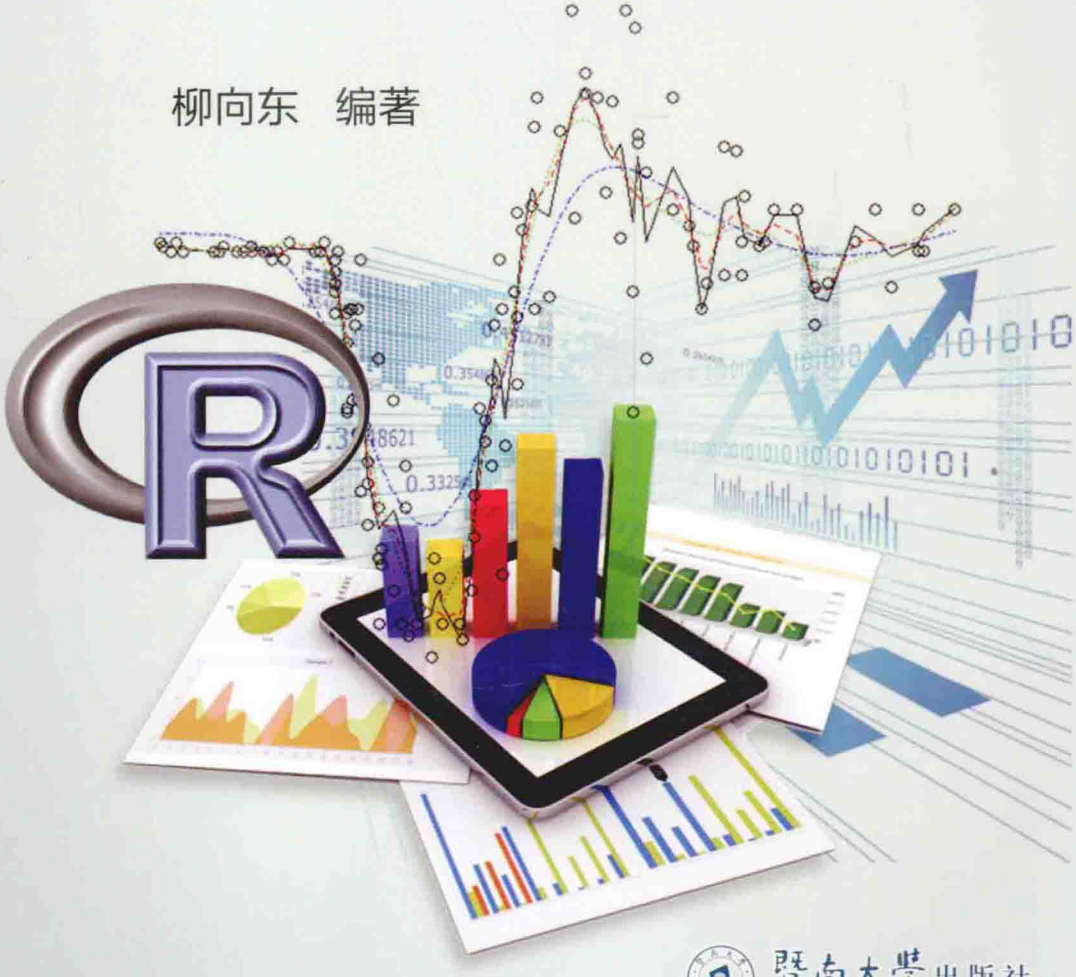
非参数统计

——基于R语言案例分析

Non-parameter Statistics

Based on the Case Analysis with R Language

柳向东　编著



暨南大學出版社
JINAN UNIVERSITY PRESS

本教材由国务院侨务办公室立项、彭磷基外招生人

非参数统计

——基于R语言案例分析

Non-parameter Statistics

Based on the Case Analysis with R Language

阜阳师范学院图书馆
藏书

柳向东 编著

暨南大學出版社
JINAN UNIVERSITY PRESS
中国·广州

图书在版编目（CIP）数据

非参数统计：基于R语言案例分析/柳向东编著．—广州：暨南大学出版社，2015.2
ISBN 978-7-5668-1320-6

Ⅰ.①非… Ⅱ.①柳… Ⅲ.①非参数统计—高等学校—教材 Ⅳ.①O212.7

中国版本图书馆CIP数据核字（2015）第011239号

出版发行：暨南大学出版社

地　址：中国广州暨南大学
电　话：总编室（8620）85221601
　　营销部（8620）85225284　85228291　85228292（邮购）
传　真：（8620）85221583（办公室）　85223774（营销部）
邮　编：510630
网　址：http://www.jnupress.com　http://press.jnu.edu.cn

