

GUDMUND R. IVERSEN
MARY GERGEN

Statistics

THE CONCEPTUAL APPROACH

统计学

基本概念和方法

吴喜之 程 博 柳林旭 译
全莉萍 钟文瑄 熊怀羽



CHEP
高等教育出版社



Springer
施普林格出版社

GUDMUND R. IVERSEN
Swarthmore College

MARY GERGEN
Pennsylvania State University, Delaware County Campue

统 计 学

基本概念和方法

吴喜之 程 博 柳林旭 译
全莉萍 钟文瑄 熊怀羽



CHEP
高等教育出版社



Springer
施普林格出版社

图书在版编目(CIP)数据

统计学 / (美) 埃维森(Gudmund R. Iversen)等著; 吴喜之等译.
—北京: 高等教育出版社; 海德堡: 施普林格出版社, 2000.3
书名原名: Statistics
ISBN 7-04-007891-0

I. 统… II. ①埃… ②吴… III. 统计学 IV. C8

中国版本图书馆 CIP 数据核字(1999)第 41555 号

图字: 01-1999-1172 号

Translation from the English language edition
Statistics by Gudmund R. Iversen and Mary Gergen
Copyright © 1997 Springer-Verlag New York, Inc.
All Rights Reserved

统计学

Gudmund R. Iversen and Mary Gergen 著
吴喜之 程 博 柳林旭 全莉萍 钟文瑄 熊怀羽 译

出版发行 高等教育出版社 施普林格出版社

社 址	北京市东城区沙滩后街 55 号	邮政编码	100009
电 话	010 - 64054588	传 真	010 - 64014048
网 址	http://www.hep.edu.cn		

经 销 新华书店北京发行所
印 刷 北京民族印刷厂

开 本	787×1092 1/16	版 次	2000 年 3 月第 1 版
印 张	31	印 次	2000 年 3 月第 1 次印刷
字 数	760 000	定 价	52.00 元

©China Higher Education Press Beijing and Springer-Verlag Heidelberg 2000

版权所有 侵权必究

译者的话

我们在看到本书时,无不有一种耳目一新的感觉。虽然我们都受到过多年的数理统计的专业教育,但是从未见过一本统计教材有如此通俗幽默的文笔及如此广泛丰富的案例和习题。一个没有高等数学背景的人可以毫不困难地读完此书,并且得到对统计学最实际的认识。我们觉得,如果不把这本书尽快地介绍给我国读者,将是一种遗憾。

无论人们意识到与否,统计存在于国民经济及每日生活的各个方面。不懂统计很可能会不知不觉地受到损失。实际上,世界上绝大多数的统计应用是非统计专业的人来施行的。没有必要,也不可能要求这些人都拿一个统计文凭。他们所需要的是能运用初步的统计知识来识别大量出现的日常统计问题,并且能够利用现成的工具自己解决其中相当一部分;当出现难题时,他们只要知道到哪里寻求统计学家帮助就行了。因此对于广大实际工作者来说,主要的问题是统计的普及。越是先进的工业化国家,统计普及率越高。在进入 21 世纪的时候,不能想象各级的决策层和广大的实际工作者中仍然还普遍存在“统计盲”。如何在统计上“扫盲”呢?我们觉得本书很好地回答了这个问题。

本书的特色是没有利用数学公式,但却通过解决实际问题来生动描述统计的基本原理和方法。使得人们可以在愉快的心情中学到被认为只有在圣殿才能得到的知识。本书能很快地使读者对统计产生兴趣,而兴趣是学习的是大的帮助。自然,为了照顾那些永不满足的精力充沛的读者,书中每章后面都附有公式和进一步的阅读指导。本书的每一章都以一些引人入胜的例子和问题开头,然后通过对这些问题的分析和解答,展示了统计世界丰富多彩的本来面目。在应用中学习是学习的最好途径。多数对应用有指导意义的理论总是产生在实际需要之后,而不是之前。但是许多初等统计教科书,无论其使用的数学工具的多少,却和人们认识事物的次序相反:先讲一般的理论,再把实践作为使用理论的例子。不能想象那些只要泡在游泳池中和其他伙伴玩耍几天就能游泳的儿童需要了解肌肉的功能以及运动力学和流体力学的概念;也不能想象使用计算机的人都要学习汇编语言以及硬件的结构。本书令人信服地感觉到在游泳中学会游泳的兴奋。这是我国统计学基本教科书的编写和教学所应该借鉴的。

我们以极大的兴趣,饱满的精力和令人愉快的团队精神完成了这项任务。翻译的初稿是按照如下分工做的。全莉萍:前言,1,2,3 章;钟文瑄:4,5 章;程博:6,7,8,11 章;柳林旭:9,10,12 章;熊怀羽:13,14 章,名词解释,奇数号练习题答案及索引;吴喜之负责全书的协调、核对和最后定稿。我们要感谢高等教育出版社徐可及其同事对此书的认真负责的工作。另外,

北京大学姜伯驹院士使得高等教育出版社 GHEP-Springer 编辑室与我们开始了愉快的工作关系;北京大学光华管理学院使得本书有可能在刚出版就成为该院的 MBA 班 2000 年春季的统计学教材;对此,我们也表示感谢。

希望本书成为你最喜爱的书之一!

吴喜之	中国人民大学
程 博	南开大学
柳林旭	University of Pennsylvania
仝莉萍	University of Chicago
钟文瑾	Bowling Green State University
熊怀羽	University of North Carolina at Charlotte

2000 年春

前 言

这本统计学教材在设计和写作上都相当独特。该书是为了满足当代学生对统计日益增长而又尚未满足的需求,使他们能够熟练地掌握统计信息的特性。对于希望他们的学生能懂点统计知识的教师们来说,这本书很有裨益。然而,仅凭这一本书,是不可能使学生们变成统计学家的。

在过去几年里,统计信息已经从政府机构积满灰尘的档案中和学术计算中心里解放出来了。从国家关于健康改革和国防的政策到对于预期寿命、婚姻、堕胎、教育和体育的态度,统计信息在很多方面扮演了重要角色。统计信息经常在报纸、杂志、广播和电视节目中出现,它们甚至偶尔会在 MTV 和卡通片中做点缀。统计也渗透到了我们的教育课程中。在小学教室里和博士生讨论班中,统计信息已成为教育的基本特征。

尽管统计有这么多的应用,但是我们很难说大家对于统计信息不仅接受而且有了较多的了解。当人们看到一个研究结果时,他们如何判断结论是否正确?他们是否会问:这个研究中的变量是如何定义的?用了什么样的统计方法?什么是“统计显著”的结果?所报告的结果有什么样的不足?这些问题正是我们在本书中讨论的一部分内容。显然,理解了统计学的主要概念以后,大家才能够明白那些专门鼓捣数字的人们都干了些什么,并对他们结果进行评价。

这本书脱胎于 Gudmund R. Iversen 开设的一门课的讲义,目的是满足人们对统计信息日益增长的需要。该课是 Swarthmore 学院为使大学文科的学生能够迎接 21 世纪的挑战而开设的一系列课程之一。开设这些课程的思想是为了使学生们能够开阔眼界,而不是拘囿于某一学科的复杂之处。这些课程试图使学生们了解一个领域的主要思想是如何联系于现实世界的。在许多方面,统计学看起来正是这类课程的理想选择之一。尽管统计学可能是一门令人困惑的、自我膨胀的、神秘莫测的学科,但它也能够成为理解许多其它学科的一把钥匙。课程《统计学 I:统计思想》就是被设计成产生这种理解力的。事实证明,这门课非常受欢迎,其规模每年都在扩大。随着时间的流逝,这门课的讲义变得越来越精练和丰富,最终构成了本书的基础。

公 式

正如大部分统计教师所敏锐地意识到的那样,统计的教学方法已经发生了戏剧性的变化。计算机与教学环境的结合,尤其是界面越来越方便友好的统计软件的使用,已经使旧的学习方法——特别是记忆并运用统计公式——不再适用于大部分学生。为了忠实于本书的目的,我们在每一章的讲解中都没有使用统计公式。尽管这看起来有些激进,但经过深思熟虑之后,我们降低了公式的地位,把它们放在每章末尾的单独一节里。

我们的经验是,统计公式就像一门外语。如果一个人理解了这种语言,那么公式会大大增进他对统计学的理解;否则,这些公式就像密码一样难以破译。我们已经看到,很多同学在学习统计时,公式反倒成了一种障碍。我们坚信,不用公式,也照样有可能获得对统计思想的深刻理解。

习 题

只通过听课和念书来学习统计是很困难的。因为通过实践能使大家学得更好,所以我们选入了大量的习题。几乎所有的例子和习题都是使用我们从书籍、杂志和报纸中选取的实际数据。这些数据被应用于实际研究和公开的报告中,它们向我们展示了统计是如何应用于广泛的人类活动中的。

这些习题分为三类:回顾问题,用于检查对本章主要概念的理解;解释问题,要求学生们了解统计信息的意义;分析问题,要求学生们分析数据并用自己的方法解决问题。回顾问题用于检查对本章的理解,并提供一个班级讨论的背景材料;解释问题着重于定性而非定量,可以增进理解和鼓励对实际问题的应用提出建议;分析问题要求同学们以组为单位或以个人为单位,熟悉统计软件包的使用。每一个习题都对撰写统计报告提供了一个可能的主题。

本书的最后,给出了编号为奇数的习题的解答以及在做习题中用到的各种统计表。

致 谢

在写作本书的漫长过程中,我们的一些学生读了手稿早期版本的全部或部分。在此我们非常感谢以下几位的建议:Megan Falvey, Reginald Tilley IV, 特别是 Maya Rao。我们感谢 Maura MacDermott 允许我们引用她在《统计学 I》中所写的一篇文章。同时,我们也从这些年来参与本课的学生们所提的问题中获益匪浅。另外, Sloan 基金会在 Swarthmore 大学施行新文科规划时为《统计学 I》讲义的最初发展提供了资助。

除此以外,我们希望表达对以下人的感谢: Mount Holyoke 大学的 George W. Cobb; Plymouth 州立大学的 Robert W. Hayden; Grinnell 大学的 Thomas L. Moore; California 州立大学 (Hayward) 的 Michael Orkin; Florida 大学的 Richard L. Scheaffer, 他们对本书的写作提出了许多富有洞察力的建议。我们也要感谢那些以其对统计的专业意见为本书大大增色的其他匿名读者。我们还要感谢我们的加工编辑, Penny Hull, 他用勤勉和激动人心的话语鼓舞我们向前; 还有施普林格出版社(纽约)的 Bill Imbornoni 和 Theresa Shields 及他们的同事, 他们为本书的出版付出了辛勤的劳动。最重要的是,我们要感谢我们的出版人, Jerry Lyons, 她坚信可以从初稿中就可以看出作品的风采。没有比她更有创造力、更支持我们和更热情的出版者了。


最后,我们感谢我们的配偶 Roberta Rehner Iversen 和 Kenneth J. Gergen 对我们的充满爱心的支持和鼓励,他们从我们开始合作到写完最后一个证明就一直陪伴着我们。我们还要感谢我们的孩子——Gudmund 的 Eric、Gretchen、John 和 Kiraten 及 Mary 的 Laura、Lisa、Michael 和 Stan——以及他们的家庭对我们的支持和祝福。最后,让我们为能在和谐融洽的氛围中给此书带来了生命而庆祝!


Gudmund R. Iversen
Mary Gergen
Swarthmore, Pennsylvania
1997 年 1 月

目 录

1 统计学:随机性和规律性	1
1.1 统计学:用一句话来说是什么?	2
1.2 懂得如何运用统计:读者的目标	3
理解什么可能出差错	6
理解统计术语	7
1.3 统计学的主要思想	7
随机性和规律性:关系密切的孪生子	7
规律性中的随机性	7
研究随机性和规律性时的两个例子	8
概率:什么是机会	9
变量:我们给事物所起的名字	10
变量、值和个体	10
理论变量和经验变量	11
常数	11
1.4 统计的使用者	11
1.5 统计学和数学、铅笔及计算机的关系	14
1.6 小结	14
补充读物	15
习题	16
2 律据统收统	19
2.1 定义变量	20
2.2 观测数据:问题和可能性	21
总体相对样本	21
样本的选择:确信锅里的汤被搅拌均匀	22
随机样本:是什么?	23
方便样本;如何产生一个“坏的”样本	23
选择合适的样本	24
用于收集观测数据的变量的选择	25
2.3 收集观测数据时的错误和误差	25
抽样误差;并非错误的“误差”	26
未响应误差;粗鲁的、匆忙的或沉默的响应者造成的结果	27
响应误差	28
2.4 实验数据:寻找造成结果的原因	29

实验组和对照组	30
选择实验组和对照组	30
对人做实验时产生的问题	31
在实验中统计的角色	32
总结:班级规模影响学校表现吗?	34
2.5 数据阵/数据文件	35
2.6 小结	36
补充读物	37
习题	38

 数据的描述:数和和	45
3.1 图:画出数据	46
生成统计图	46
图的种类	47
3.2 分类变量:和饼和和条形图	48
为一个分类变量作图	48
为两个分类变量作图	49
3.3 度量变量:点和和直方图	51
为一个度量变量作图	51
为两个度量变量作图	57
时间序列图	58
3.4 根据数据作地图	60
3.5 作和:优秀的标准	62
“最少的笔墨”:最简单的图是最好的吗?	62
“简中垃圾”:垃圾的一种新名称	63
数据密度	64
“复杂性的展示”	64
3.6 表:改变排列方式可能更合适	64
3.7 小结	67
补充读物	68
习题	69

 数述描述数:计算汇总据计量	83
4.1 各种平均数:让我们数数有几种	84
众数:“最多的”的宿主	84
中位数:数到中间那一个	86
均值:平衡跷跷板	88
众数,中位数,还是均值?	90
4.2 变差:测量生活的乐趣	91
极差:套住两个极端值	91
标准差:重要的偏差	92

4.3	均值的标准误差	96
4.4	标准得分:比较苹果和桔子	97
4.5	简单化的收益与信息的丢失	99
	用图表来代替数据	99
	用汇总值代替数据	100
4.6	房地产数据:看不见的价格	100
4.7	小结	101
	补充读物	103
	公式	103
	习题	105

5	概率	113
5.1	怎样得到概率	115
	利用等可能性事件	115
	使用相对频数的方法	116
	利用主观概率	117
5.2	概率的计算	117
	概率的加法	118
	概率的乘法	118
5.3	优势:概率的对照物	118
5.4	离散变量的概率分布	119
	二项分布	120
	Poisson 分布	121
	超几何分布	123
	用图表来表示概率	123
	概率的计算	123
5.5	连续变量的概率分布	124
	标准正态分布:钟形曲线	124
	t -分布	126
	χ^2 分布	128
	F -分布	129
	正态分布数据的需要	129
5.6	使用概率来核对假设	130
	硬币是公平的吗?	130
	是一种公平的工作环境吗?	130
	两党选民是否势均力敌?	131
5.7	决策分析:利用概率来作决策	132
5.8	小结	134
	补充读物	136
	公式	136
	习题	139

6 作出结论:估计	147
6.1 样本统计论和总体参数	148
6.2 点估计	149
什么是一个好的点估计?	150
战略中使用点估计的例子:德军有多少坦克?	151
6.3 区间估计:给结论留一些余地	152
置信区间的长度	154
差异的置信区间	156
6.4 小结	157
补充读物	158
公式	158
习题	160
7 作出结论:假设设假	167
7.1 作为一个问题的假设	168
零假设	168
备择假设	169
回答问题时的错误	169
7.2 怎样回答零假设所提出的问题	170
概率: p -值	170
假设检验的机制	171
拒绝或不拒绝零假设	172
因果关系:过犹不及	173
一些统计理论和计算游戏	173
7.3 显著水平	175
7.4 总体比例检验	177
7.5 两个总体比例的差异	178
检验零假设	179
估计差异值	179
7.6 假设检验与构造置信区间	180
7.7 统计显著和实际显著	180
7.8 应用:何时拒绝零假设	181
关于合作性与竞争性的心理测试	182
对社区的蓝领工人的研究	183
7.9 小结	183
补充读物	185
公式	185
习题	188

8 变变间的关系	197
8.1 关于两个变量的 4 个问题以及它们之间的关系	198
问题 1 变量间有关系么?	201
问题 2. 关系的强弱程度?	201
问题 3 变量在总体中的关系如何?	202
问题 4 是因果关系吗?	202
8.2 预测:从一个变量到另一个变量	202
8.3 自变量和因变量	203
8.4 不同类型的变量:分类型变量、顺序型变量和数量型变量	204
8.5 回到因果关系的问题	205
别的变量的角色	206
时间的角色	206
多元因果关系	207
8.6 小结	208
补充读物	209
习题	209
9 两个分类变量的 χ^2 分析	217
9.1 数据分析:在态度上有可靠的差异吗?	218
条形图	219
分类变量的汇总计算	220
9.2 问题 1. 变量间的关系?	221
9.3 问题 2. 关系的强度?	222
样本中的 ϕ	223
总体中的 ϕ	224
9.4 问题 3:总体中的关系?	224
提出零假设	224
检验零假设	225
从 χ^2 到 p -值	225
χ^2 分析的自由度	226
9.5 问题 4. 是因果关系吗?	227
9.6 更大的表:更多的可能性	227
问题 1. 两变量间的关系?	229
问题 2. 关系的强度?	229
问题 3. 总体中的关系?	229
问题 4. 是因果关系吗?	230
9.7 小结	230
补充读物	231
公式	231

习题	234
----	-----

10 两个数值型变量的两归分析和的关分析 251

10.1 问题 1. 两个变量间的关系?	254
----------------------	-----

作这些数据的散点图 254

了解散点图 256

线性关系 257

10.2 问题 2a. 关系的强度?	257
--------------------	-----

r 是正的还是负的? 大还是小? 257

四种不同的散点图: 关系从强到弱 258

r 的解释: 不那么严谨 260

10.3 问题 2b. 关系的形式?	260
--------------------	-----

一条通过点的中心的直线 261

怎样计算回归直线: 最小二乘原理 263

用回归分析进行预测: 从脂肪到热量 264

效果的度量: r^2 的解释 265

相关和/或回归? 多多益善 268

变化数据的回归分析 270

10.4 问题 3. 总体中的关系?	271
--------------------	-----

置信区间的方法 271

用 t 进行假设检验 271

利用 F 进行假设检验 271

10.5 警告: 所测即所得	272
----------------	-----

10.6 用虚拟变量时怎样变得聪明些	274
--------------------	-----

自变量是有两个取值的分类变量和因变量是数值变量 274

因变量是有两个取值的分类变量和自变量是数值变量 276

10.7 问题 4. 是因果关系吗?	277
--------------------	-----

10.8 小结	277
---------	-----

补充读物	278
------	-----

公式	279
----	-----

习题	281
----	-----

11 ANOVA: 一个分类型数和一个数值型数型方型分析 301

11.1 方差分析: 对比事物的平均值	303
---------------------	-----

11.2 问题 1. 犯罪率和地区之间的关系	303
------------------------	-----

散点图 303

盒子图: 更简单地了解数据 304

11.3 问题 2. 关系有多强?	305
-------------------	-----

地区变量 306

残差变量 307

地区变量和残差变量的总效应: 总平方和 307

测量关系的强度	308	
对变化量的解释程度	308	
11.4 问题 3. 这个关系是纯属偶然的吗?		311
零假设	311	
F 变量的 p -值	311	
超出 F 检验: 比较均值	313	
11.5 问题 4. 是因果关系吗?		314
11.6 方差分析: 鸟瞰回顾		314
11.7 配对分析: 每个单元两个观测		315
t -检验	315	
符号检验: 只回答是或否	316	
11.8 小结		317
补充读物		319
公式		319
习题		321

12 两个顺序变量的秩方顺 333

12.1 用词作为值的两个顺序变量		334
问题 1. 身份和兴趣间的关系?	335	
问题 2. 相关的程度?	337	
问题 3. 总体的关系?	338	
问题 4. 是因果关系吗?	339	
12.2 把数目的排序作为值: Phillies 表现如何?		339
问题 1. 数据中的关系?	339	
问题 2. 关系强度?	340	
问题 3. 相关性是由于偶然吗?	341	
问题 4. 是因果关系吗?	341	
12.3 小结		341
补充读物		342
公式		342
习题		345

13 多元分析 357

13.1 偏 ϕ : 三个分类型变量		358
控制第三个变量: 中立策略	359	
偏 ϕ	360	
13.2 数值型变量的多元回归		362
问题 1. 数据中的关系是什么?	362	
问题 2a. 这种关系的形式是什么? 偏回归系数	363	
问题 2b. 这些关系的强度有多大? 偏相关系数	365	
问题 2c. 总体关系的强度有多大? 多重相关系数	365	

问题 3 总体中的关系 ⁹	367	
13.3 用一个哑元作多元回归		368
13.4 双因子方差分析		370
仅对于时段的单因子分析	372	
仅对于路线的单因子分析	372	
时段和路线的双因子分析	373	
考虑交互效应,再进行研究	375	
13.5 建立因果关系		377
13.6 小结		378
补充读物		379
公式		380
习题		381
14 日常生活中的统计		393
14.1 通向统计精妙的基石		394
14.2 小心地处理数据		396
14.3 数据和统计方法		397
14.4 怎么会出错		398
数据收集中的危险	398	
调查研究的特殊问题	399	
分析方法的误用	401	
统计推断的误用	402	
数字的错误解释	402	
14.5 统计和专制		403
14.6 在高潮时结束		404
补充读物		404
习题		405
统计术语		406
统计符号		410
奇统计号练习奇统计案		434
索引		473

统计学：随机性

和规律性



读这本书也许仅仅是因为你觉得知道一些统计知识是非常重要的,同时你可能会猜想学习统计并不是一件愉快的事。我们已经见过太多不情愿的学生认为统计学只不过是一门讨好大众的课程罢了。我们知道你们中的一些人也许更喜欢分析一首诗、唱一支歌或者解剖一只青蛙。但是,不管你们愿不愿意学习统计,我们对你们的这种心态都有足够多的了解。

在你们中间,有一部分人意识到,在某些日常生活中,懂得如何用统计去解决问题是非常重要的;还有一部分人也许期望统计学的挑战是一项智力运动;其他人则可能把统计学看作解决棘手问题的一个手段。我们认为,统计学可激励智力,并且有很多乐趣。但我们的目的却并不是要把你领入统计学专业的密室。正如书名所示,本书仅仅是帮助你理解统计学,熟悉统计语言,并知道如何评估统计结果。如果你想研究统计学,本书只是这条漫长却又乐趣无穷的道路上的一个起点。

为帮助你漫游统计王国,我们每一章都是以一些和本章内容有关的实际问题开始。我们希望这些问题成为你对每套统计大餐的开胃菜。下面我们就以一些问题来开始第一章。

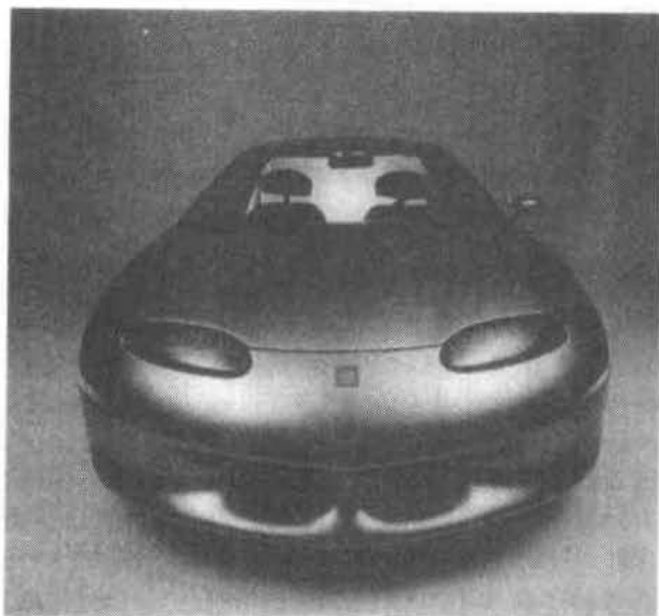
1. 作为一名未来的大学生,你会去查询 *Barron's Profile of American Colleges*^① 在“波士顿大学”这一栏中,你会看到申请者的SAT^② 语言考试的平均成绩是550分,SAT 数学考试的平均成绩是600分,这

^① 美国各大学的简介——译者注。

^② 美国大学申请入学前的一种标准考试,类似于研究生入学时的GRE考试——译者注。

些数字是什么意思?什么是平均成绩?如果你的成绩低于平均成绩,你应该申请波士顿大学吗?显然,为选择合适的学校,你需要具备一些统计知识。

2. 想象你是市场部的新任经理,一次广告活动的统计结果摆到了你面前,声称某个结果是“统计显著”的。你如何解释这份报告而又不暴露你对该术语的无知?赶快学点统计,这对你和你的事业都非常有用。



加利福尼亚州的法律要求汽车制造商们根据他们的总产量生产一定比例的电动车,以减少内燃机汽车造成的空气污染。这已开始成为整个国家的趋势。对于立法者来说,统计信息在使他们相信并检测使用电动车在改进空气质量上的有效性方面起着关键性的作用。

(Peter A. Simon, Phototoke NYC.)

3. 作为一名潜在的汽车购买者和一名有责任心的市民,你乐意为保护地球生态环境作出应有的贡献。根据最新的研究结果,消费者的行为对自然资源到底有怎样的影响呢?你应该买使用柴油发动机的汽车,还是购买电动车,或者干脆骑自行车呢?你是否应该使用喷雾器?你是否应该在你家的草坪上使用化学肥料?报纸、杂志和消费者报告中的统计结果对于你的决定很关键。这些结果到底是要建议你有怎样的消费习惯呢?
4. 当你读报时,你会看到诸如“吃生的酸奶酪可以活到 100 岁”之类的标题。这样的声明有统计支持吗?如果你讨厌生的酸乳酪怎么办?

1.1 统计学:用一句话来说是什么?

Statistics 是一个有很多意义的单词,其中一些定义较为明确。*Statistics* 这个单词源于一个德国人 Hermann Conring, 他首先于 1660 年在印刷品上使用 *Statistik*。这个单词的前半部分

是单词 *State* (德语为 *Staat*) 的变形。在三百多年前,它首次被应用,指政府部门记录人们出生和死亡信息的工作。时至今日,统计仍然是世界上各个层次的政府机构的支柱。全球统计已成为许多国际组织,像跨国公司,联合国和一些关注人口密度、生态灾难和疾病流行的组织等重点关注的对象。

除其来自国家政策的起源外,单词 *Statistics* 还有两个重要意思。首先,统计可以认为是某种形式的数字,例如德克萨斯州的平均降水量、亚利桑那州的周平均气温、波士顿 Red Sox 球队^① 的平均击球率,国际债务的规模,或者巴西的咖啡价格等等。现代社会似乎对统计有一种永不满足的需求。正是基于这种需求,统计学家们收集越来越多的统计数据。当时任人事统计局局长的 Janet Norwood,于 1989 年在美国统计联合会的演讲中曾说:“这是一个运行于数字之上的国家。”

统计数字又叫数据, *Statistics* 的一个简单意思是指数值数据。在本书中,我们超越了这种意思。我们感兴趣的是如何搜集数据和处理数据。最后,我们希望通过统计的帮助把数据中的信息变成实际的知识。

单词 *Statistics* 可以是单数,如“*Statistics is a fun course*(统计学是一门有趣的课程)”;也可以是复数,如“*These statistics indicate an upswing in the economy*(统计表明经济正在起飞)”。但是单词“*data*(数据)”却总是复数。“*datum*”,而非“*data*”,是其单数形式。(现代英语教授们已经对这一点很宽厚了,所以以前说:“*The data is impressive*(这些数据很吸引人),”被认为是犯了很大的错误,而现在却可以说只是一个小过失。)

统计学是用以
(1)收集数据、(2)
分析数据和(3)由
数据得出结论的一
组概念、原则和方
法。

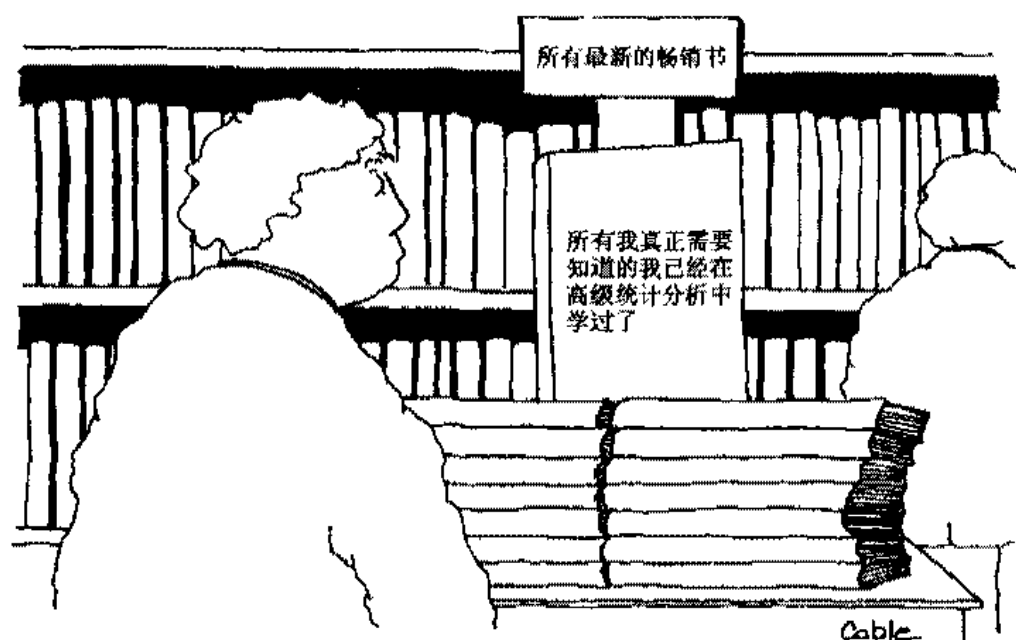
在单数形式的意义里, *Statistics* 可以被指定为一门学科——统计学。例如你正在学习统计学的一门课程,或是你的指导教师已获得了统计学的研究生学位。在统计学中,统计学家们探索、开创各种收集数据和分析数据的方法以得出结论。他们从数学式子中为统计设计出新的应用,并通过实际来检验理论模型。

在本书的最后,我们希望通过统计学的帮助,你能够欣赏到数据是如何被转化为比数字本身更为复杂的知识。不像化学、社会学或心理学,它们都是研究已经定义得很好的某种现象,统计学没有自己的基于实验或观察的经验研究对象。然而,统计学为化学家、社会学家、心理学家以及其他人提供了一套研究对象的方法。

1.2 懂得如何运用统计:读者的目标

由于统计已被应用于如此多的学科之中,统计分析结果总是包围着我们。比如说,学术研究杂志就依赖于统计结果。在许多学科中,一篇文章是否能够发表在主要杂志上,在很大程度上依赖于该文章是否能正确地使用统计方法。

① 波士顿 Red Sox 球队是美国的一支知名棒球队——译者注。



甚至漫画中也充满了统计。(Reprinted with permission of the artist, Carol Cable.)

在学术领域之外,统计也被大量使用着。没有一份报纸、周刊或杂志不刊登以统计为基础的文章。统计更在工业中被大量使用,尤其用于研究新产品、质量控制和市场开发中。统计同样构成了其他印刷媒体的素材。在杂志 *Playboy*, *Cosmopolitan*, *Vanity Fair* 和 *the New Yorker* 中,我们读到诸如有婚外情的人的比例、为慈善事业捐款的人的比例、在百老汇演出业失败的人的比例等统计数据。电视节目中我们看到的节目主持人和出现的广告都依赖于统计,因为只有排名靠前的才能生存。

民意调查要利用统计。在当今时代,很难想象一次不需要民意调查来获知选民对于各种问题及候选人看法的选举。总统的形象设计及党派竞选纲领,都依赖由统计得到的选民的反馈信息。统计能够为获悉将是谁以多大的优势赢得选举提供依据。更为戏剧化的是,统计能够在选举结束之前,甚至在真正的选举发生之前,极为准确地预测选举结果。候选人在所有选票被统计之前已经声称胜利或承认失败的事实说明,人们对统计很有信心。

现在你开始思考统计了,在一定程度上,我们的文化在统计上欠的债也许会在你身上得到补偿。这里有两个例子可以激发你的想象力:

有时飞机票由于卖出的多于实际座位而发生预定过多的现象,其实这并不是令人遗憾的疏忽。航空公司做出如上安排,是极据了每次航班中通常有多少人不能如期来乘机的统计结果。这样,如果航空公司赢了,飞机票将正好被预定满,而如果航空公司输了,就得给出几张免费票。

退休社区^① 根据精细的价格规划来吸引客户。当确定一个退休中心的总费用时,全体居民的预期寿命就成为必须估计的一个重要因素。人们活得越长,花费就越多。统计分析能够帮助经营者们确定一个既有竞争力,又能带来收益的价格。

^① 美国的退休社区相当于我们的养老院,由私人经营,以赢利为目的——译者注。



这些人仍然在主持节目吗?在电视新闻广播力争第一的竞争中,这要依赖于时时进行的调查的统计结果。(UPI/Bettmann; Corbi-Bettmann)

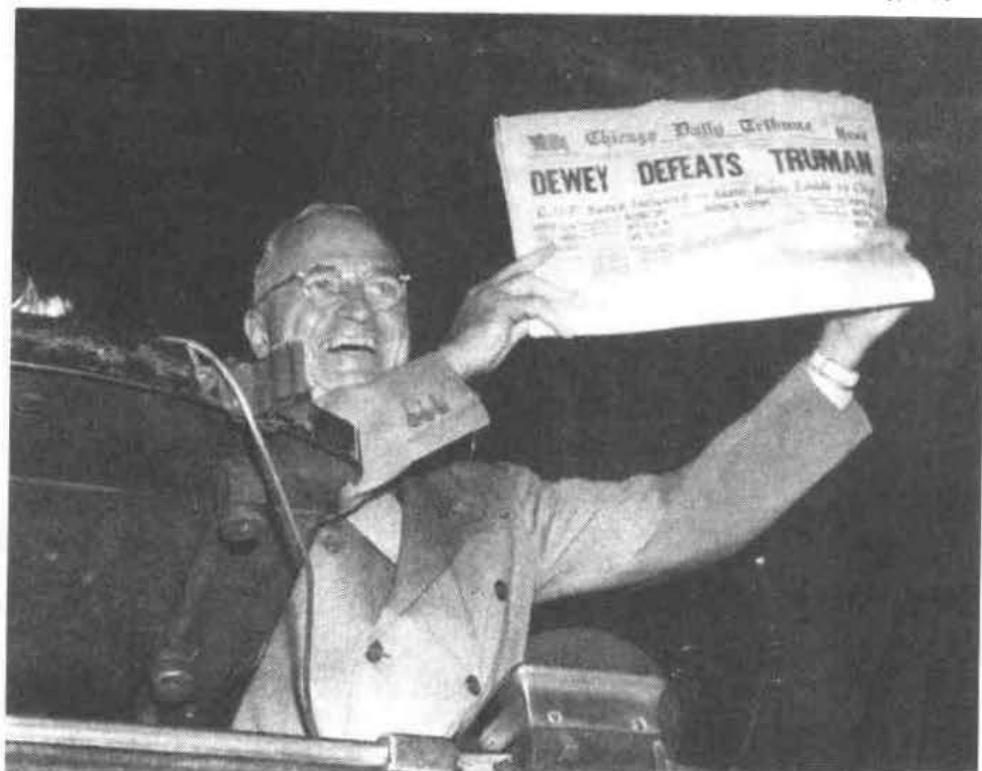
停下来想一想 1.1

标志为“停下来想一想”的习题点缀在各个章节中,主要是鼓励你在统计时加入自己制定的有创造性的“调味剂”。我们仅仅是邀请你在这道菜中加入少许调味剂而已,我们并不包揽所有的例子、应用和细节,而邀请你也加入例子。我们想,如果你肯花上一点儿时间把书中讨论的话题和你自己的实际经验结合起来,你将会更清楚地认识到,你对统计理解了多少,哪里存在问题,你该如何把课堂内的体验和课堂外的生活联系起来等等。这种行为将引起你对统计更持久的学习。

第一个习题:请从你的实际经验中想出一些以统计分析为决策依据的例子,按你的观点,正面的、负面的都可以;并思考出现这种统计结果的原因。或许,最近某电视节目被取消也许会是一个不错的选择。

理解什么可能出差错

作为消费者,如果我们希望能完全理解统计的广泛应用,那么我们就应知道为得到所读到的结果而需要的原则和方法。对统计的了解可以帮助我们评价这些结果,并使我们持批评态度并警觉到,从开始叙述问题到完成最终报告的过程中,随时可能出现一些错误。



预测选举结果并不总是容易的! (UPI/Bettmann)

下面的故事有关一次著名的失败的统计调查,它一直是一个统计传奇。在1936年美国总统选举前,一份名为 *Literary Digest* 的颇受人尊重的杂志曾进行了一次民意调查。调查的焦点当然是谁将成为下一届总统——是挑战者,堪萨斯州州长 Alf Landon,还是现任总统, Franklin Delano Roosevelt。为了了解选民意向,民意调查专家们根据电话簿和车辆登记簿上的名单给一大批人发了简单的调查表(电话和汽车在1936年并不像现在这样普遍,但是这些名单比较容易得到)。尽管发出的调查表大约有一千万张,但收回的比例并不高。在收回的调查表中, Alf Landon 非常受欢迎。于是,该杂志预测 Landon 将赢得选举。

如果读者有一些统计知识,他们会对这个声称 Alf Landon 将赢得选举的预测结果有疑问。正如你所怀疑的,在经济大萧条时期调查拥有电话和汽车的人们,并不能很好地反映全体选民的观点。此外,只有少数的调查表被收回,这一点也是值得怀疑的。事实表明,最终是 Franklin Roosevelt 而不是 Alf Landon 赢得了这次选举。由此可见,那次的调查结果有多么错误了。当前大多数应用统计不会像上一例子错得那样厉害,但即便在今天,我们也很容易发现统计被误用的情况,尤其在需要考虑选择正确的样本时。(来源: Jeffrey Wüner, *DATA Analysis: An Introduction*, Englewood Cliffs, NJ: Prentice Hall, 1992, p. 97.)

理解统计术语

如果我们不理解那些统计术语的话,统计分析结果就不可能给我们提供太多帮助。例如,统计报告中最典型的一个术语是“统计显著”;在给出选民偏爱某一候选人的比例时,可能会有“样本误差等于 $\pm 3\%$ ”或“边际误差等于 $\pm 3\%$ ”等术语。两个变量可能“高度相关”。以上是三种最常见的统计术语。对于知道它们的人来说,这些术语包含了有用的信息。而不知道它们意义的人则有可能根本不知道它们在说什么,甚至得出错误的结论。

1.3 统计学的主要思想

随机性和规律性:关系密切的孪生子

当我们不能预测一件事情的结果时,随机性就和这件事联系起来了。例如,当掷硬币时,我们并不能够确定硬币将正面朝上还是反面朝上。类似地,当我们去旅游时,我们也不能够确定我们是否会发生意外。

同时,当我们把随机的事件放在一起时,它们表现出令人惊奇的规律性。甚至当我们考察掷硬币这样的随机事件时,模式和趋势也变得很明显。如果你将同样的硬币掷 100 次,你知道它将差不多 50 次正面朝上,50 次反面朝上。类似地,尽管某一车祸在几次不大可能的事故中是唯一的,但当你考察所有的事故时,你会发现其中有带有令人不安的规律性。一年又一年过去了,美国差不多每年都有 40000 人死于车祸。尽管某一特定车祸的发生概率很小,但这个数字却令人难以置信地稳定。举一个更“个人”些的例子,每年当 Mary Gergen 在她的《心理学引论》的课堂上调查 100 名学生时,她总发现差不多有 50% 的人在过去一年中遇到过车祸。她发现随机事件的确能够表现出规律性。

通过对看起来随机的现象进行统计分析,我们开始认识这个世界。统计思想的基础知识能够帮助把随机性归纳于可能的规律性中。统计思想从我们如何观察事物和事物本身如何真正发生两方面,帮助我们理解随机性和规律性的重要性。因此,统计可以看做是一项对随机性中的规律性的研究。

规律性中的随机性

然而甚至规律也表现出某种随机性。如果你再掷 100 次硬币,正面朝上的次数几乎不会和前 100 次完全一样。在第一个 100 次中,也许有 48 次硬币的正面朝上,然而在第二个 100 次中,也许就有 53 次正面朝上。这表明了统计的一个重要的本质特征。

不管我们是否再进行一次或一组新的观察,大部分时候我们并不能够得到和上次观察一模一样的结果。

这种偏差不仅仅发生于掷硬币时,而且发生于调查、实验和其它任何一种方式的数据收集。如果在某次调查中,人们被问到他们如何看待当今的某一重要问题,某一比例的人会有某一特定的观点。如果对不同的人再做同样的调查,则有不同于上述比例的人支持这一观点。

这两个比例之间的差异主要是由于数据本身的随机性引起的。在这种意义下来说,统计就成了对数据中的偏差问题的研究。

根据作为统计基础的数学理论,我们可以确定一项调查中的某一比例有多大的随机性,以及在下一次的重复调查中,这个比例可能有多大的偏差。我们甚至可以指出,这两个比例之间的差异,是否大到了随机性本身所不能解释的地步。我们将在以后章节中引申和详细讨论这些思想。

在规律性中,变化趋势时有出现。比如,随着逐渐增加的汽车安全带的使用和保险气袋^①的安装,发生车祸的比例正在下降。统计把单独的、随机的事件置于规律性中,并揭示其变化趋势。如果在不同时期,交通事故发生次数的(两种规律性的)差异超出了随机性本身可以解释的地步,那么变化趋势就发生了。

研究随机性和规律性时的两个例子

作为一个说明两个数字之间的差异是否不能仅归因于随机性的例子,让我们回顾一下20世纪50年代小儿麻痹症疫苗的投入使用。小儿麻痹症是一种可怕的疾病,通常能使患者(大部分是儿童)瘫痪或死亡。在这种病经过多年流行之后,一种疫苗最终被研制出来。科学家们希望该疫苗能够预防这种可怕的疾病,但是没有人清楚这种疫苗是否真能像人们期待的那样起作用。尽管实验室和动物实验的结果很使人兴奋,然而唯一检验这种疫苗是否起作用的方法还是人体实验。因为小儿麻痹症是一种较罕见的疾病,疫苗必须试用于相当一大批孩子的身上,所以研究者们决定在200000个孩子身上做实验。此外,研究者们还决定用另外相同数目的孩子作为对照组。对照组的孩子仅仅得到安慰剂——一种看起来像疫苗的替代品——为观察疫苗是否真的起作用。

当孩子们被注射了疫苗或安慰剂以后,研究者们开始在下一个“小儿麻痹症季节”^②中观察实验结果。在对照组中,有138个孩子感染了此病。这个数字当然有一定的随机性,研究者们并不能确定它意味着什么。如果另外一组的200000个孩子也被注射安慰剂,那么不一定会同样多的孩子感染此疾病。根据随机性的大小,可能有130或140或其它数目的孩子们染上小儿麻痹症。

在被注射了疫苗的那一组中,有56个孩子患了小儿麻痹症,这个数字当然也有随机性。一个重要的问题是,56和138的差别是否超过了随机性所能解释的程度。如果是的话,那么研究者们就能够有把握说,疫苗起作用了。利用后面第七章介绍的方法,我们可以看到,138和56的差别超出了随机性本身所能解释的范围,因此疫苗被宣布为是成功的。从此以后,这种疫苗在许多国家根除了小儿麻痹症。全世界的健康组织所做的进一步的努力,将使不发达国家的孩子们,在不远的将来,也有可能不再遭受小儿麻痹症所带来的痛苦。在某种重要的意义上说,统计推理为发展和检验疫苗的研究者们提供了有力的支持。

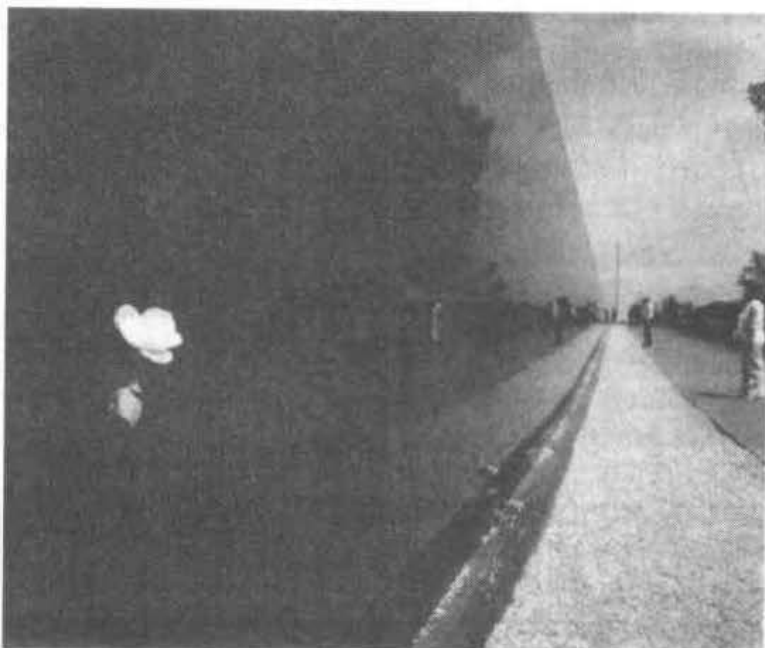
另外一个著名的随机性的例子——或者说缺乏随机性,正如这个特定的例子中的情况——发生于军事中。在美国对越南的战争中,为使前线有足够多的士兵,美国政府制定了一个“抓阄”的征兵计划。该计划打算把1到366的号码随机地分配给一年中的每一天,然后由军事部门按分配的号码顺序把生日与之对应的年轻人分批征召入伍。这种方法的目的是为了给

^① 保险气袋,又称气囊,是一种在汽车受冲撞时自动充气以免驾驶者撞伤的安全装置——译者注。

^② 指小儿麻痹症流行的季节——译者注。

大家相等的机会卷入这场不受欢迎的战争中,因为被征召的可能性应该是随机决定的。

在第一年的征兵计划中,号码1被分配给了9月14日,分配方法是随机抽取一个大容器中的366个写上了日子的乒乓球。结果所有年满18岁且生于9月14日的合格青年将作为第一批被征召入伍。生日被分配为号码2的青年则在第二批被征召入伍,以此类推。我们知道,并不是所有人都被征召入伍,因此,生日被分配的号码较大的人也许永远轮不上到军队服役。



由抓阄决定的越战死难者。(FPG International.)

