



面向21世纪课程教材

高等学校经济学类、工商管理类核心课程教材

统 计 学

(第四版)

袁 卫 庞 皓 贾俊平 杨 灿 主 编

高等教育出版社



面向 21 世纪 课程教材

高等学校经济学类、工商管理类核心课程教材

统 计 学

(第四版)

袁 卫 庞 皓 贾俊平 杨 灿 主 编

Tongjixue

高等教育出版社·北京

内容简介

本书是教育部“面向 21 世纪课程教材”，是高等学校经济学类、工商管理类核心课程教材。

本书在第三版基础上进一步强化“统计学是数据和数据分析的科学”这一主线，突出应用统计的特点。在基本保留原有框架的基础上，更新了案例和数据，精简了较为烦琐的数学推倒和统计软件可以取代的数学计算，删去了第 10 章国民经济统计基础知识，力图使数据和数据分析的核心内容更加简洁。在基本数据分析的方法上，增加更多的实际案例，从正反两个方面强化正确选择和使用分析工具的能力。

本书可作为高等学校经济学类、工商管理类专业本科生的教材，也可作为人文、社会科学以及其他学科的本、专科教材或参考书，还可作为相关管理人员和研究人员的参考书。

图书在版编目（CIP）数据

统计学 / 袁卫等主编. --4 版. --北京：高等教育出版社，2014.7

ISBN 978-7-04-039898-4

I. ①统… II. ①袁… III. ①统计学-高等学校-教材 IV. ①C8

中国版本图书馆 CIP 数据核字（2014）第 105020 号

策划编辑 施春花	责任编辑 施春花	封面设计 杨立新	版式设计 范晓红
插图绘制 杜晓丹	责任校对 刁丽丽	责任印制 尤 静	

出版发行 高等教育出版社
社 址 北京市西城区德外大街 4 号
邮政编码 100120
印 刷 北京四季青印刷厂
开 本 787 mm × 1092 mm 1/16
印 张 20.25
字 数 460 千字
购书热线 010-58581118
咨询电话 400-810-0598

网 址	http://www.hep.edu.cn http://www.hep.com.cn
网上订购	http://www.landraco.com http://www.landraco.com.cn
版 次	2000 年 7 月第 1 版 2014 年 7 月第 4 版
印 次	2014 年 7 月第 1 次印刷
定 价	39.00 元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换
版权所有 侵权必究
物 料 号 39898-00

总 前 言

高等学校经济学类核心课程和工商管理类核心课程是在高等教育面向 21 世纪教学内容和课程体系改革计划“经济学类专业课程结构，共同核心课程及主要教学内容改革研究与实践”和“工商管理类专业课程结构及主要教学内容改革研究与实践”两个项目调研基础上提出、经经济学教学指导委员会和工商管理类教学指导委员会讨论通过、教育部批准的必修课程。其中，经济学类各专业的核心课程共 8 门：政治经济学、西方经济学、计量经济学、国际经济学、货币银行学、财政学、会计学、统计学；工商管理类各专业的核心课程共 9 门：微观经济学、宏观经济学、管理学、管理信息系统、会计学、统计学、财务管理、市场营销学、经济法。这些课程确定后，教育部高教司组织有关专家制定了各门课程的教学基本要求，并组编了相应的各门教材。各门课程的教学基本要求及相应教材由高等教育出版社 2000 年秋季出齐，供各高等学校选用。

教育部高等教育司

2000 年 3 月

第四版前言

我们已经进入了大数据时代。作为数据和数据分析科学的统计学，在大数据环境中应用越来越广泛而深入。但与此同时，大数据的分析越来越依赖于计算机技术，具有熟练计算机技能和数据分析能力的毕业生成为社会最紧缺的人才。基于这样的社会需求，我们在第三版基础上进一步强化“统计学是数据和数据分析的科学”这一主线，突出应用统计的特点；在基本保留原有框架的基础上，更新了案例和数据，精简了较为烦琐的数学推倒和统计软件可以取代的数学计算，删去了第10章国民经济统计基础知识，力图使数据和数据分析的核心内容更加简洁；在基本数据分析的方法上，增加了更多的实际案例，从正反两个方面强化了正确选择和使用分析工具的能力，比如第2章中增加了辛普森悖论的内容等。

