



新鲜有趣，

就这样爱上统计学！



生活中的 统计学

李倩星◎著



好玩
实用

从各种实际案例中
学习统计学



内容
全面

一网打尽数种常用的
统计分析方法



由浅
入深

从零开始，引领你
一步步走入统计学殿堂

清华大学出版社



生活中的 统计学



清华大学出版社
北京

本书封面贴有清华大学出版社防伪标签,无标签者不得销售。
版权所有,侵权必究。侵权举报电话:010-62782989 13701121933

图书在版编目(CIP)数据

生活中的统计学/李倩星著. —北京:清华大学出版社,2017
ISBN 978-7-302-48269-7

I. ①生… II. ①李… III. ①统计学—通俗读物 IV. ①C8—49

中国版本图书馆 CIP 数据核字(2017)第 207295 号

责任编辑:刘志彬
封面设计:汉风唐韵
责任校对:王荣静
责任印制:宋 林

出版发行:清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址:北京清华大学学研大厦 A 座

邮 编:100084

社 总 机:010-62770175 邮 购:010-62786544

投稿与读者服务:010-62776969, c-service@tup.tsinghua.edu.cn

质量反馈:010-62772015, zhiliang@tup.tsinghua.edu.cn

印 装 者:三河市金元印装有限公司

经 销:全国新华书店

开 本:148mm×210mm

印 张:8.125 字 数:188 千字

版 次:2017 年 9 月第 1 版

印 次:2017 年 9 月第 1 次印刷

定 价:36.00 元

产品编号:074258-01

前言：就这样爱上统计学

统计学是一门与实际生活十分贴近的学科，它起源于研究社会经济问题，早在古希腊时期，亚里士多德就发明了以记录各希腊城邦的历史、行政、科学、艺术、人口、经济等数据的城邦纪要，除去这些日常记录外，统计学也很早就被应用在卫生监察和健康等方面。

约翰医生平息霍乱是一个著名的统计学例子。霍乱是 19 世纪最令人害怕的流行疾病，它在伦敦曾肆虐多次，夺走了数千万人的生命。约翰·斯诺统计了一些霍乱死者的生活情况，发现霍乱的发生与水源有明显关联，凡是在百老大街的水泵取水的伦敦居民，发病率明显要高很多。由此，约翰医生证明水源是霍乱传播的一大途径，提出了实用的几种预防措施，有效降低了霍乱的致死率。

另一个类似的例子发生于克里米亚战争期间。1845—1856 年，南丁格尔为克里米亚交战双方的伤兵服务，将后方医院的死亡率由 42.7% 下降至 2.2%，以人道、慈善之心挽救了许多人的生命。同时，她还是英国皇家统计学会的第一名女性会员，她发明了极区图，反映了不同时间段内战斗死亡的士兵人数与受伤而缺乏治疗死亡的士兵人数，清晰地说明前者的数量少于后者，从而使英国当局意识到改善伤兵医疗工作的必要性。

在现代社会,统计学逐渐推广到社会科学、自然科学和工程技术科学等越来越多的领域,应用例子更是多如繁星、数不胜数。美国零售巨头沃尔玛归纳分析了消费者的购物小票,发现年轻爸爸们在购买尿布时,常常会搭两瓶啤酒,好在晚上看电视时过过酒瘾。于是,沃尔玛将两者放在一起销售,使尿布和啤酒的销量均大幅增加,这就是著名的“啤酒与尿布”案例。

而1936年美国总统大选则是一个统计学家津津乐道的反面案例。著名的《文学摘要》杂志社按照电话簿目录和汽车俱乐部成员名单向外派发了1 000万份调查问卷,调查结果显示兰登将获胜,但实际结果恰好与调查结果相反。

这是由于当时电话和汽车十分昂贵,几乎是中产阶级的代名词,《文学摘要》忽略了这点,它的调查结果只能说明中产阶级更支持兰登,而实际上人数更多的贫困阶级则是罗斯福的忠实拥护者。这一疏忽直接导致《文学摘要》威信扫地,关门停刊。

在亚里士多德之后,英国的威廉·配弟使用数字、重量和尺度将社会经济现象数量化,并于1672年出版《政治算数》一书,这是近代统计学成立的标志。从此,统计方法与数学计算和推理方法开始结合。

统计学的两个核心理论是大数定理和中心极限定理,前者证明了一小部分样本即可代表全体,后者解释了样本量和结果可靠性之间的联系。这两个定理共同保证了抽样分析的合理性。本书开篇便介绍了这两个定理,以及如何挑选样本,确保分析结果的可靠性。

