



创新系列 · 管理学

统计学 Statistics

◆ 王凌峰 主编



电子工业出版社

PUBLISHING HOUSE OF ELECTRONICS INDUSTRY

<http://www.phei.com.cn>



创新系列·管理学

内容新

统计学

Statistics

◆ 王凌峰 主编

◆ 郑延智 覃义 成党伟 王顺武 刘洋 副主编

电子工业出版社

Publishing House of Electronics Industry

北京·BEIJING

内 容 简 介

本书在知识安排上借鉴了国内代表性的经济管理类本科统计学教材的特点,同时考虑了实际工作的需要,适当扩充了知识点,如介绍了回归共线性问题、目前国内外具有代表性的统计软件、统计指数评价、费雪检验三原则等。本书强调统计方法的应用,配以较多实例,并提供了比较丰富的现实性、时效性较强的阅读材料,用 Excel 以及 SPSS 软件完成实例的计算处理。

本书可作为本、专科经济管理类专业统计学课程教材,也可作为社会相关人员学习统计知识的参考读物。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有,侵权必究。

图书在版编目(CIP)数据

统计学/王凌峰主编. —北京:电子工业出版社,2014.11

(华信经管创新系列)

ISBN 978-7-121-24637-1

I. ①统… II. ①王… 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 246850 号

策划编辑:王二华

责任编辑:王二华

印 刷:涿州市京南印刷厂

装 订:涿州市京南印刷厂

出版发行:电子工业出版社

北京市海淀区万寿路 173 信箱 邮编:100036

开 本:787×1092 1/16 印张:14 字数:360 千字

版 次:2014 年 11 月第 1 版

印 次:2014 年 11 月第 1 次印刷

定 价:29.00 元

凡所购买电子工业出版社图书有缺损问题,请向购买书店调换。若书店售缺,请与本社发行部联系,联系及邮购电话:(010)88254888。

质量投诉请发邮件至 zltz@phei.com.cn,盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线:(010)88258888。

前 言

目前国内流行的一些统计学教材,内容安排有一些待完善之处,如概率论一般在大学一年级已学过,没有必要再涉及;线性回归共线性问题没有介绍;抽样在工作实践中使用较多,但只简单介绍等。针对这些存在的问题,本教材做了一些补充调整。

本书在编写上具有以下特色:(1)贴近教学实际,本书体例结构及各章节内容,剔除了不必要的烦琐内容;(2)例题、习题丰富,使读者在学习理论知识的同时,加强练习,学会实际操作,巩固所学知识;(3)本书提供习题答案及 PPT 课件,需要者可以登录华信教育资源网(www.hxedu.com.cn)下载使用。

当然,不足是不可避免的,一是因为参编者分散,时间进度与思路协调不容易;二是出于多方面考虑,本教材仍然沿用目前国内流行的统计学教材框架与多数内容,对统计学的新方法、新进展介绍还不够,如统计功效、Bootstrap 自助法、统计检验方法优劣的蒙特卡洛数值仿真检验等都没有涉及,关于时间序列目前国内流行的统计学教材都介绍历史上最先提出的移动平均法,而实际上国内外政府统计机构已经普遍使用 X-11 等方法;三是对统计思想的介绍还不够多,如相关系数只是给出表达式,没有介绍定义的思路等,留待以后完善。

本书编写分工如下:王凌峰负责第 1~2 章和第 7~9 章的编写,以及全书框架设计与全书统稿;覃义负责第 4~6 章的编写;刘洋负责第 3 章的编写;郑延智负责第 10 章的编写;成党伟负责第 11 章的编写;王顺武负责第 12 章的编写。此外,硕士研究生吕健发承担了第 8~12 章算例输入及习题答案的整理工作,张泽玺承担了第 7 章算例输入及习题答案的整理工作。2012 级人力资源管理专业黄雨、姜金柳两位同学承担了全书的公式录入工作,在此一并致谢!

本书在编写过程中,参考了大量国内外文献,由于时间和条件限制未能与各著作者一一联系,在此谨向各位著作者表示诚挚的谢意!