统计学教学和研究已经离不开计算机和统计软件。在学习过程中，学生通过大量实际的数据分析才能真正掌握统计方法的精髓，才能将书本的知识转化为自己的能力。因此，我们将与教材配套的习题也作了相应的较大的修改，原则就是少而精，重点是把握方法背后的思想和基本原理。考虑到各地教学条件的差异，本书统计方法的实现仍然以Excel为主。对于有条件、有能力的教师和读者，建议使用SAS、SPSS、R等更专业的统计软件来处理和分析数据。

第四版修订仍由中国人民大学、厦门大学和西南财经大学合作完成。各章的执笔人分别是：第1、2章袁卫（中国人民大学）；第3章袁卫、贾俊平（中国人民大学）；第4、5、6章贾俊平；第7、8章庞皓（西南财经大学），第9章杨灿（厦门大学）。全书由袁卫、贾俊平修改定稿。

本书可作为高等学校经济学类、工商管理类以及人文、社会科学和其他学科的本、专科教材或参考书。感谢广大读者和任课教师，你们的鼓励和建议是我们修改完善的动力。感谢高等教育出版社相关编辑，你们的支持和帮助使得这一版能和读者尽快见面。由于统计方法和科技发展日新月异，我们受知识能力所限，这一版中仍有不少疏漏乃至错误之处，恳请同行和读者继续提出批评和建议，以便进一步修改完善。

编者

2014年3月

第一版前言

在人类迈进 21 世纪的今天，知识的更新速度正不断加快，社会对新知识的需求也日益增加。无论是国民经济管理和公司、企业的经营及决策，还是科学研究都越来越依赖于数量分析和统计分析方法。统计方法已经成为理、工、农、医、人文、社会、管理、军事等所有学科领域科学研究的基本方法。因此，社会对统计方法的应用以及对统计教材的编写也就提出了更高的要求。近年来，我国高校经济管理类专业的统计教材在内容和体系上都有了较大的改进，但还缺少普遍适用的统编教材。本教材就是为了适应这一需要而编写的。它总结了我们长期的教学经验，参阅了国内外同类的优秀教材，既可作为高校经济管理类专业的教材，也可作为其他专业和广大实际工作者的参考书。

构建统计学教材的内容体系，关键是要对统计学科有一个全面科学的认识。我们认为，统计学是一门收集、整理和分析统计数据的方法科学，其目的是探索数据内在的数量规律性，以达到对客观事物的科学认识。取得统计数据是进行统计分析的基础和前提，离开了统计数据，统计方法就失去了用武之地。如何取得准确可靠的数据是统计学研究的重要内容之一，通常需要对调查人员和试验人员进行专门的培训。统计数据的整理是通过对统计数据的加工处理使其系统化、条理化，符合统计分析的需要，是介于数据收集与数据分析之间的一个必要环节。统计数据的分析是统计学的核心内容，它是通过统计描述和统计推断的方法探索出数据内在的数量规律性的过程，也是本教材的重点。本教材在编写中力求简明易懂，强调应用实例阐明统计方法的基本原理和思想，并结合 Excel 软件进行教学，以提高读者学习统计的兴趣和应用统计方法分析解决实际问题的能力。

本教材由中国人民大学、厦门大学、西南财经大学合作编写而成。由袁卫教授（中国人民大学）、庞皓教授（西南财经大学）、曾五一教授（厦门大学）担任主编。各章执笔人分别是：第一章袁卫，第二章、第三章贾俊平（中国人民大学），第四章周惠彬（西南财经大学），第五章、第六章王青华（西南财经大学），第七章金勇进（中国人民大学），第八章曾五一，第九章庞皓，第十章曾五一，第十一章杨灿（厦门大学），第十二章刘文卿（中国人民大学），第十三章杨灿，附录林飞（厦门大学）。全书最后由袁卫、贾俊平总纂定稿。中国科技大学的陈希孺院士、中国人民大学的倪加勋教授和天津财经学院的肖红叶教授认真审阅了书稿，并提出了宝贵的修改意见。在本书付印之际，谨向所有帮助和支持本书编写和出版的同志表示衷心的感谢。

