第 1 章　基 本 概 念

1.1　非参数统计概念与产生

1. 非参数统计的概念

回顾数理统计基础知识可知, 分布是回答不确定性问题的基本统计工具, 对数据的分布做出推断是统计推断的根本任务. 典型的统计推断过程是从假定分布族开始的, 从数据到结论通常由 5 个步骤组成：分布族假定, 抽样, 统计量和抽样分布, 推估和检验, 评价模型. 假定分布族是对实际问题的数学描述, 它是统计推断的基础. 比如, 研究某类商品的市场占有率, 假定在平均的意义之下, 每个消费者是否占有待研究商品来自两点分布 $B(1,p), 0<p<1$；在研究保险公司的索赔请求数时, 可能假定索赔请求数来自 Poisson 分布 $\mathcal{P}(\lambda), 0<\lambda<\infty$(当然还可能有其他类型的分布假定); 在研究肥料对农作物产量的影响效果时, 假定平均意义之下, 每测量单元 (可能是) 产量服从正态分布 $N(\mu+x\beta,\sigma^2)$, 其中 x 是肥料的用量. 数据样本被视为从分布族的某个参数族抽取出来的总体的代表, 未知的仅仅是总体分布具体的参数值, 这样推断问题就转化为分布族的若干个未知参数的估计问题, 用样本对这些参数做出估计或进行假设检验, 从而得知数据背后的分布, 这类推断方法称为 **参数方法**.

然而在许多实际问题中, 要对数据的分布做出具体的假定常常需要很多背景知识, 特别是在探索性的问题研究中, 人们往往对总体的信息知之甚少, 很难对总体的分布形式和统计模型做出相对比较明确的假定. 甚至在有些情况下, 能够对问题尝试数学描述本身就是问题的核心. 比如在人为控制因素不多的大部分经济和社会问题中, 数据的分布形态和数据之间的关系常常是不能任意假定的, 最多只能对总体的分布做出类似于连续型分布或者关于某点对称等一般性的假定. 这种不假定总体分布的具体形式, 尽量从数据 (或样本) 本身获得所需要的信息, 通过估计而获得分布的结构, 并逐步建立对事物的数学描述和统计模型的方法称为 **非参数方法**.

问题 1.1(见光盘数据 chap1student.txt)　我们想比较两组学生的成绩是否存在差异, 传统的方法如 t 检验可以帮助我们分析问题. 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, 应用附录 A 中介绍的探索性数据分析方法绘制两组数据的分布, 如图 1.1 所示, 很难看出数据的分布是对称的. 这样, 应用 t 检验会有怎样的问题？我们将在第 2 章回答这个问题.

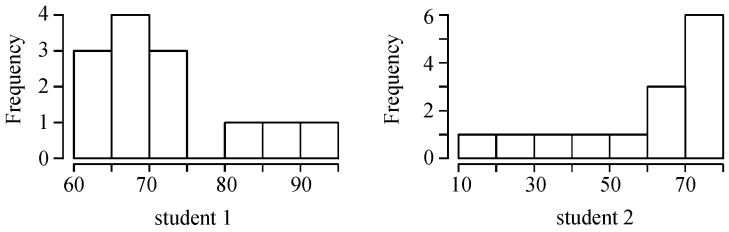

图 1.1　两组学生成绩的直方图

问题 1.2(见光盘数据 iqeq.txt)　我们希望比较两组被试的 IQ 成绩和 EQ 成绩之间是否存在着相关性, 传统的方法如 Pearson 相关系数检验可以帮助我们分析问题. 应用附录中介绍的探索性数据分析方法绘制两组数据的散点图, 如图 1.2 所示, 很难看出数据之间是否存在相关性. 这样的数据分布, 应用 Pearson 检验能测量出真实的关系吗？我们将在第 5 章回答这个问题.

