我们还可以计算对数经验生存函数的方差：

$$\begin{aligned}\mathrm{var}\{\ln[1-F_n(t)]\} &\approx \frac{\mathrm{var}[1-F_n(t)]}{[1-F(t)]^2}\\ &= \frac{1}{n}\frac{F(t)[1-F(t)]}{[1-F(t)]^2}\\ &= \frac{1}{n}\frac{F(t)}{[1-F(t)]}.\end{aligned}$$

例1.4 对上例中的数据, 图1.9展现的是负对数经验生存函数随时间t的变化. 从曲线的斜率我们可以看到危险函数随时间t 的变化。开始危险率是比较小的, 随着剂量的增加，猪的死亡率增加得很快。而且就早期危险率而言，高剂量组的相比于低剂量组增加得更快. 从图上可以看出, 当t 值很大时,负的对数经验生存函数会变得不稳定, 因为此时$1-F(t)$的值变得很小. 所以画图时, 每组的最后几个点被忽略了。

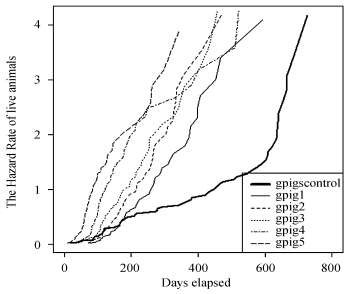

图 1.9 几内亚猪负对数经验生存函数随时间t的变化情况

§1.4 检验的相对效率

正如1.3节所述, 一个好的检验, 在达到犯I类错误α的水平下，势应该越大越好, 当对一个检验问题有许多检验可以选择时, 用怎样的标准选择检验函数是一个自然的问题. 这一节将给出选择检验函数的一些理论评价结果.

对同一个假设检验问题而言, 选择不同的统计量, 得到的势函数也不同. 一般一个好的检验应有较大的势, 因而可以通过比较势大小选择较优的检验. 然而直接比较势是困难的,转而考虑影响势大小的因素：总体的真值、检验的显著性水平和样本量. 在这些因素中, 真值未知对我们的帮助不大, 在显著性水平固定的情况下, 势的大小依赖于样本量, 样本量越大势越大。考虑势的大小问题可以转化为对样本量的比较：在相同的势条件下, 比较不同检验所需要的样本量的大小, 样本量越小的检验认为是更优的统计量, 于是依赖于该统计量所做出的检验也认为是较优的或是更有效率的. 渐近相对效率(Asymptotic Relative Efficiency, 简称ARE)给出了该问题的一个可行的答案, Pitman渐近相对效率是ARE的代表. 针对原假设只取一个值的假设检验问题, 在原假设的一个邻域内, 固定势, 令备择假设逼近原假设, 将两个统计量的样本量比值的极限, 定义为渐近相对效率.

具体而言, 对假设检验问题

$$H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta \neq \theta_0,$$

取备择假设序列$\theta_i (i = 1, 2, \cdots), \theta_i \neq \theta_0$, 且$\lim\limits_{i\to\infty} \theta_i = \theta_0$. 在固定势$1 - \beta$之下, 我们考虑两个检验统计量$V_{n_i}$和$T_{m_i}$. 其中$V_{n_i}$和$T_{m_i}$分别是备择检验为$\theta_i$所对应的两个检验统计量序列, n_i和m_i 是两个统计量分别对应的样本量. 势函数满足：

$$\lim_{i\to\infty} g_{V_{n_i}}(\theta_0) = \lim_{i\to\infty} g_{T_{m_i}}(\theta_0) = \alpha,$$

$$\alpha < \lim_{i\to\infty} g_{V_{n_i}}(\theta_i) = \lim_{i\to\infty} g_{T_{m_i}}(\theta_i) = 1 - \beta < 1.$$

如果极限

$$e_{VT} = \lim_{i\to\infty} \frac{m_i}{n_i}$$

存在, 且独立于θ_i, α和β, 则称e_{VT}是V相对于T的**渐近相对效率**, 简记为ARE(V,T). 它是Pitman于1948年提出来的, 因此又称为Pitman渐近相对效率.

下面的Nother定理给出了计算渐近相对效率应满足的5个条件.

定理1.2 对假设检验问题$H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta \neq \theta_0$：

(1) V_n和T_m是相容的统计量. 也就是说：当$n, m \to +\infty$时, $\forall \theta \neq \theta_0$,

$$g(\theta_i, V_{n_i}) \to 1, \quad g(\theta_i, T_{m_i}) \to 1.$$

(2)如果记$E(V_{n_i}) = \mu_{V_{n_i}}$, $\mathrm{var}(V_{n_i}) = \sigma^2_{V_{n_i}}$, $E(T_{m_i}) = \mu_{T_{m_i}}$, $\mathrm{var}(T_{m_i}) = \sigma^2_{T_{m_i}}$,则在$\theta = \theta_0$的邻域中一致地有①

$$\frac{V_{n_i} - \mu_{V_{n_i}}(\theta)}{\sigma_{V_{n_i}}(\theta)} \xrightarrow{\mathcal{L}} N(0, 1),$$

① $\mathcal{L}$表示依分布收敛.

$$\frac{T_{m_i}-\mu_{T_{m_i}}(\theta)}{\sigma_{T_{m_i}}(\theta)} \xrightarrow{\mathcal{L}} N(0,1).$$

(3)存在导数$\left.\frac{\mathrm{d}\mu_{V_{n_i}}(\theta)}{\mathrm{d}\theta}\right|_{\theta=\theta_0}$, $\left.\frac{\mathrm{d}\mu_{T_{m_i}}(\theta)}{\mathrm{d}\theta}\right|_{\theta=\theta_0}$; 而且$\mu'_{V_{n_i}}(\theta), \mu'_{T_{m_i}}(\theta)$在$\theta=\theta_0$的某一个闭邻域内连续, 导数不为0.

(4)

$$\lim_{i\to\infty}\frac{\sigma_{V_{n_i}}(\theta_i)}{\sigma_{V_{n_i}}(\theta_0)}=\lim_{i\to\infty}\frac{\sigma_{T_{m_i}}(\theta_i)}{\sigma_{T_{m_i}}(\theta_0)}=1;$$

$$\lim_{i\to\infty}\frac{\mu_{V_{n_i}}(\theta_i)}{\mu_{V_{n_i}}(\theta_0)}=\lim_{i\to\infty}\frac{\mu_{T_{m_i}}(\theta_i)}{\mu_{T_{m_i}}(\theta_0)}=1.$$

(5)

$$\lim_{i\to\infty}\frac{\mu'_{V_{n_i}}(\theta_0)}{\sqrt{n_i\sigma^2_{V_{n_i}}(\theta_0)}}=C_V,$$

$$\lim_{i\to\infty}\frac{\mu'_{T_{m_i}}(\theta_0)}{\sqrt{m_i\sigma^2_{T_{m_i}}(\theta_0)}}=C_T.$$

则V相对于T的Pitman渐近相对效率等于

$$\mathrm{ARE}(V,T)=\lim_{i\to\infty}\frac{m_i}{n_i}=\frac{C_V^2}{C_T^2}.$$

这意味着计算Pitman渐近相对效率只要用到$\mu'_{V_{n_i}}(\theta_0), \mu'_{T_{m_i}}(\theta_0)$ 和$\sigma^2_{V_{n_i}}(\theta_0)$, $\sigma^2_{T_{m_i}}(\theta_0)$, 而这四项都不难计算.

定义1.2 假设检验问题：$H_0:\theta=\theta_0 \leftrightarrow H_1:\theta=\theta_1(\theta_0\neq\theta_1)$,上述定理中定义的极限为

$$\lim_{i\to\infty}\frac{\mu'_{V_{n_i}}(\theta_0)}{\sqrt{n}\sigma_{V_{n_i}}(\theta_0)},$$

称为V_n的**效率**, 记为$\mathrm{eff}(V)$.