由于编者水平所限，书中难免有疏漏或错误之处，恳请同行和读者多提宝贵意见，以便我们进一步修改和完善。

编者

2000 年 3 月

序

近年来,我国的统计学教材建设取得了不小的成绩,出版了一批国人自著的教科书和引进了一些国外的优秀著作。但这些教材中,多数是偏于社会经济统计学或数理统计学一方的,能将这两方面熔合于一炉的著作尚不多见。袁卫教授等的新著《统计学》是朝着这个方向努力的一个尝试,而具有其鲜明的特色。这不仅表现于在内容上包罗了上述两个方面,也表现于在讲述数理统计方法时,采用了大量的富于社会经济统计内涵的例子,标示了数理统计方法在解决社会经济统计问题中的作用,使这两个方面达到了一定程度的有机的融合。

在以往相当长一段时间内,甚至直到如今,在我国统计学界有一种看法认为数理统计学和社会经济统计学是两个不同性质甚至对立的学派。笔者认为这是一种误解,其根源盖出于对数理统计方法的“工具性”注意不够。数理统计学作为一门自然科学,是一种“价值中立”的工具。这个属性决定了它不对任何现实问题采取特定的、先入性的立场。一般讲,凡是其中涉及带随机性的数据的问题,都是数理统计方法这个工具的可能用武之地。当然,在一个具体问题中是否该采用随机化的模型因而引入数理统计方法,要由问题的性质和条件等诸多方面的考虑来决定。如在一项调查工作中,究竟是采取普查或抽样调查方法,取决于许多因素,不能武断地说某种方法必然是最好的方法。

数理统计方法这种“工具性”和“价值中立”属性看来属于自明,其所以被误解,一定程度上与其易被滥用或误用有关。统计方法之正确使用有赖于三个方面:一是数据的取得必须符合随机性的要求,而这“随机性”的含义,要依所考察的问题的性质而定,这在医药试验中的所谓“双盲法”中看得很清楚。简单一点的例子,如在抽样调查中要避免为迎合调查所希望的结论而有意偏向地选择样本等。这个道理说起来容易,真正做到不易。即使是抱着负责的、公正的态度,也有可能无意中掉入不易觉察的陷阱。这一点做得如何,恐怕也是衡量一个统计工作者职业训练的一个重要方面。二是模型和方法的选择。是选择线性的还是非线性的模型,选择多少个自变量和哪些自变量,使用哪种类型的相关指标,问题适合于用方差分析还是因子分析来处理为好,等等,有时不易做出合适的判断。而这种判断不当可以使尔后的统计分析形式上合乎规范,但实质却是个“伪问题”。三是在统计分析结果的解释上。在这一点上的考虑不周可以从一个正确的统计分析引申出不当的结论。如不合理的外推,在数据量不大时,对接受或否定一个统计假设的含义做出过度的解释,以及轻率地将统计相关关系解释为因果关系之类。这些问题的妥善处理并无一定之规,在很大程度上取决于经验以及对所研究问题的背景的了解。不可否认,上述种种问题的存在使对数理统计方法的滥用和误用开了方便之门。这是在学习、理解和使用数理统计方法时必须留意

的地方。《不列颠百科全书》把统计学定义为“收集和分析数据的科学和艺术，其中标出统计方法的艺术”属性，当然不能拘泥于其字面意义去理解，其含义从上文所论可以有所领悟。