抽取出样本数据后,探索性分析是不能跳过的一个分析步骤。本书的第2章介绍了各种探索性分析方法,第3章则用图表的形式

来表现分析结果。这两个章节的意义在于初步了解样本数据的特点。

概率分布是现代统计学的理论基础。从15世纪开始,数学家就对赌博问题产生了浓厚的兴趣,最知名的赌博问题有分赌注问题和三门问题等。传统的赌博问题引出的是离散概率,由离散概率进一步研究,又得到连续概率。本书第4章讨论了几种常用的概率分布。

在学习统计学的理论基础后,一个自然而然的想法是使用样本中的信息去估计总体中的信息。例如,灯泡厂抽样检查时,抽取的灯泡寿命和全部灯泡的寿命是什么关系?参数估计和非参数估计所关心的就是这样的问题。第4章详细地讨论了这些内容,即如何用样本中的信息来表达总体中的信息。

有了对总体的估计后,我们还关心这个估计是否可靠。同样的估计方法在不同情况下的可靠性是不同的,第5章总结了如何判断估计值的可靠性,即假设检验的内容。第6章的主题是方差分析,它是假设检验的进一步延伸。方差分析与实验设计紧密相关,它最早用于解决农业问题,即改变各个培养条件,通过观察产量找出最佳的麦子种植方法。

第5、6、7三章的内容彼此相关,第8章则较为独立。它研究了一些变量是如何决定另一些变量的,相关分析和回归分析是统计学中十分重要的部分,也是实际生活中应用最广的两种分析方法。

总之,统计学是一门发展时间较长、较成熟的学科。随着大数据的崛起,统计学也被注入了新的元素。越来越多的人激发了对统计学的兴趣。本书较全面地介绍了常见的统计学分析方法,包括描述性统计分析、参数估计、非参数估计、相关分析和回归分析等部分。此外,本书还注重与实际案例的结合,生活中的许多常见现象都可

以在本书中找到踪影。在写作本书过程中,我们也注意坚持以下特色。

本书特色

1. 案例贴近生活,语言生动有趣,实用性强

统计学出现在生活中的方方面面,一些大家常见的生活现象背后蕴含着深刻的统计学原理。本书通过讨论这些活生生的案例,使用生动活泼的语言,形象地讲解了10余种统计分析方法,这些案例能帮助读者较容易地领略统计分析背后的原理,而这些方法反过来又可以解决许多生活中的实际问题。通过阅读本书,读者能够深切地体会到统计学在生活中的各种用处。

2. 内容深入浅出,层层递进,适合各个层次的读者

本书从最简单的描述性统计分析入手,由易到难,依次讲解了统计图表绘制、概率分布等基础知识,以及参数估计、方差分析、相关分析和回归分析等分析方法。而在每一章节中,作者也精心安排知识点分布,以达到层层递进的效果。本书内容深入浅出,读者既可以按照顺序依次阅读,也可根据自己的实际水平,单独阅读每一章节。

3. 文章脉络清晰,构建出一个完整的数据分析知识体系

统计学分析的各个分析方法之间存在递进关系,如方差分析是在假设检验的基础上发展得来的,而假设检验又是对参数估计的拓展与延伸。本书按照各个知识点的关系合理地组织了文章结构,各个章节间彼此关联,构建出一个完整的数据分析知识体系,可帮助读者对统计分析形成一个全面的认识。

本书体系结构

本书共由8章构成,每一章都有一个独立的主题,分别为数据的

收集、描述性统计分析、绘制统计图表、常用的几种概率分布、参数估计和非参数估计、假设检验、方差分析、相关与回归问题。这8个章节又进一步可分为两大部分。

第一部分为前4章。其中第1章讨论了大数定律、中心极限定理和几种常用的抽样方法,解释了统计分析中样本数据通常是从何而来,以及统计分析方法为何有效。第2章讨论了一些常见的统计量,如均值、方差、四分位差等,并从统计学角度严格地阐述了它们的不同,以及它们的特点,这一章节有助于读者初步构建统计思想,从统计学的角度理解问题。第3章的主题是绘制图表,这一章与第2章节前后呼应,向读者展示了统计学的一个基本职能,即将分析结果简洁明了地表达出来。第4章则是有关概率分布的基本知识,这是比较抽象而难以理解的一章,通过阅读这一章节,读者将正式踏入统计学的大门。

第二部分为后4章。在学习基础知识后,读者将开始接触更高深的统计学分析方法。第5章包含两个主题,即参数估计和非参数估计,它们研究的是如何用样本中的信息去估计总体中的信息。第6章讨论了样本对总体的估计是否可靠。而第7章则是对第6章的拓展,将两个样本的情况拓展到多个样本中去。