例1.5 考虑总体为正态分布,$\{X_j, j=1,\cdots,n\}$是独立同分布的样本，

$$p(x,\mu,\sigma)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},\ -\infty<x<+\infty,$$

假设检验问题：$H_0:\mu=0 \leftrightarrow H_1:\mu=\mu_i(i=1,2,\cdots)$, $\lim\limits_{i\to\infty}\mu_i=0$, 考虑检验统计量$T_n=\sqrt{n}\bar{X}/S$和$\mathrm{SG}_n=\sum\limits_{j=1}^{n}I(X_j>0)$, 其中, $\bar{X}=\frac{1}{n}\sum\limits_{j=1}^{n}X_j$是样本均值, $S^2=\frac{1}{n-1}\sum\limits_{j=1}^{n}(X_j-\bar{X})^2$是样本方差, $I(X_j>0)$是示性函数, 计算$\mathrm{ARE}(T,\mathrm{SG})$.

解 根据t分布的性质有

$$E_\mu(T_n)=\frac{\mu}{\frac{\sigma}{\sqrt{n}}},\quad \mathrm{var}_\mu(T_n)=1;$$

$$E_\mu(\mathrm{SG}_n)=np,\quad \mathrm{var}_\mu(\mathrm{SG}_n)=np(1-p).$$

因而$\mathrm{eff}(T_n)=\dfrac{1}{\sigma}$.其中

$$p=\int_0^\infty \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}\,\mathrm{d}t.$$

容易证明它们满足Nother定理的条件(1) ~ (5),而且：

$$[E_\mu(T_n)]'=\frac{\sqrt{n}}{\sigma},$$

$$\begin{aligned}[E_\mu(\mathrm{SG}_n)]'&=\frac{n}{\sqrt{2\pi}\sigma}\int_0^\infty \frac{1}{\sigma^2}(t-\mu)\mathrm{e}^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}\,\mathrm{d}t\\&=\frac{n}{\sqrt{2\pi}\sigma}\int_0^\infty \mathrm{d}\left(-\mathrm{e}^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}\right)=\frac{n}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{\mu^2}{2\sigma^2}},\end{aligned}$$

$$\begin{aligned}\mathrm{eff}(\mathrm{SG}_n)&=\lim_{n\to\infty}\frac{[E_0(\mathrm{SG}_n)]'}{\sqrt{n\mathrm{var}_0(\mathrm{SG}_n)}}\\&=\lim_{n\to\infty}\left[\frac{n}{\sqrt{2\pi}\sigma}\Big/\frac{n}{2}\right]=\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}.\end{aligned}$$

于是, T相对于SG的渐近相对效率为

$$\mathrm{ARE}(\mathrm{SG},T)=\left[\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\Big/\frac{1}{\sigma}\right]^2=\frac{2}{\pi}.$$

$$\mathrm{ARE}(T,\mathrm{SG})=\frac{\pi}{2}$$

从结果看, 在正态分布下, T相对于SG的渐近相对效率还是不错的. 后面我们会给出其他分布下的结果, 在偏态分布下T相对于SG的渐近相对效率可能会小于1。

§1.5 分位数和非参数估计

1. 顺序统计量

定义1.3 假设总体X有样本量为n的样本$X_1,X_2,\cdots,X_n$,将$X_1,X_2,\cdots,X_n$按从小到大排序后生成的统计量

$$X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)},$$

则称统计量$\{X_{(1)}, X_{(2)}, \cdots, X_{(n)}\}$为顺序统计量。其中$X_{(i)}$ 是第i个顺序统计量。顺序统计量是非参数统计的理论基础之一, 许多非参数统计量的性质与顺序统计量有关。

定理1.3 如果总体分布函数为$F(x)$, 则顺序统计量$X_{(r)}$的分布函数为

$$\begin{aligned} F_r(x) &= \mathrm{P}(X_{(r)} \leqslant x) = \mathrm{P}(\text{至少} r \text{ 个} X_i \text{ 小于或等于} x) \\ &= \sum_{i=r}^{n} \binom{n}{i} F^i(x)[1-F(x)]^{n-i}. \end{aligned}$$

如果总体分布密度$f(x)$存在, 则顺序统计量$X_{(r)}$的密度函数为

$$f_r(x) = \frac{n!}{(r-1)!(n-r)!} F^{r-1}(x) f(x)[1-F(x)]^{n-r}.$$

解： 注意到第r个顺序统计量的随机事件$\{X_{(r)} \in (x, x+\Delta x]\}$ ，该随机事件等价于如下随机事件：在n 个样本点里，有$(r-1)$ 个点比x 小，有1 个点落在区间$(x, x+\Delta x]$里；上一句话等价于在n个样本点里，有$(r-1)$ 个点比x 小，剩余的$(n-r+1)$ 个点不能都比$x+\Delta x$ 大。

$$\begin{aligned} &\mathrm{P}\left\{X_{(r)} \in (x, x+\Delta x]\right\} \\ &= \tbinom{n}{r-1}\{\mathrm{P}(X \leqslant x)\}^{r-1}(1-\mathrm{P}\{X \in (x+\Delta x, +\infty)\})^{n-r+1} \end{aligned} \tag{1.11}$$

两边都除以Δx,并令Δx趋向于0，有

$$f_r(x) = \binom{n}{r-1} F(x)^{r-1}(n-r+1) f(x)[1-F(x)]^{n-r} \tag{1.12}$$

定理1.4 如果总体分布函数为$F(x)$, 则顺序统计量$X_{(r)}$和$X_{(s)}$的联合密度函数为

$$f_{r,s}(x,y) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!} F^{r-1}(x) f(x)[F(y)-F(x)]^{s-r-1} f(y)[1-F(y)]^{n-s}. \tag{1.13}$$

证明: 同理可以求出第r个顺序统计量和第s个顺序统计量的联合分布。不妨假设$r<s$，注意到，$\{X_{(r)} \in (x, x+\Delta x], X_{(s)} \in (y, y+\Delta y]\}$ ，该随机事件等价于如下随机事件：在n 个样本点里，有$(r-1)$ 个点比x 小，有1个点落在区间$(x, x+\Delta x]$里，有$(n-r)$ 个点里有$(s-r-1)$ 个点落在区间$(x+\Delta x, y]$ 里，$(n-s+1)$ 个点里，有1个点落在小区间$(y, y+\Delta y]$ 里，还有剩余$(n-s)$ 个点都比$y+\Delta y$大。因此有，

$$\begin{array}{l}\mathrm{P}\left\{X_{(r)}\in(x,x+\Delta x],X_{(s)}\in(y,y+\Delta y]\right\}\\ =\binom{n}{r-1}\{\mathrm{P}(X\leqslant x)\}^{r-1}\binom{n-r+1}{1}\mathrm{P}\{X\in(x,x+\Delta x]\}\\ \binom{n-r}{s-r-1}\{\mathrm{P}(X+\Delta x<X\leqslant y)\}^{s-r-1}\\ \binom{n-s+1}{1}\mathrm{P}\{X\in(y,y+\Delta y]\}\\ \binom{n-s}{n-s}\{\mathrm{P}(X>y+\Delta y)\}^{n-s}\end{array}\tag{1.14}$$

等式的左右同时除以Δx和Δy，并令Δx和Δy 分别趋向于0，等式左边趋向$X_{(k)}$和$X_{(s)}$ 的联合密度函数，右边趋向于：

$$\begin{array}{l}\binom{n}{r-1}\{\mathrm{P}(X\leqslant x)\}^{r-1}\binom{n-r+1}{1}f(x)\\ \binom{n-r}{s-r-1}\{\mathrm{P}(x<X\leqslant y)\}^{s-r-1}\\ \binom{n-s+1}{1}f(y)\binom{n-s}{n-s}\{1-F(y)\}^{n-s}\end{array}\tag{1.15}$$

这样就导出顺序统计量$X_{(r)}$和$X_{(s)}$的联合密度函数为

$$f_{r,s}(x,y)=\frac{n!}{(r-1)!(s-r-1)!(n-s)!}F^{r-1}(x)f(x)[F(y)-F(x)]^{s-r-1}f(y)[1-F(y)]^{n-s}.$$

由(1.15)式可以导出许多常用的顺序统计量的函数的分布. 比如极差$W=X_{(n)}-X_{(1)}$的分布函数为

$$F_W(w)=n\int_{-\infty}^{\infty}f(x)[F(x+w)-F(x)]^{n-1}\mathrm{d}x.$$

2. 分位数的定义

一组数据从小到大排序后, 每一个数在数据中的序非常重要, 给定序, 寻找对应的数据, 用分布的语言来说, 就是找分位数. 比如：分布在3/4位置的数称为3/4分位数. 中位数是分布在样本中间位置的数.