近些年来，我国统计界的一些有识之士倡导“大统计”的说法，号召我国统计工作者不分畛域，团结在“大统计”这面旗帜之下，一同为发展我国的统计事业与统计学术贡献力量。笔者很赞同这一方针。从历史上说，较早期的统计学的发展符合这种“大统计”的精神。那时，数理统计方法的研究是密切结合种种实用问题进行的，其中人口统计和社会经济统计问题是一个显著的方面。例如，格朗特在1662年发表的《关于死亡公报的自然和政治观察》一书，被认为是描述统计的开山之作，其中也包含了某些对后来推断统计发展有影响的思想，如频率稳定性、数据可靠性的检验与生命表等等。1710年阿布兹诺特考察生男生女机会是否均等的问题，其所用方法包含了近代假设检验理论的若干基本思想。拉普拉斯在19世纪初用非普查的方法估计一国的人口数，对20世纪得到大发展的随机抽样调查方法起了先驱的作用。又如，在统计学中有广泛应用的方差分析方法，公认是费歇尔在20世纪20年代的创造。事实上19世纪后期莱克西斯在研究一个与人口有关的统计问题时，已很接近这个思想。这些以及统计史上其他的例子说明，统计方法的研究必须与实际问题结合才会有活力。与其他实用部门相比，在我国数理统计方法与社会经济问题的结合要薄弱一些。这更启示了“大统计”这一提法的现实意义。

如上所说，统计学的早期历史实际上是一部大统计的历史。20世纪上半叶，数理统计学建立了严整的数学框架，逐渐形成了一门独立学科。这个事实无疑有其积极的意义，但人们也注意到，在其后续的发展中，源于学科自身的“内生性”问题占据了过大的比重，而造成了理论研究与实际应用脱节的现象。近几十年来，国际统计界一些学者对此进行了反思，如今已形成一股有影响的潮流，这一点应当引起我国统计学界的重视。笔者认为，提倡大统计，正是对此的一种呼应。应当鼓励数理统计学者多关心和参与社会经济统计方面的问题。同时，社会经济统计工作者有必要熟悉基本的数理统计方法，以拓宽视野，增强解决问题的能力，并使自己的研究工作具有更强的现代意识。

这一切不可能一蹴而就，需要一个过程。这中间统计教育是一个重要方面。而要有革新的统计教育，就必须有革新的统计教材。袁卫教授等的《统计学》，是朝着这个方向的一个有意义的努力。如前所指出，本书的一个优点是联系大量的实例来讲述统计方法，把一些较为抽象的内容用浅近易了解的语言表述而不失其科学性。本书的另一个特点是很重视“数据”。作者辟了整两章的篇幅，仔细讲述了数据的整理、表述及统计特征等方面的问题。这部分内容在统计学中一般理解为“描述统计”。以往有一种错误的理解，以为描述统计只涉及粗浅的数学，是属于低层次的统计学。实际上，描述统计是推断统计的基础。统计观念的养成很大程度上来源于对数据的“感觉”，培养这种感觉正是描述统计的一个重要目的。近来国际上有的提倡“数据分析”作为未来统计发展的可能方向时，甚至标出向描述统计回归的提法。姑不论这在科学上有其可议之处，其主张“统计学是一门关于数据的科学”这个基点，很具有启发性。

笔者相信，本书的出版，对符合新时代潮流的我国统计学教材建设，是一个有益的贡

献，同时，它也将对我国统计界同行起到一种促进作用，促进大家多关心统计学教材建设的问题，写出更多高水平的、符合新时代要求的统计学教材来。

陈希孺
2000年3月

郑重声明

高等教育出版社依法对本书享有专有出版权。任何未经许可的复制、销售行为均违反《中华人民共和国著作权法》，其行为人将承担相应的民事责任和行政责任；构成犯罪的，将被依法追究刑事责任。为了维护市场秩序，保护读者的合法权益，避免读者误用盗版书造成不良后果，我社将配合行政执法部门和司法机关对违法犯罪的单位和个人进行严厉打击。社会各界人士如发现上述侵权行为，希望及时举报，本社将奖励举报有功人员。

