高等院校应用创新教材

SHIYONG TONGJIXUE FANGFA

实用统计学方法

李兆元　刘　萍◎著

科学出版社

高等院校应用创新教材

SHIYONG TONGJIXUE FANGFA

实用统计学方法

李兆元　刘　萍◎著

科学出版社
北京

内 容 简 介

本书以非参数统计方法为主，针对不同数据类型和性质以及不同检验目的，介绍了41种常用统计方法。这些方法计算过程简单，对数据质量要求低，适用于各种数据类型。在介绍具体方法前，本书简要介绍了方法中涉及的基本概念，以帮助读者理解“数据”，去除“数据”的神秘性。本书重点不在方法的数理推导，而在如何选择正确的方法以供研究使用。全书尽量避免使用数学语言，回避繁杂的数理推演。书中增加检索表形式引导读者寻找所需方法，书末提供术语检索，方便读者随时翻阅相关定义。

本书可作为生物科学、行为科学、医学和社会科学等专业本科生、研究生的教材，也可作为研究者的案头工具书。

图书在版编目(CIP)数据

实用统计学方法/李兆元，刘萍著. —北京：科学出版社，2018
（高等院校应用创新教材）
ISBN 978-7-03-053812-3

Ⅰ. ①实… Ⅱ. ①李… ②刘… Ⅲ. ①统计方法-高等学校-教材
Ⅳ. ①C81

中国版本图书馆CIP数据核字（2017）第141178号

责任编辑：周艳萍 刘文军 / 责任校对：王万红
责任印制：吕春珉 / 封面设计：东方人华

科学出版社出版
北京东黄城根北街16号
邮政编码：100717
http://www.sciencep.com
三河市骏杰印刷有限公司印刷
科学出版社发行 各地新华书店经销
*
2018年3月第 一 版 开本：B5（720×1000）
2018年3月第一次印刷 印张：18 3/4
字数：444 000
定价：68.00元
（如有印装质量问题，我社负责调换〈骏杰〉）
销售部电话 010-62136230 编辑部电话 010-62151061

作者简介 >>

李兆元 西南林业大学教授，博士，研究生导师，中国灵长类学会理事，英国苏格兰皇家动物学会顾问。先后就读于华中师范大学、中国科学院昆明动物研究所、英国爱丁堡大学，主要从事生态学、行为学和动物地理学研究。在国内外学术期刊发表论文30余篇。长期关注旅游学、医学和社会科学等研究，并为相关科研工作提供数据采集和分析方法。参与出版的《叶猴生物学》，曾获中国科学院自然科学二等奖和新闻出版署全国优秀科技图书奖。

扫二维码，加入我们

责任编辑　周艳萍　刘文军

前　言

统计学帮助人们从或然性大、变异性高的数据中寻找规律。传统统计学是参数统计，对数据信息量、样本量、数据分布类型有较高的要求。然而，在生物学、医学、心理学、行为学、政治学、社会学、旅游学，以及其他学科中，研究者经常面临的现象重复率低、随机干扰因素多、数据信息含量低、不符合理论分布类型等问题，导致数据难以用参数统计方法进行检验。另外，传统教材大量使用数学术语和数理推演，导致使用者难以理解。

作者在多年的教学和科研工作中发现，缺乏统计学的相关知识制约着许多人的学术发展。例如，一些研究人员投稿到国际学术期刊后，因返修意见常常提到统计学方法的问题而烦恼；再者，一些已发表的文章中因统计学方法有误，致使其他人不敢引用其结论。而每年毕业生论文答辩的情况更不容乐观，论文要么完全没有统计分析，导致数据价值无法实现；要么是统计方法用错，导致结论错误。甚至还有人认为，社会科学不需要统计学，这种观点误导了大批教育者及受教育者。

以上问题，究其原因有三：第一，知道应用统计学方法处理数据，但是不会用。许多人学过统计学。但在上课时，老师大多用数学语言进行数理推导，学生往往听得云里雾里，结果仍然不会用，甚至对统计学“谈虎色变”。第二，老师介绍的是参数统计方法，对数据质量要求很高；学生采集到的数据不符合要求，但仍然使用，导致研究结论出错。第三，由于不了解统计学，以为用统计学便是用物理学、化学中的那些公式；而自己的数据变异性大，重复率低，无法使用那些公式，便认为自己的研究领域不需要统计学。