这种抓阄看起来对决定应该谁被征召入伍是一个相当不错的方法。然而,在抓阄的第二天,当所有的日子和它们对应的号码公布以后,统计学家们开始研究这些数据。经过观察和计算,统计学家们发现了一些规律。例如,我们本应预期应当有差不多一半的较小的号码(1到183)被分配给前半年的日子,即从1月份到6月份;另外一半较小的号码被分配给后半年的日子,从7月份到12月份。由于抓阄的随机性,前半年中可能不会分到正好一半较小的号码,但是应当接近一半。然而结果是,有73个较小的号码被分配给了前半年的日子,同时有110个较小的号码被分配给了后半年的日子。换句话说,如果你生于后半年的某一天,那么,你因为被分配给一个较小号码而去服兵役的机会,要大于生于前半年的

人。在这种情况下,两个数字之间只应该有随机误差,而73和110之间的差别超出了随机性所能解释的范围。这种非随机性是由于乒乓球在被抽取之前没有被充分搅拌造成的。在第二年,主管这件事的部门在抓阄之前先去咨询了统计学家(这可能会使生于后半年的人感觉稍微舒服些)。

概率:什么是机会

概率是一个取值于0和1之间的数,告诉我们某一特定的事件以多大的机会发生。

在讨论随机性时说,统计学的大部分内容根基于一个很重要的概念**概率(probability)**。概率为统计学的第三个方面,即如何从数据中得出结论,奠定了基石。我们可能永远不能十分确定,两个数字的差别是否超出了随机性本身所预期的范围,但是我们可以确定,这种差别发生的概率是大还是小。根据这个基本的思想,我们可以有很多有趣的机会,得出关于我们所处的这个世界的重要结论。至于具体做法,我们将在第五章及其后章节中详细阐述。

变量:我们给事物所起的名字

变量是指一个可以取两个或更多个可能值的特征、特质或属性。

统计的又一块较大的基石是变量(variable)的概念。人类的性别特征是取两个值的变量,因为一个人只可能是男性或女性。宗教信仰是一个变量,在西方国家中可能取值为天主教、犹太教、伊斯兰教、新教及其它,在印度则可能取值为印度教、伊斯兰教、佛教、锡克教及其它。还有其它变量的例子,如汽车每加仑汽油所能行驶的英里数,取值范围从8到50;孩子们以公斤为单位的重量,取值是从10到70;还有一剂药的药量,等等。通常,研究者在项目开始时,就要确定他们感兴趣的变量及其取值范围。我们可以把变量的取值想象为散布在一条直线上的点,而直线本身代表了这个变量(参见图1.1)。

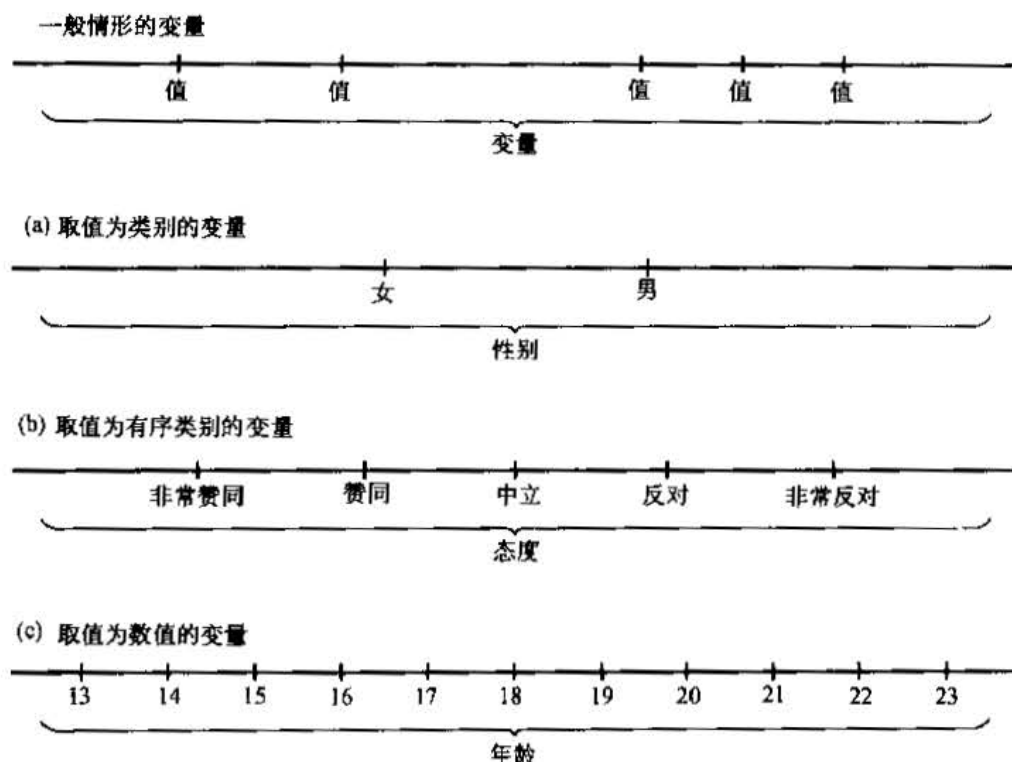


图 1.1 变量及其取值。

变量、值和个体

变量的值(value)通常是对某一特定单元的度量,这种单元常常被视为一个个体(element)。一个个体可能是一个人,一群人,一块土地,一种植物,一种动物或一个国家,只要此个体是合适的,对于使用者来说是显然的,并且在分析过程中不会变化即可。表1.1列出了一些变量、变量的取值及其所测量的个体的例子。因此,性别变量是以人为个体的观测,孩子的数目是以家庭为个体的观测。在家庭的例子中,个体是单个家庭成员的集合。

表 1.1 变量, 值, 和个体

变量	变量的值	个体
性别	男, 女	人
态度	反对, 中立, 赞成	人
失业	有工作, 无工作	人
失业	0.0%, ..., 4.6%, ...	县
玉米产量	..., 5678 lb, ...	英亩
孩子数	0, 1, 2, 3, ...	家庭
贫困程度	严重, 一般, 边缘, 没有	地区
比赛名次	第一, 第二, 第三, ...	参赛队

理论变量和经验变量

到目前为止, 我们所讨论的变量都和我们所熟悉的日常生活有关。我们称这种变量为**经验变量**(empirical variables), 因为它们处理的对象是我们周围可观测到的物质世界中的事物。除了经验变量, 我们还使用由统计学家创造出来的变量。这些可以用数学方法推导的变量称为**理论变量**(theoretical variables)。我们将在以后章节中介绍几个理论变量的例子, 其中四个为 z -, t -, χ^2 - (发音为卡方) 和 F -变量。

常数

一个常数总是有一个固定的价值。

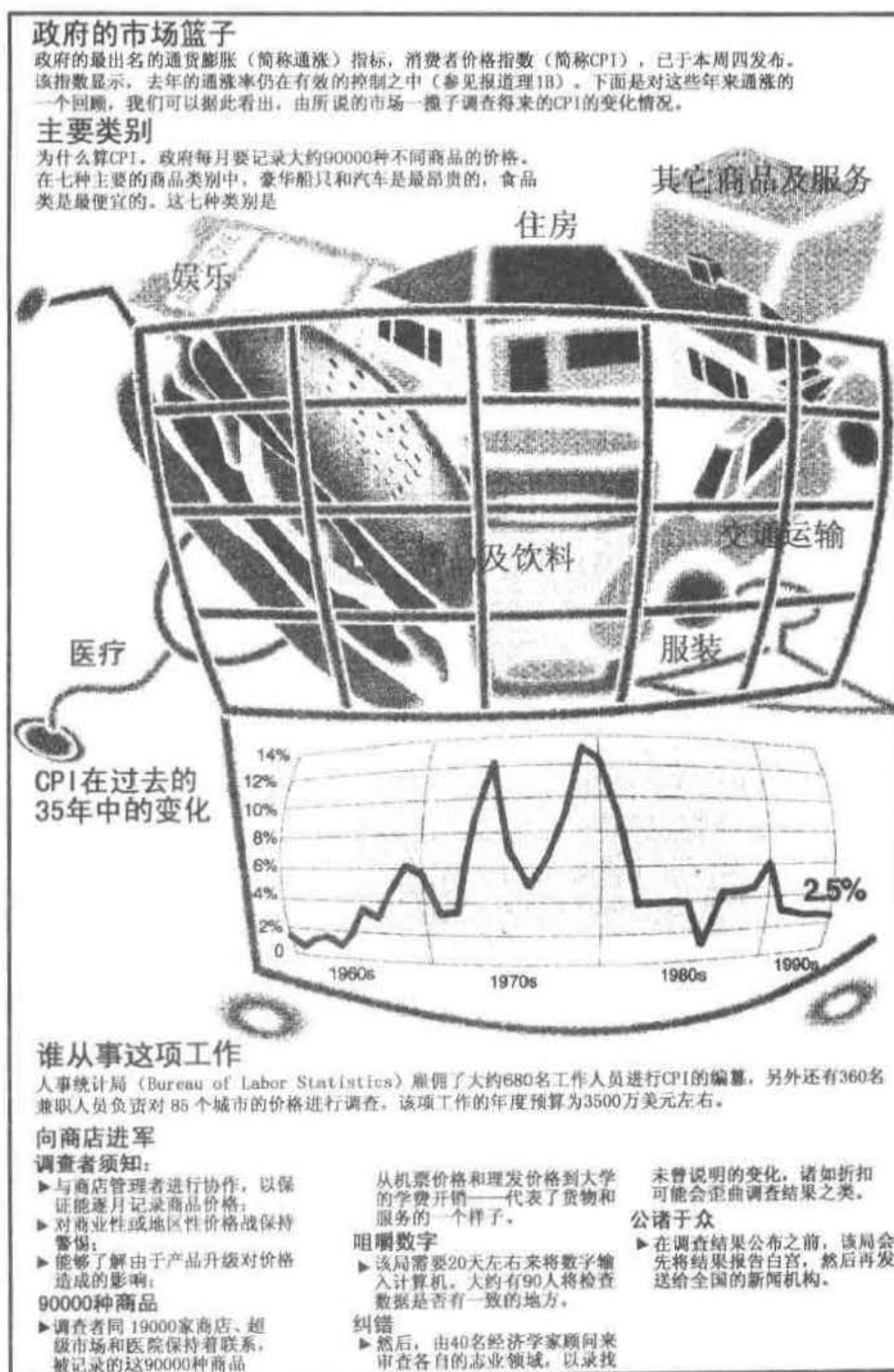
常数(constant)是变量的反义词。假设我们对统计课上的所有学生进行调查, 看有多大比例的人认为统计是有趣的。假定没有人改变主意, 则当我们重复这个小调查时, 我们将仍然得到这个比例数, 像这样的一个比例数就是常数——当实验重复时它不变。显然, 它之所以不改变是因为每次我们都是调查课堂上的所有学生。在统计中, 我们使用一种被称做**参数**(parameter)(见第 7 章及其后章节)的常数。

1.4 统计的使用者

让我们看一看某些受统计影响的领域: 政府机关、自然科学、医药、工业甚至法律。在美国, 通过人口调查局和其它的联邦统计机构像人事统计局等, 联邦政府成了最大的数据收集者和统计使用者。尽管这几年联邦统计系统的预算经费被大大缩减了, 但它始终以其杰出的表现而闻名。

联邦统计系统的两个重要活动是计算消费者价格指数和失业人数。这些结果每个月公布一次, 并在国家的经济生活中扮演了重要角色。消费者价格指数可以追溯到 20 世纪初。许多活动, 像劳动合同和社会保险支付, 在任何时候都和此数有关。失业人数这个概念形成于 20 世纪 30 年代的大萧条时期, 当时新政^① 改革家们意识到了调查美国有多少人失业的重要

^① 指美国总统罗斯福在 20 世纪 30 年代所实施的內政纲领的名称——译者注。



通货膨胀很难测量。(©1996, USA Today. Reprinted with permission.)

性。这两种报告结果，都是根据复杂的统计原理进行了大样本的统计调查之后得出的。

政府机关收集的大部分数据，都用于帮助制定针对各种问题的政策。例如，为决定税收政策，必须知道现行的税法如何影响各种收入水平的人们，并需要预测出税法变化后的影响。为

使社会福利计划能够成功,必须知道对这种计划产生需要的社会条件及这种计划如何影响与之息息相关的人们。当专家们试图确定,为学龄前儿童实施的启蒙教育计划(称为 Head Start),对参加者是否有长期的好处时,计划本身就成了需要被认真考察的对像。同样,推行一个农业补贴计划,也必须事先知道当前农业产量的情况,并须预期此计划对将来产量的影响。

各个学术领域的人们在他们的科研中都使用统计。生物学、经济学和心理学三个学科对统计的使用是如此之多,以至于它们已经发展了自己的一套统计方法,即生物统计、计量经济和心理测验。在文科方面,一大批历史学家、地理学家、语言学家和古典学者利用统计知识得出各种结论,诸如由于中世纪大鼠疫而导致的死亡人数、法语在英语国家中的普及程度等等。这表明,几乎所有的经验学术研究——报告、学术会议上的演说、杂志、文章和书——都以这种或那种方式依赖于统计。学术研究在许多方面丰富了社会生活,而统计则在此过程中扮演了不可替代的角色。没有任何一门其它学科,能对如此多的科学领域作出贡献。

一个来源于法律方面的生动的例子可以说明,统计在社会生活中日益发挥出重要的作用。当面对的除了法律问题还有统计问题时,许多律师都发觉,自己已经进入了一个新的领域。一个需要用到统计的主要领域是在共同诉讼^①中,此时需要考虑因年龄、性别、种族而造成的差异。律师必须使法官和陪审团们相信,在任何一种原告代表的选择中,年龄、性别、种族的差异是设计好的还是随机的。统计学家则需要作为专家证人向法官和陪审团解释诸如“置信区间”和“显著水平”等问题。没有统计学家的专家证明,是不可能在法庭上作出公正、合理的判决的。

DNA 检验——在这个双螺旋结构上还悬着一个故事呢

在 1995 年结束的著名的辛普森(O. J. Simpson)谋杀案的审理中,许多证词都涉及 DNA 样本及它们如何被收集、分析和确认。从证人从各种层面上收集到的血液样本证据的统计数据中,公众了解到了很多统计知识。问题的关键在于,收集的 DNA 样本有多大的可能性与受害者或被告人的血液相吻合。原告声称,血液样本不是辛普森的概率至少是非常小的,但被告律师反对该结论。

通常情况下,DNA 检验的过程是检查 DNA 链各种指标的模式并计算两个人都有这同一种模式的可能性。一旦这种方法成为可行的,公众很快开始各类案件中欢迎它的使用。在另一个审判中,一名男子因犯强奸罪已经坐了七年牢,但当他的律师证明他的 DNA 和真正的强奸者的 DNA 根本不匹配时,这名囚犯终于被释放了。

引进新的评价治疗效果的统计方法,已经改变了整个医药领域。例如,在健康保护组织提出并监督的减少医药花费的政策中,医生要想得到补偿,就必须认真遵守该组织的用药指南。而这些指南,正是通过对大量医学实践及结果进行认真的统计分析之后才发展起来的。如果有两种新药的药效相同,则健康保护组织将不对其中较贵的新药进行补偿。使用统计方法以支持保险公司的用药指南,已经产生了医疗主张上令人关注的争论,即单个病人被考虑得太少。例如,尽管新生儿及其母亲中 95% 都没有大问题,但只允许她们在医院住 24 小时的政策被新泽西州(New Jersey)的立法给推翻了。不仅统计方法本身需要认真考察,“可接受的风险”的定义也是有问题的。

^① 指一名或数名原告代表多数有共同利益关系的人提出诉讼。这因此涉及到选哪些人作为代表,以更能体现原告们的共同利益——译者注。

大公司也是统计的大量使用者。例如,为使一种新药被联邦医药总署批准,医药公司必须证明这种药是安全的。公司投入大量资金在动物和人体上做实验,以检验新药的功效。结果,这些公司雇佣了一大批统计学家。他们负责正确安排实验,分析实验结果数据,并检查此药市场化的价值,以避免医药发展中的法律纠纷及昂贵的支出。

工业中许多行业使用统计进行质量控制。从生产线出来的产品并不都是一样的,其原因一部分是由于随机误差;一部分由于在生产过程中有些地方可能会出错。通过统计方法可以研究这种差异并帮助人们精确地指出什么和哪里出错了。好的质量控制系统可保证消费者对他们所购买的产品满意,并保证他们在下次不买竞争厂家的产品。美国统计学家 Edward Deming 是统计方法中质量控制论发展的先驱。然而,出乎意料的是,他的许多方法是在美国以外的其它地方,尤其是日本的公司被首先采用的。日本工业在二战后迅速发展的其中一个原因,就是日本的公司领导人早早地采用了 Deming 的方法。(来源:W. Edward Deming, *Out of the Crises*, Cambridge, MA: MIT Center for Advanced Engineering Study, 1986.)

1.5 统计学和数学、铅笔及计算机的关系

统计学的基础是数学。虽然今天,许多著名大学的统计系独立地培养统计学家,然而刚开始,统计只是数学系的一部分。统计推断牢牢地建立于数学基础之上,结果,我们很容易发现一些统计课本看起来像充满了定理和证明的数学课本。但是,没有数学知识也是有可能学会统计的,这正是本书的讲述方式。目前,大多数统计分析都在计算机上处理,因此,理解计算机的输入和输出的内容比知道计算机软件如何计算重要得多。

这里我们所强调的,正如前边所说,是学会基本的统计思想——某些专业术语,数据如何被收集、演示、分析,结果意味着什么,及它们何时该或不该应用于实际生活——而不至于深陷于公式和计算细节的泥潭中。对今天的大部分人来说,对统计思想的理解对于成为一名有知识、有能力的市民是至关重要的;能够自己做出色的统计分析是高专业化职业的一部分。

另外,还有其它原因使我们进行统计分析,如为了纯粹的乐趣或为了满足对于技巧的嗜好。对于那些希望用传统的纸和笔或更先进的计算机来做统计分析的人来说,每章后面的习题提供了很多机会。习题分为三部分,第一部分检查你的基本概念,第二部分检查你解释数据和把统计结果应用于日常生活的能力,第三部分则要求你能够运用公式和数学技巧。各种统计计算的公式全部放到每章末尾。理论变量的统计数表则可在本书末尾找到。

1.6 小 结

1.1 统计学:用一句话来说是什么?

统计起初是作为一个和政府有关的词被提出来的。后来,统计开始指单个数据点的集合。作为一个研究领域,统计学可以被定义为一组由(1)收集数据(2)分析数据(3)由数据得出结论而组成的概念、原则和方法。

1.2 懂得如何应用统计:读者的目标

统计没有自己的研究对象,而是研究来自各个领域的数据。由于当今社会中统计的流行和魅力,我们无法回避这些分析结果。

1.3 统计的主要思想

随机性和规律性是统计的两个重要概念。随机性是指不能够预测某一特定事件的结果。规律性是指我们从许多事件中收集数据时发现的模式。规律性本身包含随机性。统计可被定义为在随机性中寻找规律性。当两种规律之间的差异超出了随机性本身的影响时,变化趋势就发生了。

概率为我们从数据中得出结论提供了基础。概率是一个0到1之间的数,它告诉我们某一事件发生的机会有多大。统计学家们利用概率判断数据间的差别是否超出了随机性本身的影响。

变量可定义为一个特征或属性,例如一个人的年龄,可取两个或多个可能的值(例如,0到100多岁)。变量的值总是用于描述某一特定个体,如一个人、一群人、一块地、一种植物、一种动物或一个国家。许多变量是我们日常生活中所熟悉的,这些变量叫做经验变量。另外,我们还使用统计学家们创造出来的一些变量,即理论变量,它们可用数学公式推导出来。其中的四个理论变量是 z -、 t -、 χ^2 -、和 F -变量。

变量的反义词是常数。常数是不可变的数值。统计中,某种常数也称为参数。

1.4 统计的使用者

统计方法对于政府部门在政策的形成和评估上非常重要。它们对所有科学领域的发展也是必要的。统计方法在很多专业领域,如法律、医学及各类商业领域,也不断赢得地盘。

1.5 统计和数学、铅笔及计算机的关系

统计的基础是数学,但本书的主旨是使你熟悉基本的统计思想,而不是使你变成统计分析专家。

补充读物

Gani, J. (ed.). *The Making of Statisticians*. New York: Springer-Verlag, 1982. 十六个统计学家谈他们为什么和怎样成为统计学家的。

Gonick, Larry, and Woolcott Smith. *The Cartoon Guide to Statistics*. New York: HarperPerennials, 1993. 如果通常的统计教科书不能引起兴趣的话。

Huff, Darrell. *How to Lie with Statistics*. New York: W. W. Norton, 1954. 一个经典的关于可能误用统计的书。

Peters, William S. *Counting of Something: Statistical Principles and Personalities*. New York: Springer-Verlag, 1987. 通过历史的上下文来教统计。

Tufte, Edward R. *Data Analysis for Politics and Policy*. Englewood Cliffs, NJ: Prentice-Hall, 1974. Chapters 1 and 2 five food introductions to various statistical issues. 其第一章和第二章对各种统计问题给出了很好的介绍。

习 题

回顾(习题 1.1 - 1.14)

- 1.1 为什么 *statistics* 的词根来源于单词 *state*?
- 1.2 为什么统计可以被称为“协助者”科学?
- 1.3 a. 定义随机性。
b. 定义规律性。
c. 在统计研究中, 随机性和规律性扮演着什么样的角色?
- 1.4 给出三个日常生活中的例子, 说明随机性中包含规律性。
- 1.5 概率如何帮助研究者确定两组数据之间的差异超过了随机波动?
- 1.6 定义“变化趋势”, 并给出包含它的例子。
- 1.7 a. 定义术语“变量”。
b. 经验变量和理论变量之间有何差别?
c. 本章中提到的四种理论变量是什么?
- 1.8 假如你想研究世界上的飓风, 指出五个你可能会用到的经验变量。
- 1.9 给习题 1.8 中的每一个变量赋一些值。
- 1.10 a. 哪句话是正确的, “The data is interesting” 还是 “The data are interesting”?
b. 解释你的选择。
- 1.11 常数和变量有何不同?
- 1.12 说明法律和医学实践如何受统计影响。
- 1.13 举出两个对国民经济起关键作用的联邦统计系统。
- 1.14 说明统计分析怎样导致更好的生产过程。

解释(习题 1.15 - 1.18)

- 1.15 讨论如下问题: 一个住在 Cape Hatteras 的房主在一次飓风中被掀掉了房顶, 这是否属于随机事件? 然而看起来飓风袭击 Cape Hatteras 却有一定的规律性。一次飓风中的财产损失, 例如被掀掉房顶, 是如何与随机性和规律性相联系的?
- 1.16 指出本章中所描述的一个生活领域, 其中统计分析恰好影响了你的生活, 并简要描述之。
- 1.17 当今的体育报道大量使用计算机产生的统计数据, 内容从每一个专业网球运动员在每一赛季的收入, 到棒球历史上的某一队员完成的三垒打的总次数。
a. 你认为为什么近几年的电视体育节目充满了统计?
b. 你认为统计如何影响了观众对体育运动的喜好?
c. 你认为文化的其它方面, 像音乐、电影、政治、业余体育运动已经或者将要像专业体育运动一样被统计数据所“入侵”吗? 举例论证你的观点。

- 1.18 多数统计学家的目的是从实际数据得到比数据本身更复杂的信息;借此来评论下面的论述:“我们的样本就像我们不会进去的洞穴入口处的影子”。

分析(习题 1.19—1.23)

- 1.19 本题的目的是要描述变化和随机性。闭上你的眼睛,随便翻开一页书,任意指一个地方,选择最接近你手指的完整的一句话。
- 这句话中有多少单词?
 - 同样选择另外一句话,数一数其中的单词数。
 - 为什么这两句话中的单词数不相同?
 - 如果班上的每位同学都随机选择两句话并计算其长度,那么你就可以估计本书的句子的平均长度,比较此长度和莎士比亚的《哈姆雷特》中的句子长度。
- 1.20 拧松水龙头刚好到只有水滴下来,计算并记录 5 分钟内每个 20 秒里的水滴数。利用你的数据,你如何描述水滴,即什么方面是随机的,什么方面是有规律的?

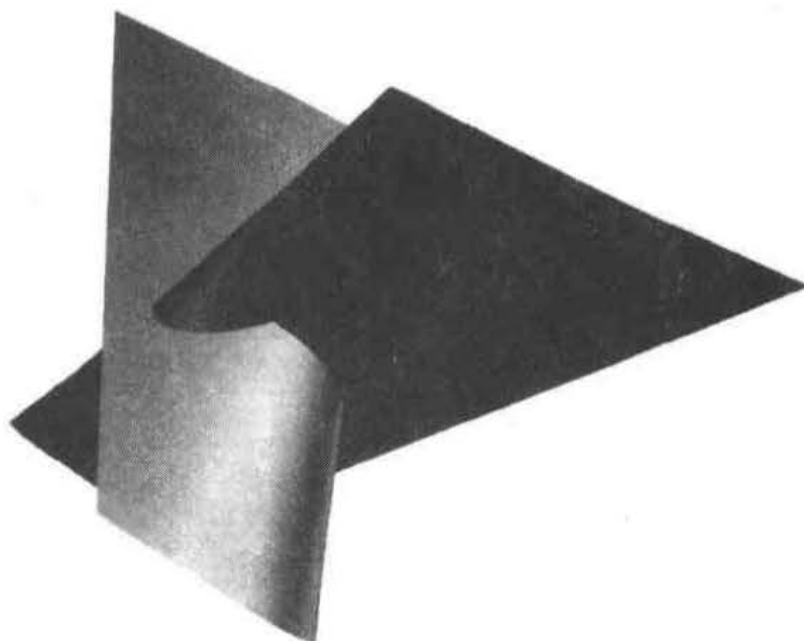
表 1.2 1915—1945 年新生儿和母亲的死亡率*(习题 1.21)

年数	新生儿死亡数		母亲死亡数	
	白人	非白人	白人	非白人
1915	98.6	181.2	6.0	10.6
1920	82.1	131.7	7.6	12.8
1925	68.3	110.8	6.0	11.6
1930	60.1	99.9	6.1	11.7
1935	51.9	83.2	5.3	9.5
1940	43.2	73.8	3.2	7.7
1945	35.6	57.0	1.7	4.5

* 死亡率是指每 1000 名婴儿和母亲的第一年的死亡总数。来源:Data compiled by U.S. Bureau of the Census.

- 1.21 根据表 1.2, 做下面的习题:
- 观察 30 年来新生儿死亡率的趋势,你能得出两个什么样的主要结论?
 - 观察此段时间内成人死亡率的趋势,你又能得出两个什么样的主要结论?
 - 哪组数据看起来更易于描述,新生儿死亡率还是成人死亡率?
 - 关于新生儿死亡率和种族的关系,从这些数据可得出什么结论?
 - 如果这些数据的收集有失准确性,那么可能存在什么样的问题?哪组数据看起来更不准确?
- 1.22 找一篇包含统计信息的报纸或新闻杂志上的文章,
- 指出文章中使用的变量;
 - 确定每一个变量的取值;
 - 什么样的读者会对这篇文章特别感兴趣?
 - 此文描述了某种变化吗?
- 1.23
- 联系 1.22 中你选择的文章,如果变量以不同的形式报道出来,文章是否会更加准确、有趣或者有价值?
 - 你认为根据本章提供的材料,有方法使那篇文章改善吗?

C H A P T E R 2



2.1 定义变量

2.2 观测数据:问题和可能性

2.3 收集观测数据时的错误和误差

2.4 实验数据:寻找造成结果的原因

2.5 数据阵/数据文件

2.6 小结

数据的收集



去年在洛杉矶有多少人被性伙伴传染了艾滋病？去年在纽约市有多少垃圾被回收了？是什么导致 19 世纪的水手在长期的海上航行中患了坏血病？班级的大小影响学生们的成绩吗？总统干得好吗？

为回答这些以及其它的许多问题，必须收集相关的信息。在这些例子中，我们需要知道很多事情，从性生活习惯到垃圾回收的实践。一眼看去，好像得到这些信息很容易，只需走出去询问一下或做一做实验即可。但是，现在开始有问题了：谁去问——是你，我，无工作的大学生，还是退休职员？应该问谁？我们有足够的钱去问涉及到该问题的每一个人吗？对第一个例子来说，那可是要问遍洛杉矶所有的人！然而，如果不是问所有的人，可以只问某周六下午路过购物中心的人吗？或者只问在某个网球馆买啤酒的人？或者你可以想一个相对来说更好的方法？

一旦这些问题被解决之后，就要考虑该问些什么了。某些问题至少来说是比较“微妙”的。如果我们直接问某人有多少性伙伴，我们能得到直接的答案吗？我们应该期望有答案吗？我们是否根本不该问？有多少人只会告诉我们一些他们认为我们想听的或使他们面子好看一些的答案呢？如果发问人是医务工作者，或警察，或垃圾工，或登记赌注的人，答案会有差别吗？“干得好”又是什么意思？每一个问题都该有一个深思熟虑的答案，尽管没有一个答案看起来是完全正确的。

一位睿智的统计学家说过，世上有两种数据：好数据和坏数据。当然还有其它划分数据类型的方法，不过这样开始也挺好。好数据是指根据合理、正确的统计原理收集到的数据。坏数据是指通过其它方法收集到的数据。本章将介绍一些统计学家和另外一些人提出的提高数据质量的方法。



数据依赖于很多因素。(“Sally Forth” reprinted with special permission of King Features Syndicate.)

2.1 定义变量

数据有各种各样的类型,是通过各种各样的方法收集到的(此时,笔者正在品尝某一市场调查公司提供的速溶咖啡的样本)。在最普遍的情形中,收集数据也包括测量数据。例如,研究者问人们有关他们的睡觉习惯,计算赌场收入,称出有多少垃圾被回收,给某一植物一定量的水并测量它长了多少。研究者们用各种各样的方法称重、测量、询问和计算他们的研究对象。

数据收集的第一个准则是要清楚测量的是什么。换句话说,变量必须有一个仔细斟酌过的定义。这有时是听起来容易做起来难。



这个家庭中有多少个孩子?(Bruce Coleman, Inc.)

假定我们对家庭生活感兴趣,在一次调查中问如下问题:“在这个家庭中有多少个孩子?”我们也许认为我们自己知道想要了解的是什么,但是,没有理由期待回答问题的人(一般称为

响应者)和我们有同样的想法。我们也许不加考虑地认为,“孩子”应定义为一个不满 18 岁并和他(她)的亲生父母一起生活的人。但如果一个家庭中包含大于 18 岁的亲生子女、前妻或前夫的孩子、养子或养女、过继子女或者其他年轻的亲戚,那该怎么办呢?对于不和亲生父母生活在一起的孩子怎么算呢?对于父母离了婚而共同抚养的孩子怎么算呢?有很多种发生混淆的可能。如果作为研究者,我们对这些问题考虑得不全面,那么就没有理由指望响应者能把它们区分开。如果我们的思想模糊,响应者的回答又不一致,那么我们的数据在含义上将会是极端不平衡的。教训就是,在我们做研究之前,对变量必须要有一个清晰、详尽的定义。在上面的例子中,就需要我们首先明确“孩子”的定义。

2.2 观测数据:问题和可能性

观测数据是指仅凭观测而非通过操作或控制事物所得到的数据。

数据收集有两种主要方法,其中一种是当我们观测现实世界时收集到的数据,如在不同城市回收的铝罐的平均重量。当我们只是观察周围的世界时,就会产生观测数据。收集观测数据的研究者们试图不干涉研究对象的行为模式。数一下洛杉矶有多少人被确诊为艾滋病病毒^①携带者就是一个收集观测数据的例子;将某一政治性调查的结果列成表则是另一个例子。

观测研究(observational study)的内容是多种多样的。它们考察地方组织和企业的运作,人类和动物在正常状态下的行为,可在图书馆找到的历史证据,互联网上的相互作用,以及生理学的、心理学的、社会学的,或是环境的数据,例如在血液抽样、“墨渍”测验^②、股票市场价格指示器、质量控制研究、一氧化碳污染读数等级,或其它任何一种你可想到的现象的测量结果!在所有观测研究中,统计在如何收集数据和如何分析数据这两个方面扮演了重要角色。

总体相对样本

收集数据是为了从收集的个体中得出结论。社会学家们收集有关人的数据以了解人类行为;植物学家们收集有关植物的数据以了解它们如何生长;工程师们收集有关滚珠轴承的数据以确定它们的尺寸是否适用于他们制造的发动机。所有我们感兴趣的个体就组成了总体(population)。2000 年 1 月 1 日加拿大的所有居民是一个总体的例子;除夕之夜,纽约时代广场上的所有的香槟酒瓶塞是另一个例子。

有时我们能够收集到总体中所有个体的数据。在这种情况下,我们就是对总体做了普查(census),如美国每十年对所有居民进行人口调查。然而,在苛刻的现实生活中,由于资金、时间有限以及不断变化的环境条件,做普查通常是不可能的。这时,我们需要把收集数据限制在总体的一个样本(sample)上。

① 应为 HIV 病毒,艾滋病是由 HIV 病毒感染而导致的疾病——译者注。

② 广泛用于精神病临床诊断的一种方法,由瑞士精神病专家 Rorschach 发明——译者注。

总体 包含所有需研究的个体。

普查 是指基于收集整个总体的数据的过程。

样本 是总体的一个被选中的部分。

下面让我们看一下如何选择样本,什么导致一个样本好或坏,以及为什么一个好的样本比一次平庸的普查更好。

样本的选择:确信锅里的汤被搅拌均匀

统计研究者所面临的一个关键问题是如何选择样本。一个研究者希望确认,由研究样本而得出的结论能够适用于该样本所属的较大的总体。然而,没有一个“好”的样本,这是不可能实现的。用烹调作例子,可帮助我们理解为什么一个好的样本如此重要。当我们品尝一勺我们做的汤时,我们关心的不是这勺汤怎样,而是整个锅里的汤味道如何。如果锅里的汤被充分搅拌了,我们只需品尝一勺即可知道整锅汤的味道。我们品尝的这一勺汤无论是来自家庭厨房中的一个小锅,还是来自汤羹工厂的一个大锅,我们都可以窥一斑而知全豹。这正如我们从总体中选择一个样本——从某种意义上来说,选择一个来自“搅拌均匀”的总体的样本。如果总体能被搅拌均匀,那么一个包含 1000 个个体的样本,不管它是以整个国家为总体,还是以—一个城镇或乡村为总体,都可以告诉我们同样多的内容。

停下来 想一想 2.1

收集数据有许多方法。你还能想到其它方法吗?



Mitch Reardon, Tony Stone Images.

我们可以把这勺汤的例子应用于样本调查。某一选举前的民意测验表明,有 57% 的人喜欢某一候选人。如果样本选择正确,这个比例将和整个选民中的比例大致相同。类似地,在质量控制研究中,检查灯泡的某个样本的目的,不是要看这些抽查到的灯泡能否达到预期的寿命,而是要看生产的所有灯泡组成的总体能否正常工作。选择的这个样本应该能够很好地反映总体,因此也能够很好地反映生产过程本身。

如果不能正确地选择样本,那么对于“整锅汤”的判断可能导致错误的结论。如果民意调查专家们只对他们的家人和朋友提问,那么将可能产生坏样本。如果检查员只查标有“易碎”字样箱子中的上层灯泡,而没有看到由于垫塞物不够已经被压碎了的下层灯泡,那么此样本将导致错误的结论。由于样本选择对于结果的可信度有重要作用,所以根据正确的统计原理选择样本是非常必要的。在第一章中我们提到,在对越战争中,选择士兵的征兵计划,就是一个不好的选择样本的例子。

随机样本:是什么?

随机样本指来自于总体的这样一个样本,即总体的每个个体有一个已知的(有时是相等的)机会被包含在该样本中。

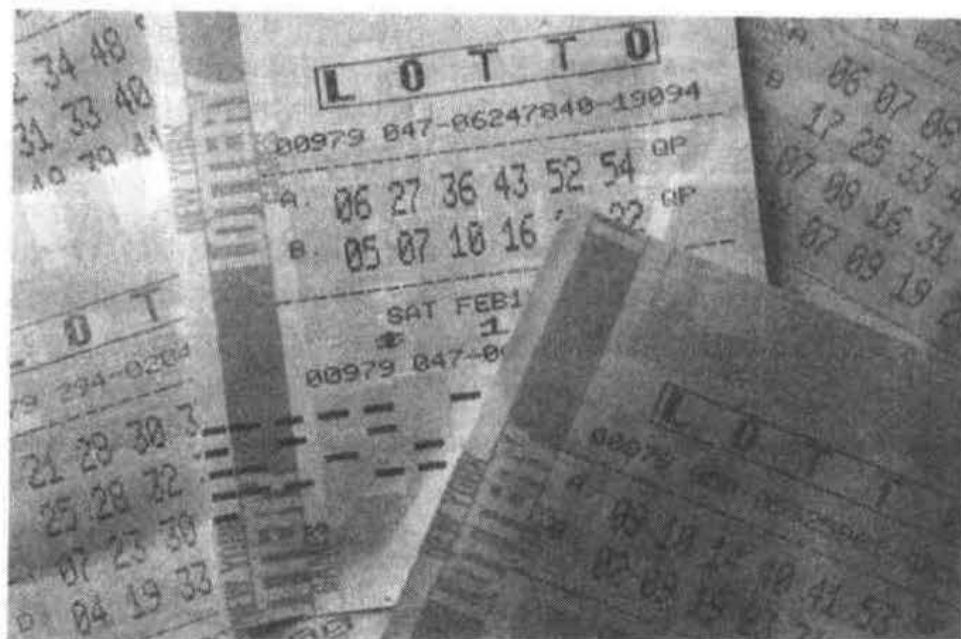
随机样本(random sample)指一个合适的、能够被推广应用于更大的总体的统计样本。从一个帽子中抽签决定名字,是选择随机样本的最简单的例子。叠好的纸签是组成整个总体的个体,每一个个体都有相等的被选中的机会。从这种意义上说,完全可能实现总体中所有的群体在某样本中的代表的数量比例大致等于这些群体在总体中的比例。因此,如果 Dubrovnik 城有 10000 个塞尔维亚人和 100000 个克罗地亚人,那么在该城市的一个随机样本中,每 100 个克罗地亚人应该对应着大约 10 个塞尔维亚人。

方便样本:如何产生一个“坏的”样本

研究者们经常习惯于研究手头方便的总体中的个体。例如,许多心理学杂志上的研究经常以那些被要求报名参加实验的人为研究对象,通常是一些上《心理学介绍》课程的学生;医学研究者和物理治疗家经常对他们自己的病人做研究;市场调查者研究被他们说服进行合作的消费者。能够很容易很经济地得到的样本称为方便样本(convenience sample)。尽管某些时候方便样本对研究来说已足够了,然而大部分时候却不行。从方便样本中得出的结果推广到整个总体的程度是很有限的。

有时杂志会要求读者回答某些问题并寄回答案,由此得到一些样本。根据随机取样的原则,我们可以对此提出疑问:不买杂志的人显然不包含在样本中,而寄出答案的人成了自选的团体,从他们那儿得到的数据是不能够作为自选团体以外的其他人的推广的;即使对于该杂志的读者总体本身它们也不一定是典型的。这些数据只是很好地描述了那些花了时间和精力寄回答案的读者,仅此而已。

社会调查和自助图书中报告的结论也存在同样的问题。Shere Hite, 一名自由作家,以专写女性爱情生活闻名,曾描述了几千名妇女对婚姻、性生活和丈夫的不满。也许她的最有名的统计论断是,结婚五年以上的女人中 70% 有婚外性生活(来源:Shere Hite, *The Hite Report: Women and Love: A Cultural Revolution in Progress*, New York: Knopf, 1987, p. 360)。显然,这是从 Hite 的方便



这些数据有怎样的随机性? (Patti Mcconville, The Image Bank.)

样本中得出的结论。我们并非说这个论断不合理,而是说由于这一样本不是被随机抽取,因此它并不代表整个的美国妇女总体。因此,把此结果推广于所有结婚五年以上的妇女总体是不正确的。

选择合适的样本

简单随机样本 当一个总体中的名字或电话号码被“放进一个帽子里”,搅拌均匀,并随机抽取,其结果就是一个简单随机样本(simple random sample)。本书中每一章末尾的所有公式都基于简单随机样本的使用。

得到随机样本的一个方法是,使用计算机随机生成的来自某一总体的电话拨号。没被列出的号码和列出的号码有同样被选中的概率,这一点比从电话号码本中随机选取样本要好。然而这种收集数据的系统意味着公司的电话也有可能被选中。因此,同时有公司电话和家庭电话的人被选中的概率是只有家庭电话的人被选中的概率的两倍。此外,电话调查把没有电话的那一小部分人也给漏掉了,这也是电话调查的一个明显的缺点。

抽样的其它形式 抽取比简单随机样本更复杂的样本也是可能的。其中一种抽样方法是从投票单位清单上随机选取若干小的地区,然后随机选取居住在这个地区的一些人进行直接调查。这是得到样本的一种有效途径。通过调查每一地区居住相邻的一些人,研究者们就避免了走很远的路从一个居住区到另一个居住区。

任何一种抽样程序的一个普遍的困难是,很少有哪份名单能完全包含属于某一特定总体的所有样本。比如,我们没有关于古柯碱吸毒者、犯轻罪者、第三次结婚的男人,或是有覆咬合毛病的孩子的完全名单。即使这样的名单存在,它们也永远不可能是完全的,甚至在制作名单时就可能有人又进入或离开这个总体了。(即便是活着的美国前任总统的名单也有可能由于某位突发心脏病而被改变。)美国并没有全国范围的人口注册机构。尽管这对于调查研究很不

方便,但这种考虑是为了保障公民自由。虽然的确存在公民的社会安全号码的一览表,但任何一个人都不能为抽样目的而得到它。

用于收集观测数据的变量的选择

为使他们研究的问题有一个结论,研究者们必须清楚应该测量什么样的变量。然而,仅使用观测数据来判断哪些变量对其它变量有因果作用,而哪些没有,即便不是不可能的,也是非常困难的。有时,观测各种现象的研究者们可能把原因归于某一变量,却忽视了其它更有影响的变量。例如,选举研究表明,女人倾向于支持民主党,而男人倾向于支持共和党。这是否意味着,一个人支持民主党是因为她是一个女人?更正式地说,性别变量对于投票变量有因果影响吗?抑或是其它变量有影响?

一个人的性别是由染色体的某种模式决定的,从这个意义上来说,很难想象染色体对一个人投票会有什么影响。也许是某些经济变量扮演了这个角色。比如,如果女人的薪水比男人少,而民主党又比共和党更关心这个问题,那么毫无疑问,女人会受此影响而支持民主党。研究者们可能没有注意到,实际上是经济上处于不利地位的人——而非女人本身——支持民主党,而经济上处于有利地位的人则支持共和党。

区分观测数据中的这些混淆着的作用比区分实验数据中的更困难。在以正确方法得到的实验数据中,其它变量的作用已经被关于试验组和对照组的随机设计抵消了。然而不幸的是,为了统计的纯洁性,试验数据并不总能够被收集到,因为实验设计的要求也许会冒犯习俗、法律甚至道德标准。(比如,随机安排新生婴儿到某些家庭中去,以研究婴儿哭的差别,就是不能够被社会所接受的,当然也没有哪位当代统计学家会认真考虑这个主意。)

从某种意义上说,从来不可能确定一个最好的方式去识别因果变量。例如,如果在导致人们支持某一政党时,收入水平比性别更重要,人们也许会问,是收入水平的什么东西导致人们在选举中的差异呢?是害怕失去吗?是怕失去所拥有的一切,还是怕失去购买所需物品的能力,或是令人骄傲的社会地位,又或是和一个人的收入水平相联系的任何一种其它的特性?在某种重要的意义上,对研究变量的选择是下列因素的函数:研究者的兴趣及目标,是谁为该研究付钱,结果的某些解释通常具有什么作用,以及对结果的某些解释相对于其它的解释有什么影响。

2.3 收集观测数据时的错误和误差

研究抽样技术使我们意识到,很多因素可使样本中的数据产生错误并导致错误结论。若仅凭某一样本中有60%的人赞同总统的作法,我们还不能够得出结论说,全国人口中的60%赞同总统。从刚开始决定调查到最后报告结果,任何一件事情都有可能出错。大部分调查也的确犯了这样或那样的错误。

为衡量某一调查的结果,我们必须知道:

- 样本是否是数据的合适的统计样本。
- 响应率(response rate)。

- 提问题时所用的实际措辞。
- 在调查程序中,该问题被安排在什么位置。
- 访员是谁。

抽样误差:并非错误的“误差”

如果已经选取了许多不同的样本,那么抽样误差可以告诉我们,20个不同样本的结果中的19个偏离总体真值有多远。

调查中的有些误差纯粹是统计上的,当然也有一些超出了统计误差的范围。主要的统计误差即所谓的“抽样误差”(sampling error)。这并不是某件事出错造成的误差,而是指这样的一个事实:如果研究被再做一遍,结果未必会和上次一模一样。例如,在下一次的抽样中,也许并不是60%的人赞同总统,而是59%——或62%或其它相近比例的人——赞同总统。

但是,即便不同的样本会产生不同的答案,大部分答案仍都位于总体中的真正比例的某一变化范围内。例如,通过每次大约1000个响应者的多次抽样,大部分样本(95%)得出的比例和实际的比例至多相差3个百分点。也就是说,抽样误差等于加或减三个百分点($\pm 3\%$)。

这种结果仅仅是每一个统计研究所固有的随机性的反映。别忘了这些比例是来自不同的样本,我们没有理由相信一个样本的结果会和另一个样本的结果一模一样。并且,没有理由相信某一特定样本的结果恰好等于从整个总体可能得到的结果。

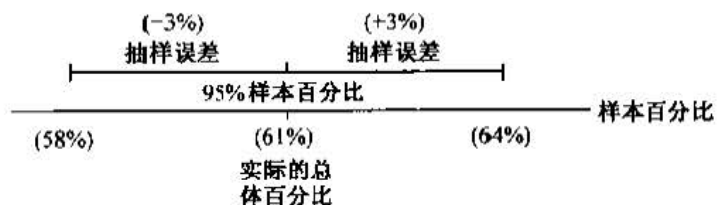


图 2.1 实际的总体比例和抽样误差为 $\pm 3\%$ 的例子。

图 2.1 说明了这一点。它表明了,当实际的总体比例为 61% 时,计算机产生的数据的情况。其中,100 个不同样本中的 95 个样本比例位于 58% 和 64% 之间。在这种情况下,我们说有 $\pm 3\%$ 的抽样误差——64% 比 61% 多 (+)3 个百分点,58% 比 61% 少 (-)3 个百分点。

这一例子是基于实际的总体比例为 61% 的基础之上的。而现实中,我们几乎从来不知道这个数字;实际上,我们之所以做调查,就是要估计总体比例。然而,通过样本,我们依然可以计算出抽样误差有多大。这个非凡的成果是由于数理统计学家推算出了计算抽样误差的公式。其中的一些公式将在第六章至第十三章中给出。

抽样误差的大小依赖于得到样本的方式和样本中包含的观测的个数。样本越大,误差越小。如果样本等于整个总体,则样本比例就等于总体比例。在总体变化以前,对整个总体做重复研究,就会得到相同的结果。在这种情况下,抽样误差是 0。

在公布任何一次抽样调查的结果时都应说明抽样误差的大小,不管是比例、均值还是其它形式。抽样误差告诉我们,样本离总体的实际值可能有多远。我们将在第六章和第七章的估计和假设检验中再次提到抽样误差。

未响应误差 (nonresponse error): 粗鲁的、匆忙的或沉默的响应者造成的结果

未响应误差是指, 由于包含在样本中的一部分人未回应调查而造成的误差。

另外一种影响抽样调查结果的误差是未响应误差 (nonresponse error)。这可能是由于某一选定的电话号码即使拨了多次也没有反应。或接通后那人拒绝回答问题。邮寄调查通常比电话调查有更多的未响应误差; 不理睬一封信比不理睬一个响着的电话容易得多; 而信被寄错地址的可能性也要比拨一个无人使用的电话号码的可能性大得多。有时, 一次好的电话调查, 通常会有 85% 至 90% 的响应率; 而一次邮寄调查的响应率很少有到达 50% 的。

拒绝参加所有形式调查的人的比例正在上升。人们越来越不愿回答问题, 因为他们怀疑某一调查是推销产品或服务的一种伪装。名声好的调查组织现在也经常不超过 60% 的响应率。

对研究者来说, 高拒绝率是一个很大的问题, 因为他们对于被选中但未参与调查的人了解很少。于是出现许多无法回答的问题: 是什么使得人们有了不响应和响应的区别? 相对响应者来说, 未响应者是富有还是贫穷? 保守还是自由? 有影响力还是缺乏影响力? 如果他们响应的话, 他们的回答会怎样影响研究结果?

一个最坏的假设情况表明, 未响应误差的影响可能有多大。假定我们计划调查 1200 个人, 却只有 1000 人接受了调查, 这意味着我们缺了 200 人的数据。在 1000 个我们调查的响应者中, 我们发现 600 人 (或 60%) 赞成某事物而其余人反对它。如果我们假定另外 200 人也赞成, 那么在 1200 人中就有 800 人赞成, 比例为 67%。但另一方面, 如果我们假定那 200 人反对, 那么 1200 人中只有 600 人赞成, 比例为 50%。因此, 仅仅由于未响应误差, 观测样本中 60% 的赞成比例有可能实际只是 50% 和 67% 之间的一个随机数。这就可能给我们的研究结果造成很大的差别。

一些经验表明, 在大部分情况下, 未响应者和响应者并无多大差别。如果我们开始时有一个高的响应率, 那么可假定未响应者也依同样的比例作出回答。但是如果响应率很低, 例如不超过 50%, 那么不响应的影响可能会很大。

研究者如何处理无人接电话的情况呢? 似乎可以用另外一个电话号码代替它, 但是这样对事情的改变会超出你可能的想象。在一次电话调查中, 替换意味着不经常在家的人比整天在家的人有较小的被调查的机会。这就违背了每人都有固定的被调查的机会这一原则, 并且我们有很好的理由认为, 不经常在家的人和经常在家的人有区别。对不在家的人, 唯一的办法是以后再打电话。但是, 这需要时间, 也许会花上好几天打好几个电话才能得到一个响应。

由于没有时间再打电话, 所以只靠一夜时间的民意调查的数据, 就不如可再打电话的调查得到的数据好。某一重大事件发生后马上调查人们态度的民意测验是很吸引人的, 但这类调查的结果会包含大量的未响应误差。总统候选人辩论之后的夜间调查是此类情形的典型例子。

停下来 想一想 2.2

如果很少在家或从不接电话人的观点不被考虑在内的话, 那么大部分时候在家的人的政治观点有可能使某次电话调查的结果偏差很大。你能够想到这样的一个例子吗?

响应误差

响应误差是指在调查过程中,由于问题的提问方式、问题所处的位置或访员的影响而使响应者在回答问题时产生的偏差。

如果研究者小心一点的话,由调查得来的数据是有可能避免响应误差(response error)的。我们在这里讨论其中的一部分(而不是全部)情况。即使所有的问题都有了回答,我们所知道的也仅仅是调查时人们告诉访员的,而未必是他们实际上做的、感觉的或想的。当我们在报纸上读到,在最近的一次调查中有60%的人赞同总统目前的工作,那么我们应该在大脑里把这句话严格为:被调查并回答了问题的人在当时有60%对访员说他们赞同总统的处理方法。

停下来想一想 2.3

我们中的许多人在购物中心、电话中或通过邮件被调查过。你是否能记起这样的例子,即当你试图缩短调查时间时,你就马马虎虎地应付,或者试图说访员想听到的内容,而不管你自己实际怎么想的? 作为一名响应者,你怎样评价你自己? 市场研究人员想要得到好的数据时,是否能够指望你这样的人?

问题的措辞 调查中问题的措辞影响着人们的回答。题目很微妙地划定了人们必须给出答案的问题。有时,题目使人们困惑,因而导致了非自本意的结果。例如,1992年由Roper协会做的调查发现,22%——五个中有一个——的响应者说他们怀疑大屠杀^①曾经发生过。经过对这一统计结果的最初反应之后,读者们开始把注意力转向问题本身:“在你看来,‘纳粹对犹太人的灭绝从未发生’是可能的还是不可能的?”这个题目包含了双重否定,这很可能使响应者困惑。两年以后又做了新的调查。这一次,提问方式变成了:“在你看来,‘纳粹对犹太人的灭绝从未发生’可能吗? 还是你确信它发生过?”这样,只有1%的人认为大屠杀从未发生过,和最初的22%很不一样。

除了修改措辞,统计学家们还经常遇到这样的问题:响应者是否一开始就没有观点,或者是措辞通过选项给了响应者一个观点? 如果一对夫妇让你给他们的孩子起名,你可能有些不知所措。但是如果他们说:“我们正在讨论三个名字,Maria, Gertrude 和 Maud。”这样你可能发现你已经有想法了。在大屠杀问题上,可能的选择项有两个,即这事件从未发生过或确定它发生过。没有想过这个问题的人和对这个问题不发表意见的人就没有合适的答案去选择了。当持有中立观点的人也不得不选择两个答案中的一个时,中立的态度就没有代表了,同时这两种观点就有可能被夸大(来源:The New York Times, July 8, 1994, P. A10.)。

防止问题的选择项产生倾向性答案的一个方法是,开始时只是提出筛选性问题,并不给备选答案。问题“你对大屠杀是否发生这个问题有什么看法吗”也许可以在大屠杀的民意测验中被提出。对于回答“没有看法”的那些人就没有必要问他们下一个关于具体看法的问题了。一般来说,调查所用的问题应和结果一起公布,否则我们很难对测量人们态度的那些调查结果作出评价。

^① 指第二次世界大战前及大战期间纳粹对欧洲犹太人进行的大屠杀——译者注。

问题所处的位置 调查中问题所处的位置也有可能影响结果,这就更增添了问卷设计的复杂性。在调查的初期,访员和响应者之间还不能很好地沟通,响应者对于表达某些观点也许较犹豫。随着调查的深入,响应者也许会感觉自如一些,因而有可能说话更直率且减少了些套话。响应者也许更会作出有偏见的评论和“政治上不正确”的意见并陈述个人的观点。到调查结束的时候,响应者也许会感到疲劳或厌倦。如果响应者希望尽快结束这次调查,比起调查中间的回答,最后的回答就有可能较短,较不准确,较不认真。研究者试图通过如下方法来使响应者放松,即在开始时间问较容易、不涉及个人的问题,而在关系融洽时间问较难的和涉及个人的问题。例如大部分在美国的调查中,收入问题在较靠后的地方才被问到。结束调查时的问问题经常短且简单。

响应者有可能想在被问到的这一点和那一点上保持一致。如果他们在一个问题上支持某一观点,那么他们有可能认为应该在其它地方维护此观点,尽管并没有义务要求他们这样做。比如,在某一问题中支持死刑的人也许会在战争问题上对称他是和平主义者比较犹豫。通过一个调查,确定地表达自己的观点是响应者们不变的要求,调查者也试图通过把问题放在合适的位置以使人们能很好地给出代表他们本意的观点。

通过电话进行调查时的响应偏差

全国黑人政治研究(National Black Politics Study)是一个关于1,204名非裔美国人的调查。在这个调查中,为使响应者在回答问题时感到舒服,访员都安排为非裔美国人。但是,因为这是一个电话调查,响应者并不能看见访员。Lynn Sanders,一位政治学家,研究了响应者对于访员种族的判断对他们回答问题的影响。在问题“访员是哪族人”时,14%的响应者认为他的访员是白人。

响应者还被问到他们是否同意“美国社会对每个人都是公平的”这个说法。在认为访员是非裔美国人的那些响应者中,有14%同意此说法;而在认为访员是白人的那些人中间,则有31%的人同意此说法(来源:Chance, vol. 8(1995), no. 4, P. 5.)。

停下来想一想 2.4

响应者对于访员种族的判断对其答案的影响是很清楚的。还有多少没有发现的访员因素影响了调查结果呢?