由于水平有限,书中难免有错误、疏漏之处,恳请读者不吝赐教提出宝贵意见,以便使本书不断完善,更好地服务于广大读者。联系邮箱:wangsir@guet.edu.cn。

编者

目 录

第 1 章 统计学概论	1
1.1 统计学的定义	1
1.2 统计学的典型应用	1
1.2.1 经济：经济普查与宏观经济监测	1
1.2.2 人口：人口普查	2
1.2.3 管理：企业管理	2
1.2.4 文学：《红楼梦》作者之谜	3
1.2.5 体育：NBA 的数据分析	3
1.3 统计学的主要分类	4
1.3.1 数理统计与应用统计	4
1.3.2 描述统计与推断统计	4
1.3.3 参数统计与非参数统计	4
1.4 统计软件介绍	6
1.4.1 国外综合性统计软件：R、SPSS 和 SAS	7
1.4.2 国内综合性统计软件	8
1.5 统计软件的比较	8
1.5.1 算法可靠性与丰富性	8
1.5.2 学习便利性	9
1.5.3 可定制性	9
1.5.4 并行计算与多核支持	9
1.5.5 总结	10
1.6 阅读材料	10
1.6.1 统计学简史	10
1.6.2 本福特法则及其应用	13
1.6.3 阅读资源拓展	14
国内外代表性统计学学会与网站	14
思考讨论题	15
第 2 章 统计学的基本概念	16
2.1 基本概念	16
2.2 数据的分类	17
2.2.1 定类、定序、定距与定比数据	17
2.2.2 离散数据和连续数据	18
2.2.3 截面数据、时间序列数据与面板数据	18
2.2.4 小样本、大样本与大数据	18

2.3	数据的间接来源与直接来源	20
2.4	调查数据误差类型	21
2.4.1	随机误差	21
2.4.2	系统误差	21
2.5	调查数据系统误差的常见产生原因	22
2.5.1	调查员误差	22
2.5.2	敏感性问题回答失真	22
2.5.3	无回答误差	22
2.5.4	被调查者的回答偏差	23
2.5.5	调查问题与回答项设计不合理	23
2.6	调查数据系统误差的减少方法	24
2.6.1	调查中的脱敏技术	24
2.6.2	无回答误差的减少对策	25
2.6.3	强化调查员培训与管理	26
2.7	数据质量评价	26
2.7.1	定性评价	27
2.7.2	逻辑检验	28
2.7.3	统计检验	29
	思考讨论题	30
	练习题	30
第3章	描述统计量与数据可视化	31
3.1	代表性描述统计量	31
3.1.1	集中趋势测度值	31
3.1.2	离散程度测度值	38
3.1.3	数据的分布形态	41
3.2	数据可视化	43
3.2.1	统计表	43
3.2.2	统计图	44
	思考讨论题	46
	练习题	47
第4章	统计量及其抽样分布	49
4.1	样本均值及其抽样分布	49
4.2	样本方差与样本标准差	52
4.3	正态总体的抽样分布	54
4.3.1	χ^2 分布(卡方分布)	54
4.3.2	F 分布	55
4.3.3	t 分布	57
4.4	一些重要结论	58
	练习题	60

第5章 经典参数估计	61
5.1 参数估计的基本原理	61
5.1.1 估计量与估计值	61
5.1.2 估计量的评价标准	62
5.1.3 点估计与区间估计	63
5.2 单正态总体参数的区间估计	65
5.2.1 单正态总体均值的区间估计	65
5.2.2 单正态总体方差的区间估计	69
5.3 双正态总体参数的区间估计	71
5.3.1 两正态总体均值之差的区间估计	71
5.3.2 两正态总体方差之比的区间估计	74
5.4 一般总体均值和比率的大样本区间估计	75
5.4.1 一般总体均值的大样本区间估计	75
5.4.2 总体比率的大样本区间估计	77
练习题	79
第6章 参数假设检验	82
6.1 基本思想和概念	82
6.1.1 小概率原理	83
6.1.2 假设检验的思想及步骤	83
6.1.3 利用 P 值进行决策	84
6.2 单正态总体的假设检验	85
6.2.1 总体均值的假设检验	85
6.2.2 总体方差的假设检验	89
6.3 双正态总体的假设检验	91
6.3.1 双正态总体均值差的假设检验	91
6.3.2 双正态总体方差比的假设检验	94
练习题	98
第7章 非参数假设检验	100
7.1 分布检验	100
7.1.1 χ^2 拟合优度检验	100
7.1.2 K-S 检验	102
7.1.3 χ^2 检验与 K-S 检验比较	103
7.2 符号检验与秩检验	104
7.2.1 符号检验	104
7.2.2 符号秩和检验	106
7.3 列联表检验	108
7.4 等级相关分析	110
7.4.1 Spearman 等级相关系数	110
7.4.2 Kendall 等级相关系数	112