编写本书的目的在于引导读者正确选择和使用统计方法，因而写作过程中坚持的第一个原则是易读性和易操作性。在不影响对统计学思想的理解及方法运用的前提下，尽可能少地使用数学术语，使晦涩的统计学变得通俗易懂；同时，回避公式推导过程，只陈述如何选择和使用。第二个原则是实用性。为此，书中介绍的方法大多是非参数统计方法，对各种质量的数据都适用。

本书主要参考了相关文献（Siegel et al.，1988；Fowler et al.，1998），并结合

了作者的以及文献中的研究案例。

作者在写作过程中，得到西南林业大学胥辉教授、西北大学李保国教授、西南林业大学赵龙庆教授的特别指导，以及合肥师范学院李进华教授、四川省民族研究所罗凉昭研究员、西南民族大学马尚林教授、云南师范大学杨士剑教授、西南林业大学杨晓军教授和西华师范大学周材权教授的支持，研究生西南林业大学郭凯飞和刘地为本书作图，地理学院也给予了支持，谨此致谢！

由于编写时间有限，书中难免存在不妥和疏漏之处，恳请读者批评指正。

目　录

第1章

绪　论

1.1　本书为谁写

一位研究生撰写毕业论文时，眉头紧锁 3 个月。当我们协助他在 2 小时内完成了全部数据的统计检验后，他突然眉开眼笑道："老师，原来统计学如此简单啊！我心烦好几个月呢！"他与这些年来我们遇到的很多学生一样，对统计学心存畏惧，原因是在学习统计学课程时，老师无论是进行公式的数理推导，还是介绍公式的来龙去脉，使用的都是数学术语，使学生听得云里雾里，以致学完课程仍然不得要领。考试前，担心不及格夜不能寐，刻苦死记硬背，最多考个及格；考试结束，再也不愿意提及"统计学"这 3 个字。如果捧着本书的您就是这样一位，那这本书就是为您写的。

我国学术界目前有一种观点：社会科学不需要数学，因为社会科学中的现象重复率很低（有些人甚至认为不能重复）。这种观点与国际学术界的普遍观点非常不同。很显然，持此观点的人不了解统计学的奥妙和功能。低重复率、数据变异、规律的或然性是生物学和行为学的数据特征。这些学科可以用数学（Fowler et al.，1998；Siegel et al.，1988），社会科学也可以用数学，而且应该用。关于数学在社会科学中的地位，可以参阅相关的文献著作（威尔逊，1988；巴比，1987；Stern，1979）。现在广泛使用的统计学软件包 SPSS 的全名是 Statistical Program for Social Sciences，即社会科学的统计学程序。SPSS 最早就是提供给社会科学界使用的统计学工具。

如果您是一位在读研究生，正在写开题报告，建议您浏览一下本书，它可能

会帮助您在数据采集地中某个不起眼的角落里发现有价值的数据。当然，读了这本书，肯定能帮助您设计数据采集表格，确保外业的数据采集的操作过程符合规范，避免后期发现数据不符合要求时因无法弥补而导致时间和经费的浪费。

随着计算机的普及和软件的广泛使用，人们开始高度依赖计算机进行数据统计检验，而且认为因此可以免修统计学。但是，由于总是照葫芦画瓢，看着文献上别人用什么方法，自己就用该文献中的方法，因此许多人不知道如何选择恰当的方法。每种方法都有其使用的前提条件，采集数据时的客观条件不同，样本数据就不同，简单照搬别人的方法风险极高。例如，在进行因子相关分析中，因数据达不到参数统计要求而应该用斯皮尔曼相关检验（非参数统计方法）时，学生常常使用皮尔逊积距相关系数检验（参数统计方法）。当被问及用什么方法时，他们回答："相关性检验。"再问："哪种相关性检验？"便回答不上来，因为他们不知道有适用于不同情况的两种方法。另外，皮尔逊积距相关系数检验的结果输出为 r 和 p 值，学生普遍明白，p 等于/小于 0.05 表明有显著意义，但不明白为什么等于/小于 0.05。而知道 r 的含义的人就更少了。相应地，研究者从皮尔逊积距相关系数分析中获益甚少。因此，如果您有计算机和相关软件，建议先读本书；如果您没有计算机或相关软件，则可以借助本书完成统计处理。本书介绍的大多数方法，只需一个科学计算器（多数智能手机的普通计算器横过来使用，就变成了科学计算器）、一张纸和一支笔便可完成统计计算。如果利用 Microsoft Excel，则更容易完成。因此，本书可作为您的案头工具书。