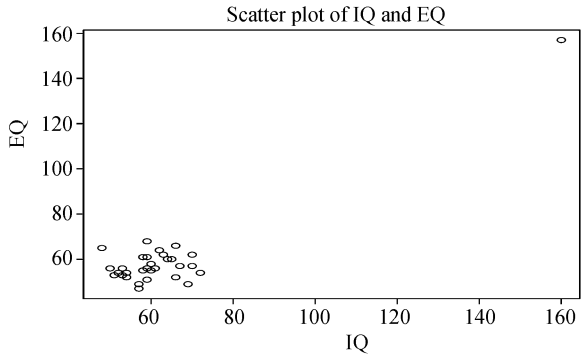

图 1.2　两组学生 IQ 成绩和 EQ 成绩相关散点图

问题 1.3　我们希望从光顾超市的用户购买清单数据中分析出哪些物品可能会被客户同时购买, 传统列联表分析能够给我们提供一些思路, 但是当物品数量很大的时候, 传统方法很难出现有效的结果. 我们将在第 5 章回答如何解决类似的问题.

以上这些问题, 并不总是能够在参数统计的框架结构中找到对应的方法, 数据驱动的方法会带领数据分析的实践者突破传统的框架, 思考如何对数据进行合理的运用. 总而言之, 非参数统计学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定性, 即不会由于总体分布与数据之间不一致导致发生大的结论性错误. 在经典的统计框架中, 正态分布一直是最引人瞩目的, 可以描述许多相对而言更为确定的问题.

比如：自动生产链处于稳定状态下的产品的质量. 然而, 正态分布并不是神话, 用于探索性问题时并不总是合适的, 随意对数据做出假定可能方便了计算和解释, 但可能产生错误的判断. 在某些推断问题中, 当数据不能支持显著性的结论, 常常表现为模型没有通过检验, 一些分析人士往往将原因归为信息量太少. 样本量不足可能是结论不显著的一个原因, 然而追加样本量在很多行业中代价是巨大的. 另外一个可能的解决方法是尝试更为宽松的模型假设, 即换用更有效的方法取代一味地增加样本量, 在节约成本和降低资源环境代价的条件下, 有效率地解决问题.

(2) 非参数统计可以处理所有类型的数据, 有广泛的适用性. 我们知道, 统计数据按照数据类型可以分为两大类：定性数据 (包括类别数据和顺序数据) 和定量数据 (包括等距数据和比例数据). 拿检验来说, 一般而言, 参数统计主要针对定量数据, 原因是理论上容易得到比较好的结果, 然而实践中, 我们所收集到的数据常常不符合参数统计模型的假定. 比如：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, 尝试非参数方法是自然的. 即便对于定量数据而言, 也常常出现数据测量误差问题、不同分布数据混合问题, 此时传统的统计推断未必适用于噪声密集的数据环境, 如果将这些数据转化为顺序数据, 有可能弱化颗粒噪声的影响, 尝试用非参数方法分析, 甚至可能获得理想的结果.

(3) 非参数思想容易理解, 计算容易. 作为统计学的分支, 其统计思想非常深刻, 很多原理与参数统计思想平行, 容易发展生成算法. 特别是伴随计算机技术的发展, 最近的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. 非参数统计方法在小样本的时候, 可能涉及更多不常见的统计表, 过去会对一些非专业的使用者造成不便. 如今很多统计软件, 如 R 中都已提供现成的函数供人们计算和使用, 一些统计量的精确分布或近似分布都可以轻松地从软件中更为精确地得到, 取代纸质编制的粗糙且不精确的表.

当然, 非参数方法也有一些弱点, 如果人们对总体有充分的了解且足以确定其分布类型, 非参数方法就不如参数方法具有更强的针对性, 有效性可能会差一些. 所以非参数统计并非要取代参数统计, 它作为参数统计的一个有力的补充, 符合人类认识问题、解决问题的认知过程.

2. 学习非参数统计的意义

统计学研究的是从数据到结论的数据研究方法, 数据分析工作常常不是一个单纯方法的简单应用, 而是一个对数据内部规律认识的从无到有、从局部到整体的判断、推断和下结论的过程. 在这个过程中, 常常需要数据分析的综合思考, 如数据的收集、选择、分布推断等, 这个过程对于培养学生解决问题的能力非常必要.