访员的影响 响应者的答案受到他们所认为的访员身份及访员的观点等因素的影响。调查设计者们总是尽量使访员和响应者在人口统计特征像年龄、性别、种族等方面差不多。尤其对于敏感的问题,例如对其他群体的态度,伦理或法律行为,性生活等,在调查时双方最好能有共同语言。

2.4 实验数据:寻找造成结果的原因

收集数据的另一种办法是在实验中控制一个或多个变量并测量操纵的结果。例如,如果我们给一组植物施肥,另外一组不给施肥,那么我们就是在控制植物土壤的成分。我们可以测

量像增长率、成活率等变量。实验是检验变量间因果关系的一种方法。在实验中,研究者试图控制某一情形的所有相关方面,操纵少数感兴趣的变量,然后观察实验结果。

实验最早的例子发生在 17 世纪初,当时英国海军试图发现坏血病的起因。该病症状为牙龈肿大出血,皮肤上有青灰斑点;在海上长期航行的水手经常患有此病。英国海军部怀疑是因缺乏柑橘类水果而导致此病。当这个想法被提出时,四艘海军军舰正要离开英国做长期航行。为调查是否因缺乏柑橘类水果而导致了坏血病,海军部安排其中一艘船上的水手每天喝柑橘汁,而其它三艘船上的水手则没有柑橘汁供应。

航行还未结束,在其它三艘船上就有如此多的水手染上了坏血病,以至于不得不把每天喝柑橘汁的水手们分配到这些船上以帮助他们进港。尽管实际操作实验计划有可能从各方面得到改进,但该实验显然成功地证实了最初的想法。

实验组和对照组

实验数据是指在实验中控制实验对象而收集到的变量的数据。

在坏血病的例子中,喝柑橘汁的水手构成了实验组(experimental group),而不喝柑橘汁的水手则构成了所谓的对照组(control group)。实验组是指随机选择的实验对象的子集。实验组中的个体要接受对照组所没有的某种特殊待遇。几乎所有设计较好的实验(有时也包括观测研究)都有一个对照组和一个或多个实验组。

需要对照组的原因是:没有对照组,就没有办法确定这样的操作或是某些其它变量(或几个联合变量)是否产生了作用。如果所有四艘船上的水手都喝柑橘汁,那么,没有感染坏血病也许可归因于特别好的食物配给或水手们在航行中接受的其它治疗。但是实验组和对照组的唯一差别是一组喝柑橘汁而另外一组没有。因此,可以得出结论说,柑橘汁使水手们避免了坏血病是合乎逻辑的。在第一章实验设计的例子中,20 世纪 60 年代小儿麻痹症疫苗的检验也说明了这一点;没有对照组,就没有比较疫苗效果的基础。

选择实验组和对照组

对照组是指实验对象中的一个被随机选择的子集,其中的个体没有特殊待遇。

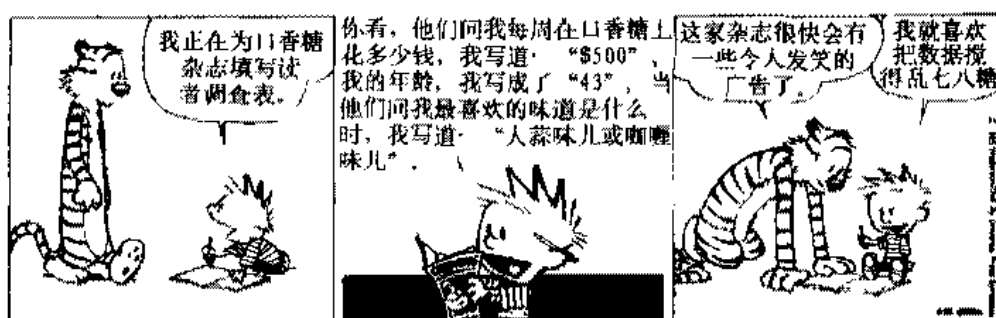
安排实验的另外一个问题是谁应该在实验组,谁应该在对照组。坏血病的例子不是一个完善的实验,因为我们对于为什么一艘船上的人没有得坏血病可以有其它解释。也许这三艘船——不包括第四艘——本身有什么导致得坏血病的東西。尽管这一般不可能,但是此类现象还是有发生的可能性。因此,如果对每一个水手随机地决定谁喝或不喝柑橘汁,而不是考虑按船来分的话,结果会更有说服力。通过随机决定待遇,有关于船的其它变量的影响相互抵消,就不会影响最终结果了。

一个人也许想知道,如果使用的是自愿而非随机分配的方式产生实验组和对照组,结果会怎样呢?例如,喜欢喝柑橘汁的人在实验组,而喜欢喝酒的人在对照组,这样行不行呢?这种方法产生的问题就是:不能确定在研究开始以前,两组的人是否有相同的身体状况。如果实验对象是随机安排的,那么健康和不健康的水手在每一组中的数目差不多,因而身体状况对导致坏血病的影响就被抵消了。

随机选择实验对象的原则是伟大的英国统计学家 Ronald Fisher 爵士的一个主要贡献。他在 20 世纪 20 年代曾致力于农业实验的研究。从那以后,这个原则就成了所有好的实验所遵循的一个原则。

对人做实验时产生的问题

以人为对象的实验,目标依然是随机地把人安排到实验组和对照组。但是,这是非常困难的,甚至是不可能达到的。安排给一株马铃薯一块贫瘠的土壤比安排给一个人低于常人的生活条件容易多了。



对数据总是应该采取怀疑态度。“Calvin and Hobbes” copyright 1995 Watterson. Dist. by Universal Press Syndicate. Reprinted with permission. All rights reserved.

组织问题 我们都能想得出很多原因,为什么研究人要比研究种土豆困难得多。首先也是最重要的,人们都有其自己的计划和兴趣,未必会服从科学家们的研究兴趣。他们在满足研究条件、定时约会、听从指挥和完成安排给他们的部分等方面,同样可能有困难。在讲述通过电话和访员调查的方式从人们中获取好的数据时,我们已经提及到这样一些问题,在实验研究方面也会遇到同样的限制。

心理学问题 在实验研究中,人们对被研究非常敏感。这使他们注意自我,从而对他们的行为产生了很多约束。首先记录这种影响的例子之一是 1924 年到 1933 年间,对通用电器公司的工人生产率的系列调查。在一次调查中,一组社会学家和公司人事部门的成员研究了各种照明程度对生产灯泡的工人生产率的影响。研究者们增大照明度,发现产量增加。但奇怪的是,当他们减少照明度时,发现产量也增加。看来无论研究者们做什么,工人的产量都会增加。这些工人看来是对研究者的关注有了反应,而不是对照明度。

后来,这种工人们研究者的注意,而不是对预想中的控制产生反应的现象,称做 Hawthorne 效应,取自所做实验的灯泡工厂的名字。一些措施,例如,保证对照组和实验组受到同样多的关注可以避免这种效应。(来源:可参看 Robert K. Merton, Social Theory and Social Structure, New York: The Free Press, 1957, P 66)

道德问题 道德问题使得对人和动物做的实验复杂化了。当某种道德上的困境和收集观测数据联系在一起时,例如袖手旁观的负作用产生时,操纵和控制事件的实验者更容易陷入道德困境中。比如,使人们接受某种不可预知结果的药物治疗是正确的吗?设想人们也许会遭受到出乎意料的负面作用?考虑负作用有可能使一个人在检验和介绍新药时采取保守态

度。然而,从另一方面说,如果新的治疗方式是有益的,结果又会怎样呢?有致命的疾病的人应该等多长时间去试一种新药?他们能等多长时间?

对于只接受安慰剂的对照组来说,他们缺少了有益治疗,结果会怎样?在小儿麻痹症疫苗的实验中,对照组中得小儿麻痹症的人比实验组中多得多。如果对照组中的儿童也接受疫苗,那么我们完全可以相信会有多至 100 名孩子不会得小儿麻痹症了。

在人们研究阿司匹林对心脏病的作用时,同样的困境发生了。实验设计允许一组男医生每天吃阿司匹林以看这种治疗是否会减少对心脏病的威胁。实验进行了五年以后,吃阿司匹林的治疗组比吃安慰剂的对照组有较少的人得心脏病。结论是如此清楚以至于实验在计划结束以前就停止了,对照组的人们开始被鼓励吃阿司匹林。在其它项目中,结论就未必如此清楚,长期的副作用可能会抵消短期的治疗效果。反应停(thalidomide)^①的实验就是这种情况。该药在 20 世纪 50 年代被用于防止孕妇流产,然而吃这种药的母亲生下的有畸形四肢的婴儿出奇地多。

几乎在美国做的所有研究,尤其涉及到健康后果的研究,都需经过专家筛选,无道德问题才可进行。例如,设想检验一种有望治疗艾滋病的新药。如果该药是有效的,那么在对照组的人们如果得不到新药就会冒着死的危险。然而,如果发现这种药有副作用而导致使用该药的人两年后有更高的死亡率,那么对照组则可能避免了使用这种致命的药。道德问题该怎么考虑?没很简单的答案可以给出。道德问题必须被时常地考虑及重新考虑。幸运的是,大多数研究没这么戏剧化,后果也没这么严重。



许多因素影响工人的生产力。
(Michael Rosenfeld, Tony Stone Images.)

停下来想一想 2.5

近来你读过有关这样的研究吗,即公司被起诉在对待顾客时侵犯了他们的权力?这些事件里有道德困境吗?鉴于这些事件的性质,在你看来,这些公司在道德上应负责还是不应负责?

在实验中统计的角色

大多数做实验的研究者接受统计建议。统计对实验过程的贡献集中体现在三个方面:确定合适的观测数以得到显著的结果;设计实验以使之符合统计分析的标准;发明尽可能最有效地同时研究几个变量影响的方法。

有多少观测? 对需要多少观测以得出达到预期的精确度的结果,统计学家们可以给出建议。通常来说,数据越多越好。但相对于较少的观测值,观测值多时,花费较高,收集数据所

^① 酞酰亚胺基戊二酰胺,一种镇静剂及安眠药——译者注。

RONALD FISHER 爵士



伟大的英国统计学家 Ronald Fisher 对统计方法的重要贡献之一是在如何做实验方面的工作。他意识到,当几个变量影响一个结果变量时,把所有变量的作用放在一起研究要比单个研究好得多。Fisher 发明出来的同时研究几个变量的实验方案之一是众所周知的“拉丁方设计”。

Fisher 的一部分时间是在英国剑桥大学的 Caius 学院作为评议员(Fellow)度过的。为了纪念他,该学院在餐厅的墙上装了一扇代表拉丁方的有色玻璃窗。这扇窗是一个长 3 英尺,宽 3 英尺的正方形。该正方形被分成了 7 行 7 列共 49 个小格。每个小格是一块方形有色玻璃。玻璃的颜色总共有 7 种。颜色的设置使得正方形的每行每列都有 7 种不同的颜色。例如,黄色玻璃分别被安装在第 1 行第 6 列、第 2 行第 2 列、第 3 行第 1 列、第 4 行第 3 列、第 5 行第 4 列、第 6 行第 7 列和第 7 行第 5 列。

这扇窗户显示了 3 个不同的变量,每个变量取 7 个不同的值。窗户中行代表第一个变量,列代表第二个变量,颜色代表第三个变量。这些值的不同组合有 $7 \cdot 7 \cdot 7 = 343$ 种,而窗户上的每一种可能的小块玻璃的位置和颜色就恰好对应着变量的一种组合。窗户告诉我们,如果所做研究有三个变量,每个变量有 7 个值,那么我们只需得到其中 49 种组合的观测数据就够了。

停下来 想一想 2.6

你能够设计一个与 Fisher 设计的拉丁方性质类似的窗户吗? 你可以用不同的颜色,但请像 Fisher 的窗户那样也有黄色。或者,你可以试着设计类似的有较少的行和列(比如 4 行 4 列)的窗户。

总结:班级规模影响学校表现吗?

让我们看一个真实的实验研究的例子以结束关于实验的讨论。在这个例子中,你可以扮演一个实验者,为回答一个感兴趣的问题要面临一系列的困难。

这个例子发生在田纳西州的一个教育实验中,其结果公布在一家新闻杂志上。(来源:The Economist, August 31, 1991, p. 23.) 这里,我们仅使用新闻报道所给的信息,看能从这个实验中得出什么结论。(如果我们想得到更多的信息,就需要再咨询一些细节的描述。)

长期以来,人们有这样的一个观念,即小班比大班更有利于学习,但是,很难为这个观点找到经验证据。在 20 世纪 80 年代中期,田纳西州的学校官员决定做实验来检查班级规模是否对学生的表现有影响。该项研究的主要设计很简单,只关心单个变量——班级规模(class size)——的影响。该变量取两个值:小规模和普通规模。

操作定义 在实验开始以前,必须决定几个重要的非统计问题。也许其中最重要的一个是确定“小”班的含义是什么。在研究中,田纳西州的官员决定:包含 13 到 17 个学生的班级叫“小”班,包含 22 到 25 个学生的班级叫“普通”班。研究者们还必须确定,“有进步的表现”是什么意思。他们决定用标准化教育考试来衡量表现。

这样,研究者们就开始了实验。研究假定被具体为:小学生在小班比在大班学得好。该假定进一步具体为:5 岁的孩子,在只有 13 到 17 人的小班中学习四年以后,参加标准化考试的成绩要比花同样的时间在有 22 到 25 人的大班中学习的孩子的成绩好。明确四年的时间,是为了排除那些在研究期间转入或转出实验班的学生。这一点是很重要的。

停下来想一想 2.7

如果你住在田纳西州,你的孩子被分在了普通班而不是小班,你会怎样想?

实验学校的选择 下面一件事是要确定用哪些学校来做实验。该杂志的报道只是说,研究是在 76 所不同的小学进行的。田纳西州的小学要远远超过 76 所,我们只能希望该项研究在选择学校时是随机的。我们同样得相信,学生是被随机地安排在小班和普通班中的。

停下来想一想 2.8

为什么随机选择学校和随机安排学生到这两种类型的班级是如此重要?

实验设计 现在研究者们面临着这样的问题,对于遍及整个州的 76 所小学的所有 5 岁的孩子该做些什么。他们可以在一些学校只设小班而在其余的学校只设普通班。但是这样的话,小班和普通班之间的差异可能归因于其它变量。例如,如果所有在大学城中的学校念书的孩子都在小班,而所有在市中心和学校就读的孩子都在普通班,那么即使发现小班的孩子成绩好,也可能不是由于班的规模的差别而是由于这样的事实,即通常来自教育程度高的家庭的孩

子在标准化考试中的成绩较好。然而,如果每个学校中的孩子都是随机地被分在不同类型的班级中,那么因背景不同而可能产生的影响就会被抵消,也就不会影响总的考试成绩了。

其它变量也应被考虑进去。例如,教师也必须被随机分配。如果把所有的新教师安排在较大的班,而把所有有经验的教师安排在小班,这就是不公平的。其它自然的差异,像教室的条件等问题,也必须被平衡,以使得两组之间没有优势差别。

结果 四年后,研究者们发现小班上学生的成绩比普通班学生的成绩“显著地”好。(术语“显著地”将在第六章和第七章讲到。)实验结果表明,仅一年之后,小班学生的阅读能力领先了 1.5 个月,数学能力领先了 2.5 个月。四年之后,当实验结束时,小班的优势仍旧存在。

2.5 数据阵/数据文件

在一项研究中,数据被收集以后,不管是实验的还是观测的,它们通常以典型的表格形式被输入到计算机文件中。这意味着每一列代表一个变量,如性别;每一行代表一个个体,如人、植物、动物、组或其它我们收集数据的单元。这样的—个数据表通常叫做数据阵或数据文件。表 2.1 是一个根据抽样调查得来的数据生成的小数据阵的例子。

表 2.1 一次抽样调查的数据阵

人员编号	年龄	性别	投票	态度
1	20	女	民主党	中立
2	27	女	民主党	反对
3	19	男	共和党	反对
4	38	男	民主党	赞成
5	38	男	共和党	赞成
6	53	女	民主党	赞成
7	24	男	共和党	赞成
8	41	女	共和党	反对
9	35	女	民主党	中立
10	30	男	共和党	赞成

在计算机分析文件中的数据时,为方便起见,我们经常把数据文件中的描述性文字转化成数字。每一个人都被分配一个身份号码作为名字。年龄变量本来就是用数字测量的,因此不需要任何转化。性别变量的两个取值是“女”或“男”,因此,“女”用数字 0 代替,“男”用数字 1 代替。当然还可以使用任何两个别的数字,比如用 17 代替“女”,用 23 代替“男”。但是,由于一些实践上的原因(这将在第九章讲到),使用数字 0 和 1 会好一些。“投票”这个变量的值可类似地转化为 0 或 1;而“态度”这个值,可以用三个等级数 1、2 和 3 表示。

表 2.2 一个样本调查的数据阵

人员编号	年龄	性别	投票	态度
1	20	0	0	2
2	27	0	0	1
3	19	1	1	1
4	38	1	0	3
5	38	1	1	3
6	53	0	0	3
7	24	1	1	3
8	41	0	1	1
9	35	0	0	2
10	30	1	1	3

从计算机中打印出的数据阵正如表 2.2 所示。尽管这个表很容易读,但是一个典型的全国调查可能有 1,000 个响应者而不是这里的 10 个,并且很容易有 100 个变量而不是 4 个。如果有 1,000 行 100 列,那么数据文件中将有 100,000 个数字。这可就不是如此容易读的了!信息会都放在那儿,但数据中的趋势和模式却不清楚了。如果不用统计方法对数据进行简化和浓缩——统计分析,那么研究者就不可能得到感兴趣的信息。

2.6 小 结

2.1 定义变量

正确进行数据收集的第一步必须包括详细指明要研究的变量。

2.2 观测数据:问题和可能性

观测数据是指仅通过对世界的观察(而没有操纵或控制它)所得到的数据。

一个总体包含所有要研究的个体。普查是收集整个总体的数据的过程。一个合适的能够被用来推广到总体的统计样本叫随机样本。在随机样本中,总体中的每个个体都有已知的(经常是相等的)机会被选择到其中。“从帽子里”抽签决定样本可得到简单随机样本。

在观测研究中,要确定一个变量是否和另一个变量有因果关系通常是比较困难的。在任何一个观测研究中,必须承认可能有其它未知的变量比所研究的变量对某一变量有更直接的影响。

2.3 收集观测数据时的错误和误差

如果已经选取了许多不同的样本,那么抽样误差可以告诉我们,20 个不同样本的结果中的 19 个偏离总体真值有多远。这种一个样本的结果和另外一个样本的结果之间的偏差是由样本选择的随机性引起。抽样误差的大小依赖于样本中观测的个数和抽样方式。样本数越多,抽样误差越小。抽样误差总是应该被报告出来。

未响应误差是指样本中有缺失数据时导致结果出现的误差。缺失数据可能是由于响应者

不愿回答所有提问或联系不上某些样本成员。在最坏的假设情况中,所有未响应者都对某一问题有相同的答案,此时未响应误差就可能非常大。幸运的是,研究已经表明,在大部分问题上,未响应者和响应者并无多大差别。

数据集中响应误差的起因可能是由于提问方式使得响应者困惑或暗示了某种答案,或由于各种问题的顺序安排得不得当,或由于访员影响了响应者的回答等等。

2.4 实验数据:寻找造成结果的起因

数据收集也可以在实验中通过控制一个或多个变量,然后测量每次控制的结果而得到。实验是研究变量间因果关系的一种方法。实验时,研究者试图控制某一环境的所有相关方面,然后控制其中一小部分感兴趣的变量,并测量每次控制的结果。对好的实验设计起关键作用的是对照组,即研究对象的一个子组,它除了未受到实验控制,其余各个方面都和受到实验控制的实验组类似。实验对象随机地分配给实验组和对照组。随机分配的一个主要原因是要使无关变量的作用相抵消,不致于影响最终结果。

用实验的方法研究人通常很困难,因为他们有可能拒绝科学家们对他们进行的操作和控制。他们还有可能变得不自然,或有可能对实验感到厌倦或疲劳。实验对象在行为上有可能受实验本身性质的控制,某些情形下可能过予合作。像权衡是否给予治疗的道德困境也是进行实验的一些重要限制。

统计对于实验成功的贡献集中体现在三个方面:确定合适的观测数以得到显著的结果;设计实验使其符合统计分析的标准;发明尽可能有效地同时研究几个变量影响的方法。

2.5 数据阵/数据文件

一项研究的数据收集完以后,不管数据是观测的还是实验的,它们通常都以典型的表格的形式被输入计算机文件中。这样的数据集合叫数据阵或数据文件。

补充读物

Bartlett, M. S. "R. A. Fisher." In William H. Kruskal and Judith M. Tanur (eds.), *International Encyclopedia of STATISTICS*. New York: The Free Press, 1978. 20 世纪的一个最有名的统计学家的一生。

Cochran, William G. "The design of experiments." In William H. Kruskal and Judith M. Tanur (eds.), *International Encyclopedia of STATISTICS*. New York: The Free Press, 1978. 统计学家怎样协助计划实验。

Converse, Jean M., and Stanley Presser. *Survey Questions: Handcrafting the Standardized Questionnaire* (Sage University paper Series on Quantitative Applications in the Social Sciences, no. 07-063). Beverly Hills, CA: Sage Publications, 1986. 在调查时问问题的艺术。

Deming, W. Edwards. "Sample surveys: The field." In William H. Kruskal and Judith M. Tanur (eds.), *International Encyclopedia of STATISTICS*. New York: The Free Press, 1978. Deming 在

质量控制上出名之前是一个抽样统计学家。

Hoagling, David C., et al. *Data for Decisions*. Cambridge MA: Abt Books, 1982. 第二章:实验(pp. 18-46),第四章:比较观测研究(pp. 55-57),第五章:抽样调查(pp. 78-106),第九章:官方统计(pp. 166-178),讨论了不同类型的数据。

Hobbs, Nicholas. "Ethical issues in the social sciences." In William H. Kruskal and Judith M. Tanur (eds.), *International Encyclopedia of STATISTICS*. New York: The Free Press, 1978. 讨论了在作社会研究时我们可能遇到的伦理问题。

Kahn, Robert L., and Charles F. Cannell. "Interviewing in social research." In William H. Kruskal and Judith M. Tanur (eds.), *International Encyclopedia of STATISTICS*. New York: The Free Press, 1978. Michigan 大学社会研究所的两名作者讨论在采访时出现的问题。

Kalton, Graham. *Introduction to Survey Sampling* (Sage University Paper Series on Quantitative Applications in the Social Sciences, no. 07-035). Beverly Hills, CA: Sage Publications, 1983. 一个如何抽取随机样本的综述。

Mosteller, Frederick. "Nonsampling errors." IN William H. Kruskal and Judith M. Tanur (eds.), *International Encyclopedia of STATISTICS*. New York: The Free Press, 1978. 这个国家的一个统计大师讨论可能影响我们从数据得到的结论的误差。

Spector, Paul E. *Research Designs* (Sage University Paper Series on Quantitative Applications in the Social Sciences, no. 07-023). Beverly Hills, CA: Sage Publications, 1981. 关于如何收集实验数据的不同方案的综述。

Witmer, Jeffrey. *DATA Analysis: An Introduction*. Englewood Cliffs, NJ: Prentice Hall, 1992. 以各种不同方式收集的有趣数据集的例子。

习 题

回顾(习题 2.1-2.22)

- 2.1 某青少年杂志想借助大学生做一个调查,以吸引 17 到 19 岁的读者。列举调查小组抽取样本以前关于定义目标总体需要作出的几个决定。
- 2.2 解释什么是实验组。
- 2.3 解释什么是对照组。
- 2.4 a. 什么是随机样本?
b. 列出产生随机样本的三个困难的地方。
- 2.5 从什么意义上说,抽样误差是我们在统计分析时所遇误差中的最好的一种?
- 2.6 从杂志或书中选出一个样本数据,或普查数据,或其它数据的例子。
a. 作者令人满意地解释了数据是怎样被收集的了吗? 说明为什么。
b. 所得结果可以很好地推广到总体吗? 说明为什么。
- 2.7 学生会希望调查毕业班有关毕业典礼的观点。你自愿在学校里抽取 10% 的毕业班学生作为样本。
a. 你怎样安排抽样以保证随机性?

- b. 在抽样中你有可能遇到怎样的问题?
 - c. 这些问题可能会影响什么?
 - d. 你打算怎样解决这些问题?
- 2.8 你的几位好朋友想要帮助你完成习题 2.7 中学生会的调查。他们自告奋勇每人调查 12 个关心毕业典礼问题的朋友;这将恰好是一个 10% 毕业班学生的样本。如果班里投票赞成在湖边来一个通宵烧烤晚会,他们还情愿去买饮料。
- a. 出于什么原因你拒绝朋友的帮忙?
 - b. 如果你自己也想举行烧烤晚会,你能采取什么样的措施以防止你的观点影响你的同学?
- 2.9 解释抽样误差是否表明统计分析做得很糟糕。
- 2.10 哪些因素在决定抽样误差大小时是重要的?
- 2.11 a. 什么是总体的真实值(the true population value)?
- b. 它被认为在什么位置(Where is it supposed to be located)?
- 2.12 什么是响应误差?
- 2.13 举一个调查问题的例子,使得它和下面这个提问有同样的不可接受性:你已经停止打你的狗了吗?你为什么认为这是一个坏的问题?
- 2.14 如果目标是要确定响应者的经济收入水平,请设计一个可接受的调查问题。
- 2.15 一次民意测验发现,56%的响应者支持联邦最高法院几年前关于 *Roe v. Wade* 的决定(the *Roe v. Wade* Supreme Court decision)。公布的抽样误差为 $\pm 2\%$ 。
- a. 关于其余 44% 的响应者对 *Roe v. Wade* 的感觉如何,你能说点儿什么?
 - b. 在你利用这次民意测验的结果之前,关于数据是如何收集的,你还应该知道其它的什么东西?
 - c. 说明怎样使用抽样误差的百分比并解释结果。
- 2.16 你是否同意你的孩子参加像小儿麻痹症疫苗检验这样的实验。这里假定你并不清楚这种药有副作用,还是根本没作用,或是对所有人都有好处。说明你的理由。
- 2.17 为什么当一个实验不包括对照组时,解释其结果就比较困难?
- 2.18 a. 在什么样的条件下,你会在不知道有潜在的副作用的情况下,自愿参加一项关于牙膏的实验研究?
- b. 在什么样的条件下,你会在不知道有潜在的副作用的情况下,自愿参加一项关于改变精神状态的药的实验研究?请像考虑你自己的健康一样考虑其他人。
 - c. 你相信你的回答类似于大多数人,一部分人,还是很少人?为什么?
- 2.19 a. 能否设计一个观点看起来是中立的调查问题,内容是关于人们对堕胎赞成与否?为什么?
- b. 提问方式对给出的答案有什么重要影响?
- 2.20 在一个充满矛盾的社区,住着韩国、巴基斯坦、菲律宾、亚美尼亚、和冰岛血统的家庭。你正在进行一项关于房屋租赁的调查。
- a. 在你选择访员时,哪些考虑可能是重要的?
 - b. 描述你要雇什么样的人,包括性别、年龄、种族、教育程度或其它特征。
 - c. 什么偏差是你接受的?哪种又是你试图避免的?

2.21 什么是数据文件?

- 2.22 a. 数据文件的列通常是指变量还是个体?
b. 数据文件的行通常是指变量还是个体?

解释(习题 2.23 - 2.36)

- 2.23 睿智的统计学家宣称:“世上有两种数据:好数据和坏数据。”好数据和坏数据的区别是在收集过程中是否遵循了正确的统计原理。考虑到收集好数据时的困难,你认为统计学家是否应这样说:“世上有两种数据:坏数据和更坏的数据?”请解释为什么。
- 2.24 a. 在什么样的情形下你将(或是曾经)不愿参加一项调查?
b. 你认为你拒绝参加可能会使调查有怎样的结果(假定有其他人出于同样的原因也拒绝参加)?
c. 对于帮助或阻碍调查的赞助者们实现其目的,这将产生什么样的影响?
- 2.25 在 1991 年春季,全美国的 7-Eleven^① 商店开展了一项民意调查。在这次调查中,买饮料的顾客通过挑选杯上印着是或否的饮料来对某一问题“投票”。根据这次调查的结果,费城/特伦顿^② 地区 50.9% 的响应者“投票”说他们会因为钱而结婚,然而在全国范围内这个比例是 53.6%。(来源: *The Philadelphia Inquirer*, April 16, 1991, P B3.)
a. 这个结果是否意味着,费城和特伦顿的人比其它地方的人较少倾向于为钱而结婚呢?
b. 什么有可能解释这两个比例间的差别?
- 2.26 在地方购物中心,高矮胖瘦各不相同的顾客被一群高矮胖瘦各不相同的访员截住,并询问他们最近买的减肥饮料和减肥食品。
a. 访员和响应者的相互作用可能对这次调查的结果产生什么样的影响?
b. 更一般地,像这样的调查有可能以完全中性、无偏的方式进行吗? 解释为什么。
- 2.27 城市规划者对德拉威郡(Delaware city)志愿防火员的防火意识很感兴趣。一组关于 Garden City 的防火员的调查包含的抽样误差为 $\pm 7\%$ 。
a. 如果德拉威郡的其他防火员——Media 镇、Swarthmore 及 Rutledge 等地的防火员的样本包含在报告中,你认为会对规划者有用吗?
b. 如果对整个郡的所有志愿防火员都做调查,那么抽样误差将变大还是变小? 为什么?
- 2.28 假定你想让学生们评价他们所念过的所有的学校,
a. 说明当你定义这个变量时,可能遇到的困难。
b. 这些困难可能怎样导致对某些类型的教育机构的偏爱?(提示:结果有可能有利于那些几乎把所有精力投入到本科生教育的小学校。)
- 2.29 引用一项关于商业婚姻介绍所的研究:“一位使用电视媒介的婚姻介绍人宣称,有 40% 的人经她第一次介绍以后建立了有承诺的关系(committed relationship)。”考虑这个信息,如果你急着结婚,你会为这项服务花钱吗?(婚姻介绍所在这句话中关于介绍成功的定义有什么问题吗?)(来源: Mura B. Adelman and Aaron C. Ahuvia, “Mediated channels for mate

① 美国的一家很大的连锁店,其营业时间是从 7 点到晚上 11 点——译者注。

② 特伦顿(Trenton)为美国新泽西州首府——译者注。

seeking: A solution to involuntary singlehood?" Critical Studies in Mass Communication, vol. 8 (1991), pp. 273 - 289)

- 2.30 根据国家滥用酒精和酗酒研究所(the National Institute on Alcohol Abuse and Alcoholism)的研究,酒鬼父子比起不喝酒的父子来创造力不足。“有创造力的人可能会是个酒鬼,但酒鬼们很少是有创造力的。”研究的负责人如是说。根据这段话,在对研究结果的描述中,有创造力/无创造力这个变量是怎样被重新定义的?
- 2.31 调查表明,作弊在大学里是一个严重的问题。然而,并不总是能知道谁作弊了,谁可能被冤枉了。你对研究考试作弊对撒谎的心理指数的影响很感兴趣。作为心理学入门课程的教授,你可以在期中考试时安排一半学生在考试时的多选题答题纸中找到答案(看起来像是由于失误)。你能够知道哪些同学收到了“作弊者”考卷,而哪些没有。然后,当你由于他们作弊而质问时,你可以偷偷地把这场景录下来。最后你就能够据此观察心理指数,并看作弊的学生对你的质问如何反应。假定这个研究有科学价值并且能够被实施(从组织角度上说),你注意到其中存在的问题了吗?都是些什么问题呢?
- 2.32 由亚特兰大的 Emory 大学的 Arthur Kellermann 博士领导的小组做了一项研究,把具有相同年龄、性别、种族和居住地区的被谋杀在家里的人和没有被谋杀的人的家庭进行比较。388 名被谋杀者的家和未被谋杀者的家在很多方面不同。被谋杀者的家中更容易有枪,尤其是上了子弹的枪,易使用非法的毒品,易有被捕记录或有家庭暴力的历史。该研究声称,若不考虑其它危险,持有手枪使得人们被谋杀的危险性增大了 2.7 倍。
- 这项研究的对照组是什么?
 - 一个人是否会被谋杀的最重要的因素是什么?
 - 根据这一报告,还有其它未被控制,但有可能与被谋杀在家里有重要关系的其它的显著变量吗?
- 2.33 在一项有关数学课程发展的研究中,研究者们在一个地区随机抽取了十所学校。其中一所是 Wallingford 小学,该校的教师们被要求自愿报名参加将在感恩节假期中举办的数学进修班。在 38 名有资格的教师中,有 14 名愿意去,于是从这 14 名中随机挑选了 9 名(研究者们要求选择过程要“随机化”)。尽管很小心地试图挑出一个随机样本,然而选中的这 9 名老师中有 8 名是男性,而教师队伍中 65%的是女性。进修结束以后,教师们非常兴奋地为这项新的数学计划呈交了总结报告,其中尤其提到这一计划如何督促他们“真正地熟练了他们自己的数学技巧”。他们强烈建议在下一年实施同样的计划。研究者们得出结论说,教师们对这一数学计划的反应很热情,并建议在第二年预算时优先考虑该计划。研究者们结果非常重要,因为以前在这些学校引进新的数学计划不很成功,原因是教师们不乐意接受,并且有时在数学上缺乏准备。
- 你认为这个研究中教师的样本是此地区的一个随机样本吗?
 - 什么因素使得该样本不能够成为统计上的理想样本?
 - 导致 Wallingford 小学中选了这么多的男性是否是由于随机性?
 - 你认为下一年新的数学计划能够被很好地接受吗?解释你的回答。
 - 你认为主要是由于研究者的懒惰而不能使样本被随机选择吗?请解释原因。
- 2.34 一名班级纪念戒指公司的推销员想在你的中学调查一般的学生愿为班级纪念戒指花

多少钱。校长建议他在周五下午冠军争夺赛时调查啦啦队的学生。作为一名好管闲事的统计学学生,你要为这个调查略尽绵薄之力。

a. 对于收集数据的过程,你有什么建议?

b. 如果校长的计划被采纳的话,你认为将会给调查带来怎样的结果?

- 2.35 每当我们收集到数据,再用数字或叙述的方式在图或表中概括这些数据时,我们就丢失了信息。以下两段话是从报纸和科学报告中摘录的各种叙述,请指出有什么重要的信息被丢失了。

a. “在这个调查中,伊利诺斯理工学院(Illinois Institute of Technology)的学生最有可能说他们‘不高兴’。”(The Philadelphia Inquirer, October 10, 1993, P. B5.)

b. “根据 Alan Guttmacher 研究协会(Alan Guttmacher Institute)的报告,比起其它信仰的妇女,有较多的天主教妇女选择了堕胎。”

分析(习题 2.36 - 2.41)

- 2.36 在地方选举前夕,一家地方报纸报道说,一项选民的调查发现 Rainwater 将在市议会的选举中以 53% 对 47% 领先于 Goldthorp, 抽样误差为 4%。主编想知道她是否应该把第二天报纸的标题定为“Rainwater 大胜 Goldthorp”。你的建议是什么? 为什么?

- 2.37 下面的问题是为了调查最近在地方影院上演的电影的受欢迎程度而设计的,调查对象是去电影院的人。列出这个问题中存在的至少十个不足之处。

姓名 _____ 年龄 _____

地址 _____ 电话 _____

薪水 _____ 工作名称 _____

今天晚上你看的电影 _____

电影院的名字 _____

电影好看吗? 很好 _____ 好 _____ 不好 _____

用十分制为电影打分: 1 2 3 4 5 6 7 8 9 10

比起你看的上一场电影怎么样? 1 2 3 4 5

比起电影“Some Like It Hot”怎么样? 1 2 3 4 5

你最喜欢这部电影的什么地方?

是演员吗? _____ 是 _____ 否 _____

你买玉米花了吗? _____ 买汽水了吗? _____ 买甜饼了吗? _____

你开什么牌子的车? _____

- 2.38 在 20 世纪 90 年代,领养是一个非常重要的个人和社会问题。在过去的十年中,主管领养工作的社会部门坚持的一个政策就是,使父母和孩子的种族尽量相似。这个政策和二十年前的政策正好相反,那时鼓励不同种族间的领养。围绕家庭的社会观念形成了这两种政策的基础。假定该社会部门想做一个调查以估计社区的态度是支持还是反对恢复不同种族间的领养。要特别注意访员和被调查人之间的关系问题。请回答以下问题:

a. 你如何设计一项研究,在提出问题时尽量减少由于种族原因而产生的反应?

b. 你怎样认为种族相似和差异(例如在访员和被调查者之间)在影响研究的回答中可能扮演怎样的角色?

c. 假定有充分的资源、权力(authority)和时间,这项研究者最好怎样开展?

2.39 假定你和某一机构有联系,你将要做一个调查(为你自己,或为你的团体,或为你的上司)。你要问几个问题(也许不超过 10 个),以确定这个机构中的人对影响到他们的某个指定的政策、领导、活动或最近的变化有怎样的态度。

a. 确定了你调查的目标以后,考虑怎样开展调查:当面询问、电话调查、匿名调查、声音邮件、电子邮件或其它。

b. 对你的研究方案写一个具体的计划,包括提问方式、要问的问题和分析问题的方法。

c. 进行设计时,你面临了什么样的困境?你打算怎样处理它们?在设计中你认为你的设计有哪些欠缺?

d. 如果你有很多的资源、更多的权力和时间,那么将会有些什么不同?

2.40 从下面的信息中产生一个数据阵。一队童子军正在进行一次在外面过夜的野营。晚上,孩子们开始谈论起他们的家庭。Chris, 9 岁,有三个兄弟,三个姐妹,和父母一起住,有一个宠物沙鼠;Andy, 10 岁,没有兄弟姐妹,和父母一起住,有一只狗;Carl, 9 岁,有一个异(父)母兄弟 Sam,和父亲、继母一起住,有一只名为 Sam 的猫;Greg, 10 岁,有一个姐妹,一个异(父)母姐妹和一个异(父)母兄弟,和父母一起住,有一只名为 Rex 的狗;Alex, 8 岁,和祖父祖母一起住;Paul, 11 岁,有四个兄弟姐妹,和母亲、继父一起住,有一条名为 Wanda 的鱼。

a. 要定义一个关于兄弟姐妹的变量,你需做哪些决定?如果变量是关于父母呢?或宠物呢?你怎样使它们不同?

b. 如果你要整理一大堆童子军的数据,在兄弟姐妹、父母和宠物的变量中你愿选择哪一个?

c. 如果你仅对离婚、再婚和单亲家庭感兴趣,你如何设计数据阵?

d. 在生成数据阵时,你将舍弃哪些数据?为什么?

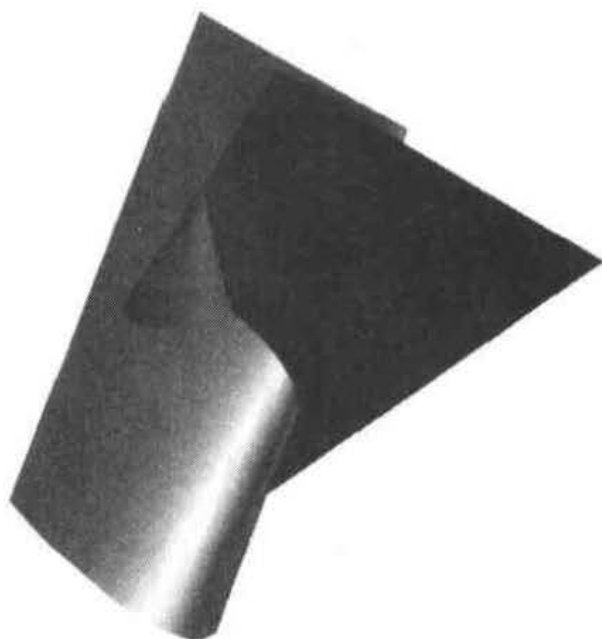
e. 如果你想研究“独生子和家中第一个孩子比其他男孩更容易参加童子军”这种可能性,上面是否丢失了你认为有用的信息呢?

2.41 1789 年,在马萨诸塞州,男性的平均寿命是 34.5 岁,女性是 36.5 岁;1850 年,男性是 38.3 岁,女性是 40.5 岁;1890 年,男性是 42.5 岁,女性是 46.6 岁;1910 年,男性是 54.0 岁,女性是 56.6 岁;1930 年,男性是 59.3 岁,女性是 62.6 岁。

a. 生成一个数据阵,把数字放到适当的行和列,使数据阵易于理解,并易于做统计分析。

b. 指出由这个数据中得出的两个显然的结论。

C H A P T E R 3



- 3.1 图:画出数据
- 3.2 分类变量:圆饼图和条形图
- 3.3 度量变量:点图和直方图
- 3.4 根据数据作地图
- 3.5 作图:优秀的标准
- 3.6 表:改变排列方式可能更合适
- 3.7 小结

数据简化有一个重要的不利之处。从简化的数据中,我们不能够再恢复最初的观测数据。因此,当我们分析数据时,几乎总会丢失某些信息。

简化的收益也包含丢失信息,好的统计学家试图在这两个相互矛盾的考虑中寻求一种平衡。

在分析统计数据时,我们为达到两个相互冲突的目标——简化和完全——而矛盾。首先,我们想简化一组数据以发现其中包含的数据模式。我们想要强调重要信息而忽略“噪音”,但同时又不想丢失感兴趣的细节。一场足球比赛可简化为最后的得分,但是这个数据并没有描述比赛过程是如何进行的。这种简化和害怕丢失细节之间的矛盾很难解决。幸运的是,实际的考虑为产生信息采取有用的形式提供了指导。我们如何描述数据通常依赖于我们要进行怎样的分析——在哪里用它,由谁用,目的是什么。另外,我们必须使自己和同事们满意这样的判断,即什么是我们能提供的最好的统计画面。

3.1 图:画出数据

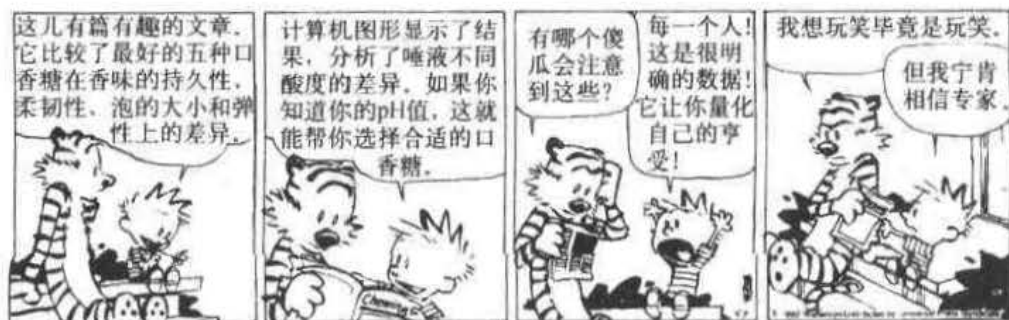
分析数据的一种方法是把它们画出来。图中包含的信息极多,因为大量数据都能概括在图中,并且一眼就能被理解。套用一句俗语就是,一幅图胜过一千个字。

作图有两个主要目的:帮助研究者从数据中提取信息和帮助把信息传给其他人。

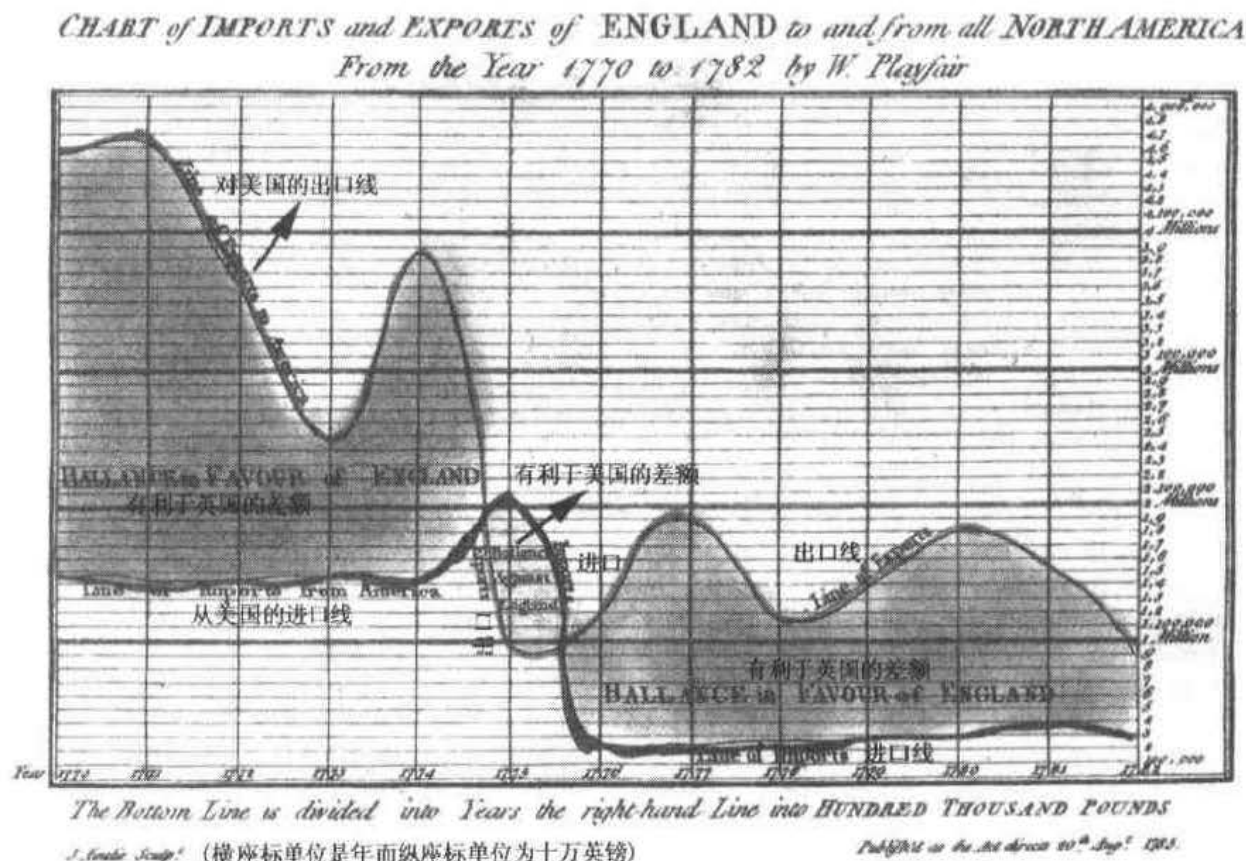
图在本质上是一种修辞工具,它是说服力的一种形式,首先是对研究者,然后是对其他人。作图是为了说明在数据中发现的特别的模式。根据某一特定的数据文件,可以作出许多其它的图,但通常只作很少几幅。我们只作那些看起来可以帮助分析家对数据进行理解和交流的图。因为有如此多的统计方法,在作结论时,有可能(有意或无意)误用了图。我们想使你能够区别好图和坏图。知道这种区别有助于防止你选择不好的图,而得出错误的结论。

生成统计图

统计图已经有二百多年的历史了。但是,出现图的历史要远远晚于其它重要的数学发现。在刚开始时,图非常罕见,而且是由人工画出来的,一般很不精确。今天,计算机软件已使得作图工作简易和准确多了,已很少有专业的研究员手工画图了。



图可能会非常重要。 (“Calvin and Hobbes” copyright 1992 Watterson. Dist. by Universal Press Syndicate. Reprinted with permission. All rights reserved.)



(图中的标题为:“1770 年到 1782 年英国和北美之间的进出口图,作者 W. Playfair”)最早的图之一——如此罕见。(来源:Edward T. Tufte, *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press, 1983.)

图形设计的计算机化有好处,也有坏处。通过计算机软件作图,图的形式在许多方面自动由编写此软件的人指定,研究者们发现很容易依赖于他们。但是如果计算机程序不够好,就会产生坏图。“坏”图意味着什么,我们在整个这一章中都有详细的描述。

在大众媒体中,统计图越来越普遍了。从计算机中打印出来的图,出现在报纸、新闻杂志和电视中。由于大众媒介充满了用图代表的信息,所以要求消费者对于图的制作有更多的了解。看懂图是 21 世纪成年人所必须具有的一种能力。

图的种类

在 3.2、3.3 和 3.4 节中,我们将讨论较常见的几种图,并向你介绍它们各自的优缺点。在 3.5 节中,我们介绍制图的原则。这些原则可用来判断一幅图是好,还是坏。

最简单的一类图是只根据一个变量概括数据,例如根据性别、年龄或智商。这样的图只包含数据文件中的一列。较复杂的图根据两个变量概括数据,包含数据的两列,如性别和年龄这两个变量。根据三个或多个变量作图较困难,但这并不是不可能的。

许多图用来显示某一变量的每一个取值中包含的观测数。例如,一张图可能是描述上个月有多少雨天和多少晴天。这张图通过显示两个值(雨天,晴天)哪个更经常发生来进行比较。

其它图则显示取值为基于一个尺度(量纲)的数量的变量。以年为单位的年龄和以一千美元为单位的收入是这种变量的两个较简单的例子。

3.2 分类变量:圆饼图和条形图

对于性别变量,其取值为男性和女性,它的任两个观测值或者相同,或者不同。这种变量叫做 **分类变量**(categorical variable)。

分类变量是指它的任两个观测值或者相同,或者不同。观测值不能够被排序,一个观测值并不比另一个多什么。

为一个分类变量作图

一个分类变量的数据分析,第一步通常是数每一个取值中包含的观测数目。例如,我们可考察有关在马萨诸塞州 New Bedford 地方法庭判决的 72 名犯人的数据。我们想知道在他们服完刑一年到两年半的时间里,他们是否又因新的罪行被判决。(来源: *The New York Times*, October 6, 1993, p. B10.) 当计算这群罪犯的观察值时,我们发现 24 个犯了新罪,而剩下的 48 个在数据收集时还没有。图 3.1 表示了描述这 72 个罪犯数据的一个圆饼图和两个不同的条形图。

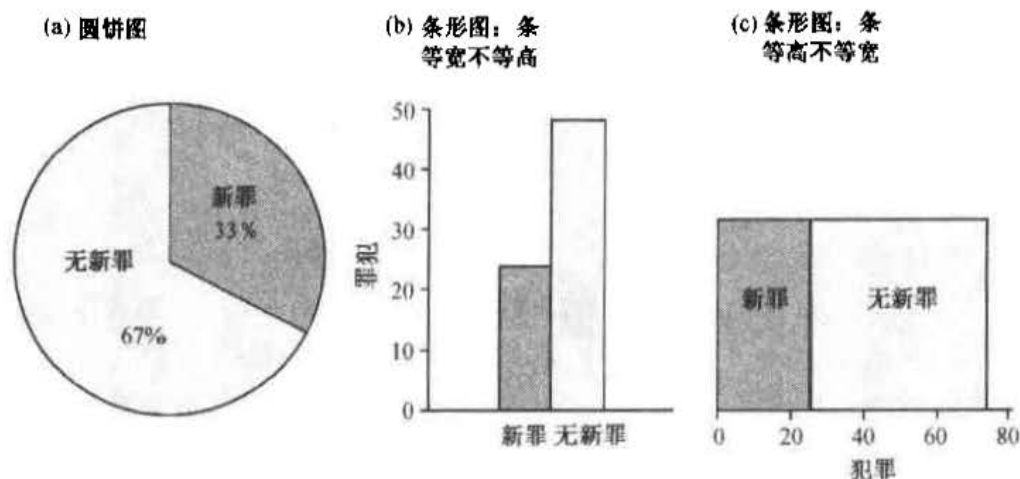


图 3.1 由分为两类(当罪犯服完刑一年到两年半的时间内是否又犯了新罪)的一个变量(犯罪)作的圆饼图和条形图。

圆饼图(pie chart) 圆饼图(图 3.1a)表明大约 $1/3$ 的犯人因新的罪行被判决, $2/3$ 没有。然而我们很难说,图被分成的 $1/3$ 和 $2/3$ 两个部分是精确的,它只是使我们迅速知道一组人大约是另外一组人的两倍。

圆饼图易于显示每一组的相对大小,可以在一个圆饼图中画出几个不同的组并进行比较。用圆饼图表示分类变量尤其好,因为它们的取值没有次序。圆饼的一小块可以被移到其它的位置而不改变图的含义,并且,相邻的组可以很容易地被合并。

圆饼图用于显示每一组有多少个观测数时不是很好。如果 240 个犯人因新的罪行被判

决,480个没有,那么其圆饼图将和图3.1a中的圆饼图一样。并且,圆饼图在组较多时就不是很有用了:分的块如此多,每一块又如此小,以至于圆饼图失去了原有的效果。

停下来想一想 3.1

举出你所熟悉的一组分类变量,根据其取值的观测数画圆饼图。例如,你可以画出你和你的朋友一天中接的电话数。所画出的圆饼图是否方便地表达了你的变量?什么可使圆饼图更好或更坏?