排　版：广州联图广告有限公司
印　刷：佛山市浩文彩色印刷有限公司

开　本：787mm×1092mm　1/16
印　张：14
字　数：349千
版　次：2015年2月第1版
印　次：2015年2月第1次

定　价：29.80元

（暨大版图书如有印装质量问题，请与出版社总编室联系调换）

前　言

非参数统计是21世纪统计理论的三大发展方向之一。标准的参数方法强烈地依赖于对数据分布的假设，而非参数统计对模型要求甚少，不假定特定的总体分布，因此更加简单、稳健和适用。随着计算工具的发展，非参数统计模型在许多领域中应用更加广泛。非参数统计不仅是统计类学科的必修课，也是统计应用工作者必须掌握的基本方法和思想。

非参数统计以概率统计这门数学学科为基础，运用了很多现代统计思想和原理，因此它的原理涉及秩统计量的渐近正态、稳定性等复杂的统计学问题，较为抽象，对学生的数学基础要求较高，教学中存在的大量的公式推导、演算，必须借助现代化的计算工具，本书正是基于广泛使用的统计分析软件——R语言进行的。

R语言是GNU系统的一个自由、免费、源代码开放且功能强大的软件，是一个用于统计计算和统计制图的优秀工具，因此开发和使用R语言对我国统计事业的发展大有裨益。笔者根据十多年的教学经验认为，初学者只要了解总体和样本、随机变量及分布、统计量、检验和估计等统计学的最基本的内容，即可看懂本书。本书的重点不在于对公式的推导和演算上，而是在对非参数思想的理解和对实例的应用以及如何读懂结果和评价结果上。一旦掌握了R语言在非参数统计研究中的运用，就会有一种游刃有余的感觉。

本书从问题背景与动机、方法引进、理论基础、计算机R语言实现以及应用实例等诸多方面来介绍非参数方法，其内容包括：基于秩检验的符号检验、Wilcoxon检验、Kendall相关、列联表、Kolmogorov-Smirnov检验、非参数密度估计和回归等。本书在强调实用性的同时，也突出了应用方法与理论相结合。

本书可以作为非参数统计的教科书，同时笔者也希望本书能够成为查询非参数统计中最有用方法的快捷参考书，读者可通过本书了解如何使用最常用的非参数方法，并从中找到清晰的说明。

本书有以下几个特色和创新点：

（1）按问题的背景与动机、原理方法、假定条件、假设、检验统计量、实例分析以及计算机实现这样的顺序来编写，使内容显得条理清晰，适用易懂，便于学生了解该方法的直观意义以及来龙去脉、适用于哪类问题以及如何去解决该类问题。选用的案例具有时代性。

（2）每章都会通过R语言结合具体案例进行统计分析，并且有一定的习题，重要的习题都有相应的解答。

（3）图文并茂，利用R语言可以绘制出高质量的图形，一图胜过千言万语。

（4）本书可以作为非参数统计的实习实践教材，所有的计算结果都是由R语言计算出来的，十分值得信赖。

（5）注重处理那些条件少、数据非正态和总体分布未知的模型，突出非参数的稳定性和效率高的特点。

（6）非参数统计是一门统计科学，同时也是一门技术，只要有一般的数学和统计学

基础，只要知道总体和样本、随机变量及分布、统计量、检验和估计等统计学的最基本的内容，学好它就非常容易。

（7）本书不提供任何统计分布表，希望读者能够使用计算机和R语言进行统计分析。强调动手能力和计算机支持，实际上，如果没有计算机支持，很难对一定规模的数据在任何统计方向进行深入的分析。