不失一般性, 对任意分布而言, 分布的分位数如下定义.

定义1.4 假定X服从概率密度为$f(x)$的分布, 令$0<p<1$,满足等式$F(m_p)=\mathrm{P}(X<m_p)\leqslant p, F(m_p+)=\mathrm{P}(X\leqslant m_p)\geqslant p$ 唯一的根m_p称为分布$F(x)$的p分位数.

例如：中位数可以定义为$\mathrm{P}(X<m_{0.5})\leqslant 1/2, \mathrm{P}(X\leqslant m_{0.5})\geqslant 1/2$. 分布的3/4分位数定义为$\mathrm{P}(X<m_{0.75})\leqslant 0.75, \mathrm{P}(X\leqslant m_{0.75})\geqslant 0.75$.

对连续分布而言, 分布的分位数可以简化如下.

定义1.5 假定X服从概率密度为$f(x)$的分布, 令$0<p<1$,满足等式$F(x)=\mathrm{P}(X<m_p)=p$的唯一的m_p称为分布$F(x)$的p分位数.

3. 分位数的估计

分位数是刻画分布的重要特征, 经验分布函数的基本思想就是建立在分位数估计上的. 如果一组数据有n个值, 分布的第$i/(n+1)$分位点的估计由第i小的数据生成. 一般而言, 对任意分位数可以构造如下估计.

给定n个值$X_1, X_2, \cdots, X_n$, 可以根据下面的公式计算任意p分位数的值：

$$m_p = \begin{cases} X_{(k)}, & \dfrac{k}{n+1} = p, \\ X_{(k)} + (X_{(k+1)} - X_{(k)})[(n+1)p - k], & \dfrac{k}{n+1} < p < \dfrac{k+1}{n+1}. \end{cases}$$

4. 分位数的图形表示

1) 箱线胡须图

箱线胡须图(boxwisker)是用分位数表示数据分布的重要的探索性数据分析方法. 箱线胡须图的基本原理是找出数据中的5个数据, 用这5个数据直观地表示数据的分布.

(1) 中位数：将数据从小到大排序后, 位于中间位置的数用粗带表示, 显示了数据的平均位置.

(2) 上四分位数和下四分位数：分别是数据中排序在3/4位置和1/4位置的数. 这两个数之间有50%的数据量, 是数据中的主体部分, 用矩形箱表示, 可以观察数据的分散程度和相对于中位数的对称情况.

(3) 异常上下警戒点：以中位数为中心, 加减3/4位置与1/4位置差的1.5倍. 1.5倍是经验值, 在R软件中可能根据情况调整. 如遇最小值或最大值, 则以最小值或最大值为限, 以W_{u}表示上警戒点, 以W_{l}表示下警戒点, 则

$$W_{\mathrm{u}} = \min\{M_{0.5} + 1.5 \times (M_{0.75} - M_{0.25}), X(n)\},$$

$$W_{\mathrm{l}} = \max\{M_{0.5} - 1.5 \times (M_{0.75} - M_{0.25}), X(1)\}.$$

这两个数之间上下四分位数以外的部分以实线段表示, 表示这是数据的次要信息. 通过次要信息可以观察到数据的特色信息：比如零散信息与主体部分两侧的对称情况, 线段相对于主体部分较长, 表示次要信息比较分散; 较短表示次要信息比较密集. 还可以表示零散信息与主体部分两侧的对称情况, 上下线基本相等, 表示分布对称; 不等表示分布不对称, 线短的一侧表示分布较密. 警戒点以外的数据表示数据主体信息以外的异常点, 常用空心点表示, 这表示这些点被诊断为异常点, 是“胡须”这个词的来源. 如果空心点数量较多而且比较集中, 说明数据有厚尾现象. 最外侧的点是最大值或最小值; 如果没有, 则上下线恰好为最大值和最小值.

例1.6 (见chap1\Airplane.txt)数据中给出了某航空公司1949—1960年每月国际航班旅客人数, 我们分别按照各年和各月为分组变量制作箱线胡须图.

从按年旅客人数各月分布图中(图1.10(a)), 可以观察到随着年代的增加, 国际旅客人数呈现明显的增长态势, 各月的旅客人数差异有逐步增加趋势, 各月人数分布大部分呈现右偏分布；从按月旅客人数各年分布图上(图1.10(b)), 容易观察到各月的旅客人数分布也呈现规律性, 一般一月和十二月是旅客人数的低谷, 七月和八月是旅客人数的高峰, 还发现均值高的月份更容易产生较高的旅客人数.

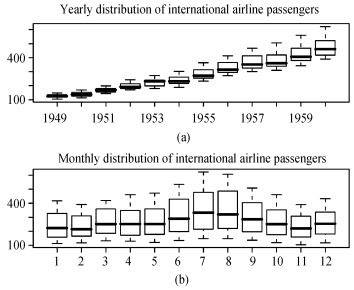

图 1.10 航空公司旅客分布线胡须图

显然, 在这个例子中发现箱线胡须图是一种直观地观察和了解数据分布的有效工具, 特别适合比较分组定量数据的分布特征.

2) Q-Q图

Quantile-Quantile(Q-Q)图是一种非常有用的通过两组数据的分位数大小比较数据分布的图形工具, 一般用于数据与已知分布的比较, 也可以比较两组数据的分布. 一般地, 如果X是一个连续随机变量, 有严格增的分布函数F, p分位点为x_p, 被比较的分布用Y表示, Y的分布是G,p分位点为y_p,满足$F(x_p) = G(y_p) = p$. 当要比较的是正态分布时, $G = \Phi$, $y_(i) = \Phi^{-1}\left(\dfrac{i-0.375}{n+0.25}\right)$, 这样如果数据服从正态分布, 数据点应该近似地分布在直线$y = \sigma x + \mu$附近, 其中μ,σ是待比较数据的均值和方差.

Q-Q图的基本原理是将两组数据分别从小到大排序后, 组成数据对$(x_{(i)}, y_{(i)})$, 描绘二者的散点图. 如果两组数据的分布相近, 表现在Q-Q图上, 散点图应该近似呈现直线; 反之, 则认为两组数据的分布有较大差异.

例1.7(见数据chap1\sunmon.txt) S.Stephens收集了墨西哥城从1986-2007年22年间空气中污染物的浓度数据,可以用每种污染物星期日的分位数和工作日的分位数制作Q-Q图(图1.11), 臭氧周日的高分位点小于工作日高分位点,极端高值更容易发生在平时而不是周日. 一氧化碳(CO)、可吸入颗粒物(PM)和氮氧排放物(NO_x)的各个分位数上, 工作日的含量都明显高于周日, 这是工作日空气污染严重的有利证据。从图上还发现随着空气污染物浓度的增加, 周日和工作日各分位点含量之间的差异

有加大的趋势, 这表示空气质量较差的周日和工作日之间的差异比空气质量较好的周日和工作日之间的差异大.

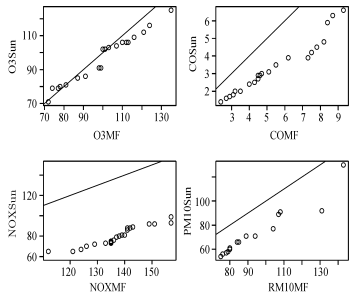

图 1.11 污染数据Q-Q图

CO(ppm), O_3(ppm), NO_x(ppb), PM10($mg{\cdot}m^{-3}$)

§1.6 秩检验统计量

§1.6.1 无重复数据的秩及性质

定义1.6 设样本$X_1, X_2, \cdots, X_n$是取自总体X的简单随机样本, $X_1, X_2, \cdots, X_n$中不超过X_i的数据个数$R_i = \sum_{j=1}^{n} I(X_j \leqslant X_i)$, 称$R_i$为$X_i$的**秩**, X_i是第R_i个顺序统计量, $X_{(R_i)} = X_i$. 令$R = (R_1, R_2, \cdots, R_n)$,$R$是由样本产生的统计量, 称为秩统计量.