条形图(bar graph) 两个条形图(图3.1b和c)显示了犯人同样的信息。图3.1b的条形图——这里的条有相同的宽度,其高度代表变量取值中包含的观测数——是最普遍的。但是,也可以使用图3.1c中的条形图,这里的条都有相同的高度,其宽度代表变量取值中包含的观测数。注意到,在这两个条形图中,每一个条的值都从0开始,然而有时也不是这样,此时条形图就代表了不同的含义(见“停下来想一想3.2”)。

图3.1b中的条形图易于显示变量每一个取值中的观测数,但用于显示总的观测数时却不好;在脑海中把一个条放到另一个的上面去看总数是非常笨的方法。图3.1c中的条形图易于显示整个的观测数和变量的第一个类别(又犯了新罪的犯人)的观测数,而用于显示其它类别(没有犯新罪的犯人)的观数时则不好。分类变量的值越多,等高不等宽的条形图就会越复杂。

停下来想一想 3.2

一个在纵轴上不从0开始的条形图,可以怎样被一名有技巧的政治家用来夸大竞争对手的加税计划?

为两个分类变量作图

继续讨论上面提到的72名犯人。马萨诸塞州 New Bedford 地方法庭的法官 Robert Kane, 在马萨诸塞大学 Dartmouth 学院的 Robert P. Waxler 教授的鼓励下,让在他的法庭上被判罪的犯人选择进监狱或上由罗伯特教授教的文学课。印地安那大学的 G. Roger Jarjoura 教授跟踪调查了选择听课的32人,发现以后其中又有6人犯了新罪;而选择去监狱的40人中,18人在被释放后又犯了新罪。(来源: The New York Times, October 6, 1993, P. B10.)

现在我们关于24人又犯了新罪,而48人没有再犯这件事,又知道了更多的细节。我们有了关于第二个分类变量,即听课或进监狱的数据。图3.2显示了三种不同的条形图,用来表示两个变量的故事。

在图3.2a中,两个竖条分别代表去听课的犯人和进监狱的犯人。每个竖条分成两部分,再次犯罪的和没有再次犯罪的。此图很清楚地表明,听课的犯人中再次犯罪的较少,尽管我们很难看出来到底少了多少,因为那一部分不是从0开始的。而进监狱的犯人中,大约有一半人又犯罪了,另外一半则没有。

在图3.2b中,图3.2a的两个竖条中的上而部分被移到了横轴上,此图中我们很容易看到

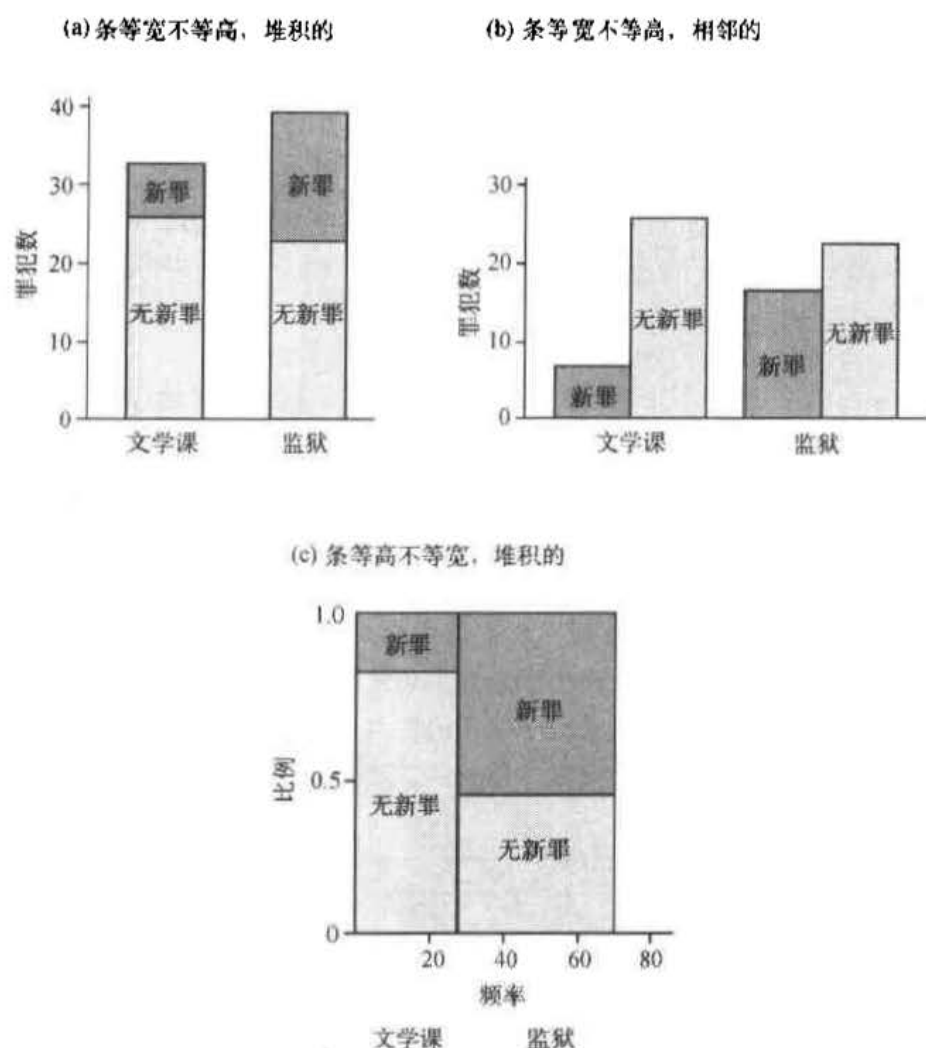


图 3.2 为两个变量(当上完文学课或坐完监狱以后犯/未犯新罪的罪犯)作的三种类型条形图。

又有多少人犯了新罪,因为这四个条都是从 0 开始的。但是,从此图很难看出来有多少犯人选择听课,多少犯人选择进监狱。

在图 3.2c 中,条等高但不等宽。条的宽度代表有多少犯人听课,多少犯人进监狱。每一条的阴影部分表明每一组中又犯新罪的人数。这是一种不同寻常的条形图,但是它比另两个图中包含了更多的信息。

所有这三个图都表明,听课的犯人比进监狱的犯人再次犯罪的比例少。尽管这项小研究很吸引人,但要想知道给犯人讲课是不是减少犯罪率的一种途径,却还有许多未回答的问题。无论如何,可能学费总比坐牢便宜。

3.3 度量变量:点图和直方图

度量变量是这样的变量,我们可根据它确定一个观测值是否和另外一个不同。我们还可以确定一个观测值比另外一个观测值多(还是少)一些,以及多多少(还是少多少)。

有一类变量我们可以用某一尺度度量其观测值。例如,我们可以用刻度为英寸的尺子测量一株植物的高度。这里我们使用的度量单位就是英寸。类似地,可用美元为单位度量一个家庭的收入;用年为单位计算一个人的年龄。像高度、收入、年龄这样可被测量的变量叫度量变量(metric variable)。度量变量中的度量不是度量系统的意思,而是指它的值可用数量表示。

因为可以收集度量变量的有含义的数值,所以可对它们进行算术运算,而对分类变量就不能进行运算。对度量变量的值可以进行加减乘除运算。

度量变量有时是区间或比例变量。本书不考虑区间和比例变量之间的区别。

停下来想一想 3.3

举一个度量变量的例子,并列它的一些取值。为什么这个变量是度量变量?

为一个度量变量作图

女性通常在多大年龄结婚?下面,是根据一家地方报纸列出的在一个星期之内申请结婚的女性的年龄(注意并不一定是第一次结婚):

30 27 56 40 30 26 31 24 23 25 29 33 29 22 33 29 46 25

34 19 23 23 44 29 30 25 23 60 25 27 37 24 22 27 31 24 26

这些数字告诉了我们什么呢?很容易发现其中最小的女性是19岁,最大的是60岁,有几个在二十多岁。但是除了这些,我们对这37位女性的年龄很难再有其它概念。事实上,观测数越多,不靠深入的分析来理解这些数据就越困难。对于年龄这样的度量变量,有几种不同的图可帮助我们更好地理解数据。图3.3显示了其中的四种类型。

点线图(lineplot) 当观测数较少时,例如上面的数据,就可以用点线图来帮助理解数据。在图3.3a中,线代表变量;线上标明的是变量的取值;线上方的每一个点代表一个观测值。点线图清楚地告诉我们,大部分女性在她们二十五岁到三十岁出头的年龄段结婚,少数在35岁到60岁时结婚。

点线图的优点之一是可以直接告诉我们,观测的分布怎样根据变量的取值变化。我们可以看到哪些地方观测密集,哪些地方观测稀少,因而年龄的数据模式看得很清楚,并且变量最初的值也没有被丢失。因此,尽管点线图简化了数据,却没有损失任何信息。

对于一个变量,当不同的观测值很多时,点线图就变得混乱起来。例如,每小时薪水的标

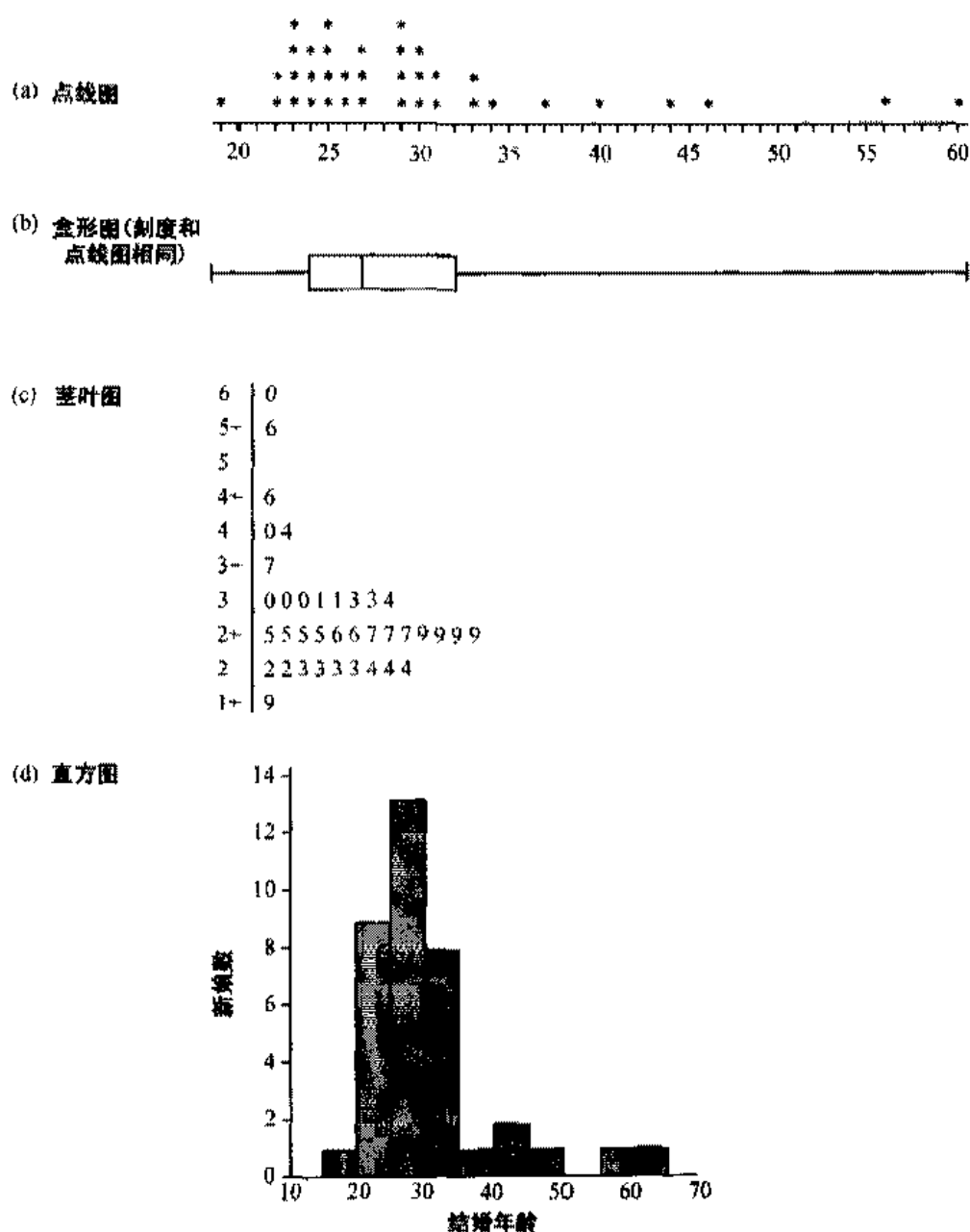


图 3.3 关于一个度量变量(女性的结婚年龄)的四种图。

准就有很多值,从婴儿看护者每小时的\$5到摇滚乐歌星每小时的演出费\$500000。类似地,大量观测——如整个一年中所有新娘的年龄——将使线上的点又多又乱。这时候,点线图看起来更像蚂蚁窝,而不是统计工具。因此,较大的数据集用其它的画图方法要比用点线图好得多。

盒形图(Boxplot) 图 3.3b 是关于女性结婚年龄的盒形图,此图和点线图有相同的刻度。盒形图在通俗出版物中并不常见,但是常出现专业性刊物上。为理解并制作盒形图,我们需要掌握较多的知识。

图 3.3b 中的盒形图的线是从 19——最小的新娘的年龄——开始,在一个矩形的盒子处

停止,盒形图的名字就源于此矩形。盒子的中间有一条垂直的线。盒子后面的线则一直延伸到 60——最老的新娘的年龄。在盒形图中,从线的最小值到盒子的开始包含了 $1/4$ 的观测 ($37 \div 4 \approx 9$),从盒子的开始(刻度为 24)到盒子中间的线(刻度为 27)包含了 $1/4$ 的观测,从这条线到盒子结束(刻度为 32)包含了 $1/4$ 的观测,从盒子结束到线的结束(刻度为 60)则包含了剩余的 $1/4$ 的观测。因此,恰有一半的数据位于盒子所在的区域。

停下来想一想 3.4

下面十三个美国最大的货币市场基金的收益率的盒形图是什么样子?

货币市场基金	收益率
Vanguard MMR/Prime Port	5.69
Schwab Value Advantage MF	5.66
Dean Witter/Active Assets MT	5.59
Fidelity Spartan MMF	5.50
Fidelity Cash Reserves	5.45
Dean Witter/Liquid Asset Fund	5.44
Kemper MMF/Money Market Port	5.40
Smith Barney Cash Port/Class A	5.29
Merrill Lynch Retirement Res MF	5.28
Merrill Lynch CMA Money Fund	5.26
Merrill Lynch Ready Assets	5.18
Dreyfus Liquid Assets	5.17
Prudential MoneyMart Assets	5.16

来源: *Money Fund Report*, USA Today, January 26, 1995, p. 3B.

看一看盒形图中的数据怎样帮助你决定投资哪个基金?

当最小或最大的观测值距盒子的距离比盒子本身的长度要好几倍时,盒形图中的线并不一定是从最小的观测开始并到最大的观测结束。在这种情况下,两端用点标上观测值即可。

盒形图是含有丰富信息的图。它们表明两端的值和中间的值的范围。在图 3.3b 中,中间一半新娘的年龄是 24 到 32 岁,其余一半新娘的年龄散布在除此之外的 19 到 60 岁之间——不失为有关女性结婚年龄的很好的一幅图。

盒形图在分析来自若干个组的数据时尤为有用。这时,每一组可作一个盒形图,然后再比较组与组之间的差别。图 3.19 中画出了来自七个地区的暴力犯罪数的条形图。从该图中,我们可以直接看到这几个地区的盒子中间的线的差异,以及哪些地区的盒子比其它地区的盒子大。根据盒形图,我们很容易看出哪些州暴力最多(比较盒子中间的线),哪些州暴力最分散(比较盒子的长度),哪些州最有可能同时包含极度安全和极度暴力(比较线的两个端点的位置)。

盒形图丢失了最初的数据并且不能够从图中恢复。然而同时,盒形图又为数据提供了简明有效的视图。

茎叶图(Stemplot) 女性结婚年龄的第三个图是茎叶图(图 3.3c)。顾名思义,此图的茎是指一条竖线,枝叶从茎的两边长出来。左边的枝叶代表年龄的第一位数字,右边的枝叶代表

年龄的第二位数字,处于该年龄的新娘有几个则列出几个。为清楚起见,每 10 年被分成了两部分;例如,2 代表年龄 20~24 岁,2+ 代表年龄 25~29 岁。

图 3.3c 的茎叶图中,最小的新娘 19 岁,有两个新娘 22 岁,有四个新娘 23 岁,等等。注意初始数据在茎叶图中被保留下来了。同时,观测根据变量的取值范围形成的分布很清楚。大部分新娘的年龄在 25~29 岁之间,在这些人中,大部分在 30 岁以前结婚。

当变量的观测数很多时,茎叶图的效果就不是很好了。因为每一个观测都会在图中占一个空间,而如果有许多观测,你可以想象枝叶该有多长!

停下来想一想 3.5

画出你的家庭成员和朋友的年龄的茎叶图。在茎的左边,用 0 代表 0~9 岁,1 代表 10~19 岁,2 代表 20~29 岁等。关于你认识的这些人的年龄的聚集程度,茎叶图告诉你什么?

直方图(Histogram) 直方图是根据度量变量的取值范围来显示观测数的最常用的图。为作直方图,变量的取值被分成了区间,通常有相同的长度(但也并不总是这样),然后,每一区间内的观测数用矩形来表示,矩形的面积代表观测的数目。(当所有区间有相同宽度时,如图 3.3d 中所示,矩形的高度就表明了观测数。但是,知道其实是矩形的面积代表观测数是非常重要的。)

直方图一词的起源

单词“histogram”(直方图)看来第一次被使用是在 1895 年,由伟大的英国统计学家 Karl Pearson 在脚注里作了定义。在他给伦敦的皇家协会发表的讲话中,Pearson 提到一些关于在 1885-1886 年英格兰和威尔士地区的房地产估价的数据:

到观测能被画成理论曲线为止,看来是不能指望有什么结果的。然而直方图*却显示了曲线末端的偏差数量。

* 在作者的统计讲座中曾被提到,指一种图示的一般形式,即在横轴上画出若干柱形,用柱的面积表示发生在柱的底边范围内观测的频率。(来源:K. Pearson, “Contributions to the Mathematical Theory of Evolution. 2. Skewed Variations in Homogeneous Material,” Philosophical Transactions of the Royal Society of London (A), vol. 186 (1895), part 1, P. 399.)

皮尔森没有解释他为什么使用这个特殊的术语。

比较语言学家(也是著者的儿子)Eric Iversen 给出了单词“histogram”一个学术上的解释:

又:histogram 从词的后面开始,可有一个比较简单的解释。“gram”当然是指图或代表什么东西,如 pictogram(彩色照片)——涂色的画像;telegram(电报)——来自远方的话语;epigram(警句)——和某种东西有关的话语。你可以看到“gram”可以代表话语或图像,其意义依赖于你怎样使用它。但是“hist”却更有趣。由于它有多种含义,因此可能有人会误解它,像“history(历史)”中,是希腊语词根“historia”;而“histology(组织学)”——关于人体组织的研究中,就是拉丁语词根“hist”,表示连接或组织。但是“histogram”中的“histo”是希腊语,表示桅杆或大梁。我想,之所以用“histogram”,仅仅是因为图中的列看起来像船上的桅杆或大梁。

在图 3.3d 的直方图中,数目最多的新娘在区间 25 ~ 30 岁之间,因为那个矩形是最大的。区间 20 ~ 25 岁和 30 ~ 35 岁之间也有一些新娘。其他新娘相当平均也相当稀疏地分布在其余年龄的区间之中。

图 3.3d 的直方图看起来非常像图 3.3c 中的茎叶图。茎叶图的每一行对应直方图的一个矩形。因为直方图表明了分布的形状,因此可以简化数据,并使我们看到在列表中看不清楚的数据模式。但同时,直方图也丢失了信息。原始数据在直方图中不能像在茎叶图中那样恢复出来。在茎叶图中,原始观测在图中由它实际的值表示,而在直方图中,每一个观测仅由矩形的一部分来表示。

因此,直方图对于简化大量观测很有用——每一个观测仅占矩形的一小部分。例如,数目是原来 10 倍的新娘,即 370 个新娘,10 个在区间 15 ~ 20 岁,90 个在区间 20 ~ 25 岁,依次类推。这样的直方图和图 3.3d 中的直方图看起来没什么区别。不同之处仅在于,纵轴刻度上的频率是 20、40、60 等而不是 2、4、6 等。然而,很难想象有 370 个观测值的茎叶图是什么样子。直方图能够很容易地表示大量数据。

我们对直方图主要的兴趣在于它们各种各样的形状。图 3.3d 中的直方图是**单峰的**(unimodal),之所以如此说是因为它仅有一个顶峰。这种形状告诉我们观测中有一组是主要的。直方图的形状也可能是**对称的**(symmetric),即左半边的分布是右半边分布的镜像。新娘的直方图不呈现出对称分布,因为右边的尾巴比左边的长。这个直方图是**不对称的**(skewed)。

许多变量,像身体特征和考试成绩,经常呈现出单峰且对称的分布。单峰对称直方图告诉我们,大部分观测集中在变量的中间值附近,只有很少的观测非常大或非常小。在现实生活中,大部分孩子都很普通,只有个别孩子优秀,不像在美国的广播节目 Lake Wobegon 中,所有

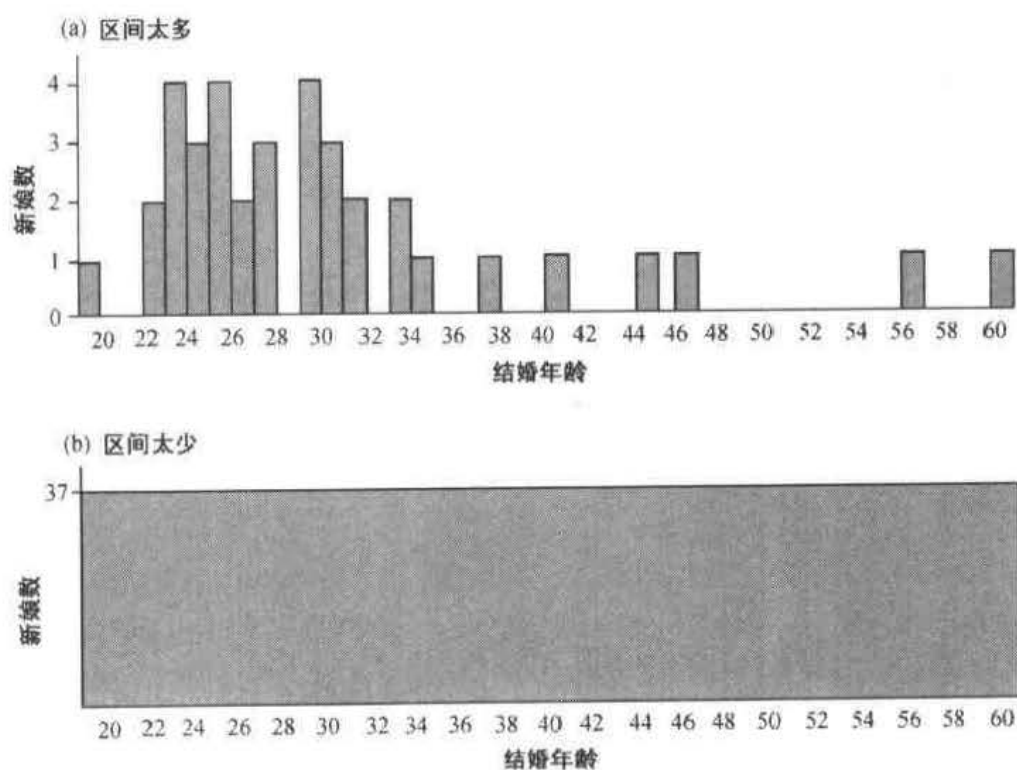


图 3.4 有太多或太少区间的直方图(女性的结婚年龄)。

孩子都优秀。“平均数”经常是变量的一个令人感兴趣的特征,在直方图中它表现得相当明显。

双峰(bimodal) 直方图有两个顶峰。为描述这种形状,想象这样的一个直方图,它表示一个包含富人和穷人而很少有中等收入的社区中人们收入的分布。直方图的形状将显示两个顶峰,告诉我们这是一个两极化的团体。

直方图的形状取决于它如何被构造,不同的形状给人以不同的印象。首先,把变量分成区间时,水平轴上区间的个数将影响直方图的形状。如果变量被分成了很多区间,则每一个区间将包含很少的观测,此时直方图看起来参差不齐(图 3.4a)。如果所有的观测在一个区间显

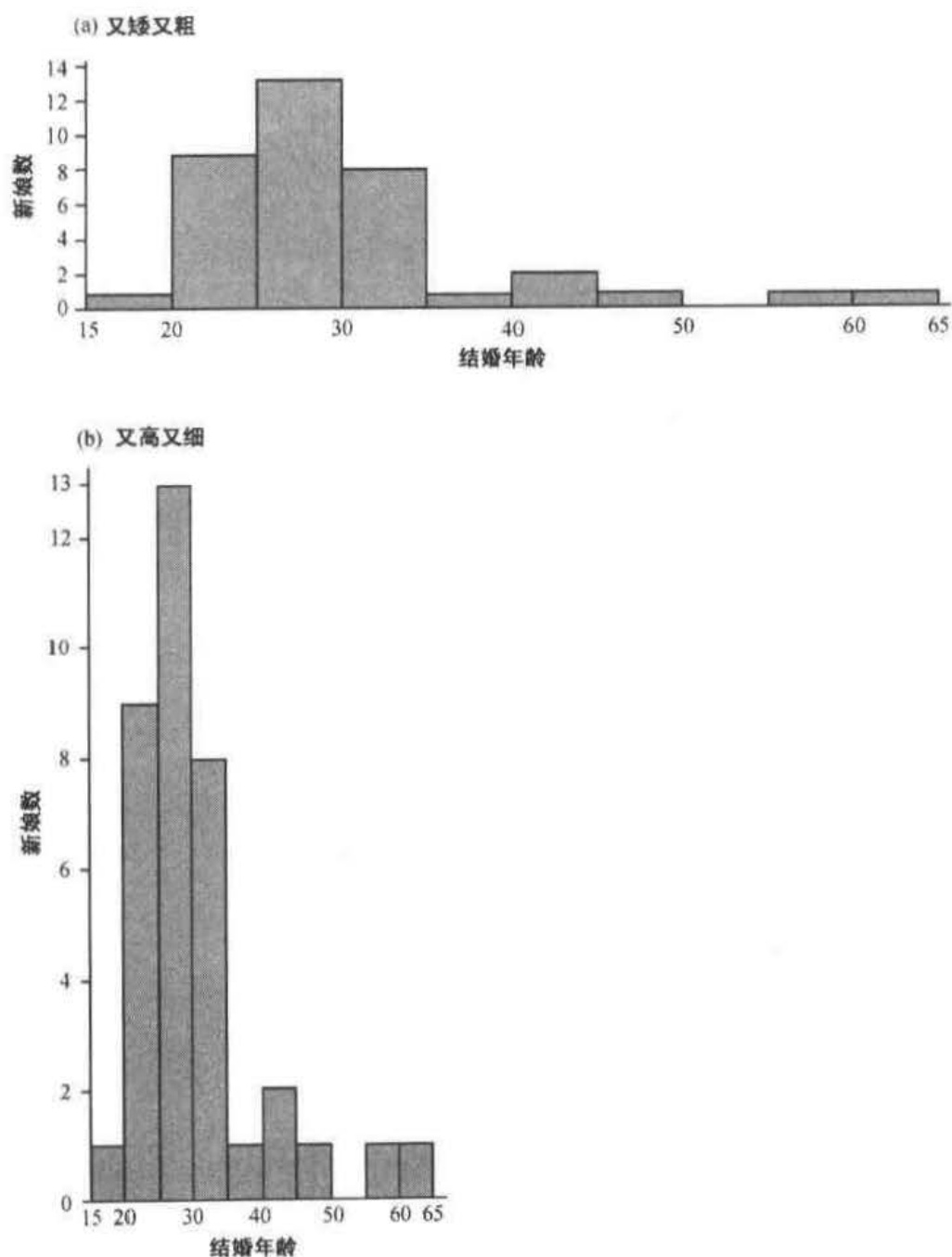


图 3.5 宽的和高的直方图(女性的结婚年龄)。

示,则直方图仅包含一个大的矩形(图 3.4b)——这不是一个有用的直方图。图 3.3d 中的直方图位于两个极端之间,它更富含信息,并且比图 3.4 中的两个直方图更吸引人。但是我们必须注意:如果一个直方图只有很少的区间,且是单峰形状,则此图可能掩盖了这样的事实,即多分几个区间可能会表明数据是双峰形状。

其次,如果矩形由又高又细变成了又矮又粗,那么直方图的形状就会改变。图 3.5 中的直方图用相同的区间表示女性的结婚年龄。但是图 3.5a 中的直方图又矮又粗,而图 3.5b 中的又高又细。由于图 3.5a 中矩形高度的差异是如此小,以至于各区间中观测数的差异也显得小了。图 3.5b 中的直方图则表示了相反的情况。

当然,不管是什么形状,这两个直方图是等价的,它们对于年龄变量包含了相同的内容。但是我们应该注意矩形又高又细或又矮又粗的直方图。这种图的制作者可能想使人们产生某种数据中实际上并不存在的印象。

停下来想一想 3.6

想象你要根据下面墨西哥 1930 年到 1990 年的人口数据作直方图。变量是从 1930 年开始,每五年中大于 12 岁的女性所生的孩子的平均数。

5.0 4.8 4.6 4.6 4.6 4.6 4.5 4.5 4.5 4.0 3.4 2.8 2.5

来源:Adapted from Zavala de Cosío (1992) in Matthew C. Gutmann, "The meanings of Macho: Changing Mexican male identities," *Masculinities*, vol. 2 (1994), P. 29.

如果你想要强调较高数字和较低数值之间的差异,你应该怎样作直方图? 如果你想尽量不强调这个差异,你应该怎样作图? 两个直方图是不对称的吗? 如果是,以何种方式偏斜?

为两个度量变量作图

统计学家们常常需要显示两个度量变量的数据。例如,新娘的年龄和新郎的年龄,或者人们的身高和体重、年龄和收入、SAT 分数和平均分数、国内非文盲率和国民生产总值等。根据两个变量显示数据最常见的方法是散点图(scatterplot)。

散点图包括两个轴,横轴和纵轴。横轴(即数学上的 x 轴)代表一个变量(如新郎的年龄),纵轴(即数学上的 y 轴)代表另一个变量(如新娘的年龄)。两个变量的一对观测值在图中用点来表示。例如,如果新郎 37 岁,新娘 30 岁,则在图中,从 x 轴的 37 这一点作一条想象中的竖线,从 y 轴的 30 这一点作一条想象中的横线,其交点即代表这一对观测值。

图 3.6 是 37 对新婚夫妇年龄数据的散点图(注意:图中的点并不够 37 个,因为有一些夫妇中新郎和新娘的年龄都相同;这些对夫妇就用一个点来表示)。观察原始数据,除了能看出年龄大的新郎配年龄大的新娘外,很难再看出其它模式。当年龄在散点图中用点表示出来以后,两个变量之间的关系就很清楚了。点从左下角开始,大体地延伸到右上角:大部分是年轻新郎配年轻新娘,年老新郎配年老新娘。从左下角到右上角的点的轨迹,用数学语言来说,具有正的斜率,它表明两个变量之间正相关。另外,这些点还表明,在某些夫妇中,新郎的年龄比新娘的大,而在其他夫妇中,新娘的年龄则比新郎的大。

在散点图中,没有丢失任何数字信息并简化了数据。散点图很容易进行制作和解释。

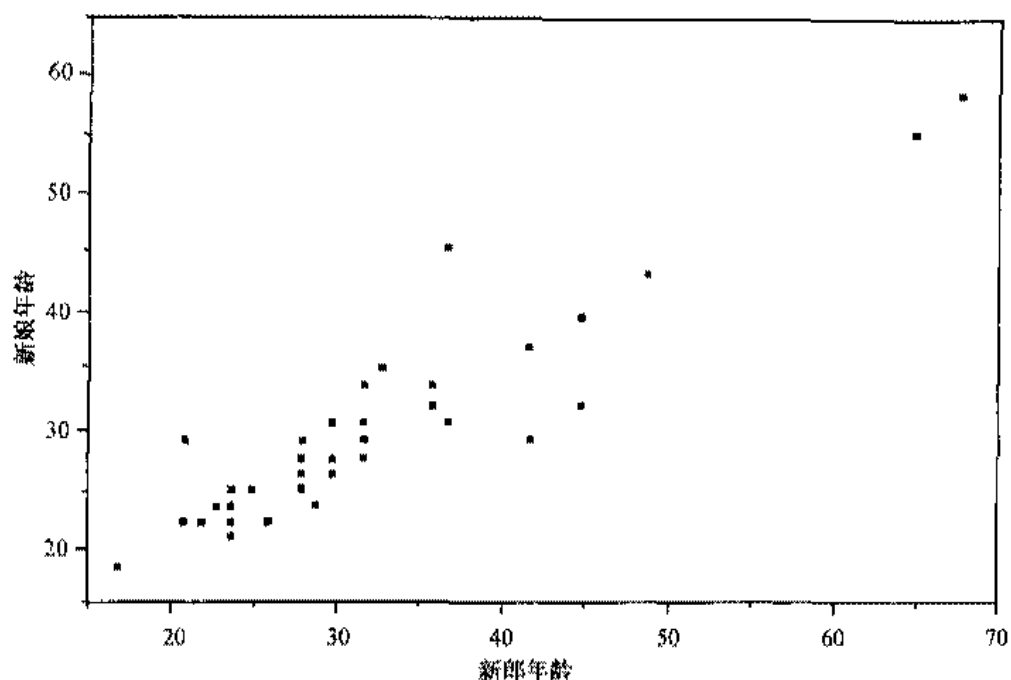


图 3 6 显示新娘和新郎年龄的散点图。

停下来想一想 3.7

根据下面 24 对夫妇的数据,变量为结婚年数和每年的争吵次数,你如何作散点图?

结婚年数	争吵次数	结婚年数	争吵次数
5	10	10	5
2	20	15	3
4	16	13	4
1	15	20	2
3	9	16	4
6	6	25	1
5	8	22	3
8	5	14	3
3	10	15	4
7	7	19	3
3	8	17	3
9	6	20	2

关于婚姻生活的特点,你能得出什么结论? 因为缺少什么信息,可能使你不能根据这些有关婚姻幸福的数据做进一步的推广?

时间序列图

变量常常包含着在一段时间内收集的数据。过去四十年的消费者价格指数是一个具有时间序列数据的变量,类似的还有二战后每年从日本进口的商品总值,1940年后棒球比赛的平均时间,1932年后裙边的平均长度等等。时间序列数据的图是一种特殊的散点图。时间作为

一个变量在横轴上被标出,另外一个变量以纵轴表示。图中的点不像结婚年龄散点图中的点那样分散,因为横轴上时间的取值通常是均匀分布的,并且横轴上的每一个值只对应纵轴上变量的一个取值。

表 3.1 1900—1936 年奥林匹克男子跳高比赛的金牌获得者跳的高度

年数	跳高高度(英寸)
1900	74.8
1904	71.0
1908	75.0
1912	76.0
1920	76.2
1924	78.0
1928	76.4
1932	77.6
1936	79.9

在横轴上被标出,另外一个变量以纵轴表示。图中的点不像结婚年龄散点图中的点那样分散,因为横轴上时间的取值通常是均匀分布的,并且横轴上的每一个值只对应纵轴上变量的一个取值。

利用表 3.1 中的数据文件,图 3.7^① 中的散点图表明了从 1900 年直到第二次世界大战,奥林匹克男子跳高比赛的金牌获得者跳的高度(以英寸为单位)。在这个数据文件中,数据有两列,表明我们要处理两个变量。从表中可知,此高度显然是增加了,图中很清楚地显示了这一点。除了 1904 年和 1928 年^②,每年都产生新的奥林匹克记录。当把 1900 年和 1904 年的点连接起来,1924 年和 1928 年^③ 的点连接起来时,线是下倾的;其余的连线则是向上延伸。(想象把此图一直延伸到现在,不仅表示年的轴要被延长,取胜高度的轴也要被延长;现在的跳高冠军认为跳 8 英尺——96 英寸——是小菜一碟。)

因为有其它许多种类的图,所以,改变时间序列图的形状并使数据给人以截然不同的印象是完全可能的。图 3.7 看起来可表明每次奥林匹克跳高冠军所跳高度有相当大的改变。但是,应注意到纵轴是从 70 而不是从 0 开始,其长度也只有 10 英寸。从 70 开始是有原因的。想象把纵轴向下延伸到 0,同时保持图所在的位置及一英寸区间的刻度,那么一年和另一年的差异看起来就不如图 3.7 那样明显了。但是,这样的图并不实用,它是图 3.7 的八倍长,而底下的八分之七都是空的!

如果使该图保持图 3.7 那样大小,纵轴的刻度是从 0 到 80 英寸,而不是从 79 到 80 英寸,那么视觉效果就很不一样(图 3.8)。点之间连线的倾斜度远不如图 3.7 那样明显。图 3.8 表明了跳高高度总的增加趋势,但是不同奥林匹克运动会跳高冠军所跳高度之间的变化看起来并不是那样富有戏剧性。

① 原书此图有误,这里由译者重画——译者注。

② 原书此处误作 1932 年——译者注。

③ 原书此处误作 1928 年和 1932 年——译者注。

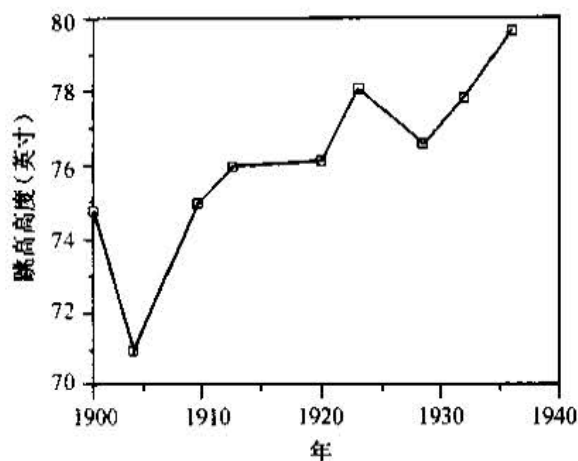


图 3.7 时间序列图,表示 1900—1936 年奥林匹克男子跳高冠军所跳的高度。

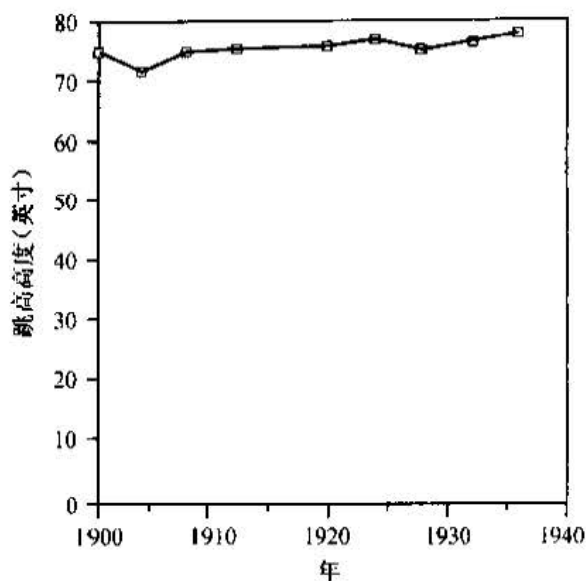


图 3.8 重画的时间序列图,显示 1900—1936 年奥林匹克男子跳高冠军所跳的高度。

显然,在根据数据得出结论以前,我们应该认真观察时间序列图,并想象如果改变作图方式,就像我们上面讨论的那样,它会变成什么样。一般来说,时间序列图具有散点图主要的优缺点:数据被简化的同时,保留了数字信息,但是图的形状可能会产生误导。

停下来想一想 3.8

从周一开始,每天给你的情绪状态打分,0 表示非常消极,7 表示非常积极。然后,用 x 轴表示日期,用 y 轴表示情绪分数,作出图来。于是一个时间序列图就产生了! 比较你的情绪分数,那么你可以对你整个一周的状态有更好的认识。也许你用不到整个数字范围,因为这种刻度通常的结果是,上半部分刻度比下半部分用得更多(见习题 3.33)。

3.4 根据数据作地图

地图不仅仅能够显示地理特征,像河流和山脉,而且能够显示统计信息。例如,美国的地图中,各个州可能会根据某些变量被涂上颜色,比如根据总统选举的投票方式等。图 3.9 中的地图是根据离婚率的大小来涂色的。

停下来想一想 3.9

关于离婚率在全国的变化,图 3.9 中的地图告诉了我们什么? 用地图显示数据会导致对数据的性质的多大程度的错误印象? 即这个地图较好还是较坏地代表了数据? 离婚率在全国的变化可能有些什么原因? 你能够想到其它像这样很好地代表数据的情况吗? 在什么条件下地图可能成为一种表示数据的非常好的方法?

地图在标明观众感兴趣的情况时可能会比较有用。(气象频道中的地图正是这样的例子!)例如,如果你考虑搬到加利福尼亚去,你可能会对哪个地方空气污染最严重感兴趣。此外,地图也用于研究地区趋势,像某种昆虫可能在这个地区泛滥,而在那个地区则没有。



图 3.9 显示美国的离婚率的分布的地图(每 1000 人)。(来源:Adapted from a map based on data from the National Center for Health Statistics by Paul O. Pugliese in Time, December 6, 1993, p. 23.)

美国大概有 3000 个县可以被列出来,涂上颜色以标明患癌率(每个县报告上来的癌症病人人数除以该县的人口数)。可确定低、中、高三类患癌率,并根据各个县的患癌率的水平涂上三种颜色中的一种。这样,一个地图将显示地域性模式。

尽管地图很有用,但它们可能产生误导。这是因为被涂色的是地理区域,而地理区域的大小可能变化很大。例如,一个具有高患癌率的东部小县在地图上将不如一个具有低患癌率的西部大县那么醒目;在新泽西一个具有高患癌率的地理上较小的县会影响到许多人,而在内华达州一个具有低患癌率的较大的县却只影响到少数人。

3.5 作图:优秀的标准

本章向你介绍一些标准的统计图。它们中的大部分是用计算机统计软件画的。点一下鼠标,以前需要花费很长时间的图现在可以在瞬间内完成并进行修改。制图者可以试验很多种形式,它们中的大部分都是公众所不熟悉的。每一种形式都可能在揭示数据的某一方面时有用,但同时又掩盖了其它特性。每一次视图上的新突破也产生了新的未知的弊端。为衡量一个图,我们必须对什么构成一个“好”图有一些概念。

对于“好”图和“坏”图的相当好的介绍可以在 Edward R. Tufte 的著作 *The Visual Display of Quantitative Information* (Cheshire, CT: Graphics Press, 1983) 中找到。Tufte 是数据的视觉展示专家;他使用术语 **图优性** (graphical excellency) 来描述一个“好”图。在他看来,一个好图是这样的图,即复杂的思想能够在图中清楚、精确、有效地被表达出来。

图优性是指图能够

在最短的时间内

用最少的笔墨

在最小的空间里

给观众最多的思想。

“最少的笔墨”:最简单的图是最好的吗?

图 3.10 和图 3.11 的数据都是 19 世纪以二十年为单位的美国人口规模 and 价格指数的关

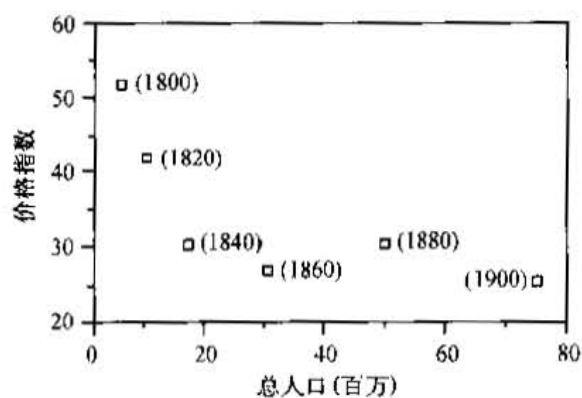


图 3.10 1800—1900 年人口和价格指数的散点图。(来源:U. S. Bureau of the Census, Historical Statistics of the United States, Colonial Times to 1970, its Bicentennial Edition, Part 1 (Washington, D. C.: U. S. Bureau of the Census, 1975). Population: Series A57 - 72, pp. 11 - 12; consumer price index: Series E135 - 166, p. 211.)

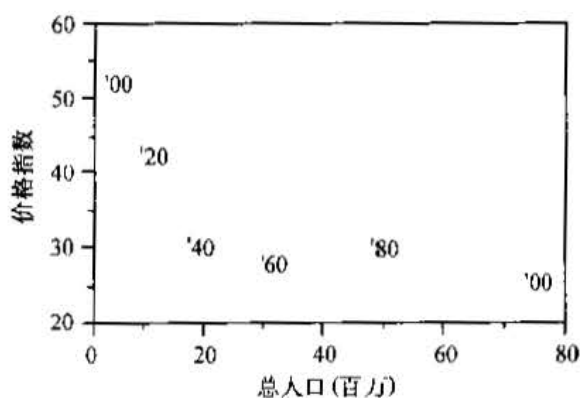


图 3.11 图 3.10 中的数据简化的散点图。(来源:U. S. Bureau of the Census, Historical Statistics of the United States, Colonial Times to 1970, its Bicentennial Edition, Part 1 (Washington, D. C.: U. S. Bureau of the Census, 1975). Population: Series A57 - 72, pp. 11 - 12; consumer price index: Series E135 - 166, p. 211.)

系。但是图 3.11 比图 3.10 用了更少的笔墨,它把时间作为点,而不是既标明了点又标明了时间(尽管图 3.10 中的点很富裕,它们没有必要在一个点外面再加上一个方块)。它去掉了年两边的括号,并且,由于题目讲数据是 19 世纪的,所以它把年的前两位数字也抛弃了(年的序列很清楚的表明左上角的'00 指 1800 年,右下角的'00 指 1900 年)。因此,图 3.11 比图 3.10 好,至少 Tufte 认为如此!

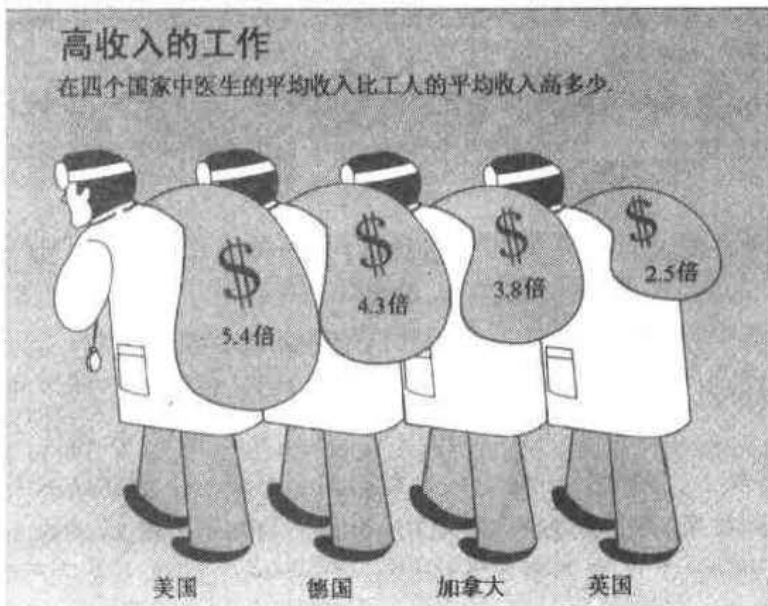
图 3.4 中新娘年龄的直方图用了太多的笔墨吗?是的,根据 Tufte 的观点。每个矩形的高度可由矩形的左边高度、右边高度、顶线的位置、阴影等任何一样来表明——有很多多余的线索。假定 Tufte 总算认可了关于矩形的想法,那么他也可能会争论阴影并不能给图增加什么东西。当然,没有阴影可以减少笔墨,但是这也使得图不清楚。阴影使得图显著地和背景分开了。没有阴影,矩形的内部和外部将有相同的颜色,图的意思——两个直方图的形状不一样——也将不那么明显。因此,多余的笔墨使得图易读了。

“图中垃圾”:垃圾的一种新名称

图中有时包含一些和要表现的数据不相干的特征——制图者添加的使图更吸引人或更有趣的特征。Tufte 把没有必要的特征,称为“图中垃圾”。图中垃圾包括矩形中的阴影,散点图中的格子,表示数量的生动的符号,以及点缀页边或图本身的说明。Tufte 在图形设计时的观点是基于“过犹不及”的假设。我们可能赞同,也可能不赞同;一个观众眼中的图中垃圾可能使图对另一个观众来说更易于理解。一个令人赏心悦目的、使人会心微笑的,或是激发一个人的好奇心或沮丧心理的图,可能不符合严格的统计图简单、有序的标准,但是它可能使观众更感兴趣。

停下来 想一想 3.10

根据 Tufte 的建议,把下图中多余的笔墨去掉。图的哪些部分是不必要的?哪些部分看起来对展示数据信息起了重要作用?想象你面对 Tufte 为下图辩护。



来源: Graph by Marty E. Mullins
From data of Stuart Altman, USA
Today, November 13, 1993, p. 1.