大部分非参数统计方法的基本思想与参数统计思想平行, 很容易参照参数统计的内容进行比对, 可以增强学生对方法比较和选择的思考和训练, 扩充学生的知

识体系. 非参数统计可以补充学生在统计计算方面的训练, 有利于培养信息创新型人才.

统计学专业在西方著名大学中的主要招生对象是本科高年级学生或研究生, 这说明统计专业主要培养学生解决问题的能力, 特别是数据的处理、比较和分析能力, 能力体现在深度和广度两个层面, 很难想像一名仅会一两种方法的学生能够比较客观、恰当地运用数据分析工具协助实际部门解决各种复杂难题, 非参数统计则在思想的层面上扩展学生的专业方法选择能力, 在数据问题中提高学生动手解决问题的能力.

3. 非参数统计的历史

一般认为, 非参数统计概念的形成主要归功于 20 世纪 40—50 年代化学家 F.Wilcoxon 等人的工作. Wilcoxon 于 1945 年提出两样本秩和检验, 1947 年 Mann 和 Whitney 将结果推广到两组样本量不等的一般情况. 继 F.Wilcoxon 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试验设计不同情况下的数据分析方法, 小样本检验和异常数据诊断方面得到了成功的应用. Pitman 于 1948 年回答了非参数统计方法相对于参数方法来说的相对效率的问题, 1956 年, J.L.Hodges 和 E.L.Lehmann 则发现了一个令人惊讶的结果, 与正态模型中 t 检验相比较, 秩检验能经受住有效性的较小损失. 而特别对于厚尾分布所产生的数据, 秩检验可能更为有效. 第一本论述非参数应用的书也是在这个时代于 1956 年由 S.Siegel 撰写, 有人记载从 1956—1972 年, 该书被引用了 1824 次. 这也说明非参数统计在 20 世纪 60—70 年代的发展相当活跃.

20 世纪 60 年代, Hodges 和 Lehmann 从秩检验统计量出发, 导出了若干估计量和置信区间. 这些方法为后来非参数方法成功应用于试验设计数据开启了一道大门. 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 60 年代中后期; Cox 和 Ferguson 则最早将非参数方法应用于生存分析.

进入 20 世纪 70—80 年代, 非参数统计获得了蓬勃的发展, 特别是 Efron 于 1979 年提出 Bootstrap 方法之后, 使得非参数方法借助计算机技术和大量计算获得更稳健的估计和预测, 因而在应用领域取得了长足的进展. 而以 P.J.Huber 以及 F.Hampel 为代表的统计学家从计算技术的实现角度, 为衡量估计量的稳定性提出了新准则. 20 世纪 90 年代有关非参数统计的研究和应用主要集中在非参数回归和非参数密度估计领域, 其中较有代表性的人物是 Silverman 和 J.Q.Fan.

20 世纪 90 年代后, 算法建模思想发展飞快, 成为非参数统计的新宠儿. Vapnik(1974) 等从学习的角度规范了预测模型选择的方法框架. Brieman L(1984, 2001) 和 Donoho(1994) 等为数据驱动的探索性构造模型敞开了大门, 大规模计算和自动化分析的需要将非参数统计引入机器学习领域, 如自适应模型选择、决策树、组合决策树 (Bagging, 装袋法;Boosting, 助推法)、随机森林以及关联分析等.

1.2 假设检验回顾

经典统计方法回答问题的基本逻辑是考察样本数据是否支持我们对总体的某种猜测. 这些没有被数据验证的猜测就是假设, 求证的过程就是假设检验. 比如研究问题：

(a) 新引进的生产过程是否优于旧过程?

(b) 几种不同的肥料哪一种更有效?

(c) 大学生的就业率与城市失业率之间是否存在关系?

从传统统计的观点来看, 这些问题都可以视为对不同分布总体的选择问题. 选择的依据首先可以从描述统计的图形和基本统计量上观察出一些现象, 但是关于分布间差异的粗略判断, 并没有揭示这种差异是本质的还是随机偶然因素造成的, 真实的情况只取其中之一, 而样本的表现则在两者之间. 检验就是希望能够找到一个合理地做出可靠性判断的临界点, 从而产生判别的条件和准则.