例1.8 某学院本科三年级由9个专业组成,统计每个专业学生每月消费数据如下：

$$300 \quad 230 \quad 208 \quad 580 \quad 690 \quad 200 \quad 263 \quad 215 \quad 520 \tag{1.16}$$

用R求消费数据的秩和顺序统计量.

解 程序如下：

```
> spending<-c(300, 230, 208, 580, 690, 200,263,215,520)
> sort(spending)
> rank(spending)
```

定理1.5 对于简单随机样本, $R = (R_1, R_2, \cdots, R_n)$等可能取$(1, 2, \cdots, n)$的任意$n!$个排列之一, R在由$(1, 2, \cdots, n)$的所有可能的排列组成的空间上是均匀分布,

即：对$(1,2,\cdots,n)$的任一排列$(i_1,i_2,\cdots,i_n)$有

$$P(R=(i_1,i_2,\cdots,i_n))=\frac{1}{n!}.$$

上面定理1.5给出的是$R_1,R_2,\cdots,R_n$联合分布. 类似地, 每一个R_i在空间$\{1,2,\cdots,n\}$上有均匀分布；每一对(R_i,R_j)在空间$\{(r,s):r,s=1,2,\cdots,n;r\neq s\}$上有均匀分布.以推论的形式表示如下.

推论1.2 对于简单随机样本, 对任意$r,s=1,2,\cdots,n;r\neq s$及$i\neq j$, 有

$$P(R_i=r)=\frac{1}{n},$$
$$P(R_i=r,R_j=s)=\frac{1}{n(n-1)}.$$

推论1.3 对于简单随机样本,

$$E(R_i)=\frac{n+1}{2},$$
$$\mathrm{var}(R_i)=\frac{(n+1)(n-1)}{12},$$
$$\mathrm{cov}(R_i,R_j)=-\frac{n+1}{12}.$$

证明:

$$E(R_i)=\sum_{r=1}^{n}r\cdot\frac{1}{n}=\frac{n+1}{2}.$$

$$\begin{aligned}\mathrm{var}(R_i)&=\sum_{r=1}^{n}(r^2)\cdot\frac{1}{n}-[E(R_i)]^2\\&=\frac{n(n+1)(2n+1)}{6}\cdot\frac{1}{n}-\frac{(n+1)(n+1)}{4}\\&=\frac{(n+1)(n-1)}{12}.\end{aligned}$$

$$\begin{aligned}\mathrm{cov}(R_i,R_j)&=E[R_i-E(R_i)][R_j-E(R_j)]\\&=\sum\sum_{r\neq s}\left[\left(r-\frac{n+1}{2}\right)\left(s-\frac{n+1}{2}\right)\cdot\frac{1}{n(n-1)}\right]\\&=\left[\sum_{r=1}^{n}\sum_{s=1}^{n}\left(r-\frac{n+1}{2}\right)\left(s-\frac{n+1}{2}\right)-\sum_{j=1}^{n}\left(s-\frac{n+1}{2}\right)^2\right]\cdot\frac{1}{n(n-1)}\\&=-\frac{n+1}{12}.\end{aligned}$$

这些结果说明, 对于独立同分布样本来说, 秩的分布和总体分布无关.

§1.6.2 带结数据的秩及性质

在许多情况下, 数据中有重复数据, 称数据中存在结(tie). 结的定义如下.

定义1.7 设样本$X_1, X_2, \cdots, X_n$取自总体X的简单随机样本, 将数据排序后,相同的数据点组成一个“结”, 称重复数据的个数为结长.

假设有样本量为7的数据：

$$3.8 \quad 3.2 \quad 1.2 \quad 1.2 \quad 3.4 \quad 3.2 \quad 3.2$$

其中有4个结, $x_2 = x_6 = x_7 = 3.2$, 结长3；$x_3 = x_4 = 1.2$, 结长2; $x_1 = 3.8$和$x_5 = 3.4$的结长都为1. 如果有重复数据, 则将数据从小到大排序后, $(R_1, R_2) = (2, 3)$, 也可以等于$(3, 2)$,这样秩就不唯一. 一般常采用秩平均方法处理有结数据的秩.

定义1.8 将样本$X_1, X_2, \cdots, X_n$从小到大排序后, 如果$X_{(1)} = \cdots = X_{(\tau_1)} < X_{(\tau_1+1)} = \cdots = X_{(\tau_1+\tau_2)} < \cdots < X_{(\tau_1+\cdots+\tau_{g-1})} = \cdots = X_{(\tau_1+\cdots+\tau_g)}$, 其中$g$是样本中结的个数, τ_i是第i个结的长度, $(\tau_1, \tau_2, \cdots, \tau_g)$是$g$个正整数, $\sum_{i=1}^{g} \tau_i = n$,称$(\tau_1, \tau_2, \cdots, \tau_g)$为结统计量. 第$i$ 组样本的秩都相同, 是第i组样本原秩的平均, 如下所示：

$$r_i = \frac{1}{\tau_i} \sum_{k=1}^{\tau_i} (\tau_1 + \cdots + \tau_{i-1} + k) = \tau_1 + \cdots + \tau_{i-1} + \frac{1 + \tau_i}{2}.$$

例1.9 样本数据为12个数, 其值、秩和结统计量(用τ_i表示, 为第i个结中的观测值数量)如表1.2所示：

表1.2 结的计算

观测值	2	2	4	7	7	7	8	9	9	9	9	10
秩	1.5	1.5	3	5	5	5	7	9.5	9.5	9.5	9.5	12

其中有6个结, 每个结长分别为2, 1, 3, 1, 4, 1.

1. 结数据秩与秩平方和的一般性质

在一个由n个已排序好的数列中，其中有一段由τ个数组成的结数据，如果这个结的第一个数的秩$R_{r+1} = r + 1$, 考虑以下两种情况：

(1).当这τ个数完全不同时，每个数的秩都一样，取$r + \frac{\tau+1}{2}$这些数的秩和为

$$(r + 1) + (r + 2) + \cdots + (r + \tau) = \tau r + \frac{\tau(\tau + 1)}{2} \tag{1.17}$$

这些数的秩的平方和为

$$(r + 1)^2 + (r + 2)^2 + \cdots + (r + \tau)^2 = \tau r^2 + \tau r(\tau + 1) + \frac{\tau(\tau + 1)(2\tau + 1)}{6} \tag{1.18}$$

(2).当这τ个数完全相同时，这些数的秩和为

$$(r+\frac{\tau+1}{2})+(r+\frac{\tau+1}{2})+\cdots+(r+\frac{\tau+1}{2})=\tau r+\frac{\tau(\tau+1)}{2} \tag{1.19}$$

这些数的秩的平方和为

$$(r+\frac{\tau+1}{2})^2+(r+\frac{\tau+1}{2})^2+\cdots+(r+\frac{\tau+1}{2})^2=\tau r^2+\tau r(\tau+1)+\frac{\tau(\tau+1)^2}{4}. \tag{1.20}$$

观察式（1.17）（1.18）和（1.19）（1.20）可以发现不论这τ个数是否全相同，秩的和都是相同的，但是秩的平方和不同，完全不同的数列比完全相同的数列的秩平方和大$\frac{\tau^3-\tau}{12}$,

2.结数为g的数据秩的一般性质

假设有n个样本，记R_i为$x_i, i=1,2,\cdots,n$的不考虑平均秩下的秩。令$\alpha(i), i=1,2,\cdots,n$为一个计分函数，当结的长度为1 时，$a(R_i)=R_i$ ，当结的长度大于1时,$a(R_i)$ 取平均秩。