本书共有7章，讲授60个学时较为合适和有效，建议每周3个学时（其中1个学时为上机实验）。本书的内容是在吸收国内外关于非参数统计的论著的基础上，根据在暨南大学十几年的教学过程中所撰写的讲义修改而成的。笔者还在最后章节尝试阐述了如何应用非参数的方法（包括非参数密度估计和回归、Kolmogorov-Smirnov检验、中位数回归等）写作本科毕业论文。感谢暨南大学经济学院统计学系的韩兆洲教授、刘建平教授、郑少智教授、王斌会教授、尹居良教授、陈平炎教授以及郭海华副教授和陈光慧副教授的大力支持。感谢陈天然、胡小娟、杨飞、钟聂等研究生全程参与了本书的编著工作，也要感谢统计学系的历届学生，特别是王文静、王琦、王思丹、范洋洋等同学，他们的参与使笔者在教学中对设计的内容越来越有体会，享受到了极大的乐趣。

由于笔者的知识和水平有限，书中难免有错误和不足之处，恳请读者批评指正！

柳向东

2014年9月于暨南大学经济学院统计学系

试读结束：需要全本请在线购买： www.ertongbook.com

目　录

1 统计推断

一般而言，一个典型的统计推断过程通常由5个步骤构成：假定分布族、抽样、计算统计量和抽样分布、进行估计与检验、评价模型。假定分布族是对实际问题的描述，它是统计推断的基础。然而在许多实际问题中，人们往往对总体的分布形式知之甚少，很难对总体的分布形式和统计模型做出明确的假定。甚至，有些时候对问题的数学描述本身就是问题的全部。比如，在人为控制因素不多的情况下的大部分经济和社会问题，其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，通过推断方法而获得结构关系，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。本章接下来将逐步介绍我们做统计推断时用到的一些基本概念。

1.1 总体、样本与统计量

我们对所居住的这个世界的大多数认识都来源于样本。比如，我们在某家餐馆吃过一次饭，于是会对这家餐馆的饭菜质量和服务水平有一个看法。我们结识了一些美国人，于是感觉自己差不多对所有美国人都有了一定的认识。在大多数情况下，从样本中获取的认识并不准确，但是，运用科学方法获得的样本却能够提供关于总体比较准确的信息。

科学观点的形成常常源于试验的框架。试验就是每一个步骤都规定得很明确的过程，而在试验之前，每一步的结果都是未知的。例如，检验一种新药的治疗效果的试验由以下几部分组成：选定治疗病人，按照规定的步骤服药，观察该药的治疗效果。检验产品质量的试验则包括两个部分：根据明确规定的步骤抽取和检验产品样本、记录试验结果。我们判断认知是否正确的主要方法也是通过试验验证。下面我们就研究对象——总体、样本进行一些介绍。

1.1.1 总体

总体是指一个统计问题研究对象的全体，它是具有某种（或某些）共同特征的元素的集合。总体是一个集合群体，其中每一个个体都具有已知的在样本内出现的概率。要注意总体的定义是根据所要研究的问题而定的。例如，如果要研究的是北京地区2008年长白猪的日增重，则总体由北京地区2008年所有长白猪的日增重构成。有时总体仅在理论上存在，而并不现实存在。例如，在研究某种药物对某种疾病的治疗效果（有效还是无效）时，我们将会利用一些发病个体进行药效试验，这部分个体可看成是来自一个假想总体的样本，这个假想总体由此药物对所有发病个体的治疗效果构成，它并不现实存在，因为并没有对所有发病个体用药。但是，在理论上，我们可以对所有发病个体用药。

1.1.2 样本

统计分析的目的就是要对总体的特征、不同总体间的差异等做出推断。由于总体往往很大，而且常常是无限的、动态的和假想的，所以不可能收集到总体中每个个体的数据资料，通常的做法是从总体中按一定的方法抽取部分具有代表性的个体，这部分抽取出来的个体称为样本，即我们把总体中某些元素的集合称为样本。要使一个样本含有关于总体的可靠资料，样本中的每个个体必须在随机的情况下抽取。随机抽取意味着总体内的每一个个体具有已知的在样本中出现的概率。这不是抽样者所能随意判断的，而统计分析的基本任务就是要通过对样本的分析来推断总体。

此外根据获取方法的不同，样本可以分不同的类型。比如，方便样本（convenience sample）是一些最容易获得元素的集合，在街上采访的市民或者电话调查均属此类。我们不太可能从方便样本中获得总体参数的精确估计，而概率样本（probability sample）则能够相对精确地描述总体的未知参数，因为概率样本要求总体中每一个元素都有已知的非零概率。本书中所考虑的概率样本是随机样本（random sample），这个概念我们将在本节的后面定义。