1.2 为什么要学统计学

科学是人类用实证主义对自然现象和自然过程进行解释的一种认识世界的方法和活动。实证主义，就是一切理论、观点都建立在事实基础上。在科学研究早期，用于建立科学理论的事实来源于简单的观察。随着研究的深入，观察方法已经不足以解释那些似是而非的自然现象，这时开始使用实验方法。无论是观察方法还是实验方法，科学家一开始都是对自然现象进行简单描述，即定性描述。随着研究的进一步深入，定性描述也无法解释一些复杂的自然现象，定量研究方法应运而生。在科学方法史上，与观察方法相比，实验研究方法的优势在于研究者能够控制一些条件不变，从而清楚地界定特定因素对于一个自然现象或者自然过程的影响和贡献；与定性研究方法相比，定量研究方法更能精确地解释自然现象。例如，当年牛顿在苹果树下，那只掉在他头上的苹果经过若干次往上抛，总是往下掉；无论往上抛得有多高，都无法将苹果抛到地球外，这是万有引力存在的证

明。但是，如果没有进一步的实验数据采集，从而建立起万有引力公式，那现在万有引力充其量只是人们茶余饭后的谈资，在科学领域里不会有一席之地，今天的太空旅行、宇宙飞船，甚至连飞机都无法想象。因此，数据和相应的数学工具在科学研究领域中的应用程度就标志着这门学科发展的水平。

人类在结绳记事的年代就开始使用数字。当我们以一些代码来代表自然现象的组成成分，以数字代表这些成分表现的程度后，就可以通过经验观察和反复试验获得一系列数据，并且通过数据的变化建立数学模型。在这些模型中，自然现象各组分间的关系得以通过数字间的关系体现。通过人为改变模型中自变量的数据，观察因变量的变化，来预测自然现象的变化趋势。然后进行观察/实验数据采集，来检验和修正数学模型。经过反复检验修正后的模型能够准确反映自然现象的变化模式，科学理论就诞生了。

科学发展的过程是一个不断自我否定的过程。在这个过程中，错误理论的抛弃、正确理论的诞生周而复始。一个科学理论的正确与否，很大程度上取决于数据采集方法的客观性和合理性及数据处理方法的恰当性。因此，研究方法（包括数据采集方法和处理方法）成为科学研究活动成败的关键因素。

在科学研究活动中，科学家们面对不同的自然现象，使用不同的方法。在物理学和化学领域，科学家们面对的自然现象有极高的重复性。例如，伽利略在比萨斜塔上做过的重力加速度实验，我们可以在巴黎埃菲尔铁塔上重复，也可以在上海东方明珠塔上重复，而且结果极端相似；瓦特当年在爱丁堡进行的蒸汽动力实验，可以在几百年后的上海江南造船厂获得同样的实验数据。现在，在世界各地中学的实验室中，每年都无数次地重复着氢氧化学反应的实验。在这些现象中，时间和空间维度的影响很小。因此，科学家们可以通过实验方法进行数据采集，通过线性代数或者其他复杂的数学工具进行数据分析。然而在生物学领域中，由于个体间存在着很大的遗传背景差异，个体间在各种生物现象中（形态解剖学、生理学、生物化学、行为学、生态学、心理学……）的表现存在着不同程度的差异，一些生物学过程（如演化历史）更是由于时间维度的严重影响而使其具备历史性（不可重复性）。这些特征使得生物学现象比物理学或化学现象复杂得多，研究活动中有时使用实验方法（生理学、生物化学），有时使用观察方法（形态解剖学），有时两种方法都使用（行为学）。在野外动物学研究中，大多数生物学现象和生物学过程不是确定的，而是具有概率性（或称不确定性）的。例如，长臂猿有晨鸣二重唱的习性，雌雄成对在晴朗的早上太阳上到山顶前后时分发出悦耳的鸣叫，这是长臂猿的行为的一部分。在这里，晨鸣二重唱与天气因素有显著相关性。这个显著相关性是说长臂猿的晨鸣二重唱95%在晴朗的早上会发生，5%不发生。因此，这种行为与天气的关系不是绝对关系，而是概率关系。社会科学的许

多研究领域中的现象也有这种特征。例如，在西方政党竞选体制中，假定某个选区是保守党的票仓［在各国（地区）选举中，长期固定支持某个政党的票源即被称作其票仓］，如果研究者进行 10 次数据采集以检验该选区对保守党的支持率，10 个数据不会都一样，而是在某个平均数附近波动着。由于这种概率关系，研究者们主要使用统计学方法进行数据处理，有时使用参数统计方法，有时使用非参数统计方法，具体使用哪一类方法，取决于一系列因素，详见第 2 章。