(1). 由(1.17)和(1.19),n个有结数据的秩和与无结数据的秩和是一样的：

$$\sum_{i=1}^{n}\alpha(R_i)=\sum_{i=1}^{n}\alpha(i)=\frac{(n+1)n}{2}$$

由于无结数据的秩的平方和为

$$\sum_{i=1}^{n}\alpha(R_i)^2=\frac{n(n+1)(2n+1)}{6},$$

所以结数为g的数据的秩平方和为

$$\sum_{i=1}^{n}\alpha(R_i)^2=\sum_{i=1}^{n}\alpha(i)^2=\frac{n(n+1)(2n+1)}{6}-\sum_{j=1}^{g}\frac{\tau_j^3-\tau_j}{12} \tag{1.21}$$

(2).对于$x_1,x_2,\cdots,x_n$,i.i.d的情况，$\alpha(R_i)$等可能的取$\alpha(i)$, 有

$$\mathrm{E}\left(\alpha(R_i)\right)=\bar{\alpha}=\frac{\sum_{j=1}^{n}\alpha(i)}{n}$$
$$\mathrm{Var}\left(\alpha(R_i)\right)=\frac{\sum_{i=1}^{n}(\alpha(i)-\bar{\alpha})^2}{n}$$

对于协方差$\mathrm{cov}(\alpha(R_i),\alpha(R_j))=E(\alpha(R_i)\alpha(R_j))-E(\alpha(R_i)E(\alpha(R_j)$

$$\mathrm{E}(\alpha(R_i)\alpha(R_j)) = \frac{\sum_{i\neq j}\alpha(i)\alpha(j)}{n(n-1)} = \frac{n^2\bar{\alpha}^2 - \sum_{i=1}^n \alpha(i)^2}{n(n-1)} \tag{1.22}$$

$$\mathrm{cov}(\alpha(R_i),\alpha(R_j)) = \frac{n^2\bar{\alpha}^2 - \sum_{i=1}^n \alpha(i)^2}{n(n-1)} - \bar{\alpha}^2 \tag{1.23}$$

$$= -\frac{\sum_{i=1}^n(\alpha(i)-\bar{\alpha})^2}{n(n-1)} \tag{1.24}$$

将(1.21)关于有结数据秩和和秩的平方和的结论带入，可以得到，

$$\begin{aligned}\sum_{i=1}^n(\alpha(i)-\bar{\alpha})^2 &= \sum_{i=1}^n\alpha(i)^2 - n\bar{\alpha}^2 \\ &= \frac{n(n+1)(2n+1)}{6} - \sum_{j=1}^g\frac{\tau_j^3-\tau_j}{12} - \frac{n(n+1)^2}{4} \\ &= \frac{n(n+1)(n-1)}{12} - \frac{\sum_{j=1}^g(\tau_j^3-\tau_j)}{12}\end{aligned} \tag{1.25}$$

注意到$\bar{\alpha} = \frac{n+1}{2}$,有

$$\begin{aligned}\mathrm{E}\left(\alpha(R_i)\right) &= \frac{n+1}{2} \\ \mathrm{Var}\left(\alpha(R_i)\right) &= \frac{n^2-1}{12} - \frac{\sum_{j=1}^g(\tau_j^3-\tau_j)}{12n} \\ \mathrm{cov}(\alpha(R_i),\alpha(R_j)) &= -\frac{n+1}{12} + \frac{\sum_{j=1}^g(\tau_j^3-\tau_j)}{12n(n-1)}\end{aligned}$$

(3)令$x_1, x_2, ..., x_m$为n个i.i.d数列中的任意m个数，则

$$\mathrm{E}(\sum_{i=1}^m\alpha(R_i)) = \sum_{i=1}^m\mathrm{E}(\alpha(R_i)) = m\bar{\alpha} \tag{1.26}$$

$$\mathrm{Var}(\sum_{i=1}^m\alpha(R_i) = \sum_{i=1}^m\mathrm{Var}(\alpha(R_i)) + 2\sum_{i<j}\mathrm{cov}(\alpha(R_i),\alpha(R_j)) \tag{1.27}$$

$$= m\mathrm{Var}(\alpha(R_i)) + m(m-1)\mathrm{cov}(\alpha(R_i),\alpha(R_j)) \tag{1.28}$$

$$= m\frac{\sum_{i=1}^n(\alpha(i)-\bar{\alpha})^2}{n} - m(m-1)\frac{\sum_{i=1}^n(\alpha(i)-\bar{\alpha})^2}{n(n-1)} \tag{1.29}$$

$$= \frac{m(n-m)\sum_{i=1}^n(\alpha(i)-\bar{\alpha})^2}{n(n-1)} \tag{1.30}$$

将(1.25)代入(1.26)式和(1.30)式可得，

$$\mathrm{E}(\sum_{i=1}^{m}\alpha(R_i)) = \frac{m(n+1)}{2} \tag{1.31}$$

$$\mathrm{Var}(\sum_{i=1}^{m}\alpha(R_i)) = \frac{m(n-m)(n+1)}{12} - \frac{m(n-m)\sum_{j=1}^{g}(\tau_j^3-\tau_j)}{12n(n-1)} \tag{1.32}$$

§1.7 U统计量

1. 单一样本的U统计量和主要特征

我们知道, 在参数估计和检验中, 充分完备统计量是寻找一致最小方差无偏估计的一条重要的途径, 在非参数统计中, 类似的统计量也存在. 这里我们介绍U统计量.

定义1.9 设$X_1, X_2, \cdots, X_n$取自分布族$\mathcal{F} = \{F(\theta), \theta \in \Theta\}$, 如果待估参数$\theta$存在样本量为$k$的无偏估计量$h(X_1, X_2, \cdots, X_k), k < n$, 即满足

$$Eh(X_1, X_2, \cdots, X_k) = \theta, \quad \forall\theta \in \Theta,$$

使上式成立的最小的样本量为k, 则称参数θ是k阶可估参数. 此时$h(X_1, X_2, \cdots, X_k)$称为参数θ的核(kernel).

一般地, 还要求核有对称的形式,也就是说：对$(1, 2, \cdots, k)$的任何一个排列$(i_1, i_2, \cdots, i_k)$, 有$h(X_1, X_2, \cdots, X_k) = h(X_{i_1}, X_{i_2}, \cdots, X_{i_k})$. 如果核本身不对称, 可以构造对称的核函数

$$h^*(X_1, X_2, \cdots, X_k) = \frac{1}{k!}\sum_{(i_1, i_2, \cdots, i_k)} h(X_{i_1}, X_{i_2}, \cdots, X_{i_k}).$$

$\sum_{(i_1, i_2, \cdots, i_k)}$ 是对$(1, 2, \cdots, k)$的任意排列$(i_1, i_2, \cdots, i_k)$ 共计$k!$ 个算式求和. 这时, $h^*(X_1, X_2, \cdots, X_k)$ 是满足定义1.9要求,且对称的θ 的核.

定义1.10 设$X_1, X_2, \cdots, X_n$取自分布族$\mathcal{F} = \{F(\theta), \theta \in \Theta\}$的样本, 可估参数$\theta$存在样本量为$k$的无偏估计量$h(X_1, X_2, \cdots, X_k)$, θ有对称核$h^*(X_1, X_2, \cdots, X_k)$, 则参数$\theta$ 的U 统计量如下定义：

$$U(X_1, X_2, \cdots, X_n) = \frac{1}{\binom{n}{k}}\sum_{(i_1, i_2, \cdots, i_k)} h^*(X_{i_1}, X_{i_2}, \cdots, X_{i_k}).$$

其中, $\sum_{(i_1, i_2, \cdots, i_k)}$ 表示对$\{1, 2, \cdots, n\}$中所有可能的k 个数的组合求和.