1.1.3 目标总体与样本总体

假如一名心理学家想要研究不停地打断一个人的睡眠对其情绪稳定的影响，他所考虑的总体应是当代的所有人。为了进行试验，他在大学校报上刊登广告来招聘所需要的有偿志愿者。他所抽取的样本很难具有代表性，因为这些志愿者都是大学生，来自同一所大学，年龄范围相当狭窄，并且有某种相似的性情促使他们回应报纸上的研究广告，并应聘成为某项人体试验的志愿者。但是，由于很多实际情况的制约，比如研究资金和时间有限，他不得不使用这种类型的样本，否则就得放弃整个试验。因此有两种总体是值得一提的，即研究的目标总体和实际的样本总体。

我们需要从中获取信息的总体称为目标总体（target population），而从中抽样的总体称为样本总体（sample population）。上面的例子中当代人类的全体是目标总体，而来应聘的志愿者是样本总体。所有的试验者都只能基于样本总体来研究问题，而试验的有效性取决于样本总体与目标总体相似的假设，至少在我们所研究的性质上是相似的。

1.1.4 随机样本

本书所讨论的统计方法通常假设样本是随机样本，所以介绍随机样本的有关概念是很重要的。我们假定总体元素的个数是有限的 N，这里 N 可以很大也可以很小。总体中每个元素的重要性相同，且等可能被抽到。容量为 n（$n<N$）的一组样本可以这样抽取：将总体中所有元素从1到 N 进行编号，从中随机抽取 n 个号码，使得出现任意 n 个号码的组合等可能，这 n 个号码对应着总体中的 n 个元素。这种抽样方法通常是无放回（without replacement）的，所有相同的元素在样本中出现的次数不会多于一次，而对于有放回（with replacement）的抽样，相同的元素可能出现两次或两次以上。

【定义1.1】从有限总体中任意抽取一组容量为 n 的样本，如果每组样本出现的可能性相等，那么称这样得到的样本为随机样本。

试读结束：需要全本请在线购买： www.ertongbook.com

上面定义中的“随机”不是针对样本本身，而是指获取样本的抽样方法，这一点看起来似乎有些奇怪。事实上，我们是通过抽样方法，而不是通过样本本身来判断一组样本到底是不是随机样本。

假如一个有限总体共有 N 个元素，那么无放回抽样得到的容量为 n 的样本共有 $\binom{N}{n}$ 种可能，有放回抽样样本共有 N^n 种可能。若每组样本出现的可能性相等，则认为这样的抽样方法是随机的，得到的样本是随机样本。

当总体有限时，前面对随机样本的定义在大多数情况下是合适的。但是，假如我们要考察某人在一个晚上做梦的个数，可能会遇到麻烦。在这种情况下，我们认为“随机样本”指某一晚做梦的个数、另一晚做梦的个数，直至比如说七个晚上做梦的个数。即使在理想的情形下，这种抽样方法也不能符合定义 1.1 中的“等可能性”这一概念。什么叫等可能性？不是针对个体，因为前面我们假设的研究对象只是个体，不是总体的一个代表（尽管这可能是我们想要研究的最终目标）。为了保证等可能性，我们不可能在这个人被期望能够活着的夜晚中，选择一些夜晚来做研究。所以，随机样本至少还需要一个其他的定义。数理统计中随机样本的标准定义如下所述：

【定义 1.2】容量为 n 的随机样本（random sample of size n）是指一组 n 个独立同分布的随机变量序列 X_1，X_2，…，X_n。

在定义 1.1 中，如果抽样方法是有放回时，则定义 1.1 和定义 1.2 是相同的，并且当且仅当在这种情形下才是独立的。无放回抽样产生的观测是非独立的，因为某个个体一旦被选中且不放回，就意味着它不可能再被抽取到。然而，如果总体容量 N 很大，有放回抽样和无放回抽样在实际应用中的差别非常小，就可以忽略这种观测间轻微的不独立性。本书中的定理和公式的推导都假设样本中的观测是独立的。对于有限总体，这些定理在其他假设下的修正是存在的，但不在本书的考虑范围之内。这种修正的效果只要在样本量 n 小于总体容量 10% 的情况下就可以被忽略。