数据密度

图的目的是把信息带给观众。例如,图 3.8 给出了 9 次奥林匹克运动会的跳高冠军所跳高度和相应的举办时间,总共有 18 个数字。该图本身相当大,因此每一平方英寸中只有相当少的数字。图中每平方英寸的数字越多,数据的密度就越大,图也更富含信息。

高密度数据图的一个例子是报纸上每天的全国气象图。这个地图表明了 50 个州每个州的轮廓、温度、气压和降水量。另一个高密度数据图的例子是 *Consumer Reports* 上的汽车修理记录。

“复杂性的展示”

Tufte 在结束他的书时,讨论了一个著名的显示拿破仑的军队在 1812 年如何在俄国惨败的图(原书第 40 页;图 3.12 复制了该图):

[这]^① 是法国工程师 Charles Joseph Minard(1781-1870)的经典[图]。该图显示了拿破仑的军队在俄国可怕的命运……这幅于 1861 年制作的数据、地图和时间序列的混合图,描述了拿破仑 1812 年的俄罗斯战役的巨大损失,似乎要通过它不容置疑的雄辩力来推翻历史学家们的叙述。从左边接近 Niemen 河的波兰—俄罗斯的边界开始,深色的条表明了 1812 年侵入俄罗斯时的拿破仑的军队(422000 人),带的宽度[侵入时的军队用灰色表示,撤退时用黑色表示]表明了地图中每个地方的军队的大小……穿过 Berezina 河时是一场灾难,最后残存的军队到达波兰时仅剩了 10000 人……六个变量被标了出来:军队的大小,军队在二维平面中的位置,军队的挺进方向,从莫斯科撤退期间的不同日期的气温。

它有可能是有史以来最好的一幅统计图。

3.6 表:改变排列方式可能更合适

表是另外一种用密集的形式归纳数据的方法。表通常是由写在行和列中的数字组成。表经常用于表明有多少或多大比例的观测落入了不同的类别中,比如,一项教育研究中不同年龄的孩子。

表用于两种广泛的目的:一种是伴随文章以支持其中的观点;另一种是组织数据。报纸、杂志和书中的表通常是第一种类型的表,官方统计机构像人口统计局作出的表,通常属于第二种类型。支持文章观点的表必须能证明观点,而仅仅是表示数据的表必须易读易解释。

表 3.2 中的数据和图 3.1 中的数据一样,都是有关犯人中谁犯了新罪,谁没有犯新罪的数据。表提供的视觉效果和图很不一样。当我们把图 3.1 和表 3.2 放在一起观察,发现在图中比较不同类别的人数比在表中容易多了。两个圆饼块或长方形的大小不同一下子就显示出了两类的差别。而在表中,首先要看到数字,然后再在头脑中进行比较。为知道没犯新罪的人比犯新罪的人多多少,我们不得不用 48 减去 24 或者用 48 除以 24。

^① 本引文中所有方括号及其中文字是著者所加——译者注。

表 3.2 在罪犯服完刑一年到两年半的时间内,又犯了新罪和没有犯新罪的人数

犯了新罪	24
没有犯新罪	48
总共	72

该表直接表达的信息是实际的频率,它说明了有多少犯人又犯了新罪。而在图中,要想知道这个数字,我们必须从矩形的顶端在想象中向纵轴画一条线,还要依赖于轴上的数字有多精确和我们的视力有多好。即便这样,可能依然难以判断数字是 24 还是 26。因此,如果精确的数字很重要,那么表就比图好;而要想对数据有一个较快的印象,图就比表好。

表通常会有一个名称,并且行和列都会被清楚地标明意义。如果需要,行和列的数据应进行合计并标出。合计提供了行和列中的细节内容。如果表中仅包含数字的合计数,像表 3.2 中那样,那么数字应该垂直排列。当然,数字在表中也可以水平排列,但是这样就不容易给人留下总的频率的印象。尽管合计已经给出了,我们不必自己做加法运算,但我们还是习惯于把每一行的数字相加,把结果放在一列中,因为比较列中的数据比较行中的数据容易。当数据被罗列起来时,表中两个数字之间的差异就更明显了。

在 Myra Chapman 的书 *Plain Figures* (与 Basil Mahon 合写;伦敦:Her Majesty's Stationary Office, 1986)中,Myra Chapman 对怎样通过重新安排来提高表的效果给出了一个指导性的例子(表 3.3 和 3.4)。其中数据来自于英格兰和威尔士。

表 3.3 是典型的统计归纳。它显示的数据包含两个变量,内容是关于已经完成义务教育的学生。一个变量是时间,从 1973 年到 1980 年共七个学年;另外一个变量是义务教育结束以后学生的去向,即学习结束以后每组学生做了什么。除了最后一行,其它所有数据都是百分数。最后一行是以千人为单位的学生数。例如,1973—1974 列表明,701000 名学生毕业后的去向按五种类别划分后各占多少比例。除了因四舍五人可能造成的一点误差以外,每一列的比例相加应该都是 100%,但是表中并没有列出这些比例的和。(每一行的比例相加不是,也不应该是 100%。)

表 3.3 义务教育结束以后学生的目标

目标(%)	学年						
	'73 - '74	'74 - '75	'75 - '76	'76 - '77	'77 - '78	'78 - '79	'79 - '80
继续在学校	25.9	26.1	27.5	28.3	27.6	27.4	27.8
全日制或 非高等的进一步教育	9.7	11.5	13.6	13.6	14.1	14.3	14.1
参加工作							
业余时间学习	17.4	16.4	12.1	10.2	14.1	12.1	12.2
没有学习	44.1	41.7	38.0	37.7	34.4	38.7	38.7
没有工作	3.0	4.2	8.8	10.1	10.0	7.5	7.2
总的学生(=100%;千人)	701	723	744	746	773	801	814

表 3.3 的目的主要是为表明各个类别的比例如何随时间变化。除了其它的要清楚表达数

据的困难外,此表要求通过行比较百分数,而不是通过列看从一年到下一年发生了什么变化。

表 3.4 表 3.3 的变形和简化

学年	目标					总的学生	
	参加工作		继续 在学校	全日制或 非高等的 进一步教育	没有工作		
	没有学习	业余学习				百分数	千人
1973 - 1974	44	17	26	10	3	100	701
1974 - 1975	42	16	26	12	4	100	720
1975 - 1976	38	12	28	14	9	101	740
1976 - 1977	38	10	28	14	10	100	750
1977 - 1978	34	14	28	14	10	100	770
1978 - 1979	39	12	27	14	8	100	800
1979 - 1980	39	12	28	14	7	100	810

在表 3.4 中,数据被重新安置以帮助观众理解这些比例随时间的变化。也许最引人注目的改进是简化:所有的比例都通过取近似值而变成了两位数。这从表中去掉了 35 个数字和 35 个小数点(作图时用最少笔墨的原则看来也适用于表)。比例数现在不是很精确,但是对于此表的目的,完全精确并不重要。学生数本身就是取的近似值,而且此表的目的本来就是为了展示学生去向的变化趋势,而不是要给出每种去向中精确的学生数。

这两个表之间另外一个很大的差异是去向分类的顺序。在表 3.3 中,类别“继续在校念书”是第一个,而表 3.4 中类别“不学习而参加工作”是第一个。这样安排的原因是把较大的数字放在表的左上角,而把较小的数字放在右下角。这样的安排容易比较数字,尤其容易比较同一列中的数字。表 3.4 还表明了每一行的比例相加是 100%。

3.7 小结

3.1 图:画出数据

图是分析数据的一种极富有信息的方法,因为一个完整的数据集可以被概括在一个图中,并且一眼就能被理解。图帮助调查者从数据中提取出了有用的结果,并帮助其他人理解这些结果。

3.2 分类变量:圆饼图和条形图

分类变量是这样的变量,它的任两个观测值或者相同或者不同。表示分类变量的主要的图是圆饼图和条形图。如果类别不是很多,圆饼图很容易理解,但是每一类中的观测数经常无法显示出来。条形图很容易读懂,但是如果矩形是由不同的群构成,那么不同类别之间的细节就很难观察到。

3.3 度量变量:点图和直方图

度量变量要求有一个度量单位,以便说明一个值比另外一个大多少或小多少。点线图、盒形图、茎叶图和直方图可以用来显示单个度量变量。点线图在一条连续的线上表明一个小的数据阵;变量原始值在图中被保留下来。

茎叶图非常适用于一个小的数据集;但是对于数值变化范围较小的数据集就不是很有用了。盒形图表明变量的两个极值和中间值的范围。盒形图在比较来自不同组的同一变量的数据时很有帮助。直方图是用矩形的面积表示变量值的观测的相对数量。单峰直方图只有一个顶峰,而双峰直方图则有两个顶峰。直方图对于显示大量的观测数据比较有用。直方图的一个缺点是无法保持观测的原始值。

经常用来画两个变量的图是散点图和时间序列图。散点图有两个轴,对每一个个体都可以画出一个点;由两个变量的观测值组成的散点图表明了两个变量之间的关系模式。时间序列图在横轴上显示时间的值,时间的值通常是均匀分布的;在纵横上显示另外一个变量的值。

条形图和时间序列图可以表示多个变量。虽然多变量图允许比较大量信息,但是如果它们包含了太多不同的变量,它们可能很难被读懂。

3.4 根据数据作地图

用象征性手法将地图染色或打上阴影可以表征统计数据并显示地区趋势。地图可能会引起误导,因为打阴影的地区代表地区的面积,而不是人口密度。

3.5 作图:优秀的标准

Edward Tufte 是数据的视觉展示专家,使用术语图优性来描述一个“好”图。在他看来,好图应能够清楚、准确、有效地表达复杂的思想。Tufte 认为作图的目的是产生“复杂性的展示”。

3.6 表:改变排列方式可能更合适

表中用行和列组织起来的数字用紧凑的方式归纳了数据。数据构成的形式能够非常大地影响观众解释数据的方式。

补充读物

Cleveland, William S. *Elements of Graphing Data*. New York: Chapman & Hall, 1993. 包含很多有趣的图。

Monmonier, M. *How to Lie with Maps*. Chicago: University of Chicago Press, 1991. 如何不用地图。

noindentTufte, Edward. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press, 1983. 一个具有历史背景的谈论如何在图中展示数据的经典。

Wainer, Howard. “how to display data badly.” *The American Statistician*, vol. 38, no. 2 (May 1984), pp. 137–147. 关于如何不展示数据的有趣文章。

Witme, Jeffrey. *DATA Analysis: An Introduction*. Englewood Cliffs, NJ: Prentice Hall, 1992 不同数据集的有趣的图。

习题

回顾(习题 3.1–3.15)

3.1 在分析统计数据时,两个相互矛盾的目标是什么?

3.2 a. 圆饼图在什么时候展示数据较有用?

b. 圆饼图主要的缺点是什么?

3.3 a. 什么是点线图?

b. 用点线图展示数据的优点是什么?

c. 点线图的不足之处是什么?

3.4 a. 什么是茎叶图?

b. 它为什么叫茎叶图?

c. 用茎叶图展示数据的优点是什么?

d. 茎叶图的不足之处是什么?

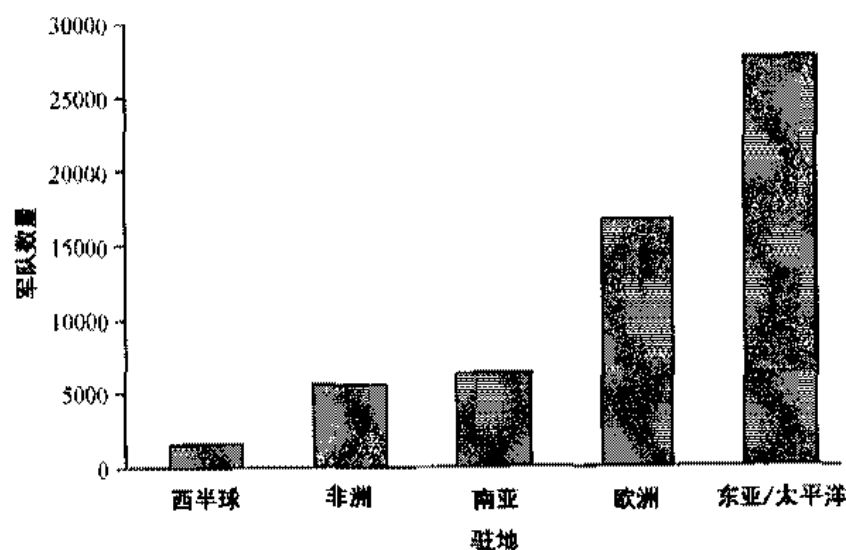


图 3.13 1993 年 6 月 30 日,美国海军舰上人员在 1993 年驻扎在世界各地的数目(习题 3.5)。(来源:Data of the U.S. Department of Defense.)

3.5 图 3.13 是美国海军(troops stationed on ships)于 1993 年在世界各地的驻扎情况的直方图。

a. 将此直方图复印下来,指出以下部分:A.横轴;B.纵轴;C.被测量的变量;D.驻扎欧洲部队的数量。

b. 图中展示的世界上的每一个地区的观测数是怎样的情况?

c. 如果用水兵面积的高度而不是用条形来表示不同的部队数量,将会带来什么错误?

- d. 从这个直方图提供的信息中可以得出什么样的结论?
- e. 这个直方图中有令人混淆的或可以改进以易于理解的地方吗?
- 3.6 在直方图中,是矩形的底、高还是面积对应落入每一区间的观测数?
- 3.7 为什么直方图是描述大数目样本的最好的图之一?
- 3.8 当直方图有两个顶峰时,这种形状的分布的名字叫什么?
- 3.9 第十一年级的 100 个男孩的身高的分布是单峰分布还是双峰分布?
- 3.10 盒形图中的哪五个数字对于展示数据是必要的?
- 3.11 什么是不对称的分布?
- 3.12 散点图可用于哪种类型的数据?
- 3.13 图 3.14 的数据展示了 19 世纪美国的人口总数和消费者价格指数。
- a. 不看课本,用你自己的话解释此图。
- b. 为什么此图是有用的?
- c. 利用该图,怎样能够较容易地说服别人相信,这两个变量之间有高度相关性?(提示:纵轴上有两种刻度)

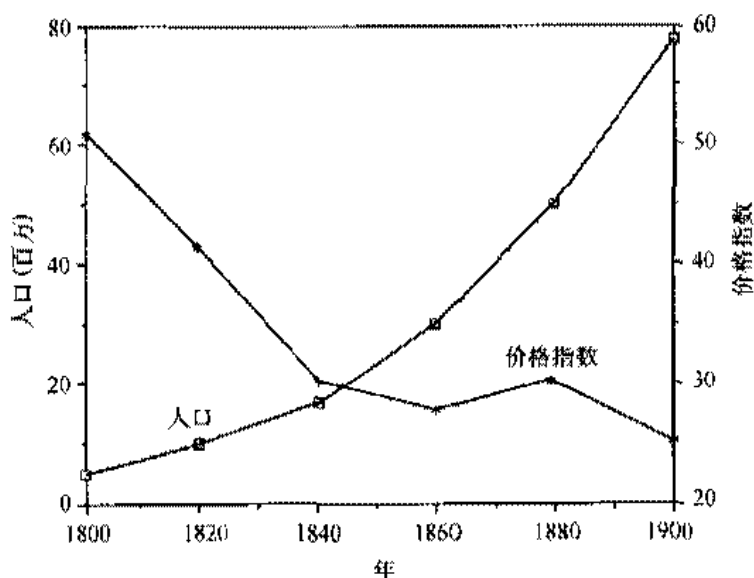


图 3.14 在 19 世纪的六次人口普查中得到的美国总人口和消费者价格指数(习题 3.13)。(来源: U.S. Bureau of the Census, *Historical Statistics of the United States, Colonial Times to 1970, Bicentennial Edition, Part 1* (Washington, D.C.: U.S. Bureau of the Census, 1975) *Population: Series A57-72, pp. 11-12; consumer price index: Series E135-166, p. 211*)

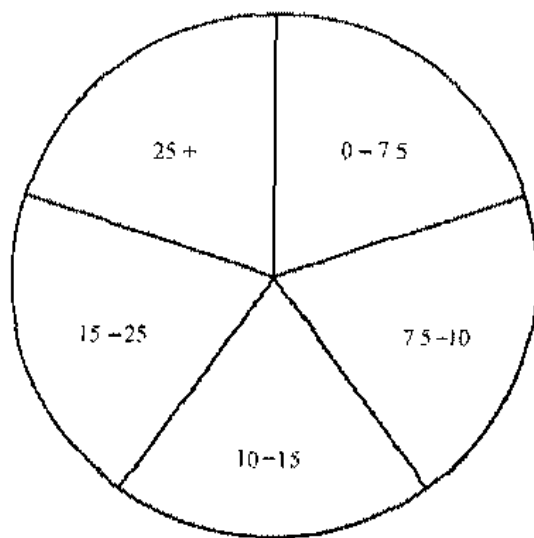
- 3.14 统计图中,图优性的关键特征是什么?
- 3.15 高质量的统计表的主要特征是什么?

解释(习题 3.16—3.33)

- 3.16 解释句于“一幅图胜过一千个字”。
- 3.17 从什么角度说,图是一种说服方式?

3.18 图 3.15 是圆饼图,它表明了在不同收入的群体中每 1000 人中犯罪的人数。

- 关于收入水平和犯罪率你能得出什么结论?
- 由于用圆饼图展示数据,什么问题是无法回答的?
- 对于分析这些数据,什么图可能会更好一些?为什么?



3.19 图 3.16 是婚姻介绍所的成功率报告的点线图,这里成功的定义是经介绍所安排后具有“长期的罗曼谛克关系”。

- 用点线图展示这些数据的一个优点是什么?
- 什么时候用点线图展示数据是一种不好的方法?

c 如果你急于寻求一种长期的罗曼谛克关系,你能从图 3.16 中得出关于投资婚姻介绍所的什么结论?

d 因为成功的比例数的分布范围很广,一位潜在的顾客可能想知道图 3.16 中哪些没有表示出来的情况?

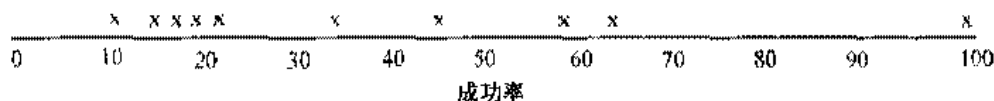


图 3.16 婚姻介绍所的成功率(习题 3.19)(来源: Mara B. Adelman, and Aaron C. Ahuvia, "Mediated channels for mate seeking: A solution to involuntary singlehood?" *Critical Studies in Mass Communication*, vol. 8(1991), pp.).

3.20 图 3.17 是 37 对夫妇结婚年龄的双茎叶图,列在地方报纸的周末版中。

- 茎叶图帮助我们观察到了数据的什么模式?
- 对于视觉吸引力和方便性,茎叶图有什么不好的地方吗?
- 有需要知道但在图中没有显示的数据的细节吗?

3.21 在报纸、新闻周刊或科学杂志上找一个统计图,并复印下来。

- 描述关于要展示的数据,该图告诉了你什么?
- 根据图优性的原理讨论该图的质量。
- 该图可以从某些方面改进吗? 解释原因。
- 为制作此图,你能想到一种方法使存在于数据阵中的信息在图中未被显示?
- 你能建议其它可能根据原始数据阵制作的图吗? 或者说,此图是数据阵唯一允许的吗?

3.22 图 3.18 中比较了几位国防部长的身高和国防经费。

- 该图告诉了我们什么?
- 该图符合好图的标准吗? 请解释原因。

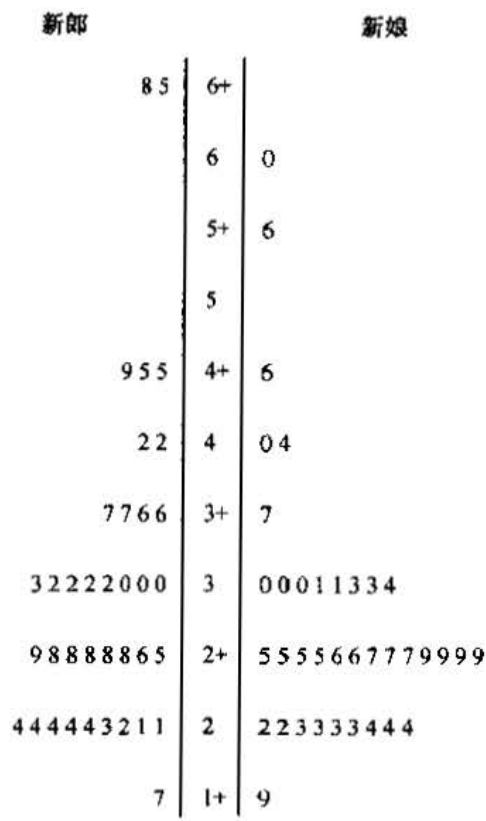


图 3.17 37 对夫妇的结婚年龄(习题 3.20)。(来源: The Philadelphia Inquirer, September 10, 1995, p. MD12 - d.)

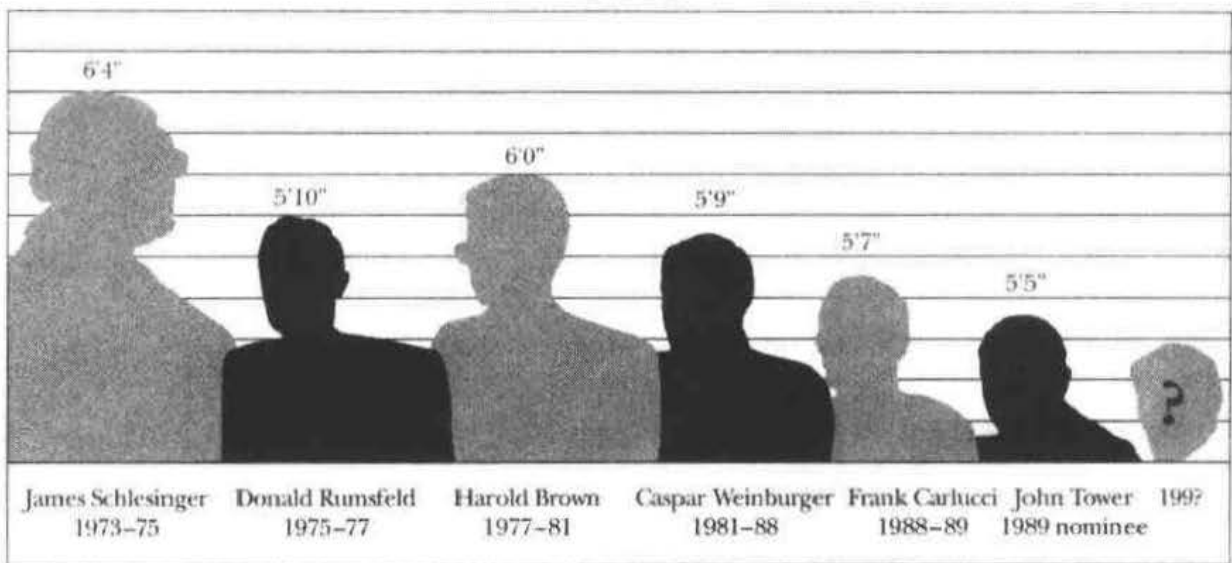


图 3.18 1973 - 1989 年国防部长的身高(习题 3.22)。(来源: Data provided by the Secretaries; adapted from the graph in The Economist, February 11, 1989, p. 20.)

- 3.23 图 3.19 中的盒形图表明了美国七个地区的暴力犯罪数。
- a. 从该图看出各地区在犯罪率上有何不同。
 - b. 该图符合好图的标准吗? 请解释原因。
- 3.24 图 3.20 表明了 20 世纪前期每十年统计一次的男性和女性的平均寿命。

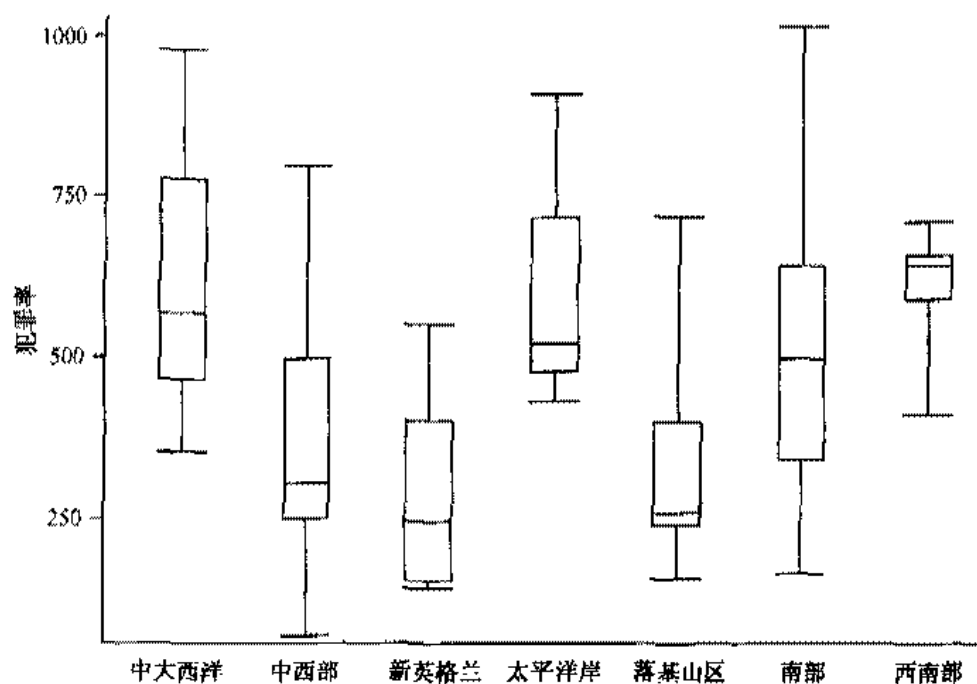


图 3.19 1968 年, 美国 48 个大陆州(指美国除阿拉斯加和夏威夷之外的 48 个州——译者注。)的暴力犯罪数(每 1000 个人)(习题 3.23)。(来源: *F. B. I. Uniform Crime Report for the United States*)

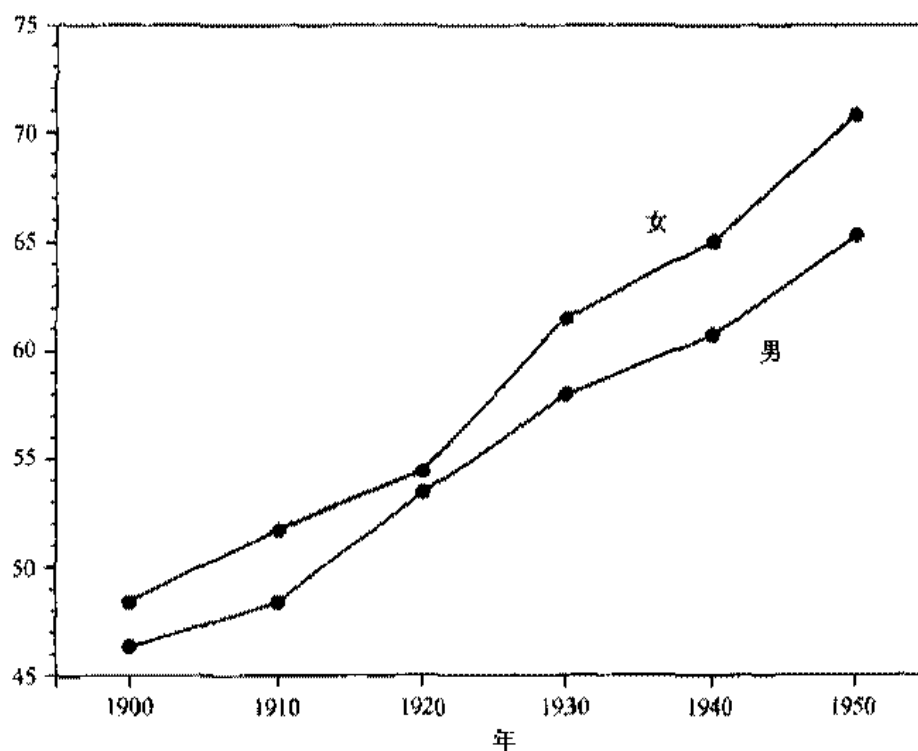


图 3.20 1900—1950 年美国人的平均寿命(习题 3.24)。(来源: *U. S. Bureau of the Census.*)

a、为什么男性和女性之间的差异有那么大?

- b. 此图如何会被某些利益集团用于政治目的?
- c. 重新画此图以符合好图的标准。
- 3.25 比较关于奥林匹克冠军所跳高度的图 3.7 和 3.8。
- a. 你认为哪一个图更好?
- b. 回答 a 时你考虑了什么问题?
- 3.26 图 3.21 中的两个条形图展示了相同的数据, 都是有关某一时间内劳动大军的比例变

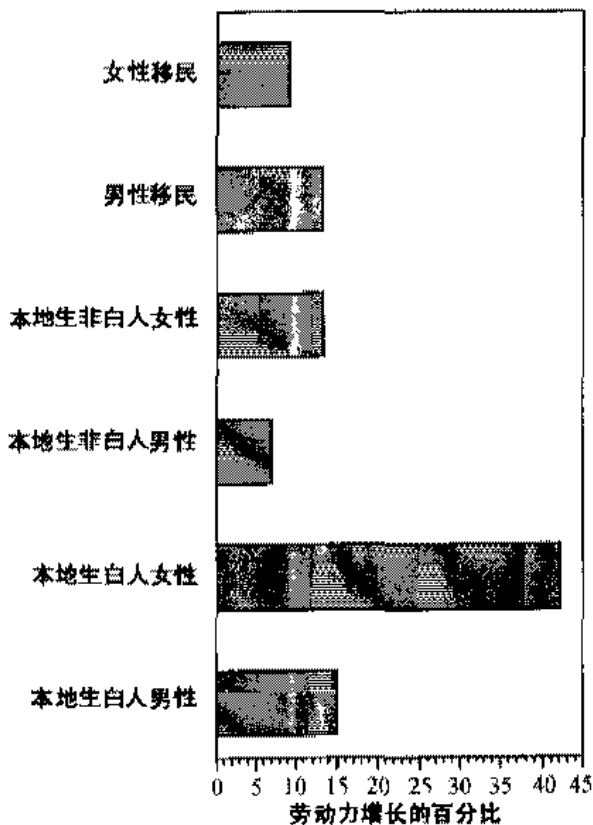
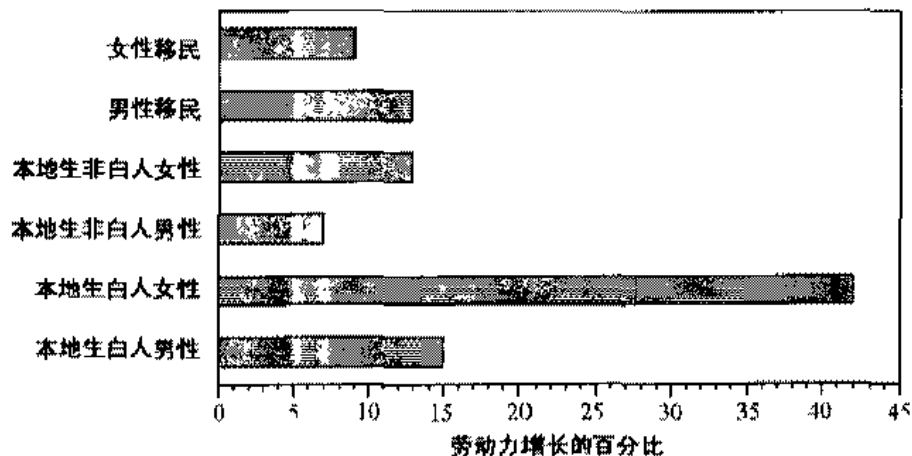


图 3.21 显示相同数据的两个条形图(习题 3.26)。(来源: *Data from Workforce 2000*, produced by the Hudson Institute, 1987.)

化,但是它们看起来彼此很不一样。

a. 为什么可以说这两个条形图展示了相同的数据?

b. 每一个图可能给读者什么样的视觉效果?

c. 如果你想尽量保持中立,你应该怎样重新做条形图?(画出或描述出你的改进。)

3.27 解释茎叶图是否能用于分类变量和度量变量。

3.28 解释盒形图是否能用于分类变量和度量变量。

3.29 比较茎叶图和盒形图的优缺点。

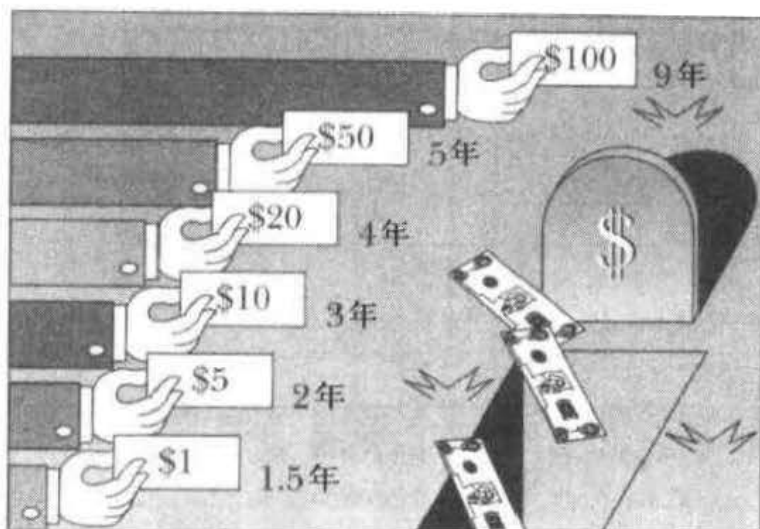


图 3.22 纸币的使用寿命(习题 3.30)。(来源: *Data of U. S. Bureau of Engraving and Printing*, adapted from the graph by Marty Baumann in *USA Today*, August 19, 1991, p. 1.)

3.30 钱币实际上会被消耗,图 3.22 表明了不同面额的纸币平均的寿命有多长。

a. 你认为图中的平均是指哪种类型的平均? 说明原因。

b. 该图符合 Tufte 图优性的标准吗?

c. 用另外的方法重画此图,并解释为什么你的图可能会更好?

表 3.5 习题 3.31 的数据

菜肴(杯子数)	热量值	脂肪(克)	脂肪中热量的百分比	钠(毫克)
蛋卷(一个)	190	11	52	463
木须肉(4 盘)	1228	64	47	2593
宫爆鸡丁(5)	1620	76	42	2608
糖醋里脊(4)	1613	71	39	818
西兰花炒牛肉(4)	1175	46	35	3146
曹将军鸡(5)	1597	59	33	3148
鲜橘牛肉(4)	1766	66	33	3135
酸辣汤(1)	112	4	32	1088
烙饼(5)	1059	36	31	3460
家常炒饭(4)	1484	50	30	2682
鸡丝炒面(5)	1005	32	28	2446
湖南豆腐(4)	907	28	27	2316
蒜茸虾(3)	945	27	25	2951
清炒小菜(4)	746	19	22	2153
川味虾(4)	927	19	18	2457

来源: *Data from Science in the Public Interest*, tabulated by the *PhiladelphiaInquirer*, September 2, 1993, page D1.

3.31 表 3.5 表明了几种中国食品中的脂肪含量。

- 说明句子“菜肴从最坏(指由脂肪导致的高卡路里含量)到最好(低卡路里含量)的排列。”对于一个想要点中国菜,并尽可能吃“最好”的食品的人来说,会产生什么样的误导。(该表中什么地方——如果有的话,是有问题的?)
- 你能把此表用于显示脂肪含量以外的其它目的吗? 重新设计表,并说明另一种目的。

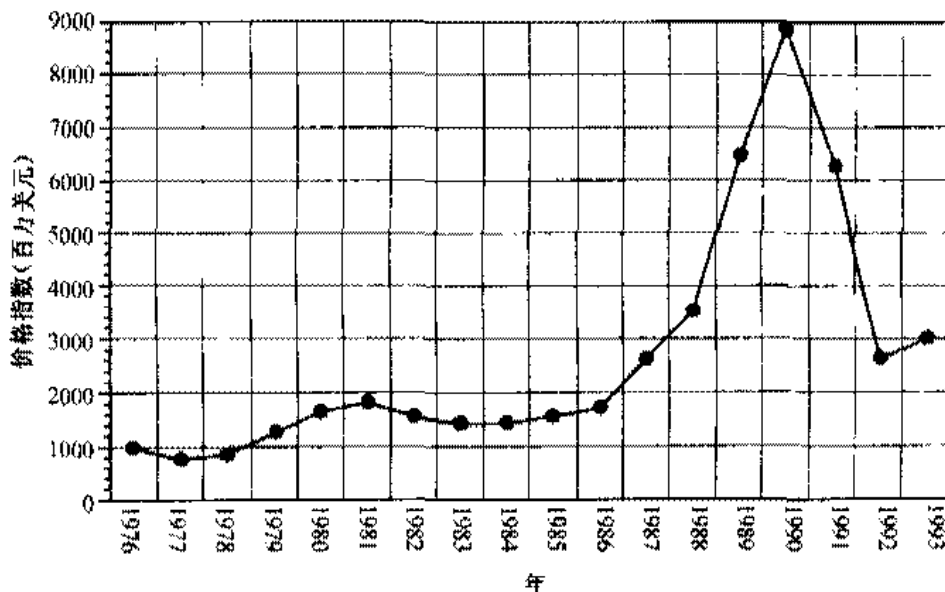


图 3.23 1976—1993 年由巴恩斯基金会(Barnes Foundation)收藏的 20 幅 Cezanne 的、16 幅 Renoir 的和 15 幅 Matisse 的图画的价值(习题 3.32)。(来源:Robin Duthy, "The boom for Barnes," *Connoisseur's World*, 1994, p. 108.)

3.32 图 3.23 复制于一个关于费城的巴恩斯基金会收藏的 Cézanne、Matisse 和 Renoir 的图画的价值报告。该图由伦敦的艺术市场研究(Art Market Research)的一名顾问制作。

- 你如何描述这种收藏最后十五年的金融历史?
- 总的来说,在图中所示的这些年中,关于这些图画的价值趋势你怎样认为?
- 该图以怎样的方式重画才能显示出更多的关于这种收藏的信息?
- 对于只想扫一眼图而不想读相应的文章的人来说,有什么办法能够使此图对他们更有帮助?

3.33 图 3.24 是一个时间序列图,它画出了研究对象给自己一周中每天的情绪(高兴程度)打的分数。

- 根据图中数据,从情绪来看,对此人来说,一周中最好和最坏的日子是什么?
- 一周中不同日子的情绪水平之间的差异是怎样在图中被强调的?
- 如果情绪得分轴上包括 0—3,那么图的效果会怎样?
- 为什么你认为 7 分制中 5 是中间水平而 4 不是,虽然 4 事实上位于 7 分制的中间?
- 该图所属文章的作者又附加了另外一个时间序列图,显示研究对象在那一周中每日记录的情绪得分。(因此,时间序列图可用于表示在同一时间内的两个不同的变最。)每日记录的情绪得分的连线比回顾的那一个更平缓,即:每日的记录表明,周末

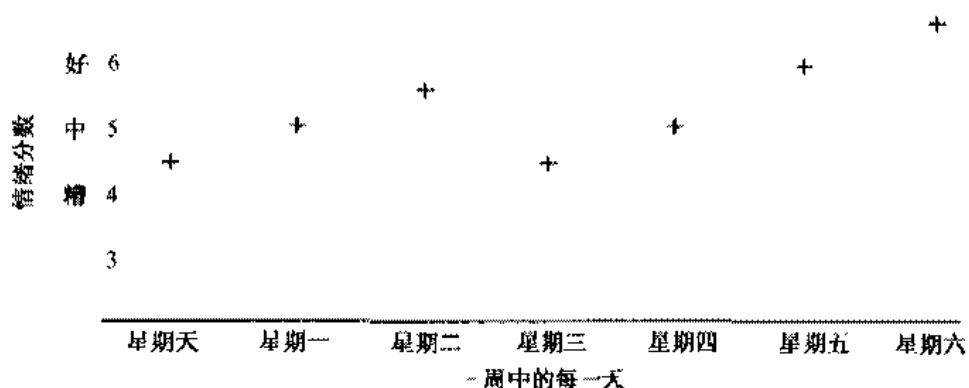


图 3.24 一周高兴程度的回顾(习题 3.33)。(来源: Jessica McFarlane, Carol Lynn Martin, and Tannis MacBeth Williams, "Mood fluctuations, women versus men and menstrual versus other cycles," *Psychology of Women Quarterly*, vol. 12 (1988), p. 214)

不那么好,周日和周三也不是那么糟。你认为为什么回顾的情绪得分不同于当日报告的情绪得分呢?

分析(习题 3.34—3.54)

3.34 社会经济学的等级分(Scale Score)的一个样本如下:

42 35 48 26 52 47
29 65 42 51 47 35

- 从这些数据的直方图中,我们可指望了解到这个例子的什么内容?
 - 选择合适的较少的区间,用这些数据作直方图。
 - 根据直方图显示的内容,你能得出关于变量的什么结论?
 - 这些数据缺失了哪些能帮助你理解的信息?
- 3.35 人们如何选择消磨时间的方式可以揭示出我们的社会的很多内容。在一个详尽的使用时间的研究中,人们发现,在工作日中,有工作的男性花费 8.1 小时在和工作相关的活动上,1.0 小时做家务,9.9 小时在个人事物如吃饭睡觉和打扮上,1.2 小时在路上,3.8 小时在闲暇活动如运动和看电视上;有工作的女性相应的数据分别是 6.5、3.4、9.8、1.1 和 3.2 小时;家庭主妇相应的数据分别是 0.0、7.8、10.3、0.7 和 5.2 小时。这些值都是平均值。(来源: J. P. Robinson, *How Americans Use Time: A Social - Psychological Analysis of Everyday Behavior*, New York: Praeger, 1977, P. 90.)
- 用两种图显示这些数据,讨论哪种图较好地展示了数据。
 - 关于这三组人如何使用时间,你的图提供了什么信息?
- 3.36 《独立宣言》(Declaration of Independence)的签名者是一群被选出来的人,我们感兴趣的问题是,这些人是否比那个时期一般的人长寿。例如,George Wythe 签名时是 50 岁,他被预期可再活 21 年,而实际上,他又活了 30 年。因此,他比预期多活了 9 年。所有签名者的实际寿命和预期寿命的差异如下所示:

24	-3	-24	2	-4	-19	21	16	-4	7	-11	-1
8	9	-6	-14	-6	2	-4	-18	14	-8	13	1
-4	22	-9	-1	13	-14	-6	1	-16	-1	-1	9
-4	19	-6	-12	-13	-1	13	4	-3	13	-14	
29	4	-9	-4	-6	-12	-13	-19	-14	-19	11	
7	9	-19	21	-9	-4	-28	-14	-21	-18	-7	

负数表示一个人比他在签字时的预期寿命少活的年数。(来源: *U S Bureau of the Census, Bicentennial Statistics Quoted in Pocket Data Book, USA 1976, Washington, DC; U S Government Printing Office, 1976, P 370*)

- 作一个表和一个直方图,表明这个变量的分布。
- 根据直方图的形状,你对签名者的寿命能得出什么结论?

- 3 37 先天愚型是发生在新生儿中的一种基因异常的病。下面的数据文件表明,1971年在瑞典出生的先天愚型婴儿的数目和他们的母亲的年龄。

母亲年龄	15 - 19	20 - 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49
婴儿数目	18	87	96	72	73	73	19

来源: *E B Hook, and A Lindsjo, "Down syndrome in live births by single - year maternal age interval in a Swedish Study," American Journal of Human Genetics, vol 30 (1978), pp 10 - 27, as reported in C J Geyer, "Constrained maximum likelihood exemplified by isotonic convex logistic regression," Journal of the American Statistical Association, vol 86 (1991), pp 717 - 724*

- 关于此数据作直方图。
- 关于先天愚型婴儿的数目和母亲的年龄,这个直方图给我们提供了什么信息?
- 当母亲非常年轻或非常老时,先天愚型婴儿的数目最少。这是否意味着,一个女人应该在非常年轻或非常老的时候生孩子?
- 如果你要给女性建议一个合适的生孩子的年龄,以减少生先天愚型婴儿的危险,你还需要其它的什么数据?

- 3 38 十六种不同快餐的热量值如下所示:

110 120 120 164 430 192 175 236
429 318 249 281 160 147 210 120

来源: *USDA dat and manufacturer's data in an advertisement in The New York Times Magazine, April 20, 1990, p 20.*

- 根据这些数据作出直方图,用 50 作为每一个矩形的宽度。
- 根据这些数据作茎叶图,在线的左边用两位数。
- 根据这些数据作盒形图。
- 每一种图的优缺点各是什么?
- 你更喜欢哪个图? 给出理由。

- 3.39 某一高中乐队的制服帽要改变。乐队指挥用一把卷尺收集 150 名乐队成员的头围,而乐队主席则让每人报上自己所需帽子的大小:小号、中号或大号。帽店卖的帽子有 10 种规格(从 $6\frac{7}{8}$ 到 $8\frac{1}{4}$)。

- 为最有效地买帽子,大致描述一下你将怎样组织数据。

- b. 在确定乐队成员所需帽子的尺寸时, 乐队指挥和主席各犯了什么错误?
- 3.40 选择一个你愿意回答的问题, 例如: 上周音乐店最受欢迎的 CD 唱片是什么, 乡村/西方音乐, rap^① 摇滚乐, 重金属音乐还是民谣?
- a. 找一组数据回答你的问题。
- b. 制成一个可以最好地反映你的数据的图。
- 3.41 对于你上个月买的 20 种商品, 作一个茎叶图来说明每一种花钱的数目。
- 3.42 a. 画一个直方图, 表示你目前所在的地区一年中每个月温度低于结冰点(用 32 华氏度或 0 摄氏度)的天数。
- b. 描述这个直方图的形状。
- 3.43 写下你二十个亲友的名字, 你认识每一个人的年数和你们发生激烈争吵的次数。
- a. 关于此数据的两个变量作散点图。
- b. 在你的样本中, 这两个变量有什么关系吗?
- c. 在你看来, 这一分析中存在什么问题吗?
- d. 研究这两个变量, 什么样的数据会更好些?
- 3.44 a. 用下面的信息画条形图, 并用左边的矩形代表男性, 右边的矩形代表女性。在具有四年或四年以上大学教育的全日制工人中, 一个收入的平均值的样本数如下: 白人男性收入 \$42000; 白人女性收入 \$29000; 亚裔美国男性收入 \$37000; 亚裔美国女性收入 \$29000(来源: *U S Census data 65 for 1990*) 为作图方便, 你可去掉数字中的“000”。
- b. 从你的图中, 你能得出什么主要结论。
- 3.45 根据由美国人口调查局在 1990 年 10 月对 60000 个家庭的调查, 当年的 5644000 个四年全日制大学生的年龄分布的百分比如下:

年龄	15 - 17	18 - 19	20 - 21	22 - 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 59
%	1.9	34.7	34.1	16.6	6.4	2.7	1.8	1.2	0.6

来源: *The Chronicle of Higher Education*, vol. XXXIX, no. 1 (August 26, 1992), p. 11

- a. 对于这个年龄分布作直方图。注意区间应具有不同的长度。这意味着你必须调整每一个矩形的高度, 以使得其面积代表对应的百分数的大小。另外, 每一个年龄组中的年龄, 是从左边最小的那个值开始, 到下一组年龄左边最小的那个值(但不包括此值)结束。例如, 20—21 年指的区间是从 20 开始到 22 结束, 但不包括 22 这一点。
- b. 这个大学生年龄分布的直方图的形状告诉了我们什么?
- c. 此分布使你感到惊奇吗? 为什么?
- 3.46 剑鱼(sword fish)的身体吸收水银, 水银含量超过 1.00 ppm(即百万分之一)的剑鱼对人体有害。在 28 条剑鱼的样本中, 我们发现其水银含量如下:

0.07	0.24	0.39	0.54	0.61	0.72	0.81	0.82	0.84	0.91
0.95	0.98	1.02	1.08	1.14	1.20	1.20	1.26	1.29	1.31
1.37	1.40	1.44	1.58	1.62	1.68	1.85	2.10		

来源: *Larry Lee and R. G. Krutchkoff, "Mean and variance of partially - truncated distributions," Biometrics*, vol. 36 (1980), pp. 531 - 536.

- a. 关于此数据作茎叶图, 用前两位数作为茎。

① 黑人音乐的一种, 节奏紧凑, 唱歌犹如急促地说话——译者注。

b. 描述水银含量的分布形状。

c. 许多剑鱼的水银含量超过 1 00 ppm 的原因在于,并不是所有的剑鱼在被出售以前都被检查过。看起来所有剑鱼的水银含量的平均水平比 1.00 大吗?

- 3.47** 在州际高速公路上,看起来很多人都在超速驾驶。得到超速罚单的概率主要依赖于有多少交警在巡逻。各个州的交警的数量都不相同。一个计算警察多少的方法是看每个州的警察在州际高速公路上所负责的平均英里数。这个数字的范围从德拉威州的每个警察 0.1 英里到怀俄明州的每个警察 7.0 英里。因此,如果德拉威州的每个警察都在路上巡逻,那么每十分之一英里将有一个警察;而在怀俄明州,每七英里才会有一个警察。

图 3.25 是 48 个大陆州每一个警察英里数的茎叶图,左边枝叶表示英里数,右边枝叶表示十分之一的英里数。

- a. 描述茎叶图中所显示的分布的英里数。
b. 根据此茎叶图作盒形图。
c. 对这些数据来说,茎叶图和盒形图的优缺点各是什么?
- 3.48** 图 3.26 显示有关在国家公园的救援任务,讨论其是否符合了 Tufte 图优性的标准

- 3.49** 当发生犯罪攻击时,大部分情况下,攻击者和受害者是同一种族。1991 年,根据美国联邦调查局(FBI)的报告,85%的黑人受害者是被黑人攻击的,75%的白人受害者是被白人攻击的,8%的黑人受害者是被白人攻击的,17%的白人受害者是被黑人攻击的,其余的则是受害者和攻击者的其它类型混合的结果。

- a. 为白人受害者和黑人受害者各画一个圆饼图。
b. 想象一个包含所有这些数据的圆饼图。从哪个方面上说,单独的一个圆饼图有可能引起误导(即单独的一个圆饼图假定了什么)?
- 3.50** 构造一个能够展示圆饼图优点的数据集。
- a. 画出此图。
b. 简述图中的结果。
c. 当你设计圆饼图时,其中有什么问题吗?
- 3.51** 构造一个能够展示茎叶图优点的数据集。
- a. 画出此图。
b. 简述图中的结果。
c. 当你设计茎叶图时,其中有什么问题吗?

- 3.52** 像在习题 1.20 中一样,打开一个水龙头直至它仅在滴水。在 3 分钟内,记录每 20 秒中的水滴数。

- a. 画一个图描述你的数据。
b. 在三分钟的时间里,水滴数的分布是随机的还是有规律的?从什么方面说是随机的,又从什么方面说是有规律的?