思考讨论题	115
练习题	116
第8章 抽样调查	117
8.1 调查内容的确定	117
8.1.1 确定调查总体	117
8.1.2 制定抽样框	117
8.1.3 确定调查变量与辅助变量	118
8.2 样本抽取的基本原则	119
8.2.1 随机化原则	119
8.2.2 代表性原则	120
8.3 抽样方法	120
8.3.1 基本抽样方法	120
8.3.2 复杂抽样方法	121
8.4 数据收集	122
8.4.1 面谈与现场调研	122
8.4.2 电话调查	122
8.4.3 网络调查	123
8.4.4 邮寄问卷	125
8.5 样本量的确定	125
8.5.1 样本量的影响因素	125
8.5.2 样本量公式	126
8.5.3 设计效应方法	128
8.5.4 经验确定	128
8.6 抽样调查软件与抽样调查发展趋势	129
8.6.1 抽样调查软件介绍	129
8.6.2 抽样调查发展趋势	129
思考讨论题	131
练习题	131
第9章 方差分析	132
9.1 问题的提出	132
9.2 方差分析的基本概念	132
9.3 基本假定与基本思路	133
9.4 单因素方差分析	135
9.4.1 单因素方差分析数据结构	135
9.4.2 分析步骤	135
9.4.3 Excel 软件操作	138
9.5 双因素方差分析	139
9.5.1 双因素方差分析及其类型	139
9.5.2 数据结构	140

9.5.3	分析步骤	140
9.5.4	Excel 软件操作	142
9.6	方差分析中的多重比较	143
9.6.1	多重比较问题的提出	143
9.6.2	单因素方差分析 LSD 法的计算过程	144
9.6.3	多重比较与方差分析的关系	145
9.7	方差分析研究与应用的趋势	146
	思考讨论题	147
	练习题	147
第 10 章	线性回归	150
10.1	相关关系与因果关系	150
10.2	相关关系的表现形态	151
10.3	线性回归实例与散点图	152
10.4	一元线性回归	153
10.4.1	参数的最小二乘估计	154
10.4.2	方程拟合优度	155
10.4.3	方程整体与系数显著性检验	156
10.4.4	Excel 软件操作	159
10.5	多元线性回归	161
10.5.1	一般描述	161
10.5.2	参数的最小二乘估计	161
10.5.3	共线性的产生	162
10.5.4	共线性的判断	162
10.5.5	多重判定系数	163
10.5.6	Excel 软件操作	163
10.6	经典线性回归的拓展	164
10.7	使用线性回归进行非线性回归	165
	思考讨论题	167
	练习题	167
第 11 章	社会经济时间序列	170
11.1	增长率分析	170
11.2	结构分析的四因素分解法	172
11.3	误差评价指标	174
11.3.1	平均误差	175
11.3.2	平均绝对误差	175
11.3.3	均方误差	175
11.3.4	平均百分比误差和平均绝对百分比误差	175
11.4	趋势因素与季节因素计算	176
11.4.1	移动平均法	176

11.4.2	函数拟合法	179
11.4.3	用移动平均法进行季节因素调整	184
11.4.4	算例	185
11.5	循环因素计算	189
11.6	时间序列软件介绍	191
	思考讨论题	192
	练习题	192
第 12 章	指数	194
12.1	基本概念	194
12.1.1	统计指数	194
12.1.2	指数体系	194
12.2	统计指数介绍	195
12.2.1	股票指数	195
12.2.2	波罗的海干散货指数	195
12.2.3	PMI 指数	196
12.3	统计指数分类	197
12.4	综合指数: 拉氏指数与帕氏指数	197
12.4.1	指数编制中的不同度量现象	197
12.4.2	拉氏指数与帕氏指数的定义	197
12.4.2	拉氏指数与帕氏指数的比较	199
12.5	统计指数优良性评价	199
12.5.1	定量评价	200
12.5.2	定性评价	202
12.6	指数体系的典型应用	204
12.6.1	因素分析	204
12.6.2	景气评价	205
12.7	综合评价指数	207
12.7.1	综合评价指数设计的一般步骤	207
12.7.2	综合评价指数计算实例	209
	思考讨论题	210
	练习题	210
	参考文献	212