统计学家们已经研究出许多统计学方法，这些方法有不同的使用前提和条件，解决的问题也不同。本书着重介绍野外动物学、行为学及社会科学中常用的方法。由于研究方法是科学研究活动成败的关键，如何选择正确的研究方法，以及如何正确操作这些方法，是撰写本书的目的之一，也是本书的主要内容。那些将读者置于云里雾里的复杂纷繁的数理推论过程不在本书范围内，留给统计学家去讨论。为此，本书在撰写过程中，尽量避免使用晦涩的术语，而是使用非数学专业人士容易理解的语言。

一个成功的研究项目，首先开始于研究方案的制订。对于行为学和心理学的学生，制订研究方案时可以参阅文献（Cherulnik，1983）；生态学的学生参阅文献（叶智彰等，1993）；社会科学的学生参阅文献（巴比，1987；Stern，1979）。在研究方案中，研究者首先要根据自己的研究目的提出一系列问题，然后进行数据采集，通过采集到的数据来解答这些问题。不同的数据如何整合在一起进行分析取决于研究者所要解答的问题。例如，当要研究某些生态因子之间的关系时，就要选择因子关系分析的方法。不同的统计学方法对数据类型、数据分布类型、样本数、样本量，以及数据的排列方式都有具体要求，因此研究者必须在数据采集前确定自己要使用的方法，并且按照这些方法制订数据采集计划（阅读本书有助于制订数据采集计划）。数据采集完成后，研究者面临较多的数据分析任务，这时可以再回到本书，参照书中相关计算过程对自己的数据进行分析和检验，这是我们撰写此书的另一目的。

第2章

统计学的基本概念和方法的选择

2.1 统计学的基本概念

2.1.1 总体、样本、采样单元

总体（population）：生物学中有物种、种群、群落及生态系统，还有包括一个物种或一个种群的行为类目；社会学中有人群、城市、国家……这些概念都有具体实体的界定。另外，有些概念没有实体界定，如医学上接受某种药物治疗的患者，心理学上某种情境中的研究对象。所有这些概念在统计学中统称为总体。总体通常是我们要研究的对象。然而，在研究中是否需要用统计学，要视总体的规模而定。例如，我们要分析某中学两个班学生的健康情况，每个班是一个总体，由于人数较少，直接测量计算学生各种健康指数并进行比较就可以了，不需要统计学检验。然而，如果我们要比较两个城市的中学生，涉及的人数太多，无法从每个学生身上采集健康数据，这时就需要用到统计学。统计学操作的第一步就是采样（sampling）。

样本（sample）：假设每个城市有 50 万中学生，我们随机抽取 2.5 万人进行数据采集，通过统计学检验，比较这两个城市中学生的健康水平。这时，总共有 2 个样本，我们称为样本数（number of samples）等于 2；每个样本有 2.5 万人，我们称为样本量或者样本规模（sample size）等于 25000。样本中，每个学生被称为一个采样单元（sampling unit），因此采样单元数就是样本量或者样本规模。不同研究者对这几个概念给予的中文名称可能不同。例如，在行为学研究中，研究者常常通过观察进行数据采集，这时他们常把采样单元称为观察单元（observation

unit)。为了方便后面的陈述，我们用样本数、样本量、采样单元来统一表述相关含义。

样本的作用在于通过在可以操作的小范围（样本自身）中采集数据来代表总体的情况。那么，如何确保样本能够代表总体呢？统计学上有两个要求：①采样单元必须是随机选出来的（保证采样单元相互的独立性），而不是主观臆断地决定从哪个学生身上采集数据；②样本量应该等于或者大于总体的5%，否则就会出现采样误差（sampling error），也就是俗语说的以偏概全。如果样本量足够大，样本中所包含的数据变异性就与总体所包含的数据变异性相似，样本对总体才具有代表性。人们常常很难达到这两个要求。例如，民意调查中，多数调查公司从互联网上进行调查，但整天在互联网上并有机会回答问题的人是总人群中喜欢互联网的人，他们可能代表社会中快速接受时尚或者新事物的人（一类性格的人群），并不能代表性格多种多样的社会人群总体。另外，调查人数常常只有几千人，对于几千万的上网人群，这个样本量还不到总体的1%，更不用谈全国人（另一个更大的总体）。这种采样误差，导致误判的可能性就很大。例如，2014 年苏格兰独立公投，公投前的调查结果是赞成与反对两派的支持率之比大约是 50∶50（一些公司的结果是 48∶52 或者 49∶51，另一些公司的结果是 51∶49 或者 52∶48，最大差值仅 4%），两派人看到这些结果都很兴奋，谁也没有做失败的思想准备。然而，实际公投结果是 45∶55，实际差值达到 10%。当然，在绝大多数研究案例中，研究者由于客观条件的限制无法达到5%的采样要求。不过，明白这个要求很重要，研究者要在条件允许的情况下尽力接近这个要求，在远离这个要求的情况下解析自己的研究数据时表述要有所保留，不能绝对化。