反盗版举报电话 (010) 58581897 58582371 58581879

反盗版举报传真 (010) 82086060

反盗版举报邮箱 dd@hep.com.cn

通信地址 北京市西城区德外大街4号 高等教育出版社法务部

邮政编码 100120

目 录

第 1 章 数据与统计学	1	2.3 分布离散程度的测度	33
1.1 统计数据与统计学	2	2.3.1 极差	33
1.2 统计学的产生与发展	4	2.3.2 内距	33
1.3 统计学的分科	5	2.3.3 方差和标准差	33
1.3.1 描述统计和推断统计	5	2.3.4 离散系数	36
1.3.2 理论统计和应用统计	7	2.4 分布偏态与峰度的测度	38
1.4 数据的种类与来源	8	2.4.1 偏态及其测度	38
1.4.1 数据的种类	8	2.4.2 峰度及其测度	39
1.4.2 数据的来源	8	2.5 统计表、统计图与辛普森悖论	40
1.5 统计数据的质量	10	2.5.1 统计表	40
1.6 统计学的基本概念	10	2.5.2 统计图	41
1.6.1 总体	11	2.5.3 辛普森悖论	51
1.6.2 变量	11	本章小结	51
1.6.3 样本	11	思考与练习	52
本章小结	11	案例分析	58
思考与练习	12	附录	60
案例分析	12	第 3 章 概率、概率分布与抽样分布	63
第 2 章 统计数据的描述	15	3.1 事件及其概率	64
2.1 统计数据的整理	16	3.1.1 试验、事件和样本空间	64
2.1.1 统计数据的分组	16	3.1.2 事件的概率	66
2.1.2 次数分配	17	3.1.3 概率的性质和运算法则	68
2.1.3 次数分配直方图	19	3.1.4 条件概率与事件的独立性	71
2.1.4 洛伦茨曲线与基尼系数	21	3.1.5 全概率公式与逆概率公式	74
2.2 分布集中趋势的测度	23	3.2 随机变量及其概率分布	76
2.2.1 众数	23	3.2.1 随机变量	76
2.2.2 中位数	24	3.2.2 离散型随机变量的概率分布	78
2.2.3 分位数	25	3.2.3 离散型随机变量的数学	
2.2.4 均值	26		
2.2.5 几何平均数	28		
2.2.6 切尾均值	29		
2.2.7 众数、中位数和均值的关系	31		

期望和方差	79	区间估计	135
3.2.4 几种常用的离散型概率分布	79	4.4 样本量的确定	138
3.2.5 概率密度函数与连续型随机变量	86	4.4.1 估计总体均值时样本量的确定	138
3.2.6 常见的连续型随机变量的概率分布	87	4.4.2 估计总体比率时样本量的确定	139
3.3 常用的抽样方法	96	本章小结	140
3.3.1 简单随机抽样	96	思考与练习	140
3.3.2 分层抽样	97	案例分析	143
3.3.3 系统抽样	97	第5章 假设检验	145
3.3.4 整群抽样	98	5.1 假设检验的基本原理	146
3.4 抽样分布	98	5.1.1 假设的陈述	146
3.4.1 抽样分布的概念	98	5.1.2 两类错误与显著性水平	149
3.4.2 $\bar{x}$ 抽样分布的形式	101	5.1.3 检验统计量与拒绝域	149
3.4.3 $\bar{x}$ 抽样分布的特征	102	5.1.4 利用 P 值进行决策	151
3.4.4 样本比率的抽样分布	106	5.2 一个总体参数的检验	152
3.4.5 样本方差的抽样分布	106	5.2.1 总体均值的检验	153
3.4.6 两个样本统计量的抽样分布	108	5.2.2 总体比率的检验	159
3.5 中心极限定理的应用	110	5.2.3 总体方差的检验	160
本章小结	112	5.3 两个总体参数的检验	162
思考与练习	113	5.3.1 两个总体均值之差的检验	162
案例分析	118	5.3.2 两个总体比率之差的检验	169
第4章 参数估计	120	5.3.3 两个总体方差比的检验	170
4.1 参数估计的基本原理	121	本章小结	173
4.1.1 估计量与估计值	121	思考与练习	173
4.1.2 点估计与区间估计	121	案例分析	176
4.1.3 评价估计量的标准	124	第6章 方差分析	178
4.2 一个总体参数的区间估计	125	6.1 方差分析引论	179
4.2.1 总体均值的区间估计	125	6.1.1 方差分析及其有关术语	179
4.2.2 总体比率的区间估计	128	6.1.2 方差分析的基本思想和原理	180
4.2.3 总体方差的区间估计	129	6.1.3 方差分析中的基本假定	181
4.3 两个总体参数的区间估计	130	6.1.4 问题的一般提法	182
4.3.1 两个总体均值之差的区间估计	130	6.2 单因素方差分析	182
4.3.2 两个总体比率之差的区间估计	135		
4.3.3 两个总体方差比的			