1.1 统计学的定义

出版于 18 世纪、英语世界最早的《大英百科全书》(*Encyclopaedia Britannica*)中对统计学的定义是:“Statistics, the science of collecting, analyzing, presenting, and interpreting data. Governmental needs for census data as well as information about a variety of economic activities provided much of the early impetus for the field of statistics. Currently the need to turn the large amounts of data available in many applied fields into useful information has stimulated both theoretical and practical developments in statistics……”

目前国内教材提及较多的统计学定义是英国统计学家肯德尔(Maurice George Kendall, 1907—1983)提出的。肯德尔在联合国经社理事会和教科文组织的资助下,与巴克兰(Buckland)合作于 1957 年出版了《统计名词词典》一书。在这本词典中肯德尔给“统计学”(Statistics)做出定义:“统计学是收集、分析和解释统计数据的科学。”一般认为,统计学是一门收集、分析、表述和解释数据的科学,具体表现为一些方法,是用来设计实验、获得数据,然后在这些数据的基础上组织、概括、分析、解释和得出结论、演示结果的科学。统计学的起源最初与政府收集的各类社会经济领域的的数据有关,现在已经拓展到所有与数据相关的学科领域,统计学所研究的现象可以是社会现象,也可以是自然现象。统计学已经发展成为由若干分支学科组成的学科体系,当代人类社会几乎所有的研究领域都要用到统计方法。

1.2 统计学的典型应用

Kotz Johnson 和 Read 等人编辑的《统计科学百科全书》(*Encyclopedia of Statistical Sciences*)是最具权威性的统计学百科全书之一,该书列举了四十多个运用统计的领域。下面列举一些统计学的典型应用。

1.2.1 经济:经济普查与宏观经济监测

经济普查是指对我国境内第二、第三产业的所有经济活动进行的国情、国力调查。全

国经济普查合并了计划经济时期的工业普查、第三产业普查、基本单位普查,新增建筑业普查、个体工商户普查,每五年一次,2013年进行了第三次经济普查。

2004年9月国务院颁布的《全国经济普查条例》指出:经济普查目的是为了全面掌握我国第二产业、第三产业的发展规模、结构和效益等情况,建立健全基本单位名录库及其数据库系统,为研究制定国民经济和社会发展规划,提高决策和管理水平奠定基础。通过经济普查,把国家家底摸清,有利于政府部门有效开展宏观调控、科学编制国家发展规划。

正如诺贝尔经济学奖获得者萨缪尔森和另外一名经济学家诺德豪斯合编的《经济学》经典教科书中写到的一样:“如果没有诸如GDP这些核算经济总量的指标的话,政策制定者们只能在杂乱无序的数据海洋中漂泊。GDP及其相关数据资料就像灯塔一样,帮助政策制定者们把经济驶向关键的目标。”

1.2.2 人口:人口普查

1790年美国进行的人口普查被认为是第一次现代意义的人口普查。美国宪法规定众议院人数按照人口普查确定的各州人口数分配。美国宪法还规定人口调查应该于美国国会第一次会议之后三年内举行,之后每十年举行一次,调查方法另以法律形式规定。联合国在第二次世界大战之后发布人口普查手册,建议各国十年进行一次人口普查。多数国家(地区)人口普查十年一次,少数国家和地区如日本、韩国、澳大利亚、中国香港五年一次。

人口普查的重要性主要体现在以下三个方面。(1)为政府服务。政府是人口普查数据最大的用户。美国人口普查的产生是因为政府要进行管理或进行选举。我国“十二五”20个规划目标里,与人口有直接联系的目标占了三分之一以上。国家对于教育、卫生、“三农”工作等政策制定,都要以人口信息为依据。(2)为科学研究服务。人口普查的数据汇总向社会公开发布,为社会各界,特别是为政府、科研机构开展社会问题研究,制定政策提供了直接的服务。(3)为企业服务。企业需要知道目标人群在哪里、有多少、年龄性别、收入等情况如何,才能更好地制订商业活动计划。