1.1.5 多元随机变量

试验者可能会测量或观测到定义 1.1 中随机样本的每个被选元素，以及定义 1.2 中的每个随机变量 X_i 的几个相互关联的特征，在这种情况下，用来描述几个特征的随机变量通常有两个脚标，比如 Y_{ij}，这里第一个脚标表示所选样本的个体，第二个脚标表示被测量或观测的某个特征。

也就是说，X_i 实际表示的是 k 维随机变量（Y_{i1}，Y_{i2}，…，Y_{ik}），X_i 仍然是独立同分布的，但是 X_i 中的每个随机变量 Y_{ij} 可以是独立的，也可以是非独立的，可以是同分布，也可以是不同的分布。

例如之前讨论的“梦”的试验，随机变量 X_i 表示第 i 个观测夜晚做的梦的个数，假设 X_i 是独立的且同分布（意思是每个 X_i 都有相同的分布函数）有一定的合理性。但是如果试验者每晚不仅记录梦的总数，还记录整个睡眠时间，我们分别用 Y_{i1}，Y_{i2} 表示，这样每晚做梦的个数和睡眠时间可能是相关的变量，所以 Y_{i1}，Y_{i2} 很可能不是独立的。但是，每个晚上的睡眠模式彼此是独立的。在数学上，这意味着 Y_{i1}，Y_{i2}，Y_{j1}，Y_{j2} 的联合概率分布函数可以分解如下：

$$f(y_{i1}, y_{i2}, y_{j1}, y_{j2}) = f_1(y_{i1}, y_{i2}) f_2(y_{j1}, y_{j2})$$

这里f_1和f_2分别是(Y_{i1}, Y_{i2})和(Y_{j1}, Y_{j2})的联合概率函数。假如连续两晚睡觉模式的联合概率分布不变，即f_1和f_2一样，那么我们可以说(Y_{i1}, Y_{i2})和(Y_{j1}, Y_{j2})有相同的分布。为了更方便地表达这种关系，即随机向量之间要求独立同分布，而随机向量内部的随机变量不必独立同分布，我们可以用Y_{i1}，Y_{i2}的联合来表示X_i，这时称X_i为二维随机变量。X_i的值实际上包括两个值，一个是Y_{i1}的值，一个是Y_{i2}的值。这样，前面所述的可以概括为“随机变量$\{X_i\}$是独立同分布的”。

类似地，我们还可以考虑每晚有k个测量，它们是Y_{i1}，Y_{i2}，…，Y_{ik}，用X_i来表示这k个随机变量，那么称X_i为k维随机变量（k-variate random variable），或是多维随机变量（multivariate random variable）。X_i是独立的就意味着所有$\{X_i\}$的联合概率分布可以分解成n个联合概率函数的乘积，并且每个都是Y_{i1}，$Y_{Y_{i2}}$，…，Y_{ik}的联合概率函数。同样地，X_i同分布是指上面提到的联合概率函数是相同的函数。

现在我们有两种随机样本的定义，第一种定义仅适用于有限总体样本并且直接与样本空间联系在一起。如果每一种可能的样本（容量为n）表示成样本空间中的一点，且样本空间中每个点被选为样本的概率相等，那么这种抽样方法是随机的，且抽得的样本是随机样本。上面的定义中，我们仅用到样本空间以及概率函数的概念，但是并没有明确或含蓄地提及随机变量这一概念。

【例1.1】一个心理学家希望选取4名研究对象来进行个体训练和考试。他登出广告，有20个志愿者应聘。他有多种从容量为20的样本总体中抽取一容量为4的样本的方法。他可能会选择最先来应聘的4名志愿者，也可能会选择那些积极主动的志愿者，但这可能就不是随机样本。他也可能严格按照定义1.1来考虑，选择容量为4的样本，即有$\binom{20}{4}=4\,845$种可能。抽取样本时，他可以用4 845张同样的纸，每张纸上写4个名字，每张纸上的组合都不同，然后把它们放到篮子里随机地抽取一张，抽取出来的纸片上的4个人则被选中。这样得到的是随机样本，但这种抽样方法是不现实的。

另外一种获得随机样本的方法是，把20个名字写在20张纸上，然后以某种随机方式一个接一个地抽取4张纸，比如可以从装满这些纸的一个帽子中抽取。这种抽样方法同样满足随机样本的定义，这个过程可以通过计算机编程来模拟。