6.2.1 数据结构	182	检验	224
6.2.2 分析步骤	183	本章小结	227
6.2.3 用 Excel 进行方差 分析	187	思考与练习	228
6.2.4 方差分析中的多重 比较	188	案例分析	232
6.3 双因素方差分析	190	第 8 章 时间序列分析与预测	235
6.3.1 双因素方差分析及其 类型	190	8.1 时间序列的描述性分析	236
6.3.2 无交互作用的双因素 方差分析	191	8.1.1 时间序列的含义	236
6.3.3 有交互作用的双因素 方差分析	194	8.1.2 时间序列的图形描述	238
本章小结	197	8.1.3 时间序列的速度分析	239
思考与练习	197	8.2 时间序列的构成因素及其 组合模型	242
案例分析	200	8.2.1 时间序列的构成因素	242
第 7 章 相关与回归分析	202	8.2.2 时间序列构成因素的 组合模型	244
7.1 相关分析	203	8.3 时间序列趋势变动分析	245
7.1.1 相关关系的概念	203	8.3.1 测定长期趋势的移动平 均法	245
7.1.2 相关系数	206	8.3.2 测定长期趋势的指数平 滑法	248
7.2 一元线性回归分析	208	8.3.3 测定长期趋势的 模型法	251
7.2.1 相关与回归分析的 联系	208	8.4 季节变动分析	257
7.2.2 总体回归函数与样本 回归函数	209	8.4.1 季节变动分析的原始 资料平均法	257
7.2.3 回归系数的普通最小 二乘估计	211	8.4.2 季节变动分析的趋势- 循环剔除法	258
7.2.4 拟合优度的度量	215	8.4.3 季节变动的调整	260
7.3 线性回归的显著性检验 与回归预测	216	8.5 循环变动分析	261
7.3.1 回归系数显著性的 t 检验	216	8.5.1 循环变动及其测定 目的	261
7.3.2 一元线性回归模型的 预测	218	8.5.2 循环变动的测定方法	261
7.4 多元线性回归分析	219	本章小结	267
7.4.1 多元线性回归模型及 假定	220	思考与练习	268
7.4.2 多元线性回归模型的 估计	221	案例分析	270
7.4.3 多元线性回归模型的		第 9 章 统计指数	274
		9.1 指数的概念与分类	276
		9.1.1 指数的概念	276
		9.1.2 指数的分类	277
		9.2 总指数的编制方法	278
		9.2.1 总指数编制的基本	

问题	278	指数	289
9.2.2 加权总指数的编制		9.4.3 股票价格指数	290
原理	280	本章小结	291
9.2.3 加权综合指数的主要		思考与练习	292
形式	283	案例分析	294
9.3 指数体系与因素分析	286	附录 常用统计表	296
9.3.1 指数体系及其作用	286	附表1 标准正态分布表	296
9.3.2 总量变动的因素分析	287	附表2 t 分布表	297
9.4 几种常用的经济指数	288	附表3 χ^2 分布表	298
9.4.1 消费者价格指数和商品		附表4 F 分布表	300
零售价格指数	288	参考书目	306
9.4.2 生产指数和生产者价格			