1.2.3 管理:企业管理

企业在市场竞争中需要进行调查分析,作为生产和营销的依据。近年来在电子商务、市场营销与数据分析领域经常被提及的一个例子是美国沃尔玛超市中“尿布与啤酒”现象,沃尔玛市场数据分析人员注意到沃尔玛公司某些店面出现啤酒、尿布销售数量高度同步的现象。原因是美国妇女经常会嘱咐丈夫在下班回家的路上为孩子买尿布,而丈夫在买尿布的同时又会顺手购买啤酒回家喝(注:后来这一发现被讹传为啤酒、尿布在沃尔玛超市中放在一起)。

财务分析方面,上市公司财务数据是股民投资选择股票、评价公司业绩与发展潜力的重要参考依据。一些投资咨询公司主要是根据上市公司提供的财务和统计数据进行分析,为股民提供投资参考。企业自身的投资也离不开对财务数据的分析。

产品质量管理方面,统计过程控制(简称SPC)应用统计技术对生产过程中的各个阶段

进行评估和监控,保持生产过程处于稳定水平,从而保证产品质量符合要求。主要有两个方面:一是分析过程稳定性,对过程存在的异常因素进行预警;二是对过程质量进行评价。通过控制生产各相关要素,避免浪费。在生产质量管理中大量、深入推行统计过程控制方法被普遍认为是日本第二次世界大战后工业崛起的主要原因之一。

保险精算是综合运用人口学、金融学、经济学知识,利用数学方法对各种保险活动未来风险进行分析、估价和管理的应用学科,常见内容如研究保险事故的出险规律、保险事故损失额的分布规律、保险人承担风险的平均损失及其分布规律、保险费率和责任准备金、保险公司偿付能力等保险具体问题。一般又分为寿险精算和非寿险精算。

哈雷彗星发现者、英国天文学家哈雷(Edmund Halley)在1693年发表世界上第一张生命表。18世纪中叶辛普森(Thomas Simpson)制作出死亡率增加而递增的保险费率表,陶德森(James Dodson)提出不同年龄的保险费计算方法。今天的保险精算以大量统计数据为基础构建出形形色色的保险模型(保险产品),以尽可能在支付保险购买者赔偿金的前提下,降低保险公司赔付风险。在目前的北美精算师资格考试和中国精算师资格考试设置的考试科目中,统计学占有相当重要的比例。

1.2.4 文学:《红楼梦》作者之谜

《红楼梦》一书共120回,一般认为前80回为曹雪芹所写,后40回为高鹗所续,然而对于这种主流观点一直存在不同的看法。一般来说,一个人在使用某些词方面会表现出一些个人习惯与特征,由此可以采用数学方法对《红楼梦》前80回与后40回作者是否为同一人进行研究,国内外有多位数学家用统计方法进行过研究。例如,1985年复旦大学李贤平教授带领学生将120回看成120个样本,将与情节无关的一部分虚词出现次数作为统计变量,运用统计分析方法分析哪些回目出自同一人。李教授团队用每个回目中47个虚词“之、其、或、亦……呀、吗、咧、罢……可、便、就……”出现次数作为《红楼梦》各个回目的量化特征,聚类后得到的结果是:120回分成前80回为一类,后40回为一类,证实了前80回与后40回不是出自同一人。统计学方法在一直采用文字分析方法进行研究的传统红学界引起了极大的震动。

1.2.5 体育:NBA的数据分析

从20世纪80年代起,所有NBA球员得分、篮板、助攻、盖帽、抢断、失误、犯规等一系列场上技术指标的数据均被详细统计,教练在大量数据基础上进行分析,改进战术,提高取胜概率。比如,数据统计发现,姚明在篮下接球、运球,右手接球时通常能运球三次,左手接球则只能运球两次。因此防守球员应该尽可能把姚明限制在远离篮筐的区域,让姚明运球三次以上才能来到篮下,然后姚明的失误概率就会大很多。