随机样本的第二个定义直接与随机变量有关，而不涉及样本空间。但是，由于随机变量是定义在一个样本空间上的函数，尽管我们没有直接引进样本空间这一概念，但是它隐含在实际背景中。随机变量所有可能取值的全体构成了样本空间，有时，为了解决出现的统计问题，将近似样本空间的点列举出来是必要的。实际上，如果所有可能的测量结果（随机变量假设的值）都是样本空间中的点，那么就不会产生混淆。我们通常认为这些测量结果是数值，但是有时测量的数值很难清楚地表达出来。所以，我们最好讨论各种不同类型的测量。

1.1.6 度量尺度

度量的类型通常被称为度量尺度（scale of measurement），各种不同的出版物都详尽地讨论过，其中包括Stevens（1946）的一篇优秀论文。我们将逐一介绍名义尺度（即

“最弱”的度量尺度）、次序尺度、区间尺度，比率尺度（即“最强”的尺度）。

第一种尺度是度量的名义尺度（nominal scale），它是使用数字将性质或元素分成不同种类或范畴的一种方法。分配到观测上的数字只是用作“名字”，以便说明观测所在的种类或范畴，因此叫做“名义尺度”。对掷硬币，我们可以定义随机变量：硬币正面朝上时，记为1，反面朝上时，记为0，这时使用了度量的名义尺度。我们也可以适当地选择7.3和3.9来分别表示正面和反面，我们选择0和1主要是因为方便计算所掷硬币中正面朝上的次数。当把12个研究对象用1到12个数字任意标号时，这时使用了度量的名义尺度，号码的分配则是随机变量的一种形式。当根据颜色将研究对象分类时，种类可以用1、2、3或蓝、黄、红或A、B、C来标记。这些号码只是类别的名字，当然只要种类保持不变，也可以用其他未使用过的号码来标记。

第二种尺度是度量的区间尺度（interval scale）。在一般的度量中，不仅会考虑度量的次序尺度，还会考虑把两个度量区间的大小，即两个度量间差别（从减法的意义上讲）的大小作为相关信息。区间尺度涉及一些单位长度的概念，任意两个度量间的距离都可以用一些单位长度的倍数来表示。用来理解区间尺度这一概念最好的例子就是我们日常生活中温度的表示法。温度的单位（度）定义为温度计中一定体积水银柱的变化量。因此，任意两个温度的差别可以用这个单位或度来衡量。温度的实际数值只是和一个任选为“零度”的点的比较。测量的区间尺度需要一个零点和一个单位长度（只有后者没有前者是不行的），但是哪个点定义为零点，哪种长度定义为单位长度并不重要。温度可以同时由华氏温标和摄氏温标来计量，它们有不同的零度和不同定义的1度或单位。区间度量的法则不会因刻度或位置或两者同时的改变而受到干扰。

第三种尺度是度量的次序尺度（ordinal scale），它用于存在诸如“更小”、“更大”、“相等”这些比较关系的度量中。度量的这些具体数字只是用来从小到大有序地排列元素的一种工具，由于它能够根据度量的相应大小对元素进行排序，所以称为次序尺度。如果其中一些元素彼此相等，我们称为结。当一个人用数字1来表示3个品牌中最喜欢的一个，3表示最不喜欢的一个，2表示剩下的那个品牌，这时，他就是在使用度量的次序尺度，数字只是他表达喜欢程度的一种方式。当然，他可以用任意三个数如16，20，75来代替1，2，3，只要这三个数的相关顺序能够表达出他相应的喜欢程度就行。

第四种尺度是度量的比率尺度（ratio scale）。当次序和区间的大小很重要，而且两度量的比率也很有意义时，我们需要引入度量的比率尺度。如果，一个量是另一个量的“2倍”是合理的话，那么引入度量的比率尺度就是合适的，如度量产量、距离、重量、高度、收入等。实际上，比率尺度和区间尺度的唯一差别是前者要求有绝对零点，而后者的零点可以是任意一点，和区间尺度一样，比率尺度的单位长度也是可以任意定义的。

我们不可能就度量本身来谈哪种度量尺度是合适的，而应该考虑被度量的量以及度量方法，然后再决定赋予度量数值的含义。

关于这四种度量尺度，科学家们没有达成一致的意见。有些科学家喜欢用其他尺度，而有些度量也不能清楚地归类于上面四种尺度的任何一种。这样看来，上面的分类显得把问题过于简单化了，但针对本书目的而言已经足够了。