第 1 章

数据与统计学

乘飞机的风险与航空意外险

在我们出差或外出旅行时，如果要买航空意外险，现在普遍的单程保费是 20 元，保额 60 万元。也就是说，如果发生空难等航空意外事故，最多可以得到 60 万元的保险赔偿。大家知道，保费费率的厘定，主要根据风险的大小，同时考虑管理费用和保险公司的利润。如果暂不考虑管理费用和利润，不难得到 20 元保 60 万元是按 $1/30\,000$ 左右的风险计算的。那么我国民航业现在的风险到底有多大？根据国际民用航空业航空风险的计算原则，一般是以 10 年作为一个周期，计算百万起降架次和百万飞行小时的事故率。从中国民用航空总局的网站上可以查到，近 10 年来（2003 年到 2012 年）我国民航每年起降架次平均为 300 万次左右，其间发生的空难有 2 次，分别为 2004 年 11 月 21 日东方航空公司包头飞上海的 MU5210 航班（55 人遇难）和 2010 年 8 月 24 日河南航空公司哈尔滨飞伊春的 VD8387 航班（44 人遇难）。不论从百万飞行小时、旅客人数或者起降架次哪个指标计算，近年来我国民航业的风险和事故率都已经大大降低，大约为 $1/8\,000\,000$ 或者更小。显然，按照 $1/30\,000$ 计算的航空意外险就有着极大的利润。这也是前些年航空意外险市场混乱、假保单满天飞的诱因之一。如果按照 $1/8\,000\,000$ 的概率计算，60 万元保额应该收多少保费？即使加上 50% 的利润和管理费，每张保单应该卖多少钱？将这一保费数额纳入机票费用是否合理？

1.1 统计数据与统计学

大数据时代,我们的工作和日常生活已经离不开各种各样的数据。例如,早上起来要关心当天气温的高低和空气质量情况;在网络“百度搜索”输入“数据”后除了列出与其相关的词汇信息外,还根据与查询“数据”的相关程度列出“相关术语”“相关网站”;球类比赛时解说员总要统计竞赛双方的进攻次数和成功率;学生考试后非常关心自己的考试成绩和名次;企业管理人员要掌握生产销售情况和利润额;报刊和电视中常提到GDP(国内生产总值)、CPI(消费者价格指数)和经济增长率的数字等。日常工作与生活中的这些数字就是我们所关心的统计数据。

统计学是一门收集、整理、显示和分析数据的科学,其目的是探索数据内在的数量规律性。正是因为统计学总是在和数据打交道,因而也可称统计学为“数据的科学”。

怎样理解统计学是数据的科学呢?下面先看看统计学英文的名词,再举几个例子。

统计学的英文是 Statistics。在英文字典中它有两个含义:当它以单数名词出现时,表示一门科学的名称“统计学”;当它以复数名词出现时,表示“统计数据”或“统计资料”。Statistics 一词的英文解释至少可以说明两件事情。

第一,由于统计数据在英文中是以复数形式出现的,表明统计数据不是指个别的单个数字,而是指同类的较多数据。因为单个数字如果不和其他数据进行比较,是不能说明问题的。例如,某个学生在某门课程的考试中得了 85 分,如果仅凭这一个数字,我们很难对这位学生的知识和能力水平作出判断和评价。因为这个 85 分可能是班上的最高分,可能是中等水平的分数,也可能是较低的分数。如果还知道这次考试的平均分数,就可以对这位学生的成绩是高于还是低于平均分数,以及是高多少或者低多少作出评价了。在生产和生活实际问题中,通常可以收集到较多的数据,进而利用统计方法对数据进行加工整理,从而发现数据中的内在联系及数量规律。

第二,作为单数的统计学和作为复数的统计数据在英文中使用同一名词,显示出二者之间的密切关系。统计学是由收集、整理、显示和分析统计数据的方法组成的,这些方法来源于对统计数据的研究,目的也在于对统计数据的研究。离开了统计数据,统计方法乃至统计学就失去了其存在的意义。这正如俗语所说“巧妇难为无米之炊”。这里的“巧妇”就是掌握统计方法的统计学家或统计工作者,“米”就是统计数据,“炊”就是统计研究或统计工作的目的,即探索数据内在的数量规律性。显然,没有统计数据或没有较好的统计数据,即使很科学的统计方法或很高明的统计学家也难有所作为。

那么,什么是数据内在的数量规律性呢?下面用几个简单的例子来说明。

【例 1.1】 一个家庭新生婴儿的性别可能是男,也可能是女。在过去没有实行计划生育时,有的家庭几个孩子都是男孩,也有的都是女孩。从表面上看,新生婴儿的性别比例似乎没什么规律可循。但如果对新生婴儿的性别进行大量观察,就会发现性别比例还是有规律的,即婴儿总数中男孩要多于女孩,大致为每生育 100 个女孩,就有 105 个左右的男孩。这个 105:100 的比例就是新生婴儿男女性别的数量规律性,古今中外都大致相同,它是由人类社会长期遗传与发展的结果。因为人类社会要延续、要发展,就要保持男女人数的大致相同。那么有人会问,新生儿男多于女,不是性别不平衡了吗?是的,新生儿时