2.1.2 因子、变量

因子（factor）和变量（variable）：研究者在每个采样单元中根据自己的研究进行相应内容的测量或者观察。例如，在上述中学生健康的例子中，研究者会测量学生的体重、心跳、血压；生物学家会测量动物的体长、植物的胸径；社会学家会调查人群的受教育年数；行为学家要观察记录研究客体的打架时间、休息时间，或者社会交往时间等。这些内容的归类称为因子（factors），也称为变量（variables）。因子是具体研究中使用的术语，变量是统计学中使用的术语，如生态影响因子，每个因子就是统计学中的一个变量。这两个术语经常被混用，不过，它们的使用场合不同，但本质是一样的。

与采样单元的选择一样，变量的选择也要注意独立性。例如，一个地区的生物多样性，我们可以用物种数作为变量，也可以用申农-维纳指数作为变量。然而，申农-维纳指数的计算本身包含了物种数（Ricklefs，1990），不能同时将这两个变

量放到统计学检验中。许多人不了解变量的独立性和相关性的关系。统计学要求变量之间独立，以确保变量 *a* 中不包含变量 *b* 的数据变异信息；而如果统计结果表明变量 *a* 与变量 *b* 相关，只表明这两个变量的数据变化存在着规律性的相同或者相反，即正相关或者负相关。这种相关是客观的相关，变量选取时的“相关”（不独立）是主观的相关，而科学需要的是客观的相关。

统计学方法中有时涉及自变量（independent variable）和因变量（dependent variable）的概念，如回归分析。自变量是一对变量中起决定性作用的变量，另一个随之而变化的是因变量。例如，在研究鱼类生态学时，水体中鱼的个体数取决于水体的生产力，而水体的生产力不取决于鱼的个体数。这里，水体的生产力和鱼的个体数相互没有包含对方，但是一个对另一个有决定性影响。总而言之，自变量和因变量也是独立变量，只是它们的变化规律存在依赖与被依赖的关系。

2.1.3　测量值、观察值、数据

在一个样本中，研究者从某个采样单元中采集到某个变量的信息。如果这个变量的信息需要借助工具来采集，如尺子、量器，这样得到的信息被称为测量值（measurement），如体长（cm）；如果不用工具，凭观察记录到的信息，则称为观察值（observation），如群体规模（个体数）。这两个概念本质上是一回事，因此研究者们经常混用，但不会造成任何影响。另外，一个常见的词是数据（data），它既可以用于表示单个的观察值或者测量值，又可以指这些测量值或者观察值的总和。当从一个采样单元中采集一个观察值时，这种数据称为单变量数据（univariate data），因为样本只有一个变量；采集两个观察值（两个变量，每个变量采集一个观察值）时称为双变量数据（bivariate data）；采集三个或者更多的观察值（有三个或更多变量）时称为多变量数据（multivariate data）。有时，研究者会在不同时间或者不同条件下从相同采样单元中获取关于相同变量的两个或多个测量值/观察值，这样的数据被称为重复记录（replicates）。重复记录组成的样本称为相关样本（related samples）或者匹配样本（matched samples），它们看似两个（或多个）样本，实则是一个样本的重复测量，但每次测量的条件不同。还有一种情况也会导致相关样本的出现，即研究者通过某些特定条件或者依据一些特定标准选择采样单元（如特定的身高、体重）后，将它们随机安排在不同的组别中。由于这些采样单元不是随机选择的，而是依据一定标准筛选出来的，它们也被视为相关样本或者匹配样本。完全随机选择的采样单元组成的样本称为独立样本（independent samples）。

研究者采集到数据后，样本中的数据反映两个重要信息，一个是样本的整体趋势（averages），它反映样本的中心趋势，是客观现象的规律性的反映；另一个