第5、6、7章的内容环环相扣、相辅相成,第8章则较为独立。但相关分析和回归分析是统计学中十分重要的部分,也是实际生活中应用最广的两种分析方法。第8章仅简要讨论了几种常见的回归方法,还有一些内容并未列出,如泊松回归分析等。

以上划分方法仅为一个参考,本书的8个章节互相联系又彼此独立,读者可按照上述顺序阅读本书,也可优先阅读某些章节,如在读完前4章后,可直接阅读第8章。

本书读者对象

- 想要学习统计学的数据分析从业人员
- 统计学、金融学、计算机技术与科学等专业的学生
- 想要提高统计分析能力的数据分析师
- 希望系统学习统计分析方法的从业人员
- 其他对统计学有兴趣爱好的各类人员

编 者

2017 年 7 月

目 录

第 1 章 如何从海量数据中筛选你需要的数据

——数据的收集 1

- 1.1 从柏拉图摘麦穗说起 2
 - 1.1.1 如何摘到最大的麦穗 2
 - 1.1.2 样本点和样本的区别 4
 - 1.1.3 37%法则 5
- 1.2 新生儿性别比例总是趋于稳定吗 7
 - 1.2.1 近似相等与相等的重要区别 7
 - 1.2.2 从婴儿性别比例推广到火柴燃烧时间 9
 - 1.2.3 大数定理在保险行业的应用 11
- 1.3 为什么一小部分人的意见就能代表全体人 13
 - 1.3.1 100 个人的收入能代表 10 000 个人的收入吗 13
 - 1.3.2 确定抽样人均收入与真实人均收入间的误差 15

1.3.3	考虑样本的最佳大小	17
1.4	一份标准试卷是怎么设计出来的	19
1.4.1	抽样时真的能做到完全随机吗	19
1.4.2	确保高考试卷的公平性	21
1.4.3	抽样调查的一些补充要点	23

第2章 小小统计量中的大奥妙

——描述性统计分析

2.1	你知道年龄属于哪一类数据吗	26
2.1.1	如何用数字表示求职者的最高学历	26
2.1.2	统计调查中专用的数据类型有哪些	28
2.1.3	新浪微博属于哪类数据	29
2.2	平均薪资的陷阱和真相	31
2.2.1	是谁拉高了平均薪资	31
2.2.2	如何计算加权平均薪资	33
2.2.3	用四分位数估计薪资区间	35
2.3	得分高就是好球员吗	38
2.3.1	用极差、四分位差和平均差衡量球员的 水平	38
2.3.2	方差和标准差的关系	41
2.3.3	消除了量纲的统计量	42
2.4	常见的社会经济统计量	44
2.4.1	国内生产总值到底是怎么核算出来的	44
2.4.2	根据国内生产总值衍生出的其他统计量	46

2.4.3 基尼系数和恩格尔系数	48
------------------------	----

第3章 使用色彩和图形传递信息

——绘制统计图表	51
----------------	----

3.1 为什么条形图比折线图更好	52
3.1.1 最基本的3种图形	52
3.1.2 条形图优于折线图的两个理由	56
3.1.3 直方图和条形图的区别	59
3.2 离散型变量适合绘制的图形	62
3.2.1 什么样的数据适合画饼图	62
3.2.2 有时候表格比图形更重要	65
3.2.3 马赛克图和茎叶图也是图形的一分子	67
3.3 为高维变量绘图	69
3.3.1 三维图一定比二维图美观吗	69
3.3.2 按照变量绘制高维图形	72
3.3.3 按照样本点绘制高维图形	76
3.4 好图形的3个标准	79
3.4.1 常见的几种绘图错误	79
3.4.2 一些优秀图表案例	84

第4章 用概率的眼光看世界

——常用的几种概率分布	89
-------------------	----

4.1 买彩票真的能发家致富吗	90
4.1.1 由福彩6+1的中奖概率导出二项分布	90

4.1.2	计算福彩 6+1 的获奖期望	93
4.1.3	超几何分布和人寿保险问题	94
4.2	几点出门才不会迟到	96
4.2.1	用二项分布逼近泊松分布	97
4.2.2	一个简单的公共汽车客流案例	99
4.2.3	如何判断一个分布是否为泊松分布	101
4.3	捕捞到金鱼的概率有多大	102
4.3.1	从水缸里捞金鱼谈到几何概率	103
4.3.2	一维均匀分布和二维均匀分布	105
4.3.3	利用均匀分布进行模拟估计	109
4.4	智商多少才是正常水平	111
4.4.1	高尔顿板实验和正态分布的联系	111
4.4.2	一元正态分布的主要性质	114
4.4.3	计算正态分布的概率	117
4.5	手提电脑的寿命到底有多长	121
4.5.1	电器寿命和指数分布的关系	121
4.5.2	从泊松过程中推导指数分布	124