例1.10 设$\mathcal{F}=\{F(\theta),\theta\in\Theta\}$为全体一阶矩存在的分布族, 则期望$\theta=E(X)$是1阶可估参数, 有对称核$h(X_1)=X_1$. 由对称核生成的$U$统计量为

$$U(X_1,X_2,\cdots,X_n)=\frac{1}{\binom{n}{1}}\sum_{i=1}^{n}X_i=\bar{X}.$$

例1.11 设$\mathcal{F}=\{F(\theta),\theta\in\Theta\}$为全体二阶矩有限的分布族, 则方差$\theta=E(X-EX)^2$是2阶可估参数. 由$E(X-EX)^2=EX^2-(EX)^2$, 可知

$$h(X_1,X_2)=X_1^2-X_1X_2$$

是参数θ的无偏估计, 显然它不具有对称性, 如下构造对称核：

$$h^*(X_1,X_2)=\frac{1}{2}[(X_1^2-X_1X_2)+(X_2^2-X_1X_2)]=\frac{1}{2}(X_1-X_2)^2,$$

相应的U统计量为

$$\begin{aligned}
&U(X_1,X_2,\cdots,X_n)\\
=&\frac{1}{\binom{n}{2}}\sum_{i<j}\frac{1}{2}(X_i-X_j)^2\\
=&\frac{1}{n(n-1)}\sum_{i<j}(X_i^2+X_j^2-2X_iX_j)\\
=&\frac{1}{n(n-1)}\left[\frac{1}{2}\sum_{i\neq j}(X_i^2+X_j^2)-\sum_{i\neq j}X_iX_j\right]\\
=&\frac{1}{n(n-1)}\left[\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i^2+X_j^2)-\frac{1}{2}\sum_{i=1}^{n}(X_i^2+X_i^2)-\sum_{i\neq j}X_iX_j\right]\\
=&\frac{1}{n(n-1)}\left[n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2\right]\\
=&\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2.
\end{aligned}$$

定理1.6 设$X_1,X_2,\cdots,X_n$是取自分布族$\mathcal{F}=\{F(\theta),\theta\in\Theta\}$的简单随机样本, θ是k可估参数, $U(X_1,X_2,\cdots,X_n)$是θ的U统计量, 它的核是$h(X_1,X_2,\cdots,X_k)$, 有

$$E(U(X_1,X_2,\cdots,X_n))=\theta,$$

$$\mathrm{var}(U(X_1,X_2,\cdots,X_n))=\frac{1}{\binom{n}{k}}\sum_{c=1}^{k}\binom{k}{c}\binom{n-k}{k-c}\sigma_c^2.$$

其中, 给定$0\leqslant c\leqslant k$,如果一组$\{i_1,i_2,\cdots,i_k\}$和另一组$\{j_1,j_2,\cdots,j_k\}$ 有c 个元素是一样的，那么

$$\begin{aligned}\sigma_c^2 &= \mathrm{cov}[h(X_{i_1},X_{i_2},\cdots,X_{i_k}),h(X_{j_1},X_{j_2},\cdots,X_{j_k})].\\ &= \mathrm{E}(h_c(X_1,\cdots,X_c)-\theta)^2\end{aligned}$$

这里，$h_c(x_1,\cdots,x_c)=E(x_1,\cdots,x_c,X_{c+1},\cdots,X_k)$, $\sigma_c^2=\mathrm{var}((h_c(X_1,\cdots,X_v))$ 是不降的，也就是说$\sigma_0^2=0\leqslant\sigma_1^2\leqslant\cdots\leqslant\sigma_k^2$.

解 U统计量的方差计算如下：

$$\begin{aligned}\mathrm{var}(U(X_1,X_2,\cdots,X_n)) &= E\left[\frac{1}{\binom{n}{k}}\sum(h(X_1,X_2,\cdots,X_k)-\theta)\right]^2\\ &= \frac{1}{\binom{n}{k}^2}\sum_{(i_1,i_2,\cdots,i_k)}\sum_{(j_1,j_2,\cdots,j_k)}\mathrm{cov}[h(X_{i_1},X_{i_2},\cdots,X_{i_k}),\\ &\quad h(X_{j_1},X_{j_2},\cdots,X_{j_k})]\\ &= \frac{1}{\binom{n}{k}^2}\sum_{c=0}^{k}\binom{n}{k}\binom{k}{c}\binom{n-k}{k-c}\sigma_c^2\\ &= \frac{1}{\binom{n}{k}}\sum_{c=1}^{k}\binom{k}{c}\binom{n-k}{k-c}\sigma_c^2.\end{aligned}$$

U统计量具有很好的大样本性质,下面的定理1.7表明当样本量较大时，U统计量均方收敛到σ_1^2, 从而U统计量是θ的相合估计(consistency)；定理1.8表明U统计量的极限分布是正态分布. 这里仅给出结果, 详细的证明参见文献(孙山泽 2000).

定理1.7 设$X_1,X_2,\cdots,X_n$是取自分布族$\mathcal{F}=\{F(\theta),\theta\in\Theta\}$的简单随机样本, θ是k可估参数, $U(X_1,X_2,\cdots,X_n)$ 是θ的U统计量, 它的核为$h(X_1,X_2,\cdots,X_k)$, 有

$$E[h(X_1,X_2,\cdots,X_k)]^2<\infty,$$

则

$$\lim_{n\to\infty}\frac{n}{k^2}\mathrm{var}[U(X_1,X_2,\cdots,X_n)]=\sigma_1^2.$$

其中$\sigma_1^2 = \mathrm{cov}[h(X_1, X_{i_2}, \cdots, X_{i_k}), h(X_1, X_{j_2}, \cdots, X_{j_k})] > 0$.其中$\{i_2, \cdots, i_k\}$和$\{j_2, \cdots, j_k\}$ 取自$\{1, \cdots, n\}$ 且没有相同元素。

定理1.8(Hoeffding定理)　设$X_1, X_2, \cdots, X_n$是取自分布族$\mathcal{F} = \{F(\theta), \theta \in \Theta\}$的简单随机样本, θ是k可估参数, $U(X_1, X_2, \cdots, X_n)$ 是θ 的U 统计量, 它的核是$h(X_1, X_2, \cdots, X_k)$, 有

$$E[h(X_1, X_2, \cdots, X_k)]^2 < \infty,$$

当$\sigma_1^2 = \mathrm{cov}[h(X_1, X_{i_2}, \cdots, X_{i_k}), h(X_1, X_{j_2}, \cdots, X_{j_k})] > 0$时, 有

$$\sqrt{n}[U(X_1, X_2, \cdots, X_n) - \theta] \to N(0, k^2\sigma_1^2)(n \to +\infty).$$

例1.12　设$X_1, X_2, \cdots, X_n$为取自连续分布族$\mathcal{F} = \{F(\theta), \theta \in \Theta\}$的简单随机样本,估计参数$\theta = \mathrm{P}(X_1 + X_2 > 0)$, 有核$h(x_1, x_2) = I(x_1 + x_2 > 0)$, 之后会知道这个核是Wilcoxon 检验统计量的核。令

$$U_n^{(2)} = \binom{n}{2}^{-1} \sum_{i<j} I(X_i + X_j > 0)$$

证明$U_n^{(2)}$是2阶可估参数$\mathrm{P}(X_1 + X_2 > 0)$的$U$统计量,当$F(\theta)$关于0对称，$\sqrt{n}(U_n^{(2)} - 1/2)$渐近服从正态分布$N(0, 1/3)$.

证明：　根据0点对称性有

$$\begin{aligned}
&\mathrm{P}(X_1 + X_2 > 0, X_1 + X_3 > 0)\\
&= \mathrm{P}(X_1 > -X_2, X_1 > -X_3)\\
&= \mathrm{P}(X_1 > X_2, X_1 > X_3)\\
&= 1/3.
\end{aligned}$$

$$\begin{aligned}
\sigma_1^2 &= \mathrm{cov}(h(X_1, X_2), h(X_1, X_3))\\
&= \mathrm{P}(X_1 + X_2 > 0, X_1 + X_3 > 0) - \theta^2
\end{aligned}$$

而第一项最大，第二项$\theta = 1/2$,，$\sigma_1^2 = 1/3 - (1/2)^2 = 1/12$. $k = 2$ 因此根据定理1.8有$\sqrt{n}(U_n^{(2)} - 1/2)$渐近服从正态分布$N(0, 1/3)$.

例1.13　设$X_1, X_2, \cdots, X_n$为取自连续分布族$\mathcal{F} = \{F(\theta), \theta \in \Theta\}$ 的简单随机样本, 固定p,假设m_p是样本的p分位数, $\forall i = 1, 2, \cdots, n$,令$Y_i = I(X_i > m_p)$, 定义计数统计量$T = \sum_{i=1}^{n} Y_i$.证明：$T/n$是1阶可估参数$\mathrm{P}(X > m_p)$的$U$统计量,$T/n$服从渐近正态分布.