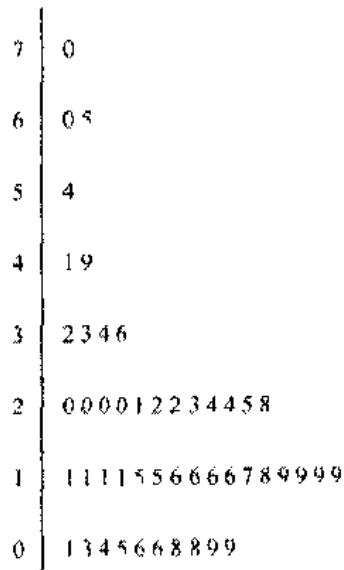


图 3.25 每一个州的警察的英里数,48 个大陆州(习题 3.47)。(来源:Autoweek, July 9, 1990, p. 37.)

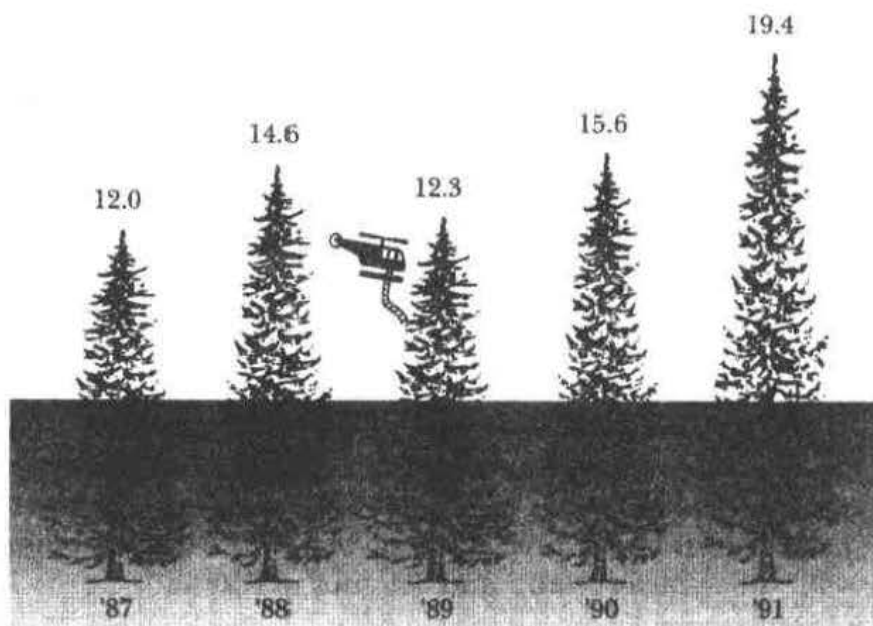


图 3.26 对于 367 个国家公园系统地区的每百万次参观中所作的研究和救援(习题 3.48)。(来源: *Data of National Park Service; adapted from the graph in The New York Times, March 25, 1993, p. A18.*)

- 3.53 表 3.6 是在选择的几个国家中,于 70 年代中期每 100000 个人中由于各种原因死亡的人数的一个简略表。重画此表,使得它更具可读性,并为你的改进说明理由。

表 3.6 习题 3.53 的数据

国家	总的死亡数	事件类型			
		交通	自然的	杀害	其它
奥地利	75.2	34.8	29.7	1.6	9.1
法国	77.8	56.8	31.0	0.9	22.1
意大利	14.2	22.8	19.2	1.1	4.1
荷兰	40.3	17.8	18.2	0.7	3.6
挪威	48.4	47.3	25.1	0.7	5.3
美国	60.6	23.4	15.8	40.0	11.4

来源: Social Indicators III, U. S. Census Bureau, December 1980, p. 252, reprinted in Howard Wainer, "Tabular Presentation," *Chance*, vol. 6 (1993), no. 3, p. 53.

- 3.54 a. 为下面的数据作一个你想作的图。当调查者访问 1000 个在私人公司工作的成人关于隐私的话题时, 61% 人说他们的老板“很好”地尊重了他们工作时间以外的隐私, 29% 的人说“有点好”, 8% 的人说“不很好”, 3% 的人说“一点都不好”。
- b. 为下面的数据作一个你想作的图。响应者认为在某种程度上雇主有权证实应聘者提供的信息。十分之八的人认为雇主检查应聘者提供的关于教育背景和犯罪记录是合适的; 93% 的人反对进行工作以外吸烟情况的检查; 69% 的人反对对喝酒情况进行尿检; 69% 的人认为对态度和社会倾向的心理测试是不合适的; 59% 的人反对进行关于爱滋病病毒的血样检查。(来源: "U. S. workers are concerned about privacy on the job, survey finds," *The Philadelphia Inquirer*, August 23, 1994, p. F6.)

C H A P T E R 4



4.1 各种平均数：让我们数数有几种

4.2 变差：测量生活的乐趣

4.3 均值的标准误差

4.4 标准得分：比较苹果和桔子

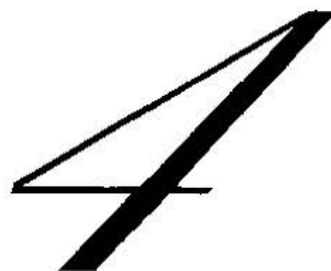
4.5 简单化的收益与信息的丢失

4.6 房地产数据：看不见的价格

4.7 小结

数据的描述:

计算汇总统计量



莎士比亚的作品是否真的是莎士比亚写的? 美国一般家庭有多少个孩子? 男性的薪水是否一定比女性高? 一颗“正常的”心脏平均每分钟跳动多少下? 新娘和新郎在结婚时年龄是多少? 平均每年有多少人死于肝硬化? 在一个街区用提高公众税收为学校发展提供更多资金的作法是否可行?

第二章我们提到,一个数据文件的原始观察值包含了该数据集的所有信息。但是只是看一看就想提取其中的信息是几乎不可能的。所有的信息都在那里,但这些信息被数据中的随机性掩盖了起来。

第三章我们提到使用图表来组织数据。图表的使用常常需要增加汇总统计量——对现有数据进行计算而得到的新数字。用一个或几个变量的值我们可以得到代表那些变量的一些新数据,这样就可以将大量的数据总结为很少的几个数值。

这一章我们将主要讨论以下的两个问题。

1. 怎样将一个变量的多个观察汇总为一个数值,并且使这个数值具有原数据的中心趋势或平均值。是否可能找到一个描述所有观察值都可能接近的一个数值呢?
2. 怎样汇总变量彼此之间的区别呢? 这些观察值是很类似的还是有很大不同? 就是说这些数据间彼此是否有很大的变差?

如图所示(第二章),计算汇总值有一个主要优点和一个主要缺点:

优点: 汇总值会使数据高度的简单化。

缺点: 任何的简单化都意味着某些信息的丢失。

在1960年美国大选前,密执根大学社会研究所的一个调查研究中心在一次问卷调查中询问被访者将在大选中支持谁? 结果在1396个准备投票的被访者中有634人计划投肯尼迪的票。这就是说在所有准备投票的被访者中有45%的人希望肯尼迪当选(他在这一年稍后的时间里确实以很接近的票数成为了总统)。这种用一个单独的百分比来代表在1396个个别的回答中的634个是一种对原始数据的高度简化。但同时,原始的变量已经无可挽回的丢失了。而且,如果我们只知道一个单独的数值,我们不可能恢复原始数据,并且许多不同的数据集也可能产生同样的平均值。

通过直观的数据表示,数据的汇总计算势必要打破简单化的获得与数据丢失之间的平衡。但是要这样做,常常并不简单,这需要我们了解对各种常用的汇总数据的优点和缺点的知识。

4.1 各种平均数: 让我们数数有几种

平均数是一个数值,是对一个变量的观察值进行计算后得到的。

在对数据计算后得到的数值中,最常用的就是某种平均数或者叫做中心值(average or central value)。我们大部分人在小学就学习了平均的概念了。现在我们常读到MBA的平均工资,平均房价,道琼斯平均股票价格,平均谋杀率等等。但我们究竟在多大程度上了解各种平均值? 或者如何仅仅计算某种平均值就能产生错误印象。

我们并非只有一种平均值,而是有好多种。为了探索其多样性,我们来认真看一看下面的句子:

当代美国的平均人是女人,平均每个女人有2.1个孩子,且这些女人住在平均价值为\$80000的住房中。

这句话中分别用了三种平均值,你能分辨出它们有何不同吗?

众数: “最多的”的宿主

一个变量的众数就是指出现次数最多的数的值。

性别变量只有两个值,男和女。在美国妇女比男子人数多。所谓平均人口是女人的说法是应用了一种叫做众数(mode)的统计的平均。

众数一般用来描述分类变量,特别是那些有许多个值的分类变量,例如:宗教,种族,社会阶级等。例如,你可能发现,在某一特别街区,宗教变量的众数是穆斯林教,人种的众数是亚裔,而社会阶级的众数是“中上等”。

同样,众数也可以被其它种类的变量使用,图4.1是对37个妇女进行结婚年龄调查而得到的直方图(与图3.3所用的样本相同)。图4.1中直方图的主峰在25~30岁之间。则我们取其中点27.5岁作为结婚年龄的众数。

有时,一个变量有两个值经常出现,这样它就有两个众数,这就叫做二众数分布(bimodal distribution)。当一个变量有两个众数时,这个变量的观察值常常是由来自两个群体的数据混合组成的。例如一个班级中学生的身高的直方图就可能是二众分布,因为这些学生中有男学生也有女学生。

众数告诉我们,这个值出现的次数比其它的值出现的次数多。但它并未告诉我们它较别的数值多的程度。一个由100人组成的群体,无论它有51个女人(和49个男人)或者99个女人(和1个男人),其性别变量的众数都是女人。这两种情况是非常不同的,但是众数并不能区分它们。因此众数掩盖的信息时常比它揭示的要多。

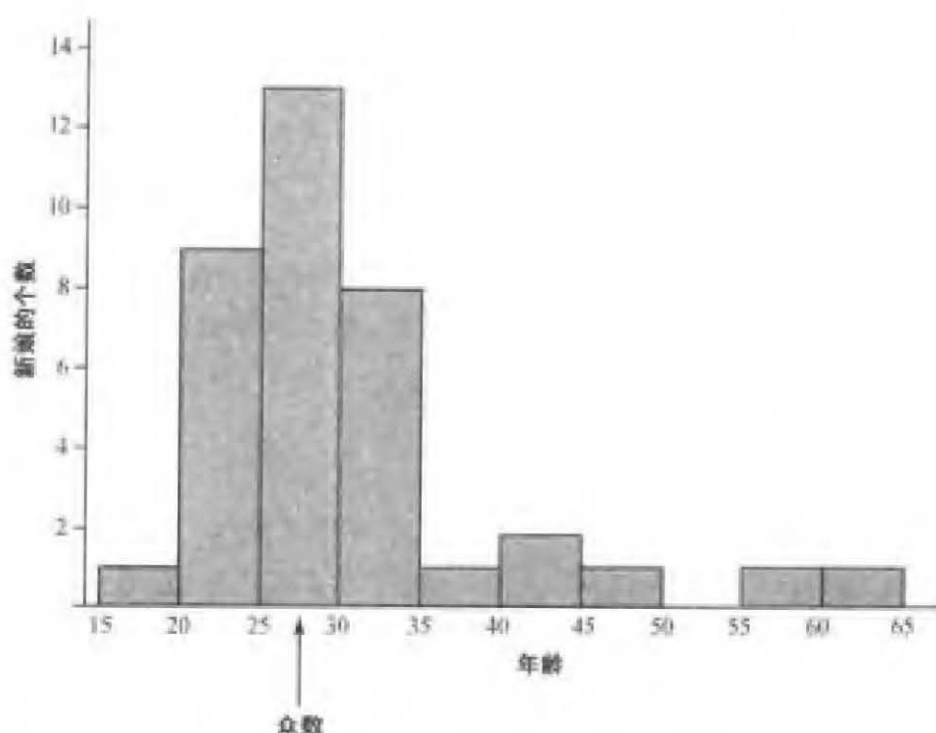


图 4.1 年龄变量的众数：最高的矩形的中点。

莎士比亚著作中的众数

几个世纪以来，学者们怀疑是否是莎士比亚写下了那许多记在他名下的精彩绝伦的戏剧和诗句。怀疑产生的原因，部分是因为人们认为莎士比亚这个历史人物是乡间的，没文化的，而且天被发现他取得如此巨大的成功。他在世时，没有人认为他是作家。他死后，没有留下任何私人信件，手稿或文学笔记。因此这些线索使人们有理由怀疑作家威廉·莎士比亚可能另有其人。一些著名的受过很好教育，且有着丰富的文学记录的人被认为写下了这些经典作品。他们中有 Francis Bacon, John Donne, Christopher Marlowe, Walter Raleigh, Edmund Spenser, 和 Elizabeth 一世女王本人。

一些专家用统计侦探方法帮助莎士比亚得到了他的应得之物，这种统计方法依赖于众数的使用。

在三年时间中，一个由 Claremont 学院本科学生组成的莎士比亚诊所 (Shakespeare Clinic)，用统计分析对 58 个与莎士比亚同时代的作家进行分析，以确定谁的写作风格与莎士比亚的作品风格最相近。他们从 58 个作家的作品中选取片段，并将其分成 500 字一段的小段。他们对区组中一些变量进行记数统计。例如，学生们考察 52 个关键字的出现情况，并找出其众数。利用各种统计策略，他们得到了各个作家的主要特征。调查结束时，

27 个备选者的诗中没有一个能通过众数检验。Thomas Heywood 是最近的一个检验，它和莎士比亚相差 2.2 个标准差。这意味着，如果 Heywood 真是莎士比亚，那么和他以自己的名字命名的诗篇不同的机会将少于 5%。John Donne 是最远的检验，与莎士比亚作品的标准差达到 36.9。

这些明显的事实证明，就是莎士比亚写下了他本人的诗篇。

(来源：Ward Elliot 和 Robert Valenza, “谁是莎士比亚全集的作者”, *Chance*, 第四卷 (1991), pp. 8 ~ 14)

对于一个度量变量,众数并不能充分利用这个变量的所有实际观测值,另外如第三章所述,选择不同的直方图的区间长度,可以获得不同的众数,众数依赖于直方图的画法。

众数的优点:一个变量的众数从图表中很容易获得。对于分类变量,它是描述平均值的一个最好办法。对于一个有二众数分布且中间值只有很少观察值的变量取两个众数比取一个仅有几个观测值的中间值含有更多的信息。众数只需要很少的实际运算,因为它的值可以很容易直接从条形图中获得。

众数的缺点:众数并不经常使用,很多统计软件包甚至没有计算众数的程序。一个变量的众数值只能传递这个数据集中的信息的很少一部分。因此只用众数,数据集中的信息就不能被很好地使用。

停下来想一想 4.1

考虑一个你曾经拥有的按周支付薪水的工作。记录你在一周中的哪一天得到薪水。对你 52 周中支付薪水的日子画一个条形图,决定哪一天是你支付薪水的日子的众数。是否这就是 TGIF^① 是一个流行短语的原因呢?

中位数: 数到中间那一个

一个变量的中位数是这样一种数字,它把观测值分成同等数目的两组数,一半观察值小于等于这个数,而另一半大于等于这个数。

前面提到,这些女人住在“平均”价值为 \$80000 的住房中。“平均”房价和许多其它的经济变量一样常常是用中位数(median)来描述的。因为价格是一个度量变量且有较高和较低之分;不像宗教等分类变量,它的价格的值可以从最小到最大排序。排序后其中间值即是中位数的值。当房价的中位数是 \$80000 时,一半的房价低于这个数而另一半房价高于这个数。

可以通过把观察值从小到大排序,并取中间的数据值就可找到中位数。对于一个很小的数据集,观察值的个数是奇数还是偶数是有区别的。但对于一个很大的数据集来说,这种区别就不重要了。另外,利用不显示原始数据的表或是直方图,也有可能找到中位数。

在这个例子中,中位数与其它房子的价格没有关系。例如:假设数据集除了价值 \$80000 的房子外还有这样两座房子,其中一座价值 \$79000,而另一座价值 \$500000,这三座房子价格的中位数仍旧是 \$80000,尽管一个房子的价格只比中位数少 \$1000,而另一个房子的价格却比中位数高出 \$420000。对于中位数来说唯一有关的是,它是数据集中处于中间位置的数。

寻找中位数:想象一个有五个孩子的家庭,他们的年龄分别为:17岁、14岁、12岁、9岁、5岁,在这个数据集有奇数个观测值,中间的数是第三个观测值,有两个观测值比它小而另两个比它大。这样其中位数就是 12 岁。(这一章末尾的公式 4.1 说明了如何在奇数个观测中寻找

^① TGIF 为英文“Thank God It's Friday”的缩写,意思是“谢天谢地又到了最后一个工作日了!”——译者注。

中位数。)

想象一个有六个孩子的家庭,其中有一对5岁大的双胞胎。这六个孩子的年龄分别是17岁、14岁、12岁、9岁、5岁、5岁,这个数据集中没有一个观测值可以将其恰好分为两个相等的部分。但对任何一个9到12岁的孩子来说,都有三个孩子的年龄比他大,三个孩子比他小。一般地,在这种情况下,我们取两个中间值的中点作为中位数。12与9的中点是10.5,这也是这六个孩子的年龄的中位数。(公式4.2表明如何在偶数个观测值中寻找中位数。)

中位数与其它的分位数:中位数也称为第五十个百分位数,因为有50%的观测值小于这个数。第二十五个百分位数是使25%的观测值小于这个数的观测值。在新娘的年龄的例子中,第二十五个百分位数是24岁,第五十个百分位数或中位数是27岁,而第七十五个百分位数是32岁。

停下来 想一想 4.2

你收集了一组做咖喱鸡的菜谱。各菜谱所用佐料的数目都不一样,分别为12、16、8、9、15、10、11、14、20、12、18种。请确定其中位数。它也是众数吗?你愿意选择哪一种菜谱?

从茎叶图中找中位数:对于处理成茎叶图的数据,中位数十分好找。在茎叶图中变量的最小值到最大值都已经按顺序排好了。中位数或其它百分位数的获得只需数到所找的那个数即可。看图3.3中新娘年龄的茎叶图。这里有37个观测值,所以中位数应是将所有新娘的年龄从小到大排序后的第十九个数。有18个新娘年轻些而另18个年老些。从底部开始,从小到大数,第十九个新娘是27岁,所以27是这个群体的中位数年龄。

从直方图中找中位数:没有原始观测值,我们可以从直方图中找出中位数来吗?如果我们假设观测值在中间的区间里服从均匀分布,即在一定条件下答案是肯定的。在图4.1的直方图中,新娘总数可以通过相加每个矩形高度代表的值来获得。在此例中有37个新娘。所以中位数是从小到大排列得到的第十九个数。直方图显示有十个新娘包含在前两个年龄区间中,我们另外需找九个新娘,因为下一个区间(25到30岁)包括了13个新娘,所以中位数一定在这个区间中。假设这个区间的13个新娘在年龄上是均匀分布的,则这第九个新娘可以由走过这个区间长度的 $\frac{9}{13}$ 而得。该区间长度为五年;其 $\frac{9}{13}$ 是 $5\left(\frac{9}{13}\right) = 3.5$ 。最后这个区间的最低年龄25岁加上3.5岁,中位数就是 $25.0 + 3.5 = 28.5$ 。

中位数在直方图上也可以显示出来。图4.2与图4.1相同。第三章强调矩形的面积意味着这个条形代表的是观察值的个数。为了找到中位数,我们把直方图中全部矩形的面积分成相同的两个部分。图4.2中划的虚线代表中位数,虚线左边的直方图的面积代表着18.5个单位,虚线右边的直方图也是18.5个单位。虚线处变量的值——中位数——是28.5岁。我们发现28.5这个中位数的估计值与真实的中位数值27岁有区别,那是因为我们假设中部区间的观测值是均匀分布的;但实际上在25到30岁的这个区间上,新娘的年龄并不是均匀分布的。

中位数的使用:当一个数据的直方图显示出是非对称(偏斜)分布时,我们常常使用中位数。房子的价格就是一个典型的非对称分布。大部分房子的价格在中部,但通常也有几个房

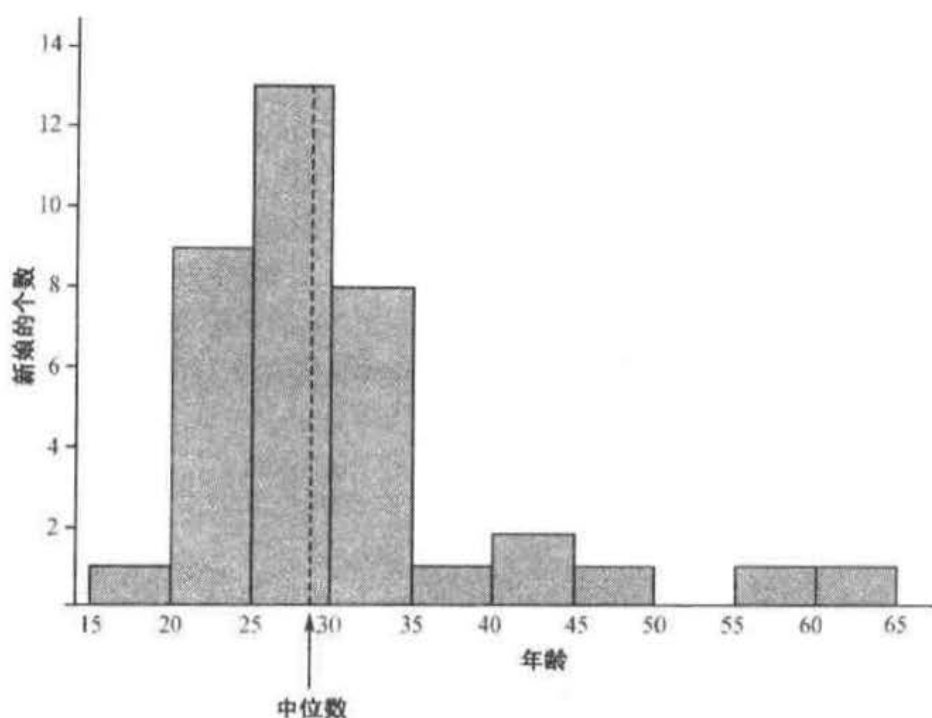


图 4.2 年龄变量的中位数: 矩形总面积的中点。

子的价格特别高。于是直方图的右侧有一个“尾巴”。

图 4.2 中的直方图说明, 两个年龄最大的新娘使数据发生偏斜。这些数据的中位数就有用了, 因为它几乎不受少数几个极端值的影响。无论那两个最老的新娘的年龄是 30 岁、40 岁、50 岁或 60 岁, 中位数的值都不会变。

中位数的优点: 中位数很好地代表了一组观察值的中点, 特别是当直方图显示出这是一个偏斜分布时。中位数只需要很少量的计算。只需将所有的观察值从小到大排序, 就可应用找中点的方法得到中位数。中位数对极端值不敏感, 在某些情况下这将是一种优点。

中位数的缺点: 除了中间值, 中位数并未利用其它观察值。这样它就没有利用数据中的所有信息。中位数对极端值不敏感, 这在某些情况下是一种缺点。

均值: 平衡跷跷板

均值是这样计算得到的: 所有观察值相加的和除以观察值的个数。

当我们说一般的美国家庭有 2.1 个孩子, 意思是说在美国每个家庭拥有孩子的均值(mean)是 2.1 个。均值是最常用的平均数。像中位数一样, 均值是一个变量值, 它大体上位于观察值中部。二者的不同在于, 均值是一个变量的值, 它可以被看作是数据集的重心。如果我们根据观察值的大小把它们放在跷跷板上, 则跷跷板会在均值处达到平衡。对于 37 个新娘的年龄数据, 均值应是 30.0 岁(图 4.3)。如果我们想象, 每个新娘都一样重, 并且按照她们的年龄值的位置站在一个水平的杠杆上, 则杠杆会在 30.0 处达到平衡。

为了找到均值,需要将所有观察值的值相加并且用观察值的个数来除。在本章末,这将用数学式子概括为公式 4.3。应用公式 4.3 找均值与找跷跷板的平衡点是相同的,也就是说均值是这个数据集的分布重心。

均值一般用于寻找度量变量数据集的中心值。和其它的平均值一样,当我们用均值代替原始变量分析时,大量信息会丢失,但是每个观测值都被用来寻找均值。只要任何一个数据的观察值变化了,均值就会改变。而中位数与众数都不具有此性质。

图 4.3 显示了均值的一个重要弱点。两个最老的新娘对均值有很大的影响,这是因为她们的年龄与均值相差太多。如果她们离开跷跷板,杠杆的平衡点会从 30.0 岁落到 28.4 岁。就是这两个新娘的年龄使均值大于中位数。如果是两个年龄与中位数年龄相差不大的新娘离开跷跷板,平衡点将不会有太大的改变。

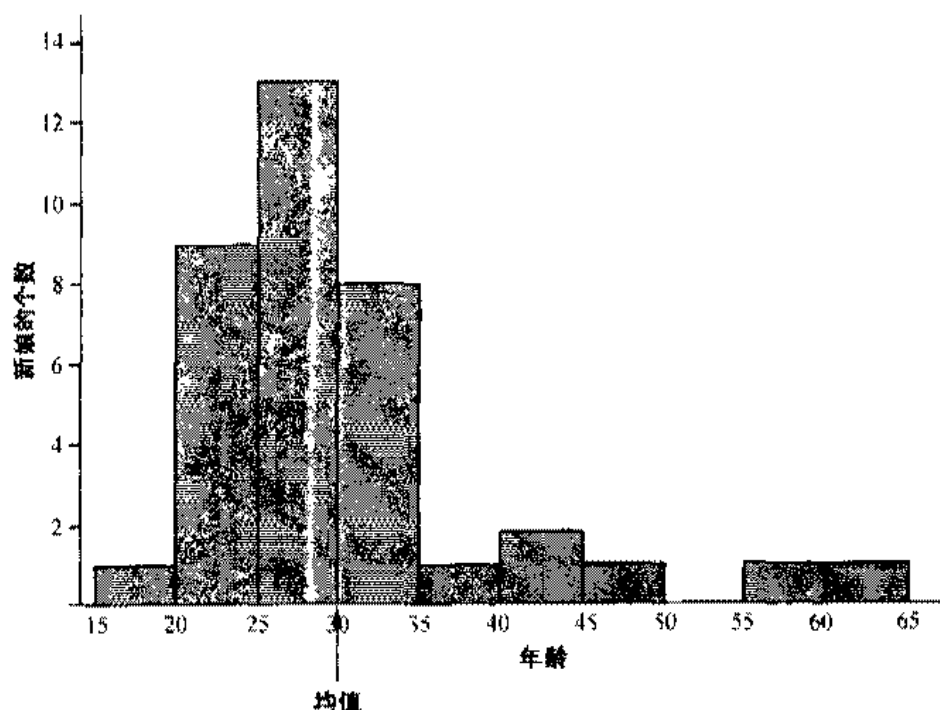
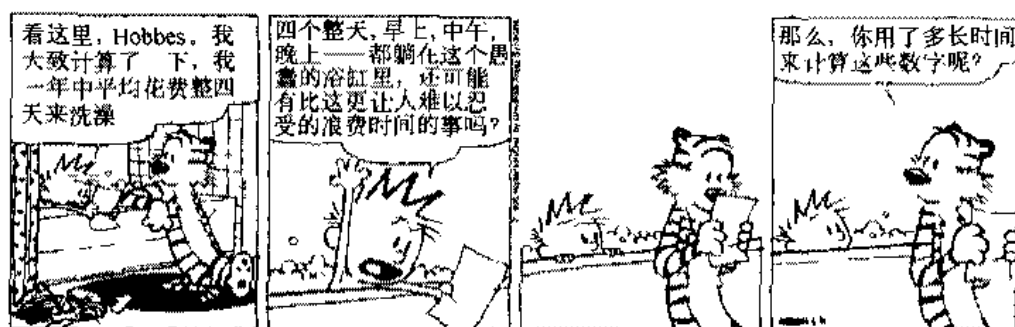


图 4.3 年龄变量的均值:直方图的重心。

因为均值对个别的极端值敏感,当数据集有极端值时,我们最好不使用均值。但小观测值与大观测值数量大致相同时,均值则是很好的统计量。小的观测值与大的观测值相互抵消了。如果数据的分布像图 4.3 一样是不对称的。我们最好使用中位数而不是均值,这因为中位数对极端值并不敏感。为了决定对一个数据集是使用均值还是中位数,最好两种都算出来。如果它们的值很接近,则我们使用均值,如果它们有很大的不同则我们使用中位数。

均值是一种最常用的平均数。所以我们给了它一个专用符号 \bar{x} 读作“ x 一杠”。使用 x 是因为它是最常用的变量符号,一杠代表平均。如果用其它字母表示一个变量,例如用 y ,则其均值同样被记作“ \bar{y} ”。



计算均值并不总是一件容易的事。(来源:“Calvin and Hobbes”版权 1995 Watterson.
Dist. by Universal Press Syndicate 授权重印,所有版权保留)

均值的优点:均值的优点是它对变量的每一个观察值都加以利用。这就意味着比起众数与中位数来,它会获得更多的信息。以后我们会说明,比起中位数与众数来,信息可以更容易地从均值中获得。

均值的缺点:因为均值使用了数据集中的每一个具体观察值,所以计算有点麻烦。均值对极端值很敏感。因此如果由于观测误差而产生的极端值会使结果不好。

众数,中位数,还是均值?

我们应该习惯于常问自己,在数据分析中我们应该使用哪种平均数?我们挑选的平均数是不是合适?偶尔某些人,也会为了使人们产生某种印象,而有意地选择错误类型的平均数,产生错误的平均值。当一个分布有很多的小观察值而仅有少数大的观察值时(例如家庭收入的分布),那么均值比中位数大。如果一个人想用尽可能大的值来概括这个分布的话,那么他会使用均值,尽管中位数比均值更合理一些。

这种歪曲在比较两组或更多组的事件时更有诱惑力。假设大标题这样写,“男人比女人挣的钱更多。”这句话的含义是什么呢?是任何一个男人比任何一个女人挣的钱都多吗?当然不是。这个标题可能是对这两组事件的平均值的比较中得出来的。如果是这样,那我们应将所用的方法说出来。也许是男人收入的中位数比女人高,也许是男人收入的均值比女人高。因此,群体之间有多大的区别依赖于用来比较它们的特别的统计方法。

停下来想一想 4.3

想象你为州的运输部门工作,你想通知州长收到的联邦政府对 26 条公路项目投资的平均值是多少。其中一条新公路收到的资助数额最大(2200 万美元)。另外的 25 个,每项投资在 20 万与 100 万之间。中位数是 25 万,均值是 100 万,众数是 20 万。你会选择哪一种平均数来代表每一条高速公路从政府获得的投资呢?你选择的这种平均数的缺点是什么呢?

4.2 变差：测量生活的乐趣

通常平均值是概括数据的一种有效方法,但有时用平均值却会使我们误入歧途。有一个很古老的笑话,有一个统计学家,他把头放在热的平底锅中,把脚放在冰箱中,然后说:“现在,在平均的意义上我感觉很好。”在计算这个统计学家的“平均”时,两个特殊的温度,平底锅的高温与冰箱的低温相互抵消了,产生了所谓的舒适的平均温度。实际上任何一种平均都会掩盖数据的极值,而极值有时正是我们感兴趣的。一个社区的 平均家庭收入可能是令人舒服的每年 10 万美元,但如果这个均值是从 200 户极穷的人家与 20 户极富的人家的收入计算得来的,它就不能代表他们中的任何一个。有时我们还需要利用平均值以外的东西来概括数据。

想象两个有相同平均值但又有区别的数据集。在一个数据集中,观察值互相很接近,而另一个数据集中,数据散布很开。没有一种平均值——众数、均值、中位数——可以表现出这种重要的不同。在这个例子中我们需要考虑数据的分散程度。

极差：套住两个极端值

一种简单的度量数据分散度的方法就是找出 极差(range),即最大与最小观察值的差:

极差是变量观察值中最大值与最小值的差。

极差 = 最大的观察值 - 最小的观察值

对于新娘年龄的数据:

$$\text{极差} = 60 \text{ 岁} - 19 \text{ 岁} = 41 \text{ 岁}$$

极差很容易计算,而且常常是一个很有用的数。数据的平均值和它的极差可以告诉我们很多被观测变量的信息。如果数据不包含一些极端的值,平均值就会更准确。极差有一个缺点就是它对极端值十分敏感,如果两个最大的观察值 56 岁与 60 岁被去掉,那么最大的观察值就是 46 岁,从而新的极差就是 $46 - 19 = 27$ 。37 个数中仅有的两个观察值就使极差增加了 50%! 去掉某些极端值会使剩余数字的极差是一种统计策略(只要去掉的观察值的个数是允许的)。

停下来想一想 4.4

学院手册中常常列出各个学院的各项指标。其中有入学班级的数学和语文的 SAT 成绩的四分位间距。例如,在 Swarthmore 学院,一个刚入学的班级的语文成绩的四分位数间距是 $690 - 580 = 110$,数学成绩的四分位数间距是 $720 - 630 = 90$ 。(极差不仅在证明中位数得分太高时有用,而且可以得出,这个学校的大部分学生,语文高于 580 分而数学高于 630 分。)在附近一个海派率不太大的私人学校中,语文的中位数分数是 530 分,数学是 597 分,语文的四分位数极差是 $579 - 483 = 96$,数学的四分位数极差是 $653 - 552 = 101$ 。利用 Swarthmore 的四分位数划分,如何找到这个学校的语文得分的中位数?

当最小的 25% 的数据与最大的 25% 的数据都被去掉后, 极差是所剩中间部分的观察值的极差, 即所谓的四分位数极差(interquartile range)。无论是全极差还是四分位数极差都可以用盒式图表示出来。在图 3.3(新娘年龄的盒式图)中, 四分位极差数覆盖了盒子的全长, 所以四分位极差为 $32 - 24 = 8$ 岁。

标准差: 重要的偏差

标准差: 重要的偏差, 是到均值的一种平均距离。

标准差是最常用的统计量, 它主要用来说明一个变量的观察值之间如何的不同。标准差(standard deviation)说明了观察值与均值相差多远(图 4.4)。离均值越远, 彼此之间离得也远, 则标准差越大。例如, 如果我们知道心脏跳动次数的标准差是 6.9, 则我们知道一个典型的观察值距离均值在 6.9 左右, 不是大些就是小些。标准差的最小值是 0.00, 这时数据集的各个观察值一样大。但没有变化的数据是极少的, 大量的数据服从具有某种分散度的分布。标准差的最大值没有限制。



图 4.4 均值作为数据的中心, 而标准差作为其分散度。

标准差一般用 s 表示。标准差是有点奇怪的名字; 什么东西可以同时是标准的又是有偏差的呢? 随着你对标准差的了解, 你会对这个名字有更多的了解。

比较数据的分散度: 以下四个数据集的直方图列在图 4.5 中:

- (a) 6 6 6 6 6 6 6
- (b) 5 5 6 6 6 7 7
- (c) 3 3 4 6 8 9 9
- (d) 3 3 3 6 9 9 9

直方图表明, 数据离均值 \bar{x} 越远则标准差 s 越大。

停下来想一想 4.5

想出日常生活中标准差很小的一些例子; 另外再想出一些日常生活中标准差很大的一些例子。(你可以任意在跑马场与杂货店之间随意畅想。)

在图 4.5a 中, 所有的观察值都是 6, 因为它们都相同, 所以这个变量的标准差是 0.00。在图 4.5b 中, 观察值在 5 与 7 之间变化, 所以标准差增加到 0.82。在图 4.5c 中, 观察值的变化

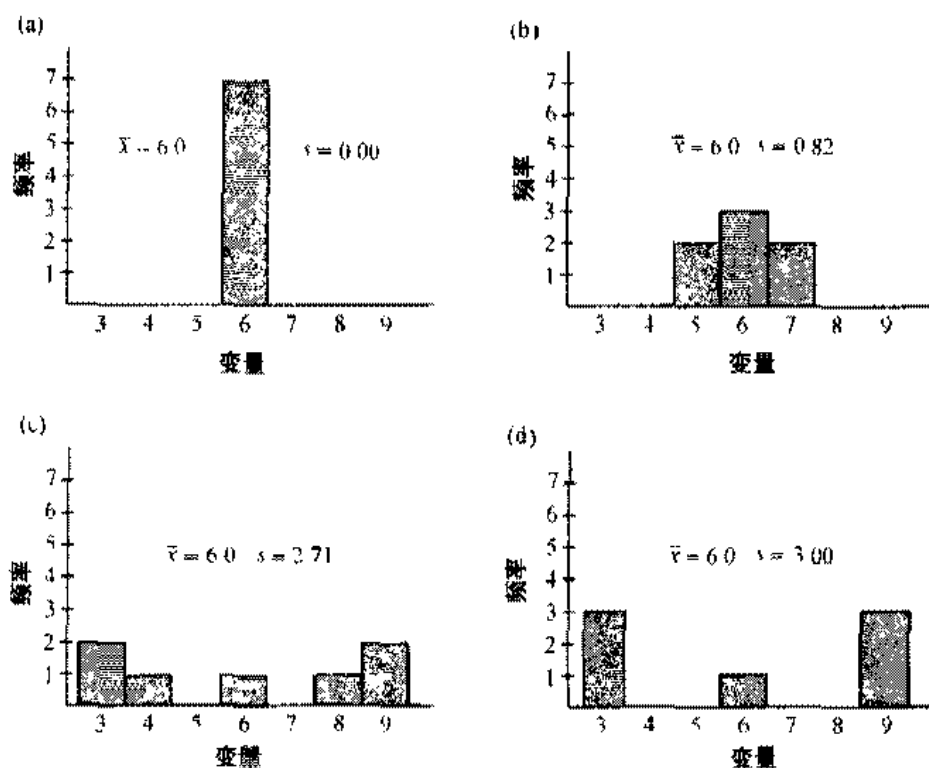


图 4.5 有同一均值不同分散度的四个数据集的直方图。

范围更大,标准差增加到2.71。在图 4.5d 中,大部分观察值都在两个极值处,标准差达到了3.00。这四个数据集都有相同的均值,如果我们只知道均值,我们就不能说出此四个数据集的区别。但是,各个数据集在观察值与均值的差与数据值变化范围方面是有区别的,所以它们有不同的标准差。

与均值的平均距离: 解剖数字 标准差 s 可以由先计算观察值与均值的差,然后将这个数字平方后求和,最后再取平均值获得。为了弄清如何计算,让我们对 4.4b 的数据,即观察值为 5, 5, 6, 6, 6, 7, 7 的数据进行逐步计算。对这些数据我们如何得到 0.82 的标准差呢? 我们知道观察值的均值是 6,按照定义我们需要知道每一个观察值与均值的差。第一个与均值的差是 $5 - 6 = -1$,第二个是 $5 - 6 = -1$,第三个是 $6 - 6 = 0$,第四个是 $6 - 6 = 0$,第五个是 $6 - 6 = 0$,第六个是 $7 - 6 = 1$,第七个是 $7 - 6 = 1$ 。这就是表 4.1 的第二列。

第二步,我们需要计算偏差的平方,这就是表中的第三列: 1, 1, 0, 0, 0, 1, 1,它们的和是 4,然后我们使这个数平均得到 0.67。最后对这个数开方得到标准差 s 是 0.82。本章末公式 4.4 与公式 4.5 表明如何求标准差。

表 4.1 中的第二列表明观察值与均值的离差的变化范围从 -1 到 1,而标准差 $s = 0.82$ 是这个变化范围内中间的某一点。它比最大离差 1 小,比最小离差 0 大。这样标准差 0.82 对于平均偏差来说并不是不合理的。

我们首先对每个观察值与均值的差作平方运算的原因是要去掉负号,平方后的单位就是原始观察值的单位的平方。例如,如果原始数据的单位是美元,那平方后的单位就是美元²。美元²也是平方的平均值 0.67 的单位。但我们很难解释这样的数:什么是美元²呢? 通过开

平方我们可以恢复原始形式的单位。

表 4.1 计算标准差 s ：与均值离差的平方的平均值的方根

观察值	与均值的离差	离差的平方
5	$5 - 6 = -1$	$(-1)^2 = 1$
5	$5 - 6 = -1$	$(-1)^2 = 1$
6	$6 - 6 = 0$	$0^2 = 0$
6	$6 - 6 = 0$	$0^2 = 0$
6	$6 - 6 = 0$	$0^2 = 0$
7	$7 - 6 = 1$	$1^2 = 1$
7	$7 - 6 = 1$	$1^2 = 1$
和	0	4
平均值	0	$\frac{4}{6} = 0.67$
平方根		$s = \sqrt{0.67} = 0.82$

我们现在还没有谈到一个小小的细节，为何在图表中有 7 个值，但我们却用 6 去除得到平方的平均值呢？这不是算错了，而是应用比应有观察值少 1 个的数去除进行计算会有更好的结果。这个细节我们会在第七章应用 s^2 进行估计时作详细的叙述。

绝对离差的平均值：最低的 CAL 选择

表 4.1 中原始离差的平均值是零，因此离差的和是零，即负离差通常和正离差会相互抵消。这样原始离差的均值就很难告诉我们什么信息。你可以通过前面的任何例子自己检验一下。也许我们应该取所有的离差的绝对值，然后找出这些数据的均值，通过这种办法我们也可以将某些离差的负号去掉。上例中，绝对值的和是 4，绝对值的均值是 $4/7 = 0.57$ ，该绝对离差的均值明显地比标准差小。

因为标准差在点估计与假设检验中非常有用，所以统计学家常用标准差代替离差均值，然而作为标准差的快速、低运算量替代品，绝对离差均值是很好的统计量。它可以很容易地猜出数的分布状态。

大部分时间里，我们避免冗长的叙述，而是仅仅使用标准差描述观察值的平均差异。如果标准差小，那观察值就相像；如果标准差大，那观察值彼此间就会很不同。

在均值上加减标准差：标准差可以用于另外一个有趣的解释。图 4.6 是 27 个人每 30 秒心跳次数的直方图，平均心跳次数是 34.4，标准差是 6.9。如我们所料，均值落在了直方图的中部，直方图的平衡点。

均值加标准差是 $34.4 + 6.9 = 41.3$ ，我们将其标在直方图下的横线上，这条横线还表明了均值加两倍的标准差，即 $34.4 + 2(6.9) = 48.2$ 。同样均值减标准差是 $34.4 - 6.9 = 27.5$ ，均值

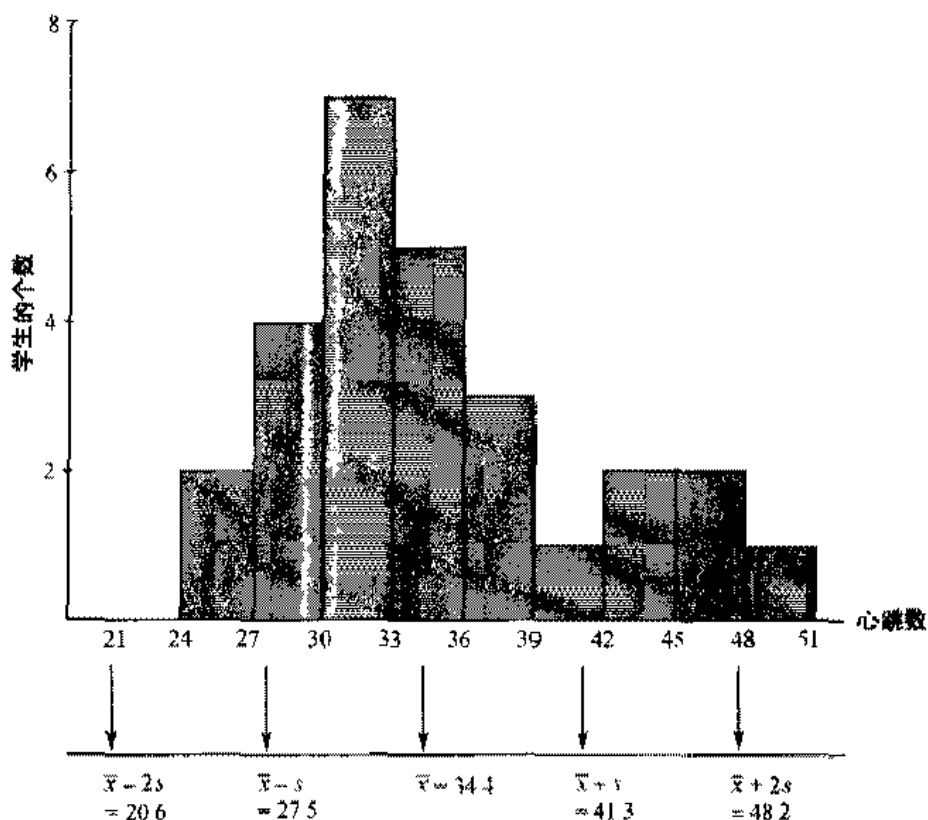


图 4.6 每 30 秒心跳数及其均值和标准差($\bar{x} = 34.4$)。(来源: Data collected from *Students in Statistics I: Statistical Thinking*, Swarthmore College, spring 1995)

减两倍的标准差是 $34.4 - 2(6.9) = 20.6$ 。这个图表明了一个区间即从均值减两倍的标准差到均值加两倍的标准差,在这个例子中即是从 20.6 到 48.2,几乎包含了所有的观察值。27 个观察值中只有一个落在了这个区间的外面。均值加减一个标准差的值(即 27.5 到 41.3)包含了所有观测值的 $\frac{2}{3}$ 。对于大多数合理的单峰对称分布,我们可以期望同样的结果。这样如果我们知道均值与标准差,我们就可以重新产生直方图,因为观察值的极差大致等于四个标准差。找到极差并且除以 4,则可以得到标准差的估计值。这个小规则常可快速地估计以下标准差的大小。

停下来想一想 4.6

你知道,过去 50 年中,Milles Lacs 湖的冰上垂钓的时间在 25 天到 73 天之间。估计冰上钓鱼的时间长短的标准差。大部分年份里,是否可以相当准确地估计你在冰棚中所呆的天数(如果你想每天钓鱼的话)? 给出可能天数的极差。

方差是标准差的平方, 是测量变化程度的一种准则。

方差: 标准差的平方 由于数学的原因, 统计学家有时愿意用方差(variance)代替标准差作为测量观察值差异的工具。方差是标准差的平方 s^2 。在新娘年龄的数据中, 标准差是 9.0, 方差为这个数的平方 81.0, 即

$$s^2 = (9.0 \text{ 岁})^2 = 81.0 \text{ 岁}^2$$

方差并没有比标准差告诉我们, 更多的东西, 而且方差很难解释, 因为方差的单位是原变量单位的平方。而标准差和均值一样, 它们的单位就是原变量的单位。什么叫 81 平方年呢?

4.3 均值的标准误差

统计分析中的一个主要准则是如果我们对事物进行第二次测量, 则通常得到不同的结果。在新娘平均年龄的数据中, 一个新娘是 19 岁, 另一个是 22 岁, 等等。如果我们注意到这个变量的所有观察值, 我们可以发现大部分观察值都彼此不同, 标准差告诉我们, 这些不同的程度有多大。

在新娘年龄的数据集中, 37 个新娘是一个样本。这个样本的均值是 30.0 岁。假设我们选择的是另外一组 37 个新娘的随机样本, 并且观测她们的年龄。再次做同样的实验, 我们一定会得到结婚年龄的不同的均值。重复此实验多次, 一定会得到多个不同的均值, 因此正如在一次研究中个体的观测值通常是不同的一样, 对不同的样本, 样本均值通常也是不同的。

重复研究产生的各均值的差异是多大呢? 他们是否比个体观测的差异小或相同呢?

停下来想一想 4.7

你能将上面的问题在我们未给答案之前回答出来吗? 你的答案是什么?

标准误差是很多不同样本的均值的标准差。

回答这个问题的一种方法是找到所有均值的标准差, 各均值只是一行数字, 就像 37 个原始观察值一样, 所以对不同变量找不同样本的均值的标准差与找一个变量的原始观察值的标准差没有什么区别。唯一的区别就是, 找均值的标准差, 我们需要先将每一个样本的均值计算出来。因此有时我们处理样本原始观察值的标准差, 有时我们处理从原始观察中得到的一系列数字的标准差, 例如均值。为了区别这两种标准差, 由原始观察值算出的叫做标准差(standard deviation), 由一组均值算出的叫做标准误差(standard error)。类似地, 标准误差同样可以由一系列中位数或一系列标准差计算得到!

停下来想一想 4.8

为什么均值的标准误差对于大的样本比对于小的样本小是不值得奇怪的?

均值的标准误差比观测值的标准差小,这就是说,均值的变化比变量原始观测值的变化小。这是不奇怪的。某些特殊样本包含有较大或较小观察值,它们在计算均值时互相抵消了,使得均值留在中部。每一个样本都会发生同样的事。样本越多,一个样本均值与另一个样本均值的变差就越小,这使得标准误差也较小。

标准差与标准误差的最大区别在于寻找标准差仅需要一个样本的数据,但是寻找标准误差需要多个样本。然而,从一个样本中的数据估计标准误差也是可能的(见本章结尾的公式 4.6)。在 37 个新娘的例子中,从多个样本中得到的均值的标准误差是 1.5 年。这个例子中的标准差是 9.0 年。显然,均值的标准误差比观察值的标准差小很多。

均值的标准误差是一个很有用的统计量。在新娘的例子中,两倍的的标准误差是 3.0 年。加减两倍的均值的标准误差可以得到一个长度为 6.0 年的区间。如果我们有足够的样本和样本均值,那么大部分的样本的均值会落在这个 6.0 年的区间之中。

4.4 标准得分: 比较苹果和桔子

不同的变量一般有不同的均值和标准差。统计上,均值和标准差不同时,一个变量的值不能与另一个变量的值相比较。在结婚年龄的例子中,新娘年龄的均值是 30.0 岁,标准差是 9.0 年,新郎年龄的均值是 32.4 岁,标准差是 11.1 岁。在这个群体中,最年轻的一对的新娘的年龄是 19 岁,新郎的年龄是 17 岁。

我们怎样比较这一对的年龄呢? 新郎明显比新娘年轻,但是否新娘的年龄与新郎的年龄一样,在 37 位新娘与新郎中都处于最年轻的位置上呢? 哪一个在统计上更破例,新娘还是新郎? 这一对新娘与新郎怎样成为最年轻的一对呢? 一个粗浅的解决办法就是将新娘与新郎的年龄变换到一个尺度下: 我们将原始得分变换为**标准得分**(standard scores)(公式 4.7)。新娘与新郎的年龄——原始得分——被变换为能表示原始得分和均值的距离的新得分(按标准差单位)。使用标准得分,一个变量的任何值都可以和任何其它变量的值相比较,因为我们知道任何一个得分与均值的相对距离。

标准得分是用某一观察值减均值所得的差除以标准差所得的值。

将年龄数据转换成标准得分的目的是建立一个新的标准得分的标度来代替老的原始得分的标度。例如那个 19 岁的新娘,原始得分是 19,样本均值是 30,标准差是 9.0。则标准得分是:

$$\frac{19.0 - 30.0}{9.0} = -1.22$$

同样,17 岁的新郎的标准得分是 $(17 - 32.4)/11.1 = -1.39$ 。通过均值和标准差,我们发



12, 相应的标准得分是 -2.00 。

任何变量的标准得分的值大部分在 -2.00 到 2.00 之间变化。如果一个观察值的标准得分大于 $+2.00$ 或小于 -2.00 , 那么这个观察值就不寻常的大或小。不寻常值常帮我们从中得出结论, 并且应用在实际生活中。标准得分常称为 t -值。

4.5 简单化的收益与信息的丢失

用图表来代替数据

画图, 制表和计算汇总统计量的目的为了更好地理解数据。每一种简化数据的方法都可以得到原始数据中并不明显的模式。但同时原始数据中的某些细节被丢失了。让我们以讨论简单化与信息的丢失之间的矛盾来结束本章。

看图 4.8。图中在方框中的数据是 30 个不同的国家中每 100000 个男人由于肝硬化的死亡率。我们从这个变量的 30 个值中可以得到什么呢? 除了最小值 1.5 与最大值 50.1, 我们很难看出这些数据是如何分布的?