第 5 章 用概率分布解决实际问题

——参数估计和非参数估计

5.1	根据月账单明细估计消费水平	128
5.1.1	估计湖中的鱼苗数目	128
5.1.2	选择账单明细的方差估计量	130
5.1.3	点估计量的其他性质	132

5.2	进一步估计消费区间	133
5.2.1	估计变量是否服从正态分布	133
5.2.2	估计消费账单的区间	137
5.3	直方图估计 VS 核密度估计	139
5.3.1	用直方图估计花萼宽度数据的分布	140
5.3.2	使用核函数使密度函数变得平滑	142
5.3.3	K 近邻估计和聚类分析	146

第 6 章 判断估计结果的合理性——假设检验 149

6.1	如何得知袋装面包有没有偷工减料	150
6.1.1	确定面包重量的分布	150
6.1.2	双侧假设检验	153
6.1.3	单侧假设检验	154
6.2	投放广告能增加销售额吗	157
6.2.1	检验两个正态分布的均值	157
6.2.2	检验两个正态分布的方差	160
6.2.3	有关假设检验的补充知识	162
6.3	由遛狗结果求解后验概率	165
6.3.1	轮盘赌和遛狗的条件概率	165
6.3.2	儿童智商问题与参数的区间估计	167
6.3.3	根据后验概率比进行假设检验	169
6.4	补习班真的能提高小孩成绩吗	170
6.4.1	参数检验和非参数检验的区别	170
6.4.2	利用样本的秩判断两组成绩数据是否 有差别	172

6.4.3	从另一种角度检验补习班问题	174
6.5	检验离散型的顺序变量和因子变量	177
6.5.1	卡方分布就是正态分布的平方和	177
6.5.2	检验历年战争次数是否服从泊松分布	179
6.5.3	检验年龄和投票结果是否相互影响	181

第7章 从稻田试验发展出的学科——方差分析 185

7.1	从 F 分布推出方差分析的基本原理	186
7.1.1	从卡方分布导出 F 分布	186
7.1.2	方差分析的一些基本知识	188
7.2	去哪家餐厅吃饭更合算	191
7.2.1	餐厅得分的组内差异和组间差异	191
7.2.2	使用 F 分布检验餐厅得分是否有所不同 ...	193
7.2.3	方差分析的多重比较问题	194
7.3	餐厅的地理位置会影响菜肴的美味程度吗	196
7.3.1	在餐厅问题中引入餐厅地理位置因素	197
7.3.2	考虑地理位置与餐厅的交互作用	199
7.3.3	从两因素方差分析推广到其他的方差 分析	201
7.4	中药和西药哪个对糖尿病更有效	202
7.4.1	配对比较实验设计问题	202
7.4.2	随机化区组实验设计问题	204

第 8 章 统计学界的明珠——相关与回归问题 207

8.1	花瓣数据和花萼数据的关系	208
8.1.1	比较 4 种花朵数据的相关性	208
8.1.2	消除其他变量对相关系数的影响	211
8.1.3	计算离散型数据的相关系数	212
8.2	姚明的儿子会比姚明还高吗	214
8.2.1	父亲身高与儿子身高的相关性	215
8.2.2	使用最小二乘估计回归参数	217
8.2.3	在回归分析中引入母亲身高	219
8.2.4	使用逐步回归筛选自变量	221
8.3	收入和支出呈线性关系吗	224
8.3.1	高收入人群与低收入人群的消费模式	224
8.3.2	多项式回归和 R^2	228
8.3.3	广义线性回归模型和非线性回归模型	231
8.4	如何计算花朵的种类	232
8.4.1	将婚姻状态处理为哑变量	232
8.4.2	花朵种类计算结果与 S 分布	234
8.4.3	逻辑回归中的优势比	237
8.5	回归分析常见谬误	238
8.5.1	使用残差项检验异常值问题	238
8.5.2	DW 检验和自相关问题	241
8.5.3	多重共线性和异方差问题	243

第 1 章

如何从海量数据中筛选你需要的数据——数据的收集



本章介绍了统计学中的一个分支——推断统计。大数定理和中心极限定理是推断统计的主要内容,随机抽样则是它们的重要前提。本章通过几个案例说明了这两个定理的有效性和重要性,还讨论了随机抽样的相关问题。