基本的假设检验原理是从两个互为对立的命题, 即假设开始的：零假设和备择假设. 对这两个互为对立的假设而言, 事先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参数的选择, 就像我们在大部分教科书中所看到的那样, 假设一般都以参数的形式出现, 这是参数统计的普遍特征. 比如：在上述问题 (a) 中, 可以如此书写假设 $H_0 : \theta = \theta_0$, 称为零假设；而新过程优于旧过程的猜测则描述成另一个假设 $H_1 : \theta > \theta_0$, 称为备择假设. 当然, 这里给出的是一个常规的单边检验问题. 类似地, 如果我们的猜测是另一个方向的或无倾向性, 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, 先假定零假设成立, 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抽样分布的中心位置附近, 不能距离中心位置太远. 这显然是零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, 因为这是在均匀假设前提下的几乎必然的抽样结果. 反之, 如果真实情况是预先未知的, 我们需要通过实验推测真实情况. 比如, 在少量的实验中, 我们发现了正面和负面出现的次数之间出现了差异, 甚至较大的差异, 一种推翻均匀假设的想法油然而生. 用逆否命题进行推断是假设检验的本质. 当然, 差距的大与小需要测量, 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, 认为有很大的把握怀疑这个离群的样本点 (outlier) 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, 从而拒绝数据对零假设的支持, 接受数据对备择假设的支持. “过远”是一个统计的概念, 我们用显著性来衡量. 几乎必然的含义是, 虽然拒绝

零假设的依据是样本偏离了零假设的分布, 然而在零假设下产生远离抽样分布中心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, 这种差异很有可能是由于数据与假设分布的不同而导致的必然结果. 对假设检验问题而言, 以下 3 个问题是值得探讨的.

(1) 如何选择零假设和备择假设：在数学上, 零假设和备择假设没有实际含义, 形式对称, 采取接受或拒绝的结论也是对称的. 但在统计实践中, 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总体的特点, 是部分的特点能否推广至整体的判断过程. 因而, 如果所建立的猜想与样本的表现相背离, 则这个猜想基本上是“空想”, 也就是说很难获得数据的支持. 比如：在研究问题 (a) 中：假定样本呈现出一种落后于旧过程的状况, 而非要将新过程优于旧过程的问题当作备择假设来处理, 这样的假设检验设置一般是不可取的, 当然也不可能有好的答案, 请参见习题 1.1. 这个例子至少揭示了一点：假设不是随意设定的, 而是根据样本的表现来设定的, 是需要花时间和精力了解和准备的. 如果数据背离理想的抽样分布, 从小概率原理来看, 似乎提出了可能拒绝零假设的证据, 接受备择假设, 认为是分布上的差异导致了样本对理想分布的偏移. 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, 零假设的设定则较为简单, 它是相对于备择假设而出现的. 我们认为, 唯有如此建立在经验基础上的假设才是对判断有意义的假设.

(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看, 起关键作用的是和检验目的有关的检验统计量 $T = T(X_1, X_2, \cdots, X_n)$ 和在零假设之下检验统计量的分布情况. 统计量的分布必须是已知的, 这样才能计算在零假设下 T 远离零假设所支持的中心, 以及和该值有关的某区间的精确概率或近似概率. 由于检验方法是由检验统计量 T 决定的, 通常也称为 T 检验. 在上面的单边检验参数问题中, 如果 T 大意味着备择假设 θ 大, 那么就要计算概率 $P(T > t)$; 这个概率称为检验的 p 值. 如果 p 值很小, 这说明统计量上的实现在零假设下是小概率事件, 这时如果拒绝零假设, 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 $\alpha > p$, 那么拒绝零假设, p 值可以认为是拒绝零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, 是单边检验的 p 值的 2 倍. p 值的概念如图 1.3 和图 1.4 所示. 对同一组数据, 同样的分布假定, 两者比较, 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.