证明: $\sigma_1^2 = \mathrm{var}(Y_i) = \mathrm{P}(X > m_p)(1 - \mathrm{P}(X > m_p))$, 因此根据定理1.8有: $\sqrt{n}\left(\dfrac{T}{n} - \mathrm{P}(X > m_p)\right)$ 渐近服从正态分布$N(0, \sigma_1^2)$. 也就是说:

$$\frac{T - E(T)}{\sqrt{(\mathrm{var}T)}} = \frac{\sqrt{n}\left(\dfrac{1}{n}T - \mathrm{P}(X > m_p)\right)}{\sqrt{\sigma_1^2}}$$

服从渐近的正态分布$N(0,1)$.

2. 两样本U检验统计量和分布

类似单一样本的U统计量的定义, 对两样本的情况, 有下面定义:

定义1.11 设$X = \{X_1, X_2, \cdots, X_n\}, X_1, X_2, \cdots, X_n$独立同分布取自分布族$\mathcal{F}, Y = \{Y_1, Y_2, \cdots, Y_m\}$独立同分布取自分布族$\mathcal{G}$, X与Y独立. 如果$h(X_1, X_2, \cdots, X_k)$待估参数$\theta \in \mathbf{F} = \{F, G\}$,存在样本量分别为$k \leqslant n$和$l \leqslant m$的样本构成的估计量$h(X_1, X_2, \cdots, X_k, Y_1, Y_2, \cdots, Y_l)$ 是θ的无偏估计, 即满足

$$Eh(X_1, X_2, \cdots, X_k, Y_1, Y_2, \cdots, Y_l) = \theta, \quad \forall \theta \in \mathbf{F},$$

上述关系成立的最小的样本量为k, l, 则称参数θ是(k, l)可估的, $h(X_1, X_2, \cdots, X_k, Y_1, Y_2, \cdots, Y_l)$称为参数$\theta$的核(kernel).

定义1.12 $X = \{X_1, X_2, \cdots, X_n\}, X_1, X_2, \cdots, X_n$独立同分布取自分布族$\mathcal{F}, Y = \{Y_1, Y_2, \cdots, Y_m\}$独立同分布取自分布族$\mathcal{G}$, X与Y独立, (k, l)可估参数θ存在样本量分别为(k, l)的对称无偏估计量$h(X_1, X_2, \cdots, X_k, Y_1, Y_2, \cdots, Y_l)$, 则参数$\theta$的$U$统计量如下定义:

$$U(X_1, X_2, \cdots, X_n, Y_1, Y_2, \cdots, Y_m) = \frac{1}{\binom{n}{k}\binom{m}{l}} \sum_{(i_1, i_2, \cdots, i_k)} \sum_{(j_1, j_2, \cdots, j_l)} h(X_{i_1}, X_{i_2}, \cdots, X_{i_k}, Y_{j_1}, Y_{j_2}, \cdots, Y_{j_l}).$$

例1.14 设总体X服从分布函数为$F(x)$的分布, Y服从分布函数为$G(x)$的分布, $X_1, X_2, \cdots, X_n$独立同分布取自分布族$\mathcal{F}, (Y_1, Y_2, \cdots, Y_m)$独立同分布取自分布族$\mathcal{G}$, X与Y独立, 待估参数是$\theta = P(X > Y)$,考虑θ的U统计量和它的性质.

解 给定i, j, 令

$$h(X_i, Y_j) = I(X_i > Y_j) = \begin{cases} 1, & X_i > Y_i. \\ 0, & \text{其他}. \end{cases}$$

容易知道: $E(h(X_i, Y_j)) = \theta$, 由$h(X_i, Y_j)$张成的U统计量定义为

$$U_{nm} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} I(X_i > Y_j), \tag{1.33}$$

这个U统计量将在第3章介绍, 它是Mann和Whitney于1947年提出的, 称做Mann- Whitney统计量,它是$\theta = \mathrm{P}(X > Y)$的最小方差无偏估计. 如果我们要检验问题：

$$H_0 : F = G \leftrightarrow H_1 : F \geqslant G,$$

则可知在原假设成立的情况下, U统计量的方差为

$$\mathrm{var}(U_n)_m = \frac{n+m+1}{12nm}.$$

因此可知, 当$n \to \infty, m \to \infty$时,

$$\sqrt{12nm} \cdot \frac{U - 0.5}{n+m} \xrightarrow{\mathcal{L}} N(0,1).$$

故在大样本情况下检验的拒绝域为

$$U \geqslant \frac{1}{2} + \sqrt{\frac{n+m}{12nm}} \cdot U_{1-\alpha}.$$

这个检验称为Mann-Whitney检验.

习题

1.1 某批发商从厂家购置一批灯泡, 根据合同的规定, 灯泡的使用寿命平均不低于1000h, 已知灯泡的使用寿命服从正态分布, 标准差是20h. 从总体中随机抽取了100只灯泡, 得知样本均值为996h, 问题是：批发商是否应该购买该批灯泡?

(1) 原假设和备择假设应该如何设置? 给出你的理由.

(2) 在原假设$\mu < 1000$之下, 给出检验的过程并做出决策. 如果不能拒绝原假设, 可能是哪里出了问题?

1.2 尝试证明，如果$X_1,\ldots,X_n$独立同分布地来自$[0,1]$上的均匀分布，则对任意的$s > k$，$X_{(s)} - X_{(k)}$服从贝塔分布，第一个参数是$(s-k)$，第二个参数是$(n-s+k+1)$。

1.3 尝试证明，例1.5中，原假设下，$\lim_{n\to+\infty} \mathrm{E}(\frac{1}{S}) = \frac{1}{\sigma}$。

1.4 思考布里斯托博士在不知道奶茶加奶顺序的前提下，将8杯奶全部猜对的可能性?

1.5 将例1.7的原假设和备择假设对调，

$$H_0 : \lambda \leqslant 1 \leftrightarrow H_1 : \lambda > 1$$

请选择$T = \sum_{i=1}^{n} X_i$作为统计量，当样本量$n = 100$时，对拒绝域分别为$W_1 = \{T \geqslant 117\}$和$W_2 = \{T \geqslant 113\}$,分别绘制势函数曲线图，令I,II 类错误相等时，给出弃权域的参数范围，比较两个检验弃权域有怎样的不同。

1.6 设$X_1, X_2, \cdots X_{(n)}$为具有连续分布函数$F(x)$的iid(独立同分布)的样本, 且具有概率密度函数$f(x)$, 如定义

$$U_i = \frac{F(X_{(i)})}{F(X_{(i+1)})}, \quad i = 1, 2, \cdots, n-1, \quad U_n = F(X_{(n)}), \tag{1.34}$$

证明$U_1, U_2^2, \cdots, U_n^n$为来自$(0,1)$上均匀分布的 iid 样本.

1.7 设随机变量$Z_1, Z_2, \cdots, Z_N$相互独立同分布, 分布连续, 其对应的秩向量为$\boldsymbol{R} = (R_1, R_2, \cdots, R_N)$, 假定$N \geqslant 2$, 令$V = R_1 - R_N$, 试证明

$$P(V=k) = \begin{cases} \dfrac{N - |\, k \,|}{N(N-1)}, & \text{当}|\, k \,| = 1, 2, \cdots, N-1\text{时}, \\ 0, & \text{其他}. \end{cases}$$

1.8 设随机变量$X_1, X_2, \cdots, X_n$是来自分布函数为$F(x)$的总体的样本, 试对下列参数确定: ① 参数可估计的自由度; ② 对称核$h(\cdot)$; ③ U统计量; 并指明④适应的分布族$\mathcal{F}$. 这些参数为:

(1) $P(|\, X_1 \,| > 1)$;

(2) $P(X_1 + X_2 + X_3 > 0)$;

(3) $E(X_1 - \mu)^3$, 其中μ为$F(x)$的期望;

(4) $E(X_1 - X_2)^4$.