当然,做好数据分析能帮助球队取胜,但是决定性因素还是球队实力,毕竟数据分析不可能代替球员在场上打球。

1.3 统计学的主要分类

1.3.1 数理统计与应用统计

数理统计学是理论统计学。根据国际统计学会 (ISI) 会刊及统计学方面的国际著名杂志近年来发表论文的主题, 中国人民大学袁卫教授将近年来国际统计界的理论研究概括为: 概率极限理论、树形概率、Banach 空间概率、泊松逼近、随机网络、马尔科夫过程及场论、布朗运动与偏微分方程、序贯分析和时序分析中的交叉界限问题、马尔科夫过程与狄利克雷表的一一对应关系、函数估计中的中心极限定理、极限定理稳定性、因果关系与统计推断、预测推断、网络推断、M 估计量与最大似然估计、参数模型中的精确逼近、非参数估计自适应方法、非线性时间序列、时间序列中确定模型与随机模型比较、极值统计、贝叶斯计算、突变点分析、函数型数据统计分析等。

应用统计学就是将数理统计应用于各个学科领域, 如社会经济统计、教育统计、心理统计、生物统计、气象统计等, 着重研究统计方法在该学科的具体应用, 主要研究内容不是统计方法数学原理的推导与证明。本教材属于社会经济统计内容。

1.3.2 描述统计与推断统计

描述统计 (Descriptive Statistics) 是研究数据收集、处理、汇总、图表描述、概括与分析的统计方法。简单地说即就事论事, 就现有的数据论数据, 不做其他推测。

推断统计 (Inferential Statistics) 是研究如何利用样本数据推断总体特征的统计方法。在很多统计问题中, 或者由于人力、物力、财力或时间限制, 或者由于取得全部数据是不可能的, 或者虽然能够取得全面数据但数据收集本身带有破坏性, 只能收集部分数据对所研究对象的数量特征或数量规律性进行推断。比如, 要了解一个地区的人口特征不可能对每个人的特征进行测量; 对产品质量进行检验不可能对每个产品进行测量, 特别是高成本产品甚至只能举行数量很少的小样本抽样。

统计推断要解决的问题分为两大类: 第一类是参数估计 (Parameters Estimation), 也就是通过一部分数据的特征来估计总体数据的相应特征; 第二类是假设检验 (Hypothesis Test), 通过一部分数据来确定总体数据是否具有某种指定的特征。

1.3.3 参数统计与非参数统计

参数统计与非参数统计的根本区别在于前者已知分布函数信息, 而后者不知。下面以常见的统计检验为例进行介绍。

参数检验 (Parametric Test) 是建立在总体数据分布已知或假设为已知基础上的, 如第 6 章中进行的均值检验就属于参数检验, 首先做出总体数据服从正态分布的假设, 然后在此假设基础上, 根据样本资料对均值进行检验。对于参数统计方法的选择, 需要注意两个方面, 即: (1) 数据是否符合正态分布; (2) 是否大样本。

正态分布在自然界与社会经济现象中广泛存在,19 世纪统计学研究最多的分布函数就是正态分布。但实践表明,在很多具体的领域,最有代表性的分布未必是正态分布,如许多电子产品元件和设备的寿命服从指数分布、威布尔分布。正态分布的检验理论相当完善,对于一些非正态分布,如二项分布或一些不知道具体分布函数形式或知道具体分布函数形式但是表达形式很复杂的情况,当样本量很大时可以转换为正态分布近似处理。例如, t 分布的极限分布是正态分布,当 t 分布样本数大于 30 时, t 分布的形态就与正态分布已十分接近,故在此情况下,当样本量大于 30 时,即可利用正态分布取代 t 分布。

数理统计研究中有大样本理论或大样本性质的概念,数理统计提出大样本概念的初衷是:讨论当某统计量的精确分布未知(或因很复杂不易获得时),当样本量趋于无穷大时是否有极限分布。如果极限分布存在,样本量足够大时,就可将这个极限分布作为该统计量的近似分布。需要注意的是,不能笼统地说样本量大于 30 称为大样本,小于 30 称为小样本。大样本与小样本的区分要看极限近似分布所要求的样本数量。不同的分布函数、不同的学科领域所说的大样本与小样本具体取值往往是不同的。

当样本量很大时可以用正态分布近似处理,但当小样本时近似处理误差可能很大。因此有必要研究非正态、小样本的参数检验和分布检验问题。目前国内本科统计学教材基本上只介绍大样本下的统计检验方法,对于小样本检验介绍很少。

根据应用的需要,学界对不同分布与不同规模样本交叉组合产生的不同情况也进行了很多研究,如本书参考文献[3]中证明了 n 个相互独立、服从 $(0, 1)$ 均匀分布随机变量的乘积 X 之自然对数值服从 $\text{Gamma}(n, 1)$ 分布的结论,进而使用该结论讨论了指数的分布、威布尔分布等分布的小样本检验问题。