大多数常用参数统计方法要求度量是区间尺度（或者比这更强的尺度），而大多数非参数统计方法通常采用名义尺度和次序尺度。当然，每种度量尺度都拥有弱度量尺度

的所有性质。因此，只需要弱度量的统计方法可能也会用到强度量。

1.1.7 统计量

到目前为止，我们已经讨论了总体、来自总体的样本，以及度量样本所感兴趣的性质的度量尺度。度量尺度涉及随机变量，因为度量样本元素的体系实际上就是一个随机变量。由于统计量（statistic）是随机变量，因此，度量尺度与统计量有关。对于数理统计学家来说，“统计量”和“随机变量”这两个术语是可以互换的。但是，统计量一词的普遍使用表明它不仅仅是一个随机变量。

统计量是描述样本特征的量，如样本平均数、样本方差、样本相关系数等。统计量可以由样本观测值计算得到，因而是样本观测值的函数。一般来说，每一个总体参数都有一个对应的样本统计量。因而由样本推断总体也可以理解为由统计量推断参数。

【定义1.3】一个统计量是将样本空间中的样本点映射到实数上的函数，其中样本空间中的样本点是一些多元随机变量的所有可能值。换句话说，统计量就是几个随机变量的函数。

作为统计量的定义，定义1.3中的每一句话都是充分的，它们清楚地阐述了这个概念。

【例1.2】用 X_1，X_2，…，X_n 表示 n 个学生的考试分数，每个 X_i 都是随机变量，令 W 等于考试分数的平均值，则

$$W=\frac{X_1+X_2+\cdots+X_n}{n}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

式中，W 是一个统计量。若 $X_1=78$，$X_2=86$，$X_3=88$，那么3个学生的考试分数的平均值为

$$W=\frac{78+86+88}{3}=84$$

统计量 W 满足定义1.3中的第二句话：它是随机变量 X_1，X_2，…，X_n 的函数。由于 W 将随机变量（X_1，X_2，…，X_n）映射到实数，这满足定义1.3中的第一句话。这时，若多元随机变量（X_1，X_2，X_3）的值为（78，86，88），那么统计量 W 的值为84。统计学中经常应用这一特殊的统计量，称为“样本均值”，下一节中将进一步讨论它。

1.1.8 顺序统计量与秩

因为非参数统计方法并不假定总体分布。因此，观测值的顺序及其性质则作为研究的对象。对于样本 X_1，X_2，…，X_n，如果按照升幂排列，得到

$$X_{(1)}<X_{(2)}<\cdots<X_{(n)}$$

这就是顺序统计量（order statistic），其中 $X_{(i)}$ 为第 i 个顺序统计量。对它的性质的研究构成了非参数统计的理论基础之一。本书并不试图在理论证明上作深入的推导，但是了解顺序统计量的基本性质对了解非参数方法的思维方法还是有益处的。

许多初等的统计概念是基于顺序统计量的，比如中位数的定义为

$$M=\begin{cases}X_{\left(\frac{n+1}{2}\right)}, & n\text{ 为奇数}\\ \frac{1}{2}\left\{X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}\right\}, & n\text{ 为偶数}\end{cases}$$

而极差（range）定义为

$$W = X_{(n)} - X_{(1)}$$

如果样本是独立随机样本，则中位数和极差常作为位置和尺度的度量。另一个位置估计量为修整均值（trimmed mean），定义为

$$T(j) = \sum_{i=j+1}^{n-j} \frac{X_{(i)}}{n-2j},\ 0 < j < \frac{1}{2}n$$

这里参数 j 为求均值之前删掉的最大的或最小的一些观测值的数目。有时也用被删除观测值的百分比 α 作为参数。注意，当修整均值所修整的部分的百分比为 0 时，就是均值，当 α 为一半（n 为奇数）或者接近一半（n 为偶数）时，修整均值为中位数。