(3) 两类错误：只要通过样本决策, 就不可避免真实情况和判断不一致的情况

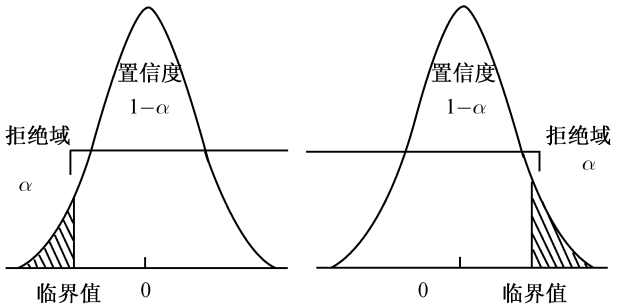

图 1.3 单边检验的 p 值

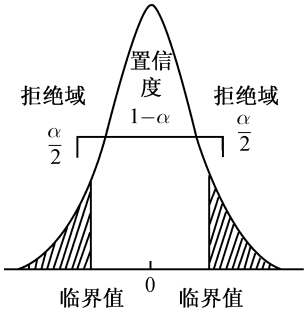

图 1.4 双边检验的 p 值

发生, 此时会犯决策错误. 在假设检验中, 有可能犯两类错误. 当拒绝零假设而实际的情况是零假设为真时, 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零假设差异显著性的 α 控制, 这表示拒绝零假设时出现决策错误的可能性不会超过 α, 因此拒绝零假设的决策可靠程度较高；另一种情况是, 当零假设不能被否定, 而实际情况是备择假设为真时, 犯第 II 类错误, 此时表现为零假设下样本统计量的 p 值较大. 不能拒绝零假设时, 如果选择接受零假设, 则会出现取伪错误. 假设检验的目的是为了给出临界值用于决策, 一个好的决策应该尽量让两类错误都小, 这在很多情况下是不现实的, 因为理论上, 第 I , II 类错误之间互相制衡, 不可能同时很小. 为了度量两类错误, 我们定义势函数如下：

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. 拒绝零假设的概率, 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, $1 - g_{T_n}(\theta)$ 是检

验犯第 II 类错误的概率. 我们用势函数将两类错误统一在一个函数中. 一个有意义的检验, 当显著性水平给定时, 检验的势函数应该越大越好, 低势的检验说明检验在区分零假设和备择假设方面的价值不大. 下面首先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 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第 I 类错误较小的检验域, 即 $\alpha(\lambda) = P\left(\sum_{i=1}^{n} X_i < C\right)$ 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 λ_0 从 0 变化到 2 的情况. 在零假设下, 我们注意到检验

$$\alpha(\lambda) = P(\text{拒绝零假设}|\text{零假设为真}) = P\left(\sum_{i=1}^{n} X_i < C|\lambda \in H_0\right),$$

$$\beta(\lambda) = 1 - P(\text{拒绝零假设}|\text{备择假设为真}) = 1 - P\left(\sum_{i=1}^{n} X_i < C|\lambda \in H_1\right).$$

检验势函数随分布参数的变化曲线如图 1.5 所示.

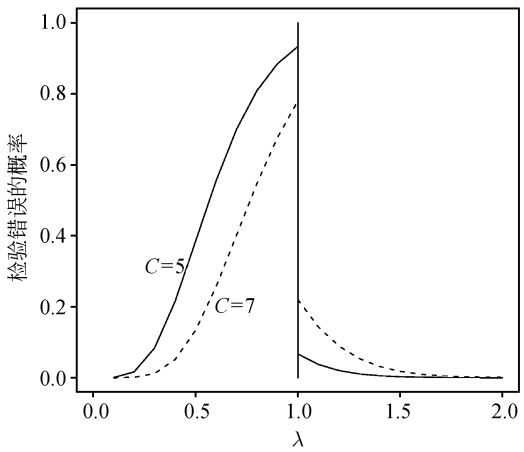

图 1.5 检验势函数随分布参数变化曲线

在图 1.5 中, 右边的两条曲线分别是 $C = 5$(实线) 和 $C = 7$(虚线) 犯第 I 类错误的概率, 我们观察发现犯第 I 类错误的概率在零假设下, 随着 λ 的增加而减小, 犯第 I 类错误在 $\lambda = 1$ 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显著的思想是一致的. 其中 $C = 5$ 的检验第 I 类错误率比

$C=7$ 的检验第 I 类错误率低, 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 类错误的可能性越小.