非参数检验(Non-parametric Tests)是相对于参数检验而言的。在许多实际问题中,人们往往对总体的分布形式知之甚少,很难有正确假定,进一步地,如果是小样本,变量的分布不清、已知不服从正态分布时,不能用参数检验方法。不考虑总体分布具体形式,直接从样本数据本身来获得所需信息,对总体分布的类型和位置进行检验的各种方法称为非参数检验(Nonparametric Statistics 或 Distribution-free Statistics)。

非参数检验的特点如下。(1)非参数检验不要求样本服从的总体符合某种预设的理论分布,适用范围更为广泛。(2)参数检验主要适用于定比数据分析,非参数检验方法可以处理所有类型的数据。(3)从学习难度看,虽然非参数检验计算方法名目繁多,但其表现形式一般比较简单并易于理解、容易学习。而参数检验都需要对统计量分布进行推导,学习难度大一些。(4)非参数检验方法结果更稳定,更不容易受到数据噪声干扰。以非参数检验普遍使用的秩序统计量为例,所谓秩序统计量就是只考虑样本数据的大小顺序,不考虑样本数据的大小,因此对样本做任一单调增的变换,样本分量的大小顺序不会变化。例如,在面试过程中,招生方往往就各方面(如相貌、表达能力、创造性等)让面试官对考生打分。假定最低分为 0 分,最高分为 10 分。当然,因为不同的面试官对考生有不同的好恶,所打的分数或多或少地具有主观性,因此评分相对意义比绝对意义更准确。在这种情况下,秩序统计量更稳定。

参数估计、假设检验两大类问题有各自对应的参数与非参数具体方法,本书后面章节会做一些介绍。

当然,非参数统计方法检验功效略低于参数统计方法。对于同一组数据,在给定的显著性水平下进行检验时,非参数统计要求的样本信息比参数统计方法少,固然提高了方法与结论的稳健性,但是也牺牲了一定的精确性。不过研究表明,常用非参数检验方法统计效能大约为参数检验方法效能的95%,也就是说,如果通过95次独立观察获得的数据能够保证参数检验所要达到的精度,那么若用非参数检验方法,需要100次的独立观察。总之,非参数检验需要更大样本容量来达到参数检验一样的检验精度,但并不需要很多,多数情况下在可接受范围内。

对于同一个问题,往往可以同时选择不同的参数统计方法与非参数统计方法,选择的基本原则是:(1)若数据具备参数统计要求的条件,则必须首先选择参数统计方法,当不知道数据分布函数时选择非参数统计方法;(2)对检验结果稳健性要求较高时优先考虑非参数方法;(3)当同一问题有多种非参数方法选择时,优先考虑利用数据信息量最多的非参数检验方法。

1.4 统计软件介绍

随着计算机的迅速发展,对于非专业统计人员的学习重心已经由过去的学习统计方法具体推导过程转移为熟悉统计方法的基本思路与统计软件的使用。只要懂得统计方法的基本前提假设、基本思路,就可以正确使用统计软件快速、高效地解决实际问题,无须将过多时间花费在数学推导上。

统计软件大致有4种来源:(1)成熟度较高的商业性统计软件,如欧美著名的SPSS、SAS,近年来崭露头角的国产DPS统计软件;(2)由个人或团队支持、持续开发的开源软件;(3)课题资助要求结题成果中包括软件的情况,往往课题完成后就束之高阁,缺乏后续开发;(4)个人出于兴趣或为完成临时任务偶尔开发的软件。

从专业程度上看,统计软件可以分为非专业统计软件与专业统计软件。

代表性非专业统计软件——4M。所谓4M也就是目前最有代表性的4种含有统计学函数与工具箱的计算机代数软件,即Matlab、Mathematica、Maple、Minitab。

Matlab软件是20世纪70年代后期美国New Mexico大学计算机系主任Cleve Moler在给 学生讲授线性代数课程时,利用业余时间为学生编写的程序。1984年Cleve Moler和工程师John Little成立Math Works公司。Matlab可以进行线性代数、数理统计、自动控制、时间序列分析、数字信号处理、动态系统仿真,计算微积分、求解各类复杂方程。Matlab在欧美高校及国内自然科学与工程界已经普及,近年来逐步向社会科学界延伸。Matlab有一些非商业的替代品,如Freemat、Scilab等,语法界面与Matlab兼容,当然功能不如Matlab强大,不过一般性应用足够。