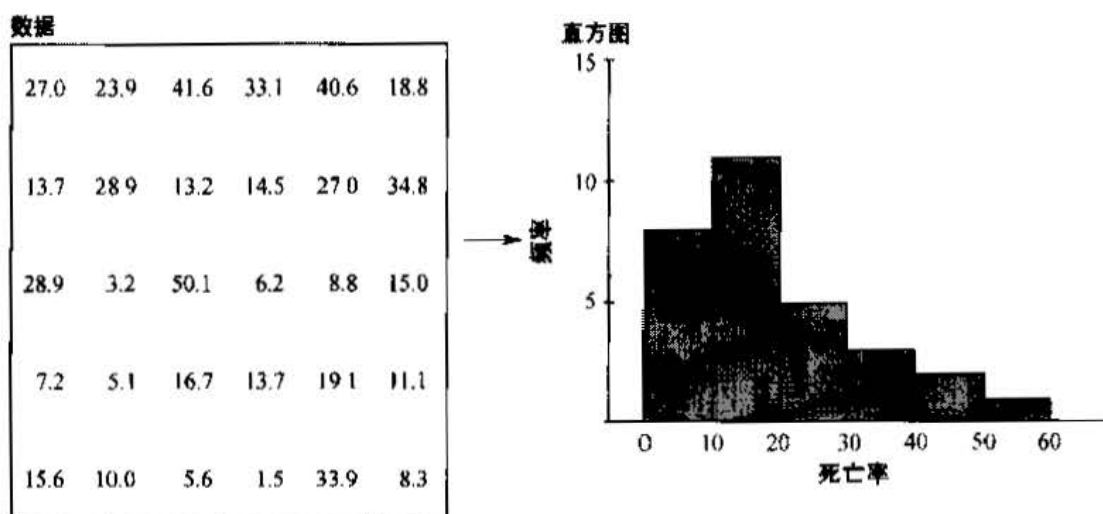


图 4.8 把选定国家的每 100000 人的肝硬化死亡率数据简化为直方图。(来源: Ann Cronin, *The t/Tipplers and the Temperate: Drinking Around the World*, The New York Times, January 1, 1995, p. E4.)

当用右边的直方图来代替左边的数据时, 就很容易理解这些数据了。这 30 个数据被简化为直方图中的 6 个矩形。直方图显示这是一个单众数的有偏斜分布。多于一半的数据分布在 10 到 30 之间。

同时, 这个数据集的某些信息——个体观测值——丢失了。例如, 直方图表示了 50 到 60 之间有一个观察值, 但是并未指出这个观察值是多少。直方图还破坏了知道某事件发生的时间。例如, 我们每月末收集每夸脱牛奶的价格的数据, 如果我们将其做成直方图则会丢失观测值的顺序。直方图只能显示某一价格出现在某一特定区间的次数。

用汇总值代替数据

图 4.9 把肝硬化的死亡率数据汇总成一个单独数——均值。平均死亡率 19.2 使我们对死亡率的大小有一个直接的概念, 这就是该组数据的中心。对一个如 19.2 这样的数的理解要比对变量的 30 个不同值的理解容易得多。知道了均值我们马上就可以知道数据的中心所在。

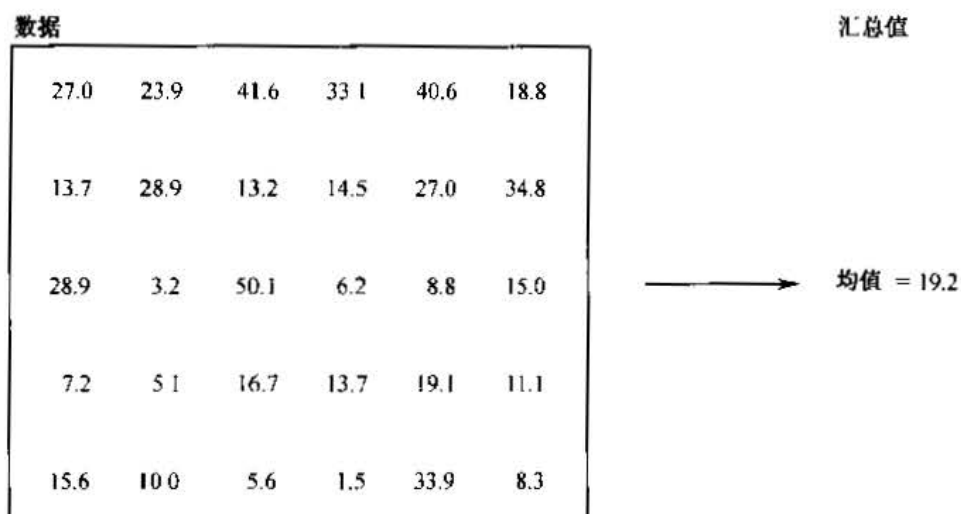


图 4.9 把选定国家的每 100000 人的肝硬化死亡率数据简化为均值。(来源: Ann Cronin, *The 1/Tipplers and the Temperate: Drinking Around the World*, The New York Times, January 1, 1995, p. E4.)

但是, 在将 30 个数据概括为一个数据的同时, 大量的信息却丢失了, 而且原始数据不可能从均值得到恢复。在计算均值时, 怎样使信息丢失与简单化的收益达到平衡依赖于数据要说明什么。一开始收集原始数据就是因为研究者对这个疾病有某些问题。

4.6 房地产数据：看不见的价格

一个学校的董事会研究提高房地产税以增加学校下一年度的收入。在税收增加之前, 董事会需要了解该校区和邻近校区域的税收状况。董事可以要求政府官员提供前一年的税收数字, 并且一页一页的察看每一个纳税人的财产评估和税额。但是, 他们从县记录得到的是他们感兴趣的数据的汇总值, 诸如平均财产价值, 平均评估值, 和每一个纳税人的平均纳税值。表 4.2 引述的是该校数据的部分节录。

除了实际房屋销售价格, 税评估, 税收额等地产数据, 表 4.2 同时显示了某些汇总数据。我们已经快要结束关于计算汇总数值的这一章了, 到底关于这三个变量的汇总数据代表的是什么意思呢?

表 4.2 1995 年 Swarthmore 地区的房子的销售价格,税评估和税收额

地址	住房销售价格	税评估	税收额
520 Cedar	\$ 335000	\$ 6400	\$ 4752
326 Cornell	220000	3300	2700
9 Cresson	183750	6500	5260
609 Elm	237000	6000	4620
60 Forest	246000	6000	4456
9 Guernsey	370000	9500	7055
624 North Chester	249000	5000	3849
513 Ogden	290500	7000	5774
310 Park	195000	4200	2800
529 Rutgers	176000	5600	4696
633 Strath Haven	272500	8000	6001
621 University	265000	6300	5132
10 Wellesley	340000	10000	7501
均值	\$ 259981	\$ 6446	\$ 4969
中位数	249000	6300	4752
标准差	61086	1890	1420
四分位数极差	105250	2200	1735

来源:感谢 D.Patrick welsh 房地产公司的 David welsh, 是他们从市长办公室获得此数据。

均值和中位数告诉我们,表中的平均房价是大约 250000 美元。因为均值比中位数稍大,这个表中一定包含几个十分昂贵的房子,正是它们把均值拉到中位数之上。税评估值与相应房子的销售价格之间有怎样的关系不能从表中的四个汇总数据中得出。要知道这个信息,我们必须应用第十章中的方法进行统计。税评估值的均值再一次比中位数高,因此一定有几个特别高的评估值使均值增高。

房子的税收额平均为 5000 元,均值也比中位数高。这样,一般说来,税收相当于房屋销售价格的 1/50;这个值可能对学校董事会有用,可以用来比较其他校区的类似数据以使他们决定如何提高税收。

均值与中位数的大小使我们知道这三个分布是偏斜的,但是标准差也使我们可以知道一些价格、评估值和税收的变化。从每一个均值上减去一个标准差表明没有几个房子的售价低于 200000 美元,也没有几个人付低于 3500 美元的房地产税。从每一个均值上加上一个标准差表明较多的房子的售价高于 320000 美元,而且房主付多于 6500 美元的税。

使用这一章中的统计方法与本校区及其它社区的类似数据,学校董事会可以了解一些本区的税收基础及是否应该提高税收;也许还得重新评估售价过高的地产。

4.7 小结

为了寻找一个数据集的规律,我们需要对观察值进行汇总。通过图表和汇总数字大大简化了数据,但同时一部分信息会丢失。

4.1 各种平均数：让我们数数有几种

三种最常用的平均数是众数、中位数、均值。众数为变量最常出现的观察值。二众数分布有两个最经常出现的值。当变量是分类变量时,使用众数是很必要的。

中位数将观察值分成相同数目的两部分,其中一部分都比这个数小而另一部分都比这个数大。当直方图显示数据服从一个非对称分布时,中位数是最常用的平均值。这是因为中位数受孤立于数据主体的极端值影响不大,中位数也叫做第 50 个百分位数。

均值——将所有的观察值都考虑进去所得的平均值——是最常用的一种平均值。将所有观察值的值相加后除以观察值的个数就可以得到均值。均值的符号是 \bar{x} 。如果均值与中位数的大小大致相等,则应选择均值作为平均值。如果它们有很大不同,那么中位数要合适些。对于有非对称的数据集,中位数更实际地描述了数据的中心。

4.2 变差：测量生活的趣味

除了知道数据的中心以外,知道数据如何散布是重要的。测量分数度的一种方法就是计算极差,即最大的观察值与最小的观察值之间的差。极差的一个缺点是对极值过于敏感。有时我们去掉最小的 25% 的数据与最大的 25% 的数据,然后求出剩下的中间数据的极差,这中间的 50% 数据的极差又叫作四分位数极差。

标准差 s 是观察值与均值离差的平方的均值的平方根。它表明了 in 平均意义上观察值与均值的偏离。这是最常用的也是统计上测量数据离散度的最复杂的方法。通常大约 2/3 的观察值落在离均值一个标准差的距离内,几乎所有的观察值落在离均值两个标准差的范围之内。标准差的平方叫方差,记作 s^2 。

4.3 均值的标准误差

均值的标准误差是多个样本的均值的标准差。均值的标准误差比观察值的标准差要小,这是因为均值的变化程度比观察值的变化程度要小。

4.4 标准得分：比较苹果和桔子

一个变量的所有观察值都可以变作标准得分。标准得分等于观察值减均值再除以标准差。它的功能是评价一个观察值相对于所有观察值的均值与标准差相比的大小。大部分的标准得分在 -2 到 2 之间;标准得分位于这个区间之外是不寻常的。

4.5 简单化的收益与信息的丢失

用图表或汇总数来简化数据意味着促进理解,但是原始数据的细节却丢失了。

4.6 房地产数据：看不见的价格

这一章的概念源于解决实际生活中的问题。

补充读物

Weisberg, Herbert F. Central Tendency and Variability (Sage University Paper Series on Quantitative Applications in the Social Sciences, no. 07 - 083). Newbury Park, CA: Sage, 1992. 这本书讨论了计算中心趋势与变差的不同种方法。

Witmer, Jeffrey. DATA Analysis: An Introduction. Englewood Cliffs, NJ: Prentice Hall, 1992. 这本书给出了从数据中计算的很多不同的量。

公 式

x 表示变量, 这个变量的 n 观察值表示为

$$x_1, x_2, x_3, \dots, x_n$$

然后将观察值从最小到最大排序。为了表示我们对观察值进行了排序, 我们在下标上加括号。这样 $x_{(1)}$ 代表最小的观察值, $x_{(2)}$ 代表第二小的观察值, 而 $x_{(n)}$ 代表最大的观察值。这样排序后的观察值可以写为:

$$x_{(1)}, x_{(2)}, x_{(3)}, \dots, x_{(n)}$$

中位数

当 n 是一个奇数时, 中位数是排序后的观察值的中间值。用符号可以表示为:

$$\text{中位数} = x_{(n+1)/2} \quad (4.1)$$

例如, 如果数据集有 $n = 11$ 个观察值, 那么 $(n+1)/2 = (11+1)/2 = 12/2 = 6$, 则中位数是第六大的观察值, $x_{(6)}$ 。有五个观察值比中位数小, 有五个观察值比中位数大。中位数就是中间观察值的值。

当 n 是一个偶数时, 中位数可以通过计算两个中间值的中点来获得。

$$\text{中位数} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2} \quad (4.2)$$

如果数据集有 12 个观察值, 那么中位数就是:

$$\frac{x_{(n/2)} + x_{(n/2+1)}}{2} = \frac{x_{(6)} + x_{(7)}}{2}$$

这就是第六个与第七个观察值的中点。有一半的观察值比这个数小, 另一半的观察值比这个数大。

均值

均值 \bar{x} 是所有观察值的除以观察值的个数后所得的值。

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum x}{n} \quad (4.3)$$

这里 $\sum x$ 代表 x 变量的所有观察值的和。

标准差和方差

标准差是数据集中一个典型观察值与均值的距离。为了找到标准差 s , 我们必须找到方差 s^2 然后计算它的平方根。每一个观察值减去均值的, 平方每一个差, 然后再将平方后的数值相加, 最后将这个数用除以 $n - 1$ 所得的值为方差:

观察值	差	平方
x_1	$x_1 - \bar{x}$	$(x_1 - \bar{x})^2$
x_2	$x_2 - \bar{x}$	$(x_2 - \bar{x})^2$
x_3	$x_3 - \bar{x}$	$(x_3 - \bar{x})^2$
\vdots	\vdots	\vdots
x_n	$x_n - \bar{x}$	$(x_n - \bar{x})^2$
和	0	$\sum (x_n - \bar{x})^2$

观察值偏差的和总是零。该和为零是检验偏差是否计算正确的核对。方差 s^2 是偏差的平方和用 $n - 1$ 除后所得的数:

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} \quad (4.4)$$

方差是与均值的均方差。用 $n - 1$ 而不用 n 除的原因我们将在第六章估计中进行详细的叙述。

标准差 s 是方差的(正)平方根, 即:

$$s = \sqrt{s^2} \quad (4.5)$$

计算标准差的三步十分烦琐。均值计算中的四舍五入误差都会带进每一步的平方计算中。但是尽管有很多计算方差的更简单和精确的公式, 这个公式是定义公式。如使用计算器或计算机则方便很多。

均值的标准误差

两个或多个均值的标准误差可以从一个大样本的一个包含 n 个观测值的样本获得。首先获得标准差 s , 然后标准误差 $s.e.(x)$ 可以通过用标准差 s 除以观察值的个数 n 的平方根来获得:

$$s.e.(\bar{x}) = \frac{s}{\sqrt{n}} \quad (4.6)$$

有时, 均值的标准误差用符号 $s_{(\bar{x})}$ 来表示。

标准得分

一个观察值的标准得分可以用观察值减去均值再除以标准差获得:

$$\text{标准得分} = \frac{x - \bar{x}}{s} \quad (4.7)$$

习题

(习题 4.1—4.30)

- 4.1 这一章中汇总数据的两个主要目的是什么?
- 4.2 为什么数据中太多的信息会为我们理解数据带来困难?
- 4.3 给出众数、中位数、均值的定义。
- 4.4 给出例子,使(众数、中位数、均值中的)每一种平均数都有比其它两种好的理由。
- 4.5 给出一个服从二众数分布的例子。
- 4.6 找出一篇利用某种平均数来说明其观点的报纸文章。
 - a. 说出文章中平均数的种类。
 - b. 说出作者使用这种平均数的目的。
 - c. 作者使用了合适的平均数吗?
- 4.7 你了解你所在社区的经济状况。
 - a. 使用你社区居民收入的均值平均还是中位数更合适? 为什么?
 - b. 你能想出一种情况使工人工资的统计众数比中位数和均值都更合理吗? 如果你能想出来,叙述这种情况。
- 4.8 解释如下陈述: 对于非对称数据,使用中位数比均值更合理。
- 4.9 想出一个汇总统计量可大大增进理解的有很多观察值的变量的例子。(例如: 这本书中每页的字数。)
- 4.10
 - a. 定义极差。
 - b. 举出一个极差的优点。
 - c. 极差是中心趋势的度量还是变化程度的度量? 为什么?
 - d. 你是否在家中可以发现极差呢?
- 4.11
 - a. 分数的分布中什么样的因素使极差不敏感?
 - b. 数据中的什么因素使得极差极其敏感?
- 4.12 观察值离均值越远,则标准差(越大,越小)。选择正确的形容词,并解释这种说法。
- 4.13 我们用什么字母来代表标准差?
- 4.14 在六次 Shaquille 自由投掷中,得分为: 5, 5, 5, 5, 5, 5。
 - a. 标准差是多少?
 - b. 为什么?
- 4.15 对于大部分的单众数 and 对称分布来说,在距离均值一个标准差的区间内,你期望发现多大比例的观察值?
- 4.16 对于大部分的单众数 and 对称分布来说,几乎所有的数据可在均值两边大约多少个标准差的范围内?
- 4.17 标准差的平方是什么?
- 4.18
 - a. 按照一个简单的规则,大部分分布的极差大约等于多少倍的标准差?
 - b. 另一方面,标准差大约是极差的几分之几?
- 4.19 假设我们请几个人对不同的比萨饼打分,分数为从 0 到 10,10 分为十分满意。为什么你希望买具有高的均值记分和低标准差的比萨饼?

- 4.20 均值的标准误差是什么?
- 4.21 为什么均值的标准误差比样本观察值的离差要小?
- 4.22 估计均值的标准误差需要知道什么信息?
- 4.23 a. 解释为什么将原始得分转化为标准得分是必要的?
b. 给出一个你经历的将原始得分转化为标准得分带来很大帮助的例子。
- 4.24 怎样将原始得分转化为标准得分?
- 4.25 一般地,标准得分在哪两个数字之间变化?
- 4.26 如果一个算命先生告诉你,你的智商的标准得分是 +15.5,你是会高兴得庆祝你是天才呢,还是认为算命先生服了迷幻药?请解释。
- 4.27 标准得分常常被称为什么值(看练习 4.26 中的双关语)?
- 4.28 如果你在说服你的合作者,将原始得分转化为标准得分是十分必要的,你认为标准得分相对于原始数据的哪一条优点是最重要的?
- 4.29 “标准差测量了数据的随机性”这样的说法意味着什么?
- 4.30 标准差和标准误差的不同是什么?

解释:(习题 4.31—4.52)

- 4.31 变量的标准差等于零,说明了变量具有什么特性?
- 4.32 直到 1992 年,国会成员可以用他们在一个国际银行的支票而且不会因为开了空头支票而受惩罚。报纸公布了每个众议院议员所开的空头支票数目。其数目的中位数为 3,而均值为 47。这些数目能给出关于空头支票的数目的分布的什么性质?
- 4.33 一篇报纸的新闻报道,1989 年美国的一般家庭收入是 \$35225(来源: New York Times, July 26, 1992, p.E5) 为什么这个数值可能是收入的中位数而不是均值?
- 4.34 习题 4.33 的文章中同样提到一般的美国人是女人,重 144 磅,住在有 2.6 个卧室的住宅中,每周看 28 小时 13 分钟的电视,家庭收入是 \$35225。
a. 这些数字中那一个是众数?
b. 中位数的特点是什么?
c. 均值的特点是什么?
- 4.35 一篇新闻报道指出一般的女人有 2.1 个孩子。你十岁的弟弟问你“这是可能的吗?不可能出现小数个小孩。”你如何回答?
- 4.36 对性别变量,众数的值是女性。说出众数作为一个汇总统计量的优缺点?
- 4.37 说出中位数作为一个汇总统计量的主要优点是什么?
- 4.38 Slippery Rock State 大学接受的语文 SAT 测验的入学分数中位数是 550 分,你的朋友得了 500 分,你是否应告诉你的朋友不要申请 Slippery Rock State 大学,还是应继续先查看其它信息?解释你的选择?
- 4.39 一个关于工人生产率的调查显示,在 1 到 100 的记分中,美国工人的得分是 100 分,法国工人是 95 分,西德工人是 89 分,日本工人是 77 分,英国工人是 75 分。文章的标题指出美国工人的生产力大于法国,德国,日本和英国工人。在文章中报道了这些群体的生产率的经济指标。“1990 年,一个全职的美国工人的产出是 \$49600,西德是 \$44200,日本是 \$38200,英国是 \$37100。”这项调查排除了政府,教育系统,卫生系统和房地产业的工作人员。(来源: Alex Domunquez, “Study say US workers are the world's top

producers,"The Philadelphia Inquirer, October 14, 1992, p. D-1.)

- a. 标题说“美国工人是世界上最好的生产者。”在什么意义上该标题是准确的,而在什么方面是误导的?
 - b. 你在这篇文章中是否发现了错误或忽略?(至少有一处。)如果你是编者,怎样纠正?
 - c. 为什么假设政府,教育,卫生,物业管理排除在这项研究之外?你能说出排除这些人的影响是什么吗?
- 4.40 在得分值相同时,你能描述出两种得分分布的不同吗?其中两个分布的均值相同,一个分布的标准差比另一个分布的标准差大一倍。
- 4.41 Atlanta Braves 队每场比赛失误的均值是 1.3,全赛季失误的标准差是 1.0。Philadelphia Phillies 队失误的均值是 2.0,标准差是 0.3。你认为以下哪一种叙述是正确的,说出理由。
- a. Braves 队比 Phillies 队打较多失误少的球。
 - b. Phillies 队比 Braves 队在失误上比较稳定。
 - c. Braves 队有时表现很差,有时又非常好。
 - d. Phillies 队很少不失误。
- 4.42 你在一份报纸中知道,一个小型四年制大学,24 岁以下的学生平均每周喝 7.0 杯酒。相对地在超过 20000 学生的校园的学生每周喝 4.6 杯酒。假设每一个样本的标准差是 2.0,使用你对标准差的知识,讨论以下叙述:
- a. 在小学校,大约 66% 的学生每周饮酒量在()和()之间。
 - b. 在大学校,大约 66% 的学生每周饮酒量在()和()之间。
 - c. 如果一个学生说她每周喝 6 杯酒,你能否确定她来自一个小学校?
 - d. 你如何描述大学校中,在 b) 中的另外 33% 的人的喝酒习惯?如何描述小学校中,在 a) 中的另外 33% 的人的喝酒习惯?
 - e. 校园中是否有很多人根本不饮酒?
- 4.43 你被告知你的孩子在阅读中的标准记分为 +1.80,数学为 2.00 分。你还被告知你的孩子在音乐理解中所得的标准记分是 0.00。
- a. 假设班级中有各种层次的学生,你的孩子在学业上达到很高水平的机会是多大?
 - b. 音乐理解的得分能否使你认为你的家庭有音乐方面的影响?你怎样理解音乐理解的得分?
- 4.44 从最初 19 个现代奥林匹克运动会的数据中,我们得知,男子跳远冠军成绩的均值是 308 英寸,中位数是 310 英寸,标准差是 19 英寸。这三个数据告诉你原始数据的什么信息?
- 4.45 某一年,明尼苏达州 Hibbing 的温度的众数是华氏 32 度(摄氏 0 度)。而明尼苏达州的 Duluth 的温度服从二众分布,众数分别是 33 和 61 华氏度。从这个数据中我们可以得到多少 Hibbing 和 Duluth 的温度差异?
- 4.46 你在申请百科全书销售的工作。征募者对你说这个工作的利润非常大;实际上,上一年中在 50 名推销员中,最好的挣得了一百万美元,推销员收入的均值是 35000 美元。
- a. 你是否能确定你可以成为该公司的一个成功的推销员?
 - b. 你希望还能获得什么附加信息?
- 4.47 在练习 4.46 中,该百科全书公司的征募者知道你需更多的信息,她告诉你,实际上

并不是所有的推销员都是成功的。赢利的变化范围是从 5000 美元到 1000000 美元。是否这个信息满足你的关于在该公司的收入前景的好奇心? 说明你可能还希望知道什么信息?

- 4.48** 习题 4.6 中的百科全书公司的会计告诉你, 推销员的薪水的四分位数范围是 10000 美元到 30000 美元。
- 你如何使用这条信息来决定是否接受此工作。
 - 为什么均值比四分位数的范围高很多?
 - 你能猜出收入的中位数是多少吗?
- 4.49** 你知道下面三片正在建筑的公寓小区的信息, 你很想买一个公寓以获得有保证的投资回报。
- Rose Valley: 所有公寓的销售价的均值去年增高了 7000 美元, 标准差是 4000 美元。
- Garden City: 所有公寓的销售价的均值去年增加了 5000 美元, 标准差是 1000 美元。
- Media: 所有公寓的销售价的均值去年增加了 6000 美元, 标准差是 800 美元。
- 你认为哪一小区最有可能使你受益? 哪一块最不可能?
 - 如果没有差错的话, 哪一小区使你获得的收益最多?
- 4.50** 等票的旅游者在 La Guardia 机场等不同的纽约到波士顿和纽约到华盛顿的飞机。对所有等票者等待时间的均值是 1 小时, 对于波士顿的旅客来说, 等待时间的标准差是 30 分钟。对于华盛顿的旅客来说等待时间的标准差是 10 分钟, 如何利用这些信息对你的不懂统计的朋友描述旅客的通行情况和他们在售票窗口的心情?
- 4.51** 在 1994 赛季棒球赛的罢工中, 报道揭示球员的薪水的均值大约是 1200000 美元, 而薪水的中位数大约是 500000 美元。从这些数据中你知道棒球运动员的薪水的分布是怎样的?
- 4.52** 考虑两个州的收入的均值。假设一个人从一个州搬到另一个州, 他的搬家能使两个州的收入均值都增加吗?

分析(习题 4.53—4.72)

- 4.53** 浏览 Springer 公司的网址(<http://www.springer-ny.com/supplements/1versen/>)找到和本书相关的文件。打开名为 Baseball Individual Scores 的数据文件。
- 找出每一列的均值、中位数、标准差和极差。
 - 对于每一个变量使用统计软件画出直方图。
 - 在直方图的基础上找出哪一个变量均值是中心趋势的最好代表, 哪一个变量中位数是中心趋势的最好代表? 哪一个变量众数是中心趋势的最好代表?
 - 为什么只对有些变量, 极差才大约是标准差的四倍?
- 4.54** 习题 3.36 给出了独立宣言签字者的寿命的数据。
- 从数据中我们是否能够看出。独立宣言的签字者的寿命作为一个整体比预料的长些还是短些?
 - 观察值的均值等于 -18 年。该均值是否告诉你这些签字者都活了多长?
 - 标准差是 13.2 年, 极差和标准差相比有多大?
 - 有多少个观察值距离均值在两倍的标准差之外?
 - 你如何描述这些人?
 - 从所有数据的直方图中你可否期望中位数与均值有很大不同吗? 说明为什么。
 - 观察值是以一个人的实际寿命与签署独立宣言后的预期寿命的差的形式出现的。

是否有分析这些数之比的必要？解释为什么。

- 4.55 习题 3.34 给出了一个社会经济得分的样本。而另一组人对于相同的变量有下面的值：55, 36, 70, 66, 75, 49。你感兴趣的是这两组人群期望存在多长；你认为，一组中的人越类似则这一组作为一组持续的时间就越长。
- 解释如何衡量每一组成员的相似性。
 - 对每一组计算这种相似性的度量值。
 - 计算这两组的一种对比：两组之间有很大不同吗？解释为什么。
- 4.56 考虑习题 3.34 和 4.55 中的社会经济得分。
- 找到这两个集合的数据的中位数。
 - 中位数告诉了你这两个数据集的什么信息？
 - 这 18 个观察值的总的中位数是什么？
 - 两组样本中各有多少个观察值数小于组合中位数，有多少个大于？
 - 从 d 问的答案告诉你两组样本的差的什么信息？
- 4.57 健康饮食的准则之一是：每日来自脂肪的卡路里的摄取不超过 30%。在一组冷冻巧克力点心中，来自脂肪的卡路里的百分比的均值是 18.9，标准差是 9.2。作为比较，数据还包含了普通巧克力冰激凌的脂肪中卡路里为 39%。（来源：“Low-fat frozen desserts: Better for you than ice cream?” *Consumer Reports*, vol 57, no. 8 (August 1992), pp. 483 - 487.）
- 将巧克力冰激凌中卡路里的百分数转化为标准得分。
 - 巧克力冰激凌与其它点心有什么不同？
- 4.58 十六中不同的零食的卡路里含量表见下：（在习题 3.38 中，你使用过这些数据来画图）

110	120	120	164	430	192	175	236
429	318	249	281	160	147	210	120

（来源：ASDA data and manufactures' data shown as an advertisement in *The New York Times Magazine*, April 20, 1990, p. 20）

- 找到数据的均值和中位数。
 - 这两种平均数哪一种对此数据更合适？
 - 求出观察值的极差。
 - 用极差来找出数据标准差的一个估计。
- 4.59 在 1995—1996 学年中，Swarthmore 学院中的数学和统计系的教师的孩子数如下：Eugene 2, Don 0, Gudmund 4, Helene 0, Charles 2, Aimee 0, Stephen 2, Michael 0, Janet 0。
- 画出描述这些数据的直方图。
 - 孩子数的众数是什么？
 - 男人的众数是多少，女人的众数是多少？
 - 这些众数告诉你什么？
- 4.60 为了得到他上的课程的课本页数的平均数，Clark 首先按课程列出了这些课本的页数：生物 657, 189；历史 348, 237, 181；英语 104, 201, 298, 87；数学 302, 99；心理学 607, 139。
- 用表将数据组织起来，使我们可以浏览此表就得到中位数。
 - 找出页数的均值。
 - 比较这两个数。什么使得它们不同。按照 Clark 的目的，哪一个答案更合适。

- 4.61 根据美国普查局的数据,右表是 1915 年到 1945 年中间每两年的医学院的数目:

1915	1916	1917	1918	1919	1920	1921	1922	1923
96	95	96	90	85	85	83	81	80
1924	1925	1926	1927	1928	1929	1930	1931	1932
79	80	79	80	80	76	76	76	76h
1933	1934	1935	1936	1937	1938	1939	1940	1941
77	77	77	77	77	77	77	77	77
1942	1943	1944	1945					
77	76	77	77					

(来源: Historical Statistics of the United States 1789 - 1945, p. 50)

- a. 画一个茎叶图来表示这些数据。
- b. 计算这些年的医疗学校的众数,中位数和均值。
- c. 每一个汇总统计量给了什么另外的汇总统计量不能给的信息?
- d. 这些数据有什么使你感到意外?
- e. 你对历史趋势怎么看?(附加信息:医学院毕业人数从 1915 年的 3500 上升到 1925 年的 4000,1930 年的 4500,从 1930 到 1945 年达到 5000 人,内科医生人数从 1915 年的大约 143000 人,上升到 1945 年的大约 181000 人。)
- 4.62 回忆你一生中做过的工作的每小时的薪水,计算极差。
- 4.63 画一个习题 4.14 中 Shaquille 自由投掷的直方图。(你将会在这个问题上浪费许多纸张!)这个直方图告诉你什么?
- 4.64 以下的数据来自高中学生吸烟、吸食大麻、喝酒的记录。为使问题简单化,先对每个分布画一个直方图。
- 一个月中吸烟的天数: 0 0 30 29 30 0 0 10 0 30 29 30 0 0 0 0 1 30 28 10 0 0 0 30 30 29 0 0 30 0 0 30 0 0 1 0 0 30 30
- 一个月中吸食大麻的天数: 0 0 0 0 0 0 1 0 0 0 0 1 2 2 1 0 0 3 0 0 2 0 0 1 0 0 1 0 0 1 0 0 4 0 0 0 0 0 1 1
- 一个月中喝酒的天数: 0 1 0 5 0 4 0 0 3 0 0 2 2 0 0 0 1 0 0 4 0 0 3 0 0 2 0 0 0 1 2 0 0 1 0 0 1 0 3 0
- a. 估计(或计算)每一个分布的均值。
- b. 哪一个分布有最大的标准差?
- c. 哪一个分布有最小的标准差?
- d. 知道了在均值的一个标准差的距离内的观测值的百分数,是否有可能找出三个分布共同的标准差?
- 4.65 使用习题 4.64 中的信息:
- a. 该样本中在最近的一个月中至少有一次吸烟、吸食大麻、喝酒的高中学生的百分比各是多少?
- b. 你的结论是否与以下关于美国学生的调查结果吻合: 46% 至少喝过一次酒, 24% 的人吸过烟, 11% 的人至少吸过一次大麻。(来源: "Teen-age drug use high," The New York Times, September 20, 1992. p. 33.)
- 4.66 考虑习题 3.46 中,剑鱼中的水银浓度数据:
- a. 找出样本中 28 条剑鱼的水银浓度的均值。
- b. 找出水银浓度的标准差。
- c. 有多少条剑鱼的水银浓度在均值加减两倍标准差的范围内。

d. 被检验并发现水银浓度大于 1.00 的剑鱼不能放入市场,但为何水银浓度的均值还是大于 1.0?

4.67 某小公司几个雇员的每小时工资如下:

\$ 6.50 \$ 6.20 \$ 6.50 \$ 7.00 \$ 10.00 \$ 10.00 \$ 11.00 \$ 15.00 \$ 21.00

- 工资的中位数是多大?
 - 工资的均值是多大?
 - 已经决定,将工资最低的四个人的每小时工资提高 \$ 4.00。新的工资的中位数是多少?
 - 新的工资的均值是多少?
 - 为何这四个人的工资增高后,工资的中位数和均值没有增加同样的水平?
- 4.68 一个变量的观察值是 1,3,3,3,3,3,3,5。另一个变量的观察值是 2,2,2,2,4,4,4,4。
- 画出这两个变量的直方图。
 - 按照直方图,两个变量是否具有相同的均值?
 - 按照直方图,两个变量是否具有相同的标准差?
 - 找出这两个数据集的均值和标准差。
 - 从 d 的结果中我们得到什么关于这两个数据的结论?
- 4.69 习题 10.34 中涉及十种不同的冰激凌中的脂肪所含热量的百分比和一些品尝者对这些冰激凌的口味的评分。
- 找出脂肪所含热量百分比的均值和标准差。
 - 有多少个观察值落在均值加减两倍标准差的区间内?
 - 找出口味变量的均值和标准差。
 - 有多少个观察值落在均值加减一倍标准差的区间内?
- 4.70
- 找出图 4.6 的数据的标准差?
 - 当知道均值和标准差时,数据告诉你什么?
- 4.71 找出习题 3.20 中结婚年龄的中位数。这两个中位数告诉了我们什么?
- 4.72 以下的结果是两个著名的体力检验的结果,采自 10 个大学游泳者。

Test	Adam	Bob	Emil	Juan	Sam	Lou	Ken	Paul	Mike	Leah
A	20	23	24	18	17	16	25	24	21	19
B	31	39	39	29	28	31	40	30	31	30

- 哪一个检验做得更好?为了解决这个问题,我们需要将原始数目转化为标准得分。对全国样本,检验 A 的均值是 20,标准差是 2;检验 B 的均值是 35,标准差是 3。
- 如果你是教练员,你想使你的队员充满信心,你应选择哪一种检验?
- 哪一个运动员是最弱的?
- 哪一个运动员是最强的?
- 哪一(几)个运动员的两个检验看起来是最不一致?
- 哪一(几)个运动员的两个检验看起来是最一致?

C H A P T E R 5



- 5.1 怎样得到概率
- 5.2 概率的计算
- 5.3 优势:概率的对照物
- 5.4 离散变量的概率分布
- 5.5 连续变量的概率分布
- 5.6 使用概率来核对假设
- 5.7 决策分析:利用概率来作决策
- 5.8 小结

概 率

5

一个家庭中的所有四个孩子都是女孩的概率有多大？同一天中有两场无安打棒球赛的机会有多大？在 daily double 中幸运数字是 71 的可能性有多大？Libby 的父母在多大程度确定她可以被 Carleton 学院录取？一个来自均值为 4.0 的总体的样本中每个家庭的孩子的数量的均值是 2.0 或更少的概率是多少？如果投票的人群分成相等的两党，一个样本中有 55% 或更多的人投(两党中的)一个候选人的概率有多大？

概率是 0 到 1 之间的一个数，它告诉了我们一个事件发生的经常程度。

有关概率的问题在日常生活和在统计课中一样经常出现。这一章中我们将讨论概率这个词在统计中的意义以及我们如何应用它来进行统计分析。从上面的问题中我们可以知道概率是与某事件发生的机会、可能性，或确定程度有关的一个词，这个词的使用已大大超出了统计的范围。

概率简单地说就是一个数。更确切地说，它是一个 0 和 1 之间的数，用来描述一个事件发生的经常性。小概率(接近零)的事件很少发生，而大概率(接近 1)的事件则经常发生。例如：一天中有两场无安打棒球赛的概率就很小，一年中至少有一场飓风袭击美国的概率就很大，因为在大部分年份中都多于一场飓风发生。

概率观念的产生可追溯到很多年以前。有关机会与概率的参考甚至在旧约全书中就有：“Saul 说，在我和 Jonathan——我的儿子之间投骰子，子是 Jonathan 被带走了。”(Samuel 1:42) 一千多年以前有这样一个传说，挪威国王(Olav the Holy)和瑞典国王是通过扔一对骰子来决定一块有争议的土地所有权的。

早在 17 世纪,就有零星的有关机会和概率的文章出现。那时,对概率的兴趣是被绅士赌徒们试图确定牌和骰子赌博中的赔率而引起的。因为小概率事件发生的可能性很小,赌博者希望这类事件应有较高的赔率;另一方面,有较高概率的事件应有较小的赔率,因为这类事件经常发生。而且概率——因此赔率——应该是公平的,即投注者既不应最后破产,也不应有过多的收益。这个问题被提到了当时的数学家的面前,我们所知的概率论就从那时发展起来。

现在,对于概率的统计兴趣已有点不同于赌博者了。为了说明统计的基本观念,如果我们对于某事物进行多次观测,大多数时候会得到不同的结果。例如:一棵树的一片树叶的长度是一个面定的英寸数,而该树另一片树叶的长度又是另外的一个数。这是因为这个(树叶长度的)变量具有随机性。同样地,在一个样本中支持现任总统政策的人数的百分比和另一个样本中支持者的百分比是不同的。这同样是由我们在随机抽样中不同样本具有的随机性决定的。



概率能为许多目的服务。(来源: Peanuts © reprinted by permission of United Feature Syndicate, Inc.)

对同一个变量的观测的差异产生了这样一个问题,无论我们是观测单个样本(叶子)还是群体样本(人群),如果重复多次测量,某一个特殊结果出现的次数是多少呢?这个问题只能用概率来回答。它显示了在一长列观察值中某一事件如何出现。例如,如果我们从投票者中抽取很多很多样本,在四分之三的样本中对现任总统政策的支持程度超过 60%,我们说对总统现行政策的支持程度超过 60% 的观察样本的百分比是 0.75,即在 100 个样本中有 75 个样本对总统现行政策的支持率超过 60%。另外的 25 个样本对现行政策的支持率低于 60%。在叶子的例子中,叶子长度大于 2.34 英寸的概率是 0.1,即 10 片叶子中只有 1 片叶子的长度是大于 2.34 英寸的,另外 9 片叶子的长度都是小于 2.34 英寸的。

概率的说法将贯穿整本书:一个变量的样本均值大于某一个数的概率是 0.023,样本标准差小于 5.67 的概率是 0.15,等等。在第四章结婚年龄的例子中,新娘年龄的均值是 30.0 岁,新郎年龄的均值是 32.3 岁,差 2.3 岁。来自两个均值相同的总体中的样本的均值差异是 2.3 或更大的概率是 0.002。这就是说,不同的 1000 个来自新娘和新郎年龄相同的总体的样本中,只有 2 个样本的新郎的均值比新娘的大 2.3 年或更多。

概率告诉我们,在样本数据的基础上,如果实验重复多次,各种结果发生的经常程度是多少。基于观察的样本,概率在开拓实际世界上是十分有用的。

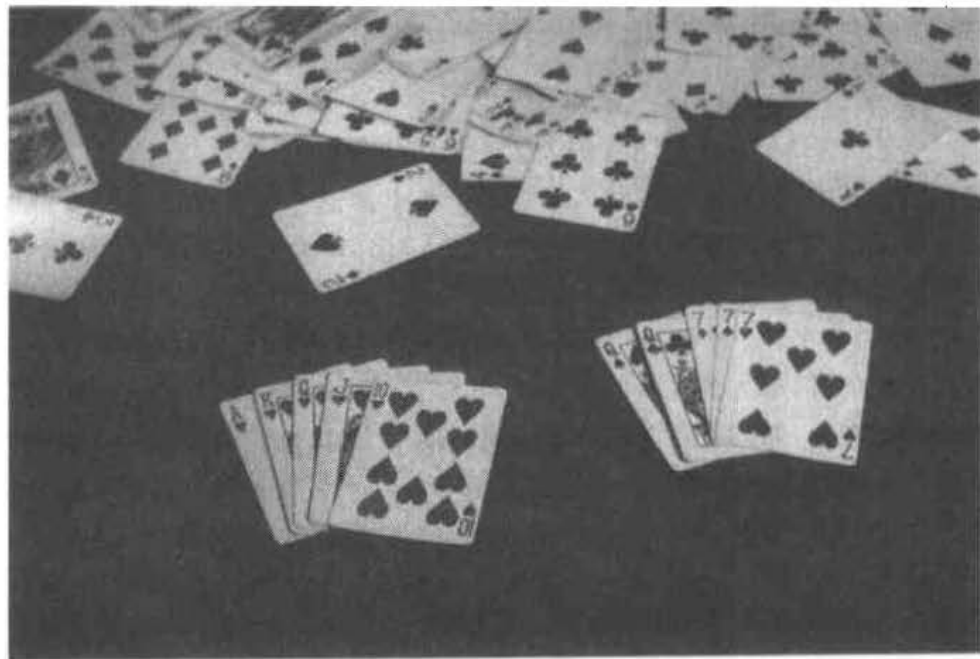
5.1 怎样得到概率

现在我们知道概率是 0 和 1 之间的一个数,那我们如何找到这个数呢?

利用等可能性事件

最早寻找概率的方法来自于扑克牌和骰子游戏。如果一个骰子有六个面,并且投掷时,每一个面出现的可能性相同,那么每一个面出现的概率是 $1/6$ 。同样的,如果一副牌有 52 张,有 13 张梅花,那么随机抽出一个梅花的概率是 $13/52$,或 $1/4$ 。

这种考虑概率的方法表明如果实验有 n 种可能的输出,一个有 k 种输出的子集为有利的,那么 k/n 就是这个有利事件出现的概率。对于骰子的问题,出现一面的输出是 $k=1$,而有 $n=6$ 种可能的输出,则出现某一面的概率是 $1/6$ 。对于一副扑克牌,有 $k=13$ 个梅花及 $n=52$ 张牌,则抽出梅花或其它某一花色的概率是 $13/52 = 1/4 = 0.25$ 。



扑克游戏者天堂——同花大顺(Royal flush,为有同样花色的五张相连的最大的牌——译者注。)和三两顺(full house,为有三张同样大小及两张同样大小的五张牌——译者注)。任何一种情况出现的概率是多少?两种情况都出现的概率是多少?
(来源:First Image West, Inc.)

这种找寻概率的方法对于扑克牌和骰子问题有效是因为可能的输出数是已知的,并且由于对称性,每一种输出的可能性都是相同的。然而,通常我们不知道是否每一种输出出现的可能性都是相同的(例如,在一场赛马中,不是所有的马赢的概率都是相同的)。有时可能的输出总数是不知道的(例如,在橄榄球投彩中赌某一数目的人数)。在这种情况下,用“等可能性”的方法来计算概率是行不通的。

使用相对频数的方法

寻找简单概率(即一个事件发生的概率)的第二种和最常用的方法是基于大量实验中一个事件出现的次数的比例。以婴儿出生为例,每一次生产都会得到一个或多个男孩、女孩,但生男、生女的概率是否一样我们就知道了。

注意我们的语言:什么是 n , 什么是 k ?

堕胎的女子是天主教徒的概率不等于天主教女子堕胎的概率。摘自 Charles F. McLaughlin 致 *The Philadelphia Inquirer* 编者的一封信, 11月8日, 1992年。

我这封信是针对于 Victoria A. Brownworth 的注释文章中对天主教妇女的描述而写的。她说: “根据 Alan Guttmacher Institute 的调查, 和其他宗教信仰的妇女比较起来, 天主教的妇女有更多的人堕胎。”

这种表达是误导的。它使人产生一种错觉, 作为一个人, 天主教女子比其他宗教信仰的女人更容易堕胎。这不是事实。罗马天主教堂在美国比其他宗教教派有更多的成员。天主教的妇女总数和美国的第二大宗教的人数之比约是 2:1。在人数上, 天主教妇女堕胎的人数较多是正确的。然而, 这并不代表天主教妇女比其他宗教的妇女更容易寻求堕胎。

在多年保留记录中, 在新生儿中女孩的概率是 0.49。这个比率是通过用女孩的个数除以所有婴儿的个数获得的。概率学家们(研究概率的人)说, 当出生的婴儿的个数接近无穷时(即有很多个观察值时)观察到的女孩的出生比率接近女孩出生概率的真值。

在这个例子中, 概率是一个长期的比率, 是长期观察某一事件的结果。这种概率的准确的数值永远是得不到的, 但是大量观察值使估计概率的数值无穷接近于真值。长期观察而计算概率的方法的问题在于, 如同著名的经济学家凯恩斯爵士(Lord Keynes)所说: “从长远来说, 我们都将会死。”没有统计学家能够希望他活得足够的长以获得概率的真值。他们通常用观察值的比率作为对概率真值的估计值。

硬币的旋转

扔硬币可以作为寻找长期概率的例子。我们现在做一种改动, 用旋转硬币代替扔硬币。用一只手指固定硬币上端, 将硬币竖起来。用另一只手指弹一下使硬币旋转, 是否硬币出现正面和出现反面的概率和扔硬币是一样的呢?

为了回答这个问题, 我们在我们的一次统计课上实验了旋转硬币。这个课上有 25 个人, 每一个人使硬币旋转 10 次, 总共有 250 次。这些实验的结果是有 97 次出现正面, 另外 153 次出现反面。这样出现正面的比率是 $97/250 = 0.396$ (或 39.6%) 而不是如果两面等可能应出现的 0.5。

是否实际概率应该是 0.5, 这个实验的结果只是随机性的作用的结果呢? 这个问题在习题 7.58 中将会得到答案。

利用主观概率

甚至使用相对频数来逼近概率有时也是无效的。在进行一次有计划的旅行时, Kaye 先生明天安全到达目的地的概率是多少? 明天只有一次旅行发生(唯一的事件)。他不可能进行旅行, 而后使时间倒转, 然后再旅行, 如此反复多次, 最后计算在所有旅行中成功到达的次数。当事件不能重复度量时, 我们没有办法知道一个特定事件发生的比例是多少。但是用概率的方法去考虑仍是有效的。Kaye 先生不能确定这次旅行是否安全, 但是从他这一类的旅行的知识来考虑, 这次旅行成功的概率足够大到能令他进行这次旅行。

一次事件的概率叫**主观概率**(subjective probability)。在这个例子中, 个人的概率仅仅是kaye先生在研究了当时所有的资料的基础上对于旅行的不确定性的一种表达。我们每个人都可以有自己的旅行的安全性的概率, 所以这里没有正确或者错误的个人概率的值。这使得个人概率是主观的。

停下来想一想 5.1

对以下情况应用哪种寻找概率的方法是合适的?

方法:

- a. 等可能性事件。
- b. 相对频数。
- c. 主观概率。

问题:

- 1. 一个有十年历史的短程航空公司继续保持无事故记录。
- 2. 玩扑克牌的人从一副牌中抽出一张 A。
- 3. Minneapolis 三月下雪的厚度大于 5 英寸。
- 4. 明天郊游时下雨。
- 5. 一个有六个孩子的家庭有一对双胞胎、三胞胎、四胞胎、五胞胎或六胞胎。
- 6. 最近使你特别感兴趣的概率问题。

个人或主观概率形成了 Bayes 统计推断的基础。在这里我们不介绍这方面的内容。本书中我们常使用长期比率作为概率。

5.2 概率的计算

概率就是一个数字, 因此可以进行加、减、乘、除。通过计算可以使我们用简单的事件的概率来得到更复杂事件的概率。

例如, 随机选择的一个新生儿是女孩的概率是 0.49。给定这个简单事件的概率, 所选的新生儿是男孩的概率是多少? 我们用 1.00 减去 0.49 得到 0.51, 这个数就是我们选择的是男

孩的概率。我们可以这样做是因为出生婴儿只有男孩和女孩两种。简单事件的概率可以帮助解决更复杂的问题;例如,一个有四个孩子的家庭有 3 个女儿 1 个儿子的概率是多少?即,一个家庭的孩子由三个女儿一个儿子组成的情况发生的频率是多大?