例 1.2 假设 X 服从概率密度函数为 $p(x)$ 的分布, 概率密度函数 $p(x)$ 有如下形式:

$$p(x)=\begin{cases}\dfrac{1}{\theta}\mathrm{e}^{-\frac{x}{\theta}}, & 0<x<+\infty,\\ 0, & \text{其他}.\end{cases}$$

考虑假设检验问题

$$H_0:\theta=2 \leftrightarrow H_1:\theta\geqslant 2.$$

简单随机抽样 X_1,X_2, 易知 $\dfrac{X_1+X_2}{2}$ 是 θ 的无偏估计, 因此构造如下拒绝域:

$$C=\{(X_1,X_2):9.5<X_1+X_2<+\infty\}.$$

计算该检验的 α 和 β.

解 在零假设之下, 由于 $X\sim\chi^2(2)$, 因此 $X_1+X_2\sim\chi^2(4)$. 计算样本落入拒绝域的概率:

$$\begin{aligned}P_\theta(\{X_1,X_2\}\in C)=&P(X_1+X_2>9.5)\\ =&1-\int_0^{9.5}\int_0^{9.5-x_2}\frac{1}{\theta^2}\exp\left(-\frac{x_1+x_2}{\theta}\right)\mathrm{d}x_1\mathrm{d}x_2\\ =&\frac{\theta+9.5}{\theta}\mathrm{e}^{-\frac{9.5}{\theta}},\quad \theta\geqslant 2.\end{aligned}$$

容易计算 $\alpha=P_2=0.05$, 因此拒绝零假设犯第 I 类错误的概率是 0.05, 继续计算可以得到: $P_4=0.31, P_{9.5}=\dfrac{2}{\mathrm{e}}$. 进一步计算可知 $P_\theta\geqslant 0.05, \theta\geqslant 2$.

$$\beta=P_\theta(\{X_1,X_2\}\notin C)=1-P_\theta(\{X_1,X_2\}\in C)\geqslant 0.05,\quad \theta>2.$$

这样, 当 $\theta>2$ 时, 犯第 II 类错误的概率至少是 0.05. 如果一个检验不犯第 II 类错误的概率不小于第 I 类错误的概率, 称这样一类检验为 **无偏检验**. 具体定义如下.

定义 1.2 设 W 表示一个检验的拒绝域, 对一般的假设检验问题, 如果

$$P(X\in W)\begin{cases}\leqslant\alpha, & \theta\in\Theta_0,\\ \geqslant\alpha, & \theta\in\Theta_1,\end{cases}$$

则称该检验为无偏检验.

因此, $C=\{(X_1,X_2):9.5<X_1+X_2<+\infty\}$ 是无偏检验. 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, 如果做接受零

假设的决策, 则可能存在着很大的潜在决策风险, 比如当实际的值与要比较的值比较接近时更应该尽量避免接受零假设. 实际上, 不能拒绝零假设的原因很多, 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, 当然也包括零假设本身就是对的. 和势有关的检验方面的问题, 我们将在 1.3 节中详细介绍. 总之, 盲目接受一个风险未知的假设不符合科学决策的基本精神.

(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：

$$(\hat{\theta} - C_\alpha,\ \hat{\theta} + C_\alpha). \tag{1.1}$$

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, 则可以拒绝零假设, 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, 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, 则表示不能拒绝零假设, 但是这没有表明 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 (左连续) 定义为：$F(x) = P(X < x), \forall x \in \mathbb{R}$. 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, $\forall x \in \mathbb{R}$, 定义

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leqslant x). \tag{1.2}$$

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 $\forall i = 1, 2, \cdots, n$, 定义变量 $Y_i = I(X_i \leqslant x), Y_i$ 服从的经验分布可以看成是在 $\{x_1, x_2, \cdots, x_n\}$ 上取值的均匀分布.

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, 于是以下结论成立：