Maple软件由加拿大Waterloo大学开发,提供了包括高等数学、线性代数、离散数学、图形学等方面5000余种数学函数,符号计算功能较佳。实际上,MathCAD和Matlab等软件的符号计算功能长期以Maple为内核。航天、航海、天文、水利、材料、建筑、电力、机

械、燃烧、爆轰、热能、原子、化学、自控、导航、通信等各个领域 Maple 都有丰富的函数供使用者调用。

Mathematica 是 Wolfram Research 公司推出的计算机软件。Mathematica 有丰富的数学函数库,可以快速解答微积分、线性代数、微分方程、复变函数、数值分析、概率统计等问题。Mathematica 长于符号计算,在物理学、工程学和数学领域影响力较大。

Minitab 软件 1972 年出现于美国宾夕法尼亚大学,是现代质量管理统计软件的领先者,其应用最广泛的领域也是质量统计。统计过程控制是现代企业生产质量、产品质量控制的基本手段。Minitab 提供了各类控制图,如 $\bar{X}$ -R 控制图、 $\bar{X}$ -S 控制图、I-MR 控制图、P 图、NP 图、C 图、U 图、EWMA 控制图等。

专业统计软件又分两种情况:第一种是单一功能的专业统计软件,围绕少数或某个统计算法设计的统计软件,如实现结构模型方程分析方法的专门统计软件有 LISREL、EQS、AMOS、MPLUS 等;第二种就是综合性专业统计软件,包含了大量代表性统计算法与模型,下面集中介绍综合性专业统计软件。

1.4.1 国外综合性统计软件: R、SPSS 和 SAS

R 软件是 S 语言的开源实现。S 语言是由 AT&T 贝尔实验室开发的用于数据分析、作图的编程语言,被认为是数据分析领域的标准语言。S 语言于 1998 年获得 ACM(美国计算机学会)的软件系统奖。S 语言的最初实现软件是商业软件 S-PLUS。后来 Auckland 大学统计系 Robert Gentleman 和 Ross Ihaka 及其他志愿人员开发了 S 语言的免费实现软件即 R 系统。R 软件与 S-PLUS 软件有一定的兼容性,故有人说 R 软件是 S-PLUS 软件的克隆。R 软件具有免费、模块(类似于 Matlab 的工具箱)丰富、更新快等特征,在欧美已经成为数理、金融、生物信息学、网络分析领域进行数据分析的标准统计软件。2009 年 1 月 7 号《纽约时报》科技版发表文章 *Data Analysts Captivated by R's Power*, 文章里有这么一句话:“The popularity of R at universities could threaten SAS Institute.”R 软件的突飞猛进,首先是因为开源免费。其次,新手学习 R 软件很快就能开发 package 软件包,放到 CRAN 网站上面供其他人免费下载,不断扩大个人影响力。R 软件的各种 package 非常丰富,而且更新速度很快。第三,内存瓶颈消失。在 Amazon、Facebook、Google 数据中心很容易从几千台机器里面集中 TB 数量级的内存用于运算。R 软件的主要官方网站是 <http://www.r-project.org>, 在该网站上可以下载 R 软件的安装程序和文档。世界各地不同的人提交功能模块可以在 CRAN 网站下载(<http://cran.r-project.org>)。The R Journal 是一本专门发表 R 语言最新动态的免费电子期刊。

SPSS 是该软件英文名称的首字母缩写,原意为“Statistical Package for the Social Sciences”,即“社会科学统计软件包”,也被开玩笑地称为“Statistic Package for Stupid Students”。1968 年美国斯坦福大学研究生 Norman H. Nie 为解决社会学研究中的统计分析问题和其他两位合作者开发了本软件。1975 年成立 SPSS 公司,2010 年该公司被 IBM 公司收购。现在 SPSS 旗下主要有 4 个产品系列:SPSS Statistics(原 SPSS),即统计分析产品;SPSS Modeler(原 Clementine),即数据挖掘产品;SPSS Data Collection Family(原 Dimensions),即数据采集产品;SPSS Collaboration and Deployment Services(原 Predictive Enterprise