男多于女会出现不平衡,但男孩的死亡率高于女孩,到了中青年时,男女人数就大致相同了。进入中老年后,男性的死亡率仍然高于女性,男性的平均预期寿命比女性短,长寿的男性要少于女性。从一个国家看,如果没有人为的B超检查后堕胎等干扰,其规律是婴幼儿时男性略多于女性,中青年时男女人数大致相同,老年时女性又略多于男性。这样既保证人类在中青年结婚生育时性别的大致平衡,又使得在人口总数上男女也大体相当,有利于人类社会的进化和发展。对人类性别比例的研究是统计学的起源之一,也是统计方法探索的最早的数量规律之一。

【例 1.2】 我们都做过掷硬币和掷骰子的游戏,都知道随机地掷一次硬币或骰子不能事先确定出现正面还是反面或某个点数,也就是说个别游戏或试验中充满了不确定性或偶然性。机会游戏或赌博正是利用了这种不确定性和偶然性才能够吸引人。但当我们进行大量观察,即不断做重复试验时,就会发现掷一枚均匀硬币出现正面和反面的比率会大体相同,即 $0.5:0.5$ 。试验的次数越多,出现正面和反面的可能性就越接近 $1/2$ 这一稳定的数值。同样,在掷一枚均匀骰子时,出现 $1\sim 6$ 点中任一点数的可能性也都接近 $1/6$ 。这里的 $1/2$ 和 $1/6$ 就是掷硬币和掷骰子出现某一特定结果的概率,也就是我们探索的数量规律性。

【例 1.3】 在进行农作物试验时,如果其他试验条件相对固定,我们会发现某种粮食作物的产量会随某种施肥量的增加而增加。当开始增加施肥量时,产量增加较快。以后增加同样的施肥量,粮食产量的增加量逐渐减少。当施肥量增加到一定数值量,产量不再增加。这时如果再增加肥料,产量反而会减少。这一施肥量与粮食产量的数量关系就是我们要探索的规律性。当我们从大量试验数据中用统计方法探索到施肥量与产量之间的数量关系,就可以考虑肥料的费用并选择最佳的施肥效果了。

【例 1.4】 北京市公安局根据多年来各区县街道刑事案件数据的积累、整理和分析,利用大数据方法在 2013 年 7 月 10 日发布了“北京市治安地图”。其中将动物园批发市场、新发地地区、锦绣大地市场等 19 处列为最不安全地区。之后,北京市公安局进一步公布治安重点整治地区、扒窃案高发地区、入室盗窃案高发地区等详细信息。这些信息的公布既有利于公安部门有针对性地整治,也让老百姓有所防范。

【例 1.5】 天津汽车制造厂与丰田公司合资生产的威驰(Vios)汽车安装的是 8 A 型发动机,排量为 1.342 L,公司声称其手动挡汽车每百公里油耗(等速情况下)不超过 5 L。要检验购买的车辆是否达到了说明书上的节油标准,就需要计算在高速路上匀速行驶的平均油耗。假定你的汽车平均每百公里油耗为 5.5 L,那么你的汽车是属于质量达标汽车中恰巧由于运气不好碰上的个例呢,还是该批产品本身就不合格呢?统计方法可以作出检验并给予回答。

【例 1.6】 某企业开发出一种新的化妆品,在正式投产之前,需要根据市场需求情况制定其价格和销售策略等。该企业委托某市场调查公司在全市 200 万户家庭中抽取 1 000 户家庭作为样本,免费赠送给这 1 000 户家庭试用,然后了解该化妆品的销售前景。如何科学地从该市 200 万户家庭中抽出 1 000 户家庭进行调查,并由这 1 000 户家庭反馈的市场信息科学地推断出全市 200 万户家庭对该化妆品的购买意愿这一数量规律,是统计工作者的任务。

为什么统计方法能够通过对数据的大量观察和处理而研究和探索出其内在的数量规律