1.9 考虑参数$\theta = P(X_1 + X_2 > 0)$, 其中随机变量$X_1, X_2$相互独立同分布, 有连续分布函数$F(x)$. 定义

$$h(x) = 1 - F(-x). \tag{1.35}$$

说明$E(h(X_1)) = \theta$. 并请回答: $h(X_1)$是对称核吗? 为什么?

1.10 设$X_1, X_2, \cdots, X_m$和$Y_1, Y_2, \cdots, Y_n$分别为具有连续分布函数的$F(x)$和$G(y)$的相互独立的 iid 样本, $\theta = P(X_1 + X_2 < Y_1 + Y_2)$.

(1) 证明在$H_0 : F = G$之下, $\theta = \dfrac{1}{2}$;

(2) 试求关于θ的U统计量.

1.11 设$X_1, X_2, \cdots, X_m$和$Y_1, Y_2, \cdots, Y_n$为分别来自连续分布的相互独立的样本, 试求$\theta = \mathrm{var}(X) + \mathrm{var}(Y)$的$U$统计量.

1.12 设$\{X_1, X_2, \ldots, X_n\}$为独立同分布的样本，服从分布$F(x)$，记最小次序统计量$X_{(1)}$的分布函数为$F_{(1)}(x)$，求最小次序统计量的分布，用geyser数据的duration变量，每次不放回抽取20个数据,计算最小值，一共重复50 次，得到最小值的观测样本50个，由50 个数据计算次序统计量的经验分布函数，请问这个经验分布函数和理论分布函数比较差距有多大，请用图示法来说明你的观察结果.

1.13 设$\{X_1, X_2, \ldots, X_n\}$为独立同分布的样本，服从连续分布$F(x)$. 证明：$h(X_1, X_2, X_3) = \mathrm{sgn}(2X_1 - X_2 - X_3)$是概率$\theta(F) = P(X_1 > \frac{X_2+X_3}{2}) - P(X_1 < \frac{X_2+X_3}{2})$的无偏估计，这里符号函数$\mathrm{sgn}(x)$是符号函数：

$$\mathrm{sgn}(x) = \begin{cases} 1, & x > 0, \\ 0, & x = 0, \\ -1, & x < 0. \end{cases}$$

1.证明：$F(x)$是对称的，那么$\theta(F) = 0$,。

2.从N(0,1)中选取随机数a，经$x = exp(a)$变换成一个新的变量x，请计算由x所形成的$\theta(F)$的U统计量的观察值，根据U统计量的观察，用图示法来观察X的分布是不是对称的。

1.14 比较图1.9中组1到组5 ～ 10%最强的组、10%最弱和中位的组的动物寿命之间的差别.

1.15 考虑一个从参数$\lambda = 1$的指数分布中抽取的样本量为100的样本.

(1) 给出样本的对数经验生存函数$\ln S_n(t)$的标准差. ($\ln S_n(t)$作为t的函数)

(2) 从计算机中产生几个类似的样本量为100的样本, 画出它们的对数经验生存函数图, 联系图补充对(1)的回答.

1.16(数据见chap1\beenswax.txt) 为探测蜂蜡结构, 生物学家做了很多实验,每个样本蜡里碳氢化合物(hydrocarbon)所占的比例对蜂蜡结构有特殊的意义, 数据中给出了一些观测.

(1)画出beenswax数据的经验累积分布, 直方图和Q-Q图.

(2)找出0.90, 0.75, 0.50, 0.25和0.10的分位数.

(3)这个分布是高斯分布吗?

1.17 考虑一个实验：对减轻皮肤瘙痒的药物进行疗效研究(Beecher, 1959).在10名20~30岁的男性志愿者身上做实验比较五种药物和安慰剂、无药的效果. (注意这批试验者限制了药物评价的范围. 例如这个实验不能用于老年人)每个试验者每天接受一次治疗, 治疗的顺序是随机的. 每个试验者首先以静脉注射方式给药, 然后用一种豆科藤类植物Cowage刺激前臂, 产生皮肤瘙痒, 记录皮肤瘙痒的持续时间. 具体实验细节可参见文献(Beecher 1959). 表1.3中给出皮肤瘙痒的持续时间(单位: s).

表1.3 皮肤瘙痒持续时间观测值

被试者	无药	安慰剂Placebo	Ⅰ Papav-erine	Ⅱ Amin-ophylline	Ⅲ Morp-hine	Ⅳ Pento-barbital	Ⅵ Tripele-nnamine
BG	174	263	105	141	199	108	141
JF	224	213	103	168	143	341	184
BS	260	231	145	78	113	159	125
SI	255	291	103	164	225	135	227
BW	165	168	144	127	176	239	194
TS	237	121	94	114	144	136	155
GM	191	137	35	96	87	140	121
SS	100	102	133	222	120	134	129
MU	115	89	83	165	100	185	79
OS	189	433	237	168	173	188	317

用经验生存函数比较不同治疗方法在减轻皮肤瘙痒的作用方面是否有差异?

案例与讨论:大学正态型成绩单与分布的力量

案例背景

2012年一则教育新闻报导，某大学正在试行一种“正态型成绩单”，正态型成绩单上许多成绩都将做正态化处理。这样一来，学生拿着成绩单无论是面试还是求学，都更容易让面试官更准确地获知学生在某门课程群体学习中的的真实水平。比如，90分表示在这项科目上低于他的学生不会少于80%（前20%），60分表示在这项科目上位于20% 较低的水平。然而在接到外界媒体询问时，该大学教务长却表示近期并不准备将这项正态型成绩单的政策推广到所有课程，仅限于本科通识课。对于部分学生质疑由此政策可能会招致规模性不及格成绩所表示的担忧，发言人表示这项政策的本意并非要扩大在校生不及格的比例，而是为了激发在校生的学习动力。

研究生面试官和企业政府人力资源管理部门最大的困扰是在翻阅学生成绩单方面缺乏可比性，不同大学不同专业不同教师给分尺度不一。一般而言，声名显赫的大学里教授对学生成绩有很大的自主权，责任心强的教授往往不轻易给学生很高的成绩；学校声誉一般且人才市场饱和度较高的专业会考虑给学生更高的成绩，以提高学生在人才市场上的竞争力；人才供不应求的热门专业则通过拉开成绩差距，实现多元化人才结构的市场布局。这表明，透过成绩单可一窥大学和专业对人才培养的理念和信心。受人才测定的不确定性、天赋导向论和社会认可度等多种因素影响，学校声誉一直是国家级高端人才选拔的硬指标，然而名校效应过度发酵也会产生负面影响。大学为社会储备大量人力资源，将哪些人力资源转化成高价值的人力资本是人力招聘中的核心问题，用人单位也需要在对人才未来能力做出评估时掌握充分的依据。大学里的成绩是一名学生专业性的基本体现，是学生掌握和学习新知识能力的基本依据，正态型成绩单增强了市场对前20% 学生的识别力。而从学校的管理来看，一旦标准化成绩单在人才选拔市场上的正面效应被培植出来，其示范效应将是不可估量的。另一方面，对于一些以考取名校作为邀功资本乃至入校后学业半途而废的学生而言，用较低水平的成绩适度降低其优越感，警示和唤醒其竞争意识也是有必要的。

综上所述，可以这么看，该大学正态成绩单的出炉，就是为了调节不确定性人才市场带给大学生学习心态的失衡。无论市场对人才的风向标如何变化，大学应始终坚持专业根基不随风摇摆。正态型成绩单的作用无疑像一道加固根基的堤坝，抵制以水分高分对人才市场的舞弊，治理教师不敢给差生低分的乱象，让优质的学生不再羞于成绩单不如名校生在身份上的尴尬，积极为人才市场营造奋进向上锐意进取的内涵型人才选择标准。

思考与讨论

1. 正态型成绩单制是不是一项好的教务策略？为什么？
2. 正态型成绩单制要解决的是一个怎样的问题？这个待解决的问题是在怎样的背景下发生的？
3. 正态型成绩单制为谁创造了价值？具有哪些特征的人群会因为这项制度的推行而获益？正态成绩单可能会使哪些人群感受到伤害？
4. 如何评价这个策略？如果你是该校教务长，你会选择怎样做？