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

$$F_r(x) = P(X_{(r)} \leqslant x) = P(\text{至少 } r \text{ 个 } X_i \text{ 小于等于 } x)$$

$$= \sum_{i=r}^{n} \binom{n}{i} F^i(x)[1-F(x)]^{n-i}$$

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

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

顺序统计量 $X_{(r)}$ 和 $X_{(s)}$ 的联合密度函数为

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

我们从此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

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

因为本书所采用的方法主要是以秩为基础的，自然要讨论介绍秩的有关分布。如果用 R_i 来代表独立同分布样本 X_1，X_2，…，X_n 中 X_i 的秩，它为小于或等于 X_i 的样本点个数，即

$$R_i = \sum_{j=1}^{n} \mathrm{I}(X_j \leqslant X_i)$$

记 $R = (R_1, \cdots, R_n)$，可以证明：对于（1，…，n）的任意一个排列（i_1，…，i_n），R_1，…，R_n 的联合分布为

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

由此可得

$$P(R_i = r) = \frac{1}{n}$$

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

$$E(R_i) = \frac{n+1}{2}$$

$$\mathrm{Var}(R_i) = \frac{(n+1)(n-1)}{12}$$

$$\mathrm{Cov}(R_i, R_j) = -\frac{n+1}{12}$$

类似地，可以得到 R_1，…，R_n 的各种可能的联合分布及有关的矩。对于独立同分布样本来说，秩的分布和总体分布无关。

上面介绍的顺序统计量和秩的一些性质可帮助认识基于秩的统计量的分布性质，以及给定具体总体分布时非参数方法对参数方法的相对效率。此外，我们还将在后面的章节中介绍很多其他有用的统计量，并进一步讨论这些统计量在分析试验结果中的作用。

1.2 估计

统计量的一个基本目的是估计总体的未知性质。这些估计出来的未知性质通常是用数字表示的，并且包括可列举的一些项目，例如未知比率、均值、概率等。事实上，估计是基于样本的（如果有概率描述，则是随机样本），并且估计是关于随机变量分布未知性质的有根据推测，这里随机变量表示对总体研究感兴趣的量。例如，我们可以用产品中样本的不合格率来估计总体的不合格率。用来做估计的统计量自然叫做估计量（estimator）。本节我们将要讨论一些估计量，例如样本均值（sample mean）、样本方差（sample variance）和样本分位数（sample quantiles）。我们首先引入一个与众不同的估计量——经验分布函数（empirical distribution function）。

1.2.1 经验分布函数

一个随机变量的真实分布函数一般是未知的，有时我们只能够推测分布函数的形式，或将推测作为真实分布函数的一个近似。根据样本的观测值作经验分布函数图，以此来作为整个未知分布函数 $F(x)$ 的估计，这是推测分布函数的一种好方法。下面将用具体的例子来介绍这种估计方法，由此我们给出定义：

【定义 1.4】设 X_1，X_2，…，X_n 是一组随机样本，经验分布函数 $S(x)$（简称为 EDF）是 x 的函数，它在 x 点的取值为小于或等于 x 的 X_i 在样本总数中所占的比例，其中 $-\infty < x < \infty$。

【例 1.3】在一项体能研究中，一高中随机抽取了 5 名男生，记录他们跑完 1 英里的时间。时间（转化成分钟后）分别为 6.23，5.58，7.06，6.42，5.20。由于经验分布函数 $S(x)$ 是小于或等于 x 的 X_i 在样本总数中所占的比例，根据这组特定样本有如下经验分布函数：

$$S(x)=\begin{cases}0, & x<5.20\\ 1/5, & 5.20\leqslant x<5.58\\ 2/5, & 5.58\leqslant x<6.23\\ 3/5, & 6.23\leqslant x<6.42\\ 4/5, & 6.42\leqslant x<7.06\\ 1, & x\geqslant 7.06\end{cases}$$

我们也可以很方便地画出该经验分布函数的图象，并且从例 1.3 中可以看出，经验分布函数总是阶梯函数，每阶的高度是 $1/n$，并且只在样本取值处有变化。我们从左到右来考虑经验分布函数的值，注意到 $S(x)$ 在样本最小值前均取值为零，在每个样本取值处会增加一阶的跃度，每个跃度是 $1/n$。在样本最大之处 $S(x)$ 取最大值 1，并且在剩下所有比样本最大值大的 x 处都取 1。$S(x)$ 很像非降的取值从 0 到 1 的分布函数。但 $S(x)$

试读结束：需要全本请在线购买：www.ertongbook.com