得到这个概率的一种方法是选定很多很多有四个孩子的家庭,计算其中多少有三个女儿一个儿子而多少不是这样。我们将会发现他们中有 0.24 的家庭(或 100 个中有 24 个)有 3 个女孩和 1 个男孩。但是这种寻找长期比率的经验方法将会花费很多时间和金钱。实际上,我们可以通过对原始的女孩出生率 0.49 这个概率进行加、乘得到答案。这个问题的答案仍是 0.24,在 5.4 节二项分布子节中将会介绍有关计算。

概率的加法

当我们想找到不可能同时发生的两个事件之一发生的概率时,我们简单地对这两个事件的概率进行相加就行了。例如,为了找到一个有四个孩子的家庭中有三个或四个女孩的概率,我们知道不可能同时存在三个和四个女孩,就可以将这两种概率相加。有三个女孩的概率是 0.24,有四个女孩的概率是 0.06,所以有三个或者四个女孩的概率就是 $0.24 + 0.06 = 0.30$ 。如果我们想知道某一事物的概率是大还是小——例如,样本均值是小于 5.6 或大于 17.8——因为这种事件不可能同时发生,我们只需要将均值小子 5.6 的概率与均值大于 17.8 的概率相加就可以了。

概率的乘法

为了寻找某一事件与另外一个事件同时发生的概率,我们需要将这两个概率进行相乘。两个事件同时发生的概率比其中任何一个事件单独发生的概率要小。这种常识产生于数学概念:两个数相乘的积小于任何一个数。例如,0.3 乘以 0.4 等于 0.12,0.12 既小于 0.3 也小于 0.4。

让我们回到有四个孩子的家庭的例子中,一个家庭先有一个女孩,接着是一个男孩,又是一个男孩,然后是一个女孩的概率是多大?只需要将每一个孩子出生的概率相乘: $0.49 \times 0.51 \times 0.51 \times 0.49 = 0.062$ 。即,1000 个有四个孩子的家庭中有 62 个家庭的孩子按先后是女孩、男孩、男孩、女孩。

在很多情况下,概率不能直接相乘。为了做乘法,我们不得不引进条件概率的说法。

5.3 优势:概率的对照物

反对一个事件的优势是表示为整数之比的一个事件没发生的可能性对其发生的可能性之比较。

1993 年,在国际奥委会决定 2000 年奥运会的举办者之前,伦敦的赌注登记经营人给出了他们认为可能是运动会主办地的城市的优势(Odds)。他们认为某些城市主办奥运会的概率高于另外一些城市,他们提供了反对这些城市举办运动会的优势是:

悉尼,澳大利亚	4 比 9
北京,中国	5 比 2
曼彻斯特,英国	10 比 3
柏林,德国	16 比 1
依斯坦布尔,土耳其	66 比 1
巴西利亚	200 比 1

这些数字看起来好像赛马的优势。因为在反对事物发生的优势中,事件不出现总是放在前面,很明显地赌注登记经营人认为悉尼有很大可能成为主办地,而巴西利亚是不大会成功的。

在使用金钱作赌注时,优势比概率更常用。反对巴西利亚的优势为 200 比 1 告诉了我们如果我们支付赌注登记经营人 \$1,而最终巴西利亚成为了主办地则我们将会获得另外 \$200 的赌金。这样,优势实际上告诉我们,我们应付下多少赌金和如果我们赢了我们应赢回多少钱。

为容易表达,优势应用整数给出,如 4 比 9。4 比 9 的优势与 2 比 4.5 的优势是相同的,但是小数点有些烦琐。这就是说在利用优势比较之前应首先熟悉它。

停下来想一想 5.2

在奥运会主办地的优势表中,哪一个城市更容易成功,是北京的优势还是曼彻斯特的优势?

巴西利亚作为一个最不可能申办成功的城市,表明了赌注登记经营人的看法:巴西利亚只有很小的概率在申办中成功。200 比 1 的优势翻译成概率就是 $1/(200 + 1) = 1/201 = 0.005$ 。悉尼的 4 比 9 的优势表明我们付给赌注登记经营人 \$9,并且悉尼赢了,那么我们将会得到 \$9 外加 \$4 的赌酬。我们不能得到更多的钱是因为赌注登记经营人认为悉尼赢得这场竞争的可能性非常大,因而很多人都会认为悉尼是赢家。

悉尼赢得这次主办权的概率,照伦敦赌注登记经营人的看法是 $9/(4 + 9) = 9/13 = 0.69$ 。北京赢的概率是 0.29,曼彻斯特 0.24,柏林 0.06,依斯坦布尔 0.015,巴西利亚 0.005。这一章末尾的公式 5.1,5.2,5.3 告诉了我们如何在优势和概率之间进行转换。赌注登记者也可以不给出反对每个城市的优势而用赞成每个城市的优势。

1993 年 9 月 23 日,国际奥委会将 2000 年奥运会的主办权授予了悉尼。这些经营人应该满意了。

5.4 离散变量的概率分布

常常既有简单的也有复杂的方法来解决事件。例如,找出从新奥尔良到芝加哥的距离的一个复杂方法是开这一段距离的车然后记下英里。简单的方法看一本道路图的后面附的城市之间距离表。当某些简单事件的概率已知时,统计学家可以用简单的方法得到一些复杂事件

发生的概率。

如果复杂事件的概率很难直接计算得到时,可以应用简单事件发生的概率来计算复杂事件发生的概率。在四个孩子的家庭的例子中,简单概率是随机选择的孩子是女孩的概率为 0.49。简单事件是出生,结果是男孩或是女孩。复杂事件是家庭中有三个女孩一个男孩。

通过预先制定的各种概率问题的解决方法,统计学家节约了大量的时间和省去了很多麻烦。两个节省精力的例子是二项分布和 Poisson 分布(Poisson 是最先引进这个方法的法国数学家)。

二项分布

相象你想知道硬币连续两次正面着地的概率是多少? 是否你需要整天坐在屋子中,不停的扔硬币,然后找出连续两次正面着地这一概率是多少呢? 也许不需要,如果你知道(1)出现一次正面的概率是 0.5;(2)只有两种可能值(正面和反面);(3)每一次扔硬币都是独立的。要找到连续两次正面着地的概率,你用 0.5 乘以 0.5 得到 0.25。即在扔两次的情况下连续两次出现正面的机率是 25%。于是可以不用计算器,不用扔一天硬币,或者高深的数学知识就可以得出结果。

再考虑这个更困难的例子:即利用生一个女孩的简单概率是 0.49 来知道一个有四个孩子的家庭中有三个女孩和一个男孩的概率是多少。在 300 年以前,数学家们就理解到,无论我们想得到的概率是男孩还是女孩,是正面还是反面,是死的还是活的金鱼,问题都是一样的。从一个简单事件的正确的概率,现在已经产生了公式,图表,和计算机软件来帮助我们找到更复杂的事件的概率。最常用的公式是二项分布公式。它是用来计算在 n 个实验中(例如,婴儿的出生)成功(如生女孩)的次数的分布。利用这个公式,通过用纸和笔,计算机软件,打印的图表,只要输入简单事件信息,我们可以得到任何结果的概率。这一章末尾的公式 5.4 就是二项分布公式。



所有随机选择的这些孩子都是女孩的概率是多少?(来源: Penny Gentieu, Tony Stone Images.)

一个仅有两个可能值的变量,如新生儿的性别的变量,是二项分布的基础。(二项的意思就是“两个数或两个名称”。)假设我们知道其中一种情况发生的概率,例如,新生儿的性别是女孩的概率是 0.49。那么新生儿是男孩的概率就是 0.51。这两个概率相加一定要等于 1。因为新生儿一定不是男孩就是女孩。

产生二项分布的下一步是对基础变量(新生儿的性别)进行几次独立观察。如果一个家庭有 4 个孩子,它们中有一定数量的女孩,则用 4 减去该数就是男孩数。这是一个新的变量: 4

个孩子中女孩的个数。这个变量可能的取值为 0, 1, 2, 3 或 4。这种变量叫做二项变量 (binomial variable)。一个二项变量指出了问题中的两个值中某一个值出现的次数。

再下一步就是找到在一个有 4 个孩子家庭中的该二项变量的每一个值 (女孩数) 的概率了。这个概率可以通过公式 5.4 计算得到, 列在表 5.1 中。这样的二项变量的值和相应的概率的总和叫做二项分布 (binomial distribution)。

表 5.1 四个孩子的家庭中女孩个数的二项分布

女孩个数	0	1	2	3	4	总数
概率	0.07	0.26	0.37	0.24	0.06	1.00

让我们回到有 3 个女孩 1 个男孩的家庭的例子中。一种顺序是先有 3 个女孩然后是 1 个男孩, 这中情况可以用序列 GGGB (其中 G 代表女孩, B 代表男孩) 来表示。第一个孩子是女孩的概率是 0.49。第二个孩子是女孩的概率也是 0.49。那么第一个与第二个孩子都是女孩的概率就是用 0.49×0.49 。将其一般化, 前三个孩子都是女孩, 第四个孩子是男孩的概率是 $0.49 \times 0.49 \times 0.49 \times 0.51 = 0.06$ 。这就是依前后顺序有三个女孩一个男孩的概率。

一个家庭有三个女孩一个男孩可以有以下几种顺序:

三个女孩一个男孩	概率
GGGB	$0.49 \times 0.49 \times 0.49 \times 0.51 = 0.06$
GGBG	$0.49 \times 0.49 \times 0.51 \times 0.49 = 0.06$
GBGG	$0.49 \times 0.51 \times 0.49 \times 0.49 = 0.06$
BGGG	$0.51 \times 0.49 \times 0.49 \times 0.49 = 0.06$
	和 = 0.24

每一种可能序列出现的概率是 0.06。将所有的概率结果相加得到有三个女孩和一个男孩的概率是 0.24。

当样本数目大于 4 时, 找可能序列的数目将会变得十分麻烦, 但是应用本章末尾给出的公式 5.4 的第一项, 将会使其变得简单。应用发表的二项概率表和计算机软件来计算二项概率避免了所有的计算。

二项分布通常只在小样本时使用, 例如有四个孩子的家庭。如果样本个数与原始概率的乘积大于 5 时, 则有较简单的方法来分析数据。(在有四个孩子的家庭的例子中, $4 \times (0.49) = 1.96$, 明显的小于 5。) 但如果在一个有 1200 个响应者的调查之中, 有 720 个人支持提出的提案, 而 480 个人反对, 更好的办法是所谓的对二项分布的正态近似 (normal approximation to the binomial distribution); 将在第六章与第七章中进行讨论。只要基础概率在 0.5 左右, 如像家庭例子, 这种逼近在样本观察值即使只有 10 或 15 个时也可进行。

Poisson 分布

在 1990 年 6 月 3 日, 体育报纸充满了对前一天出现的一种不可能现象的讨论: 两个无安打棒球赛同时出现, 一个是 Mark Langston 和 Michael Witt 为 California Angels 队打球, 另一个

是 Randy Johnson 为 Seattle Mariners 队打球的。无安打球赛在棒球赛中不常发生,因此即使一个无安打球赛也足以引起媒介的注意。两个无安打球赛同一天发生从 1898 年以后就没有过。

为了了解这个事件有多么不可能,我们应用 Poisson 分布。Siméon Denis Poisson 特别着迷对小概率事件,特别是许多情况下可能出现的事件。Poisson 研究了在那个骑兵仍旧骑马而不是用坦克的时代里普鲁士士兵被马踢死的人数的数据。他的成果发表于 1837 年。

一场无安打球赛是有两种输出值的情况之一。一场棒球赛或者是无安打球赛或者不是,所以只有两种可能性。但是不像一个孩子是男孩还是女孩的概率,无安打球赛的概率十分的小。一场无安打球赛是很不可能发生的。但是在大量比赛数目下,这种事件也有可能发生。这里 Poisson 变量是指一天中,无安打球赛发生的次数;这个变量可能的值是 0,1,2,3,等等。



Cy Young 是个著名投手,1904 年 5 月 5 日他投出一个无安打。(来源:UPI/Bettmann)

在这种情况下,当一个事件出现的可能性非常小,且有很多可能值时,Poisson 变量取不同值的概率可以用本章末尾的公式 5.7 (关于 Poisson 分布的公式)计算得到。(Poisson 分布的公式在数学上是由二项分布导出来的,但是当你检查它时,你就能知道为什么有些人认为它不像二项分布的公式那么直观明显。)Poisson 概率可以通过该公式计算得到也可以通过查表得到。当然,也仍然可通过计算机程序寻找到。

表 5.2 同一天中出现无安打球赛的次数的 Poisson 概率

无安打球赛次数	0	1	2	...	总数
概率	0.989234	0.010708	0.000058	...	1.000000

无安打球赛的数据从 1900 年开始记录,American 队和 National Leagues 队平均每年有 1.9 个无安打球赛。我们假定棒球的赛季是 180 天,每天平均有 $1.9/180 = 0.0108$ 个无安打球赛出现。对这个数字应用 Poisson 公式则我们可以得到任何一天中出现 0 次,1 次,2 次,...无安打球赛的概率(表 5.2)。任何一天中无安打球赛出现的次数是没有上界的。

在每年 1.9 个无安打球赛和 180 天赛季的基础上,在大部分的比赛中,只有一次击球,因

此没有无安打球赛出现的概率是 0.989234。同时,同一天中出现两次无安打球赛的概率是 $5.8/100000$, 或 $1/17241$ 。在 100 年中,共有 18000 天在打棒球,所以每 100 年同一天中出现两次无安打球赛能期望出现一次。第一次出现几乎是十分准确的,在保持了 90 年的记录以后又出现了一次。我们看到下一次的出现还需要等很多年。

超几何分布

第三种统计分布是超几何分布(hypergeometric distribution);当样本很少时,它能用于分析两个分类变量(看第九章)。公式 5.10 表示如何能找到超几何变量的不同值的概率。这个分布将在 5.6 节中公平工作单位(fair workplace)的例子中使用。

用图表来表示概率

我们可以像对待观察数据那样对待概率。我们可以用图表表示概率,也可以用概率计算诸如均值、标准差等数量。

任何可以表示频率的图表也可以表示概率,比如圆饼图、盒子图等等。

图 5.1 是关于四子女家庭中女孩数的二项概率直方图。很多很多有四子女家庭的女孩的数的直方图看起来应该和这个一样。利用二项分布来计算概率将会比收集数据节省很多时间和精力。

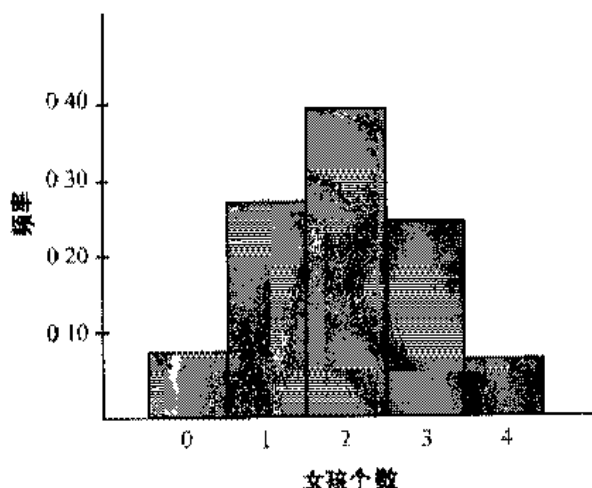


图 5.1 有四个孩子的家庭中女孩个数的二项概率。

值得重复的是,在图 5.1 中是每一个矩形的面积而不是高度是相应的概率值。因为每一个矩形的底边的长度都是 1.0,所以每一个矩形的面积就是它的垂直高度。应注意到,这些矩形的总面积等于这些概率的和 1.00。

概率的计算

我们可以对概率分布进行汇总如同我们可以对频数分布进行汇总一样。为了得到图 5.1 中的女孩个数分布的均值,我们要在水平轴上找到一个概率分布的平衡点。在使用观察值时,

我们把数据相加然后除以观察值得个数。而在这里,变量的所有可能的取值,0,1,2,3,4,都出现了;并且每个值都附有取这个值的概率或频数。即可以看作如果出现0的频数是0.07次,则出现1的频数是0.26次等等。为了得到均值不用把0相加0.07次,而仅将0乘以0.07作为该值对均值的贡献,对其它的值也类似处理得到:

$$\text{均值} = \mu = 0 \times 0.07 + 1 \times 0.26 + 2 \times 0.34 + 3 \times 0.24 + 4 \times 0.06 = 1.96$$

这个带尾巴的字母 μ 为希腊字母,用以区别概率均值和实际观察值中得到的经验均值 \bar{x} 。数字1.96告诉我们,在大部分有四个孩子的家庭中女孩的平均个数是1.96。从这个例子,我们可以看到,通过使用有一个女孩的原始概率0.49与二项分布节省了从大量家庭收集数据所花费的时间和金钱。二项分布的均值计算的公式在本章末公式5.5中可以找到。Poisson分布均值的计算可以使用公式5.8。

我们还可以找到一个变量的标准差,用希腊字母 σ 来表示,以区别用观察值计算的标准差 s 。对于女孩个数的概率分布,标准差为 $\sigma = 1.0$ 。那么均值加减两倍的标准差是 $1.96 \pm 2 \times 1.00 = -0.04$ 到 3.96 。这个区间的值包含了这个变量的几乎所有的取值。即变量值落入此区间的概率几乎为1。我们可以通过本章末的公式5.6来计算二项分布的标准差,通过公式5.9来获得Poisson分布的标准差。

5.5 连续变量的概率分布

大部分用于统计分析的数据是来自于连续变量(continuous variable),即在任意两个值之间还有其他的值。连续变量的例子包括距离、美元数量、重量、时间。

在决定某种概率时,四个理论变量是有用的。它们是标准正态 z 变量, t 变量, χ^2 变量,和 F 变量。每一种变量都有自己的特殊分布。和我们用样本数据计算均值和标准差一样,我们也可以用样本数据计算这四种变量的类似的值。这样, z 变量, t 变量, χ^2 变量, F 变量和其他样本统计量没有什么区别。在以后的章节中我们将会知道,由这四种变量算出的值对于将从样本中得到的信息推广到总体中有重要的作用。

标准正态分布:钟形曲线

标准正态分布并不是有什么“正态,”这个词估计是对其德文名称高斯(Gauss)分布和法文名称棣莫弗(DeMoivre)分布保持一种中间立场而取得的。图5.2是一个正态分布图或钟型曲线。这个分布为最容易辨认和感觉上最具有美感的曲线,以它的形状著称,看起来像钟楼上的钟。它的特征之一就是它的对称性,即中点两边曲线下的面积相等。

有一种看待正态或 z 变量的方法是想象一个变量的大量的观察值,每一个都写在一张纸上并扔在一个桶里。每一个值都叫做 z 得分(使用字母 z 并没有什么特殊的含义)。大部分的 z 变量的值在 -2.00 到 2.00 之间变动;特别是,95%的 z 变量的值在 -1.96 到 1.96 之间变动。只有很少的 z 变量的值小于 -3.00 或大于 3.00 。

z 变量的均值等于0.00,标准差是1.00。(这些数据的获得使用了很多复杂的数学方法,但是均值和标准差的获得我们仍可以视为是第四章中介绍的由大量观察值的计算获得的。)—

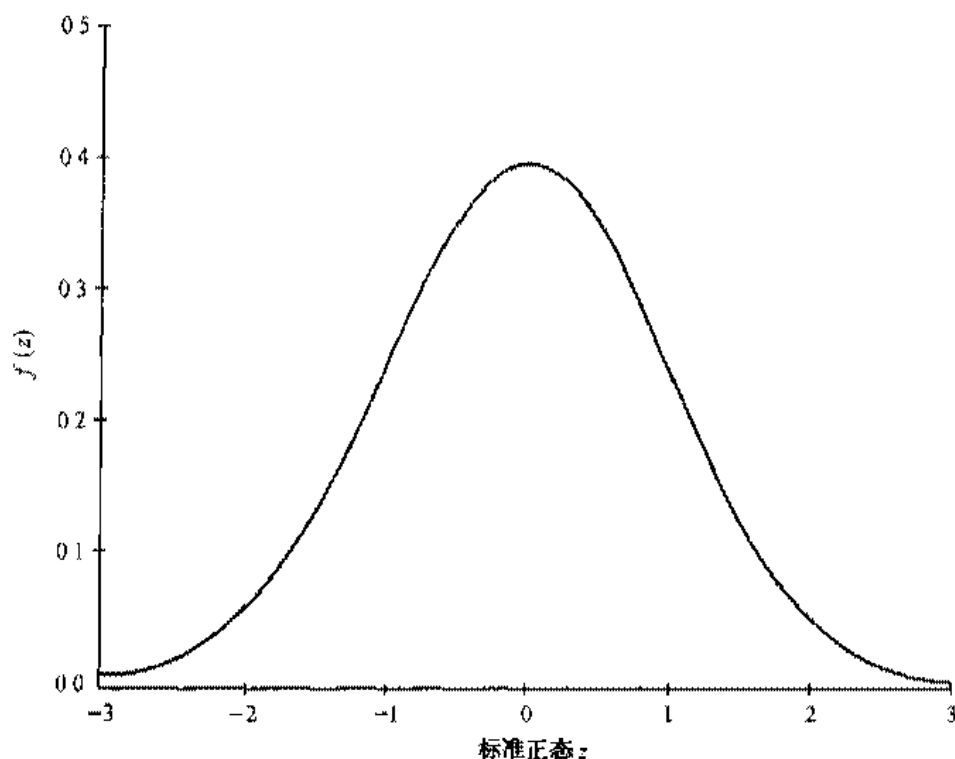


图 5.2 标准正态变量 z 的分布。

一个具有 0 均值和 1 的标准差的正态变量所具有的分布叫做标准正态分布 (standard normal distribution)。

直方图可以使我们对这些值的分布有更好的认识。这些值的范围可以被划分为一些小区间。每一个区间中 z 得分的个数用矩形的面积来代表,该面积等于落入这个区间中的观察值的比例。但是直方图很不清楚。当区间很小时,矩形就很窄。矩形的竖直线就十分接近。为了使图更清晰,去掉矩形的竖直线,只剩矩形的上边,则这个图就几乎像一条光滑的曲线,如图 5.2 所示。

一个钟型曲线描述了实际生活中的很多现象,例如,身高和体重。它还描述了许多心理检验得分;这个曲线已经成为了不同种族的智力测验的得分分布的争论焦点。

如同直方图中矩形的面积之和等于 1,钟型曲线下的面积也是 1。每一个窄矩形是相当于这个区间的 z 值的比率。这些比率的和相加为 1。从图 5.2 中我们可以看出,这个曲线的形状是单峰的而且对称于 0。由于对称性,0 左右两边的面积都是 0.5。

这个曲线中, z 变量取值在为 -1.96 和 1.96 之间的概率为 0.95。因为曲线是对称的,随机抽到的大于或等于 1.96 的 z 值的概率为 0.025。同样地,随机抽到的小于或等于 -1.96 的 z 值的概率也为 0.025。该正态分布已经被详尽地研究了,而且产生了显示这个曲线下的各种面积的表。这些表对于用 z 值来计算概率是非常有用的。在本书末尾我们列出了其中一种表 (统计表 1)。还有一个方程用来描述这个曲线。

标准正态分布的主要作用就是找到得到某一特别的值及比它更极端的 z 值的概率。例如,设 $z = 2.34$,它是否属于某一个不常出现的值的集合呢?通过查统计表 2, z 大于或等于

2.34 的概率是 $p = 0.0096$, 即 10000 个观察值中只有 96 个 z 值大于或等于 2.34。因为这个概率很小, 所以这个 z 的观察值属于某一个不常出现的 z 值的集合。在第七章假设检验中我们将会对不寻常的 z 值作讨论。

t -分布

1900 年左右, 统计学家开始觉得标准正态分布并不总是用来寻找概率的正确分布。William Gosset 是一名为爱尔兰的都柏林 Guinness Breweries 工作的化学家, 数学是他的副科; 他是对此感到怀疑的人之一。他决定经验地检验在概率问题中使用标准正态分布是否总是对的。

有些不可思议地, Gosset 以收集 3000 个犯人的身高和左手中指长度来开始他的探索。从这两个数据集(身高和手指长度), 他对每一个变量各选择了四个观察值, 因此他有了 750 个不同的样本。对于每一个样本他都计算了一个叫做 t 的值。然后他制作了两个直方图, 想看一看每一个样本的所有的 t 值的分布是什么样的? 它们与标准正态分布有多类似?

Gosset 发现他的两个直方图的形状非常接近, 但是与标准正态分布有很大不同。他将这个新分布叫作 t 分布(t -distribution), 他计算得出的值也叫 t 值。他在发表这个结果时, 因为他的雇主不愿意让他的员工发表文章, 害怕他们会将酿造啤酒的秘密泄露出去, 所以他署了一个假名叫做“学生”。因此, t 分布有时也叫做学生分布(Student's t)。

后来, Fisher 将 Gosset 的经验结论进行了数学化; 他对 t 分布的曲线导出了相应的数学函数。今天, 这个分布已经是迄今最常用的分布了。

自由度: 有不同自由度的不同的分布 t 分布有整整一族, 这一族中每一个分布都和其它的分布有所不同。将此相象为不是一个而是有一堆桶, 每一个桶都装满了写下了 t 值的纸片。为了区别这些 t 分布, 我们将桶编号 1, 2, 3, ...。这些号码就是**自由度**(degrees of freedom), 简称为 d.f. 或 df。如果处理自由度为 10 的 t 分布, 我们只需要找到 10 号桶就可以了。

统计上与这些桶等价的是 t 值的概率表。当统计学家使用 t 表时(统计表 2), 他们找到表明自由度为 10 的那一行, 样本数的大小部分地决定了应该使用哪一种 t 分布。因为自由度并不是很容易确定的, 所以读者常常被告知在进行统计分析使用的自由度是多少。

同样可以用为 z 变量而描述的有小区间的直方图的方法来找到 t 分布的图形。图 5.3 是自由度为 10 的 t 分布图。曲线下的总面积是 1.00, 与正态分布相同。分布是单峰和对称于 $t = 0$ 的。这看起来和正态分布类似, 并且我们很难看出图 5.2 与图 5.3 有什么不同。但实际上它们确实是有区别的。

正态分布和 t 分布 一种找正态分布与 t 分布区别的方法是将两种分布的曲线重叠在一张图中(图 5.4)。这两个曲线的基本形状相同, 但是正态分布的中部较高, t 分布在水平轴上的收敛不像正态分布那么快。这个区别表明 t 分布在其均值周围的聚集程度比正态分布要差一些。

例如, z 变量大于 2.5 的概率等于 0.0062, 但自由度为 10 的 t 变量大于 2.5 的概率等于 0.0152。换句话说, 10000 个 z 值中只有 62 个比 2.5 大, 但是在 10000 个 t 值中有 152 个大于 2.5。仍然, 自由度为 10 的 t 值有 95% 落在 -2.23 和 $+2.23$ 的区间内。这就意味着, 和正态分布相比, 我们必须到离中点更远的地方去获得 95% 的 t 值。面回顾正态分布, 有 95% 的取值落在 -1.96 和 $+1.96$ 的区间内。

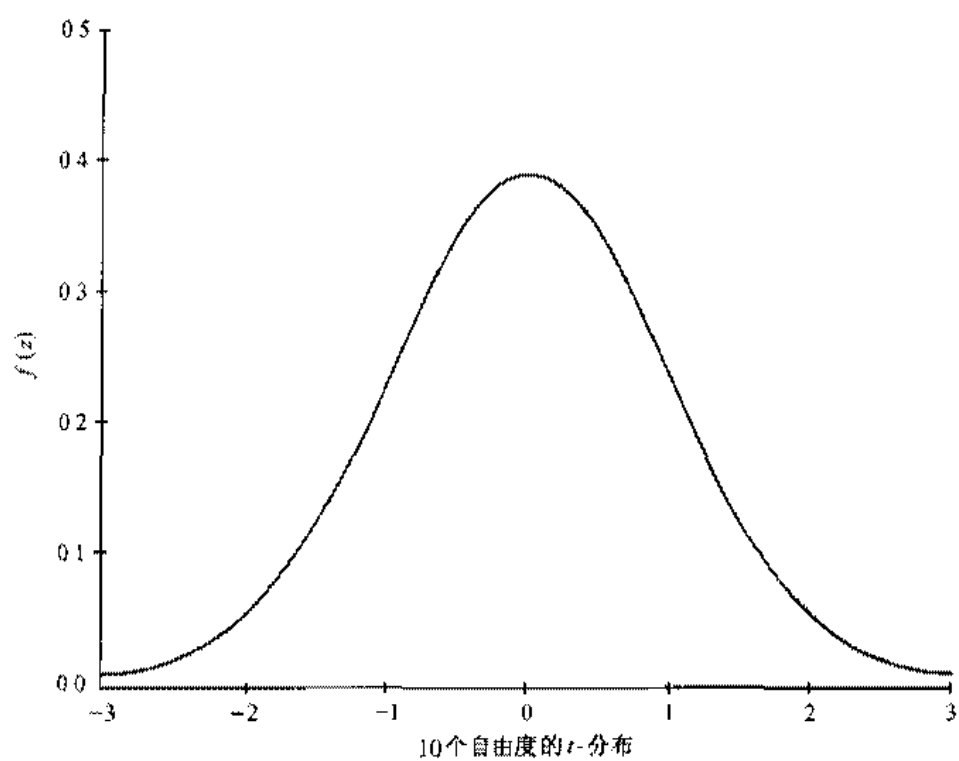


图 5.3 10 个自由度的 t 分布图形。

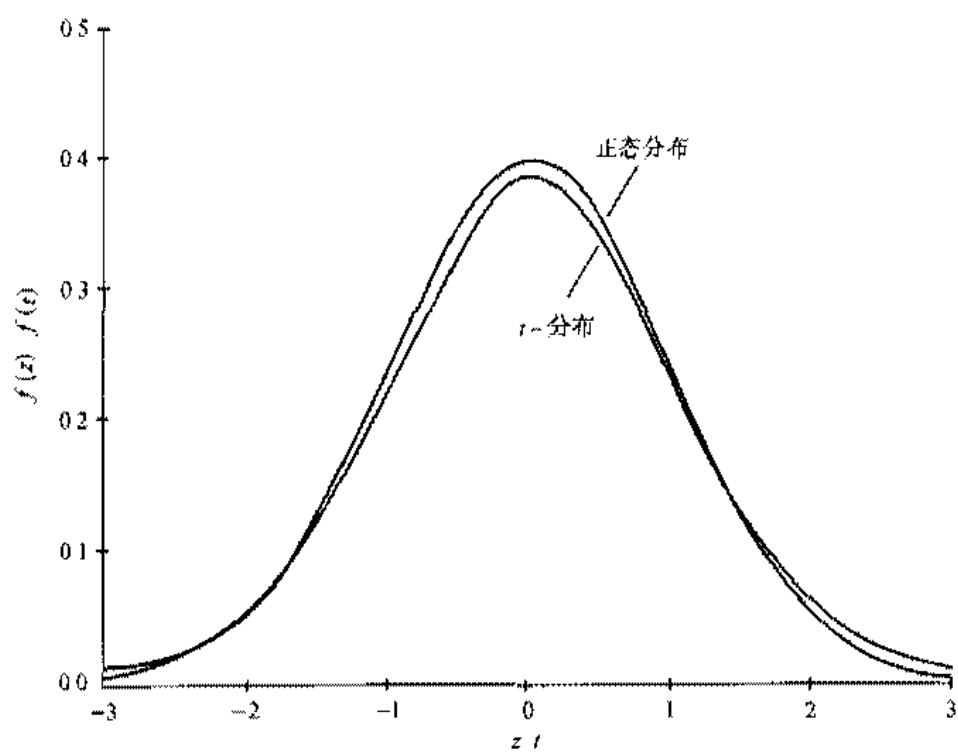


图 5.4 标准正态分布与自由度为 10 的 t -分布。

t 分布的自由度越大, 则该 t 分布的曲线就越接近正态分布。在自由度大于 30 以后就很难说出这两种曲线的差异了。在自由度等于 50 时这两种曲线就几乎相同了。这就是为什么统计表中列出的 t 分布的自由度只到 100 的原因; 此后就可以使用正态分布表来代替了。

χ^2 分布

χ^2 变量是用希腊字母 χ 来命名。(第九章中我们将会说明它在统计中的地位。) χ^2 分布和 t 分布一样, 是一族分布而不是一个单独分布。再一次相象有很多装满纸片的桶。这一次每张纸上写的是 χ^2 变量的值。这些 χ^2 分布编号为 1, 2, 3, ...。这些号码也是自由度。因此若我们想处理自由度为 3 的 χ^2 分布我们只需找到标号为 3 的桶, 即在 χ^2 统计表(统计表 3)中找到自由度为 3 的那一行就可以了。

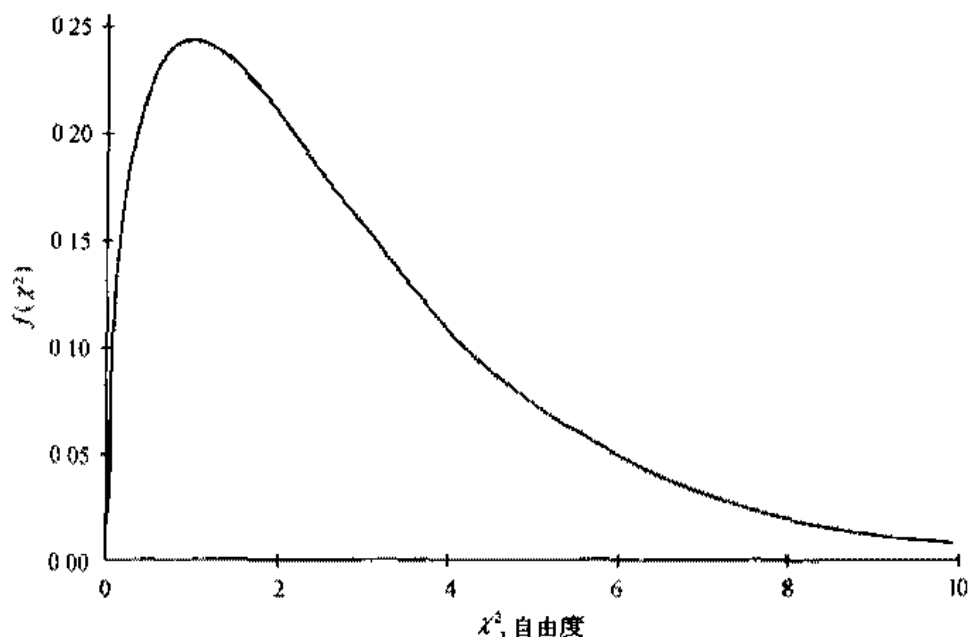


图 5.5 自由度为 3 的 χ^2 -分布。

与标准正态分布和 t 分布一样, 我们也可以用有小区间的直方图来画 χ^2 分布的图。图 5.5 表示了自由度为 3 的 χ^2 分布图。曲线下面的总面积是 1, 这一点也与正态分布和 t 分布相同。但是 χ^2 分布与 t 分布及 z 分布图的形状很不相同。这是因为 χ^2 分布没有负值; 这个图是以零为起点的。分布是偏斜的, 即它是非对称的, 大多数值小于 8。只有 5% 的值大于 7.82。换一种说法就是一个随机选择的自由度为 3 的 χ^2 值大于等于 7.82 的概率只有 0.05。该 χ^2 分布的均值等于 3, 和自由度一样。

我们利用 χ^2 分布与利用正态或 t 分布一样。如果统计问题需要我们对数据计算 χ^2 的值(在某一个自由度下)。那我们就使用 χ^2 分布来找到取到这个或更大的 χ^2 值的概率。如果这个概率很小那么这个 χ^2 值就是不寻常的; 这意味着由样本得到的结果不寻常。这个方法使我们有可能得到关于数据及产生样本的更大的总体的结论。在第七章关于假设检验的讨论中我们会详细地叙述这种观点。

F-分布

F -分布族的命名是为了纪念伟大的英国统计学家 Ronald Fisher 爵士。还是设想一些装满了写有数字纸片的桶。每一个桶都代表一种 F -分布并且有一对标号,例如 4 和 40。则这个桶代表了自由度为 4 和 40 的 F -分布。一个比较详细的 F -分布表应有 1000 种不同 F -分布的信息。

图 5.6 是自由度为 4 和 40 的 F -分布的图形。从图上我们知道和 χ^2 变量一样, F -变量也是非负的, F -分布中 F 的取值大部分在 0 到 5 之间变化。在自由度小时, F 值要大些。对这个特定的 F -分布,大多数 F -变量的取值看来都小于 3。

由 F -分布表(统计表 4), 5% 的 F -值大于 2.45, 而只有 1% 的 F -值大于 3.51。因此, 随机选择一个自由度为 4 和 40 的 F -分布的值大于 2.45 的概率等于 0.05。当我们从数据计算一个服从自由度为 4 和 40 的 F -分布的值并发现它大于 2.45 (或甚至大于 3.51), 那么我们就发现了一个不寻常的 F -值。

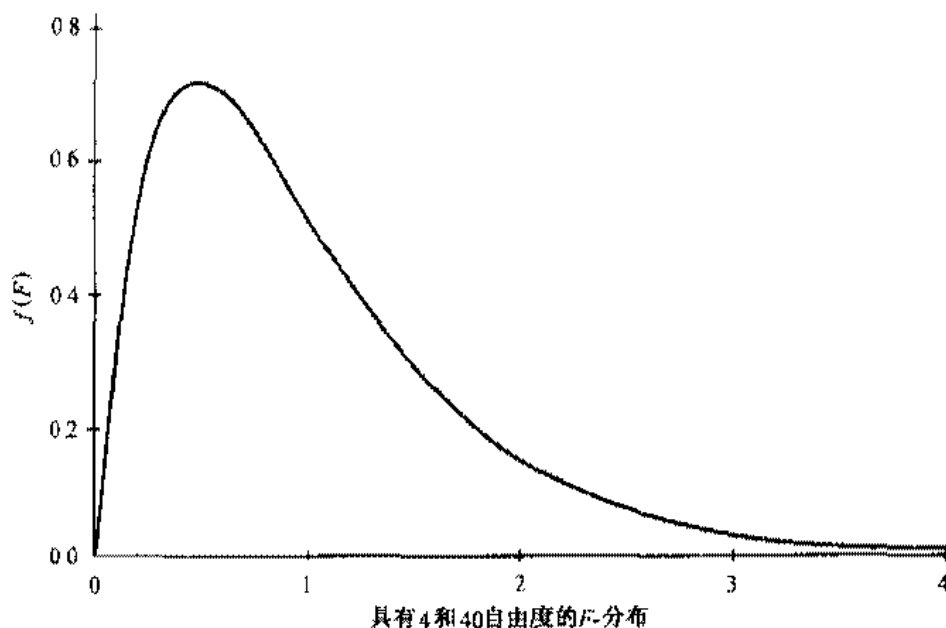


图 5.6 自由度为 4 和 40 的 F -分布。

正态分布数据的需要

我们这里说的关于四种统计变量的每一件事都可以从数学上进行研究。 t 变量, χ^2 变量和 F -变量都是从正态变量 z 中衍生出来的, 所以每次使用这三种变量时, 已经事先假定了数据服从正态分布。如果数据不服从正态分布, 那么使用这三种变量有时是不合适的。

5.6 使用概率来核对假设

硬币是公平的吗？

任何概率都是建立在某种假设为真的前提下的。如果你告诉我你手中有一枚硬币，问我如果把它扔到空中那么它落地时出现反面的概率是多少？我会回答是 $1/2$ 。我之所以这样说是因为我假设了这枚硬币是标准的美国官方铸造币。但它可能是一个伪币，两面都是正面，这样，硬币出现反面的概率就会是 0。

假设你是一个进行硬币游戏的魔术师，并且我不知你手中的硬币是真是假。如果你不想让我看你的硬币，那我可以通过收集数据来核对你的硬币是真的假设。假设你扔了硬币 10 次，每一次都出现正面。这是由 10 次正面组成的数据；现在可以在硬币是真的假设下找到出现这种数据的概率。在公平硬币的假设下，出现 10 个正面和 0 个反面的概率能由二项分布找到。这个概率等于 $(\frac{1}{2})^{10} \approx 1/1024 \approx 0.001$ ；即如果硬币是真，那么在 1024 次试验（每次试验投掷 10 次）中只有 1 次才会出现 10 次投掷出现 10 次正面的情况。但在硬币两面都是正面的假设下，上述情况出现的概率为 1。因为无论我们怎样扔都只有正面会出现。

现在这里有两种可能性：

1. 硬币是真的假设是正确的，出现上述情况的概率非常小大约是 0.001。
2. 硬币是真的假设是不正确的，出现上述情况的概率大于 0.001。

这两个可能性中一定有一个是正确的，尽管我不知道是哪一个。第一个可能性是建立在硬币是真的假设上的，我在这个假设前提下计算了观察数据出现的概率。在这个可能性下这个概率非常的小。另一种可能性是我不认为观察数据应有如此小的出现概率；最终，我观察了硬币的投掷，看起来没有什么不平常的事发生。这些就是实际出现的数据，而实际出现的事件并不常有如此小的概率。

现在我必须在这两种解释中找到一种。因为我并不知道哪一种解释是正确的，我有挑选到错误解释的风险。但是因为数据的第一种可能概率是如此的小，所以我选择第二种解释，这样观察数据出现的概率要大一些。事实上，具有高概率的事件要比低概率事件更加经常出现。作好选择后，我现在可以在观察数据的基础上说出我的结论：这枚硬币是不公平的！

作为总结，首先我们要对我们研究的事物作出某种假设。然后收集数据，并在假设的基础上计算得到该数据的概率。最后，如果这个概率非常小，如小于 0.05，则一开始的假设就是错误的。在硬币的例子中，有很强的证据表明硬币是真的这一假设是错误的。

对于科学调查中的基本准则来说这种推理方法是很重要的，它将在第七章假设检验中和其它地方作进一步讨论。

是一种公平的工作环境吗？

考虑另外一个例子，那里作了某种假设且在这种假设基础上概率也已经计算过了。你在一个有 10 个人的办公室里工作；那里有 5 个男雇员和 5 个女雇员。现在需要形成一个由 4 个人组成的委员会以研究办公环境中与性别有关的某些问题。一些人假定管理方希望有尽可能多的女性进入这个委员会。雇员们希望委员会人员是随机选择的。管理方声明人员将会是随

机选择的,但是宣布委员会时,成员有4名女性,没有男性。管理方是随机选择委员会的呢?还是有其它准则影响了选择?

对于究竟发生了什么的可能的解释与扔硬币问题的解释类似:

1. 随机选择发生了的假设是正确的,而且观察数据出现的概率很小,为0.02。
2. 管理方使用其它准则选择委员会,观察数据出现的概率大于0.02。

使用选择委员会的确是随机的假设计算观察数据出现了更极端情况的概率,如果这个概率非常小则你有理由怀疑管理方所声称的。

数据不可能再极端了,因为四个委员会成员都是女性。如果管理方有关随机选择的声明是真的,那么出现该观察数据的概率等于0.02(看图5.7)。这个概率是非常小的,即如果的确是随机选择的话,4个女性同时入选的可能性是非常不可能的。如果随机产生许多委员会,则出现这种委员会是很少见的;在100个委员会中,只有两个有这种组成。正是这种不可能的委员会会使你对随机产生的声明产生了疑问。和硬币的例子一样,由于在随机选择的假设下该数据的概率太小,你选择第二种可能性。你找到管理方并说明基于数据雇员们不相信该委员会是随机产生的。

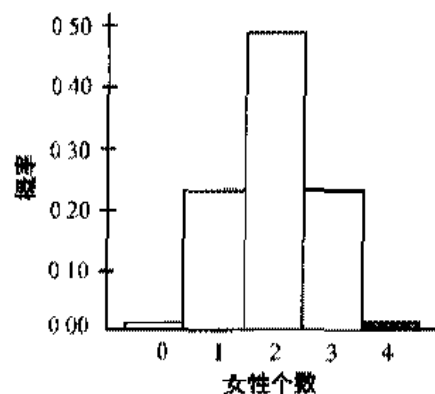


图5.7 4名委员会成员中女性的个数的概率分布。

除了有4名女性0名男性入选的概率外,图5.7还标明了3名女性与1名男性入选的概率(0.24),2名女性与2名男性入选的概率(0.48),1名女性与3名男性入选的概率(0.24),0名女性与4名男性入选的概率(0.02)。计算方法将在本章末公式5.1中进行介绍。实际观察值中女性个数是4名,该概率在图中用阴影表示出来了。这个阴影表示的区域叫做尾概率,因为它出现在概率分布的尾部。

两党选民是否势均力敌?

以下是应用概率核对检验的一个更大或也许更现实的例子。在一次选举之前,我们作了一个调查,问人们如果今天举行大选,他们将会怎样投票?我们发现650人将投民主党候选人票,550人将投共和党候选人票。我们想要用这些数据预测大选中选民将如何在这两个候选人中分野。

首先我们假设大选中,两党的选民势均力敌:即随机选定一个人,他支持任一党的概率均为0.5。在这个假设下,我们计算1200人中有650人或更多的人支持民主党的概率。我们要计算650或更多的人支持民主党的概率的原因是某一特定的民主党人数的概率是非常小的,通过计算650或更多的人我们可以知道650是属于一个小概率事件集还是属于常出现的事件的集合。

为了找到650或更多的人的概率,假设一个人是民主党人的概率是0.5,我们可以使用二项分布来计算。但是我们使用二项公式进行计算将是很费力的,因为要计算650,651,652,...等等的概率。所以我们使用标准正态分布来进行计算(在第七章中我们将会进行进一步讨论)。基于选民平分的假设,1200个人中有650个被转换为正态变量的 z 值为2.89。 z 大于或等于2.89的概率是0.002。即1200个人中有650个支持民主党的概率等于0.002。因此

在两党选民一样多的假设下,我们在 1000 次试验中仅可以得到两次在 1200 个票中有 650 以上的人支持民主党的情况。

p -值是在有关总体的某些假设下,观察值或更极端值出现的概率。

这种概率称为数据的 p -值。在图 5.8 中 p -值就是曲线下的阴影部分。如果两党的选民数目相同,那么这个数据的 p -值就会是 0.002 那么小。这或者意味着一个样本中可能的异常数据,或者是两党选民相等的假设是错误的。我们倾向于产生如此小的 p -值是由于选民相等的假设不正确。这样数据告诉我们两党的选民不相等支持率是不同的,如果今天举行大选的话,民主党候选人会赢得大选。

这种结论是典型的统计分析的结论。首先作出生成数据的总体的一个假设。然后,我们收集有限量的数据,并基于它进行某些计算。在计算结果的基础上,得出原假设是否正确的结论。因为数据个数有限,得出的结论是否正确具有某种不确定性。

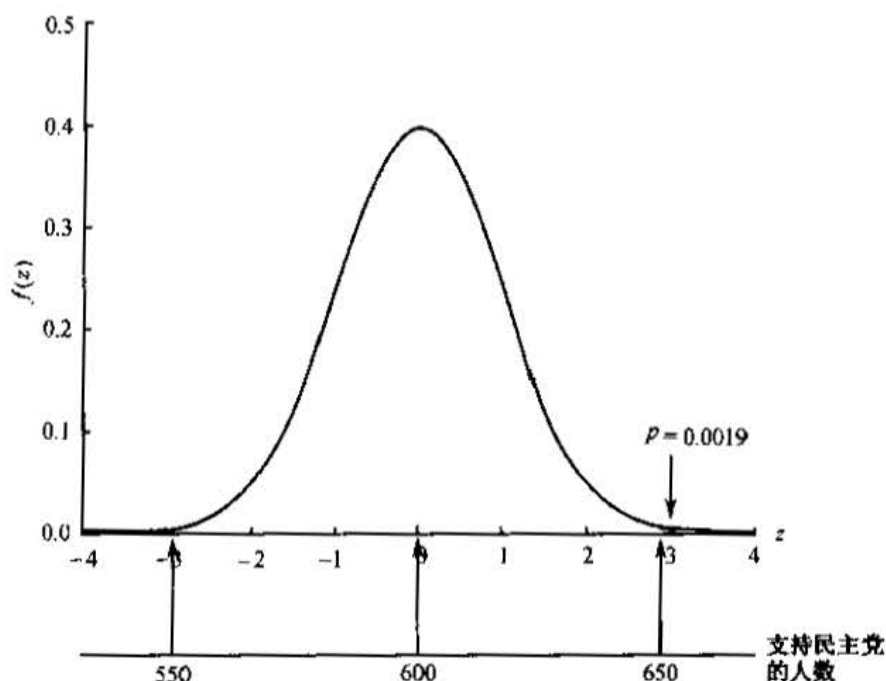


图 5.8 正态分布及 $z = 2.89$ 时的 p -值。

5.7 决策分析:利用概率来作决策

我们常常面对各种选择和决定;在很多非正式的场合,我们以概率为基础来作出大多数决定。当我们需要进行长距离旅行时,我们面临着是否乘坐飞机的选择。飞机坠毁时有发生,因此以小概率我们所乘飞机会坠毁并有生命损失。当然,我们能够不去乘飞机并躲开坠毁的概率。但是,空难发生的概率那么小,所以每天人们都乘飞机并抵达目的地。我们需要估计发生

空难的概率并对照抵达我们想去的地方所获得的利益。这种在风险和利益间达到平衡的过程叫做决策分析(decision analysis)。

面对不确定性作出决策是既基于群体水平也基于个体水平。在群体水平上,制定公共政策的决策是面对新的计划和法律的不确定性而作出的。因为空气污染对健康有害,是否社区应清除诸如汽车和工厂等以一切污染源以杜绝污染?大多数人看来不愿意如此;避免这样做是因为禁止这些行为有可能引起其它的可能更不愿意看到的后果。由于有很小的倒塌的概率一座桥就应该关闭吗?关闭这座桥将使路人绕远路还可能增加较小道路的交通事故。

科学家和统计学家常常联合起来发展基于概率的决策分析形式。决策分析曾用来形成决策以疏散困难时期的在黎巴嫩的美国人;也用来终止对前苏联出口先进计算机的禁令;它也曾用于星球大战计划。

我们怎样衡量不确定性的程度呢?如果一个事件发生的概率是1,那就是说这个事件在每一次都注定要发生。如概率是0.95则意味着发生的确定性要小一些;当概率达到这样高时,这个事件就是经常发生的但也有不发生的可能。类似地当概率是0.05意味着该事件几乎确定不会发生。最不确定出现在事件的概率等于0.5时。事件发生和不发生同等可能,而且不可能预测下次什么会发生。在这种情况下,一种衡量不确定性的方法是计算乘积 $p(1-p)$ 。当概率是0或1时,是不存在不确定性的,这个乘积都等于0。当 $p=0.5$ 这个最不确定的概率时,这个乘积达到了它的最大值0.25。

很多形成了日常决策的基础的概率并不是十分明确的。我们知道发生空难的概率很小,但并不能说出它究竟有多大。一个工程师很难说出另一名工程师同意的一座桥要坍塌的精确概率。当我们深陷爱河并准备结婚时,认为婚姻成功的可能性非常大。如果我们知道其成功概率为0.6,我们是不会举行婚礼的。当然,最近的离婚数据表明,很多人对于婚姻成功的概率有不实际的感觉,而另一些人即使知道了成功概率的精确值,仍然进行结婚。

即使在我们知道了概率之后,有时也很难理解这个数字的具体涵义。我们怎样表述由于每天服用有糖精的软饮料面死的概率是0.00001呢?将其转化为100000人中有1人的表达是有帮助的。但是就是这样,我们也很难彻底理解。有些时候,和其它概率进行比较有助于理解小概率的涵义。如果我们读到吸烟的危害是饮用软饮料的100倍,那么这种风险就很容易理解了。

有证据表明人们常常对于小概率事件产生过度反应。在80年代后期,在欧洲发生了旅行者被杀死的恐怖事件。这篇报道在当时对旅游者产生了很大的影响。实际上,恐怖主义行为的死亡数并未超过因经常发生的其它个别事故所造成的旅游者死亡数,如交通事故,自杀,心脏病突发,溺水和食物中毒等等。但是旅行的人数当年还是大大减少了,这说明人们认为恐怖主义事件发生的概率已经高的足以让他们取消旅行。

作出决策有时十分麻烦,部分因为概率并不是一个静态的数。对于一个不确定事件的个人概率常常在遇到新的事实时更新。我们对下一个周末下雨的概率随着这一周中天气的发展而越加明确。在看了周六晚间的天气预报后,我们关于第二天是否下雨的概率将十分接近于0或1。

人们在有了新的信息后更新概率的方法常常和数学上更新概率的方法不同。人们在评价概率时常常比他们应有的表现要保守,并不愿足够快地将概率移向0或1。如果一个事件发生的概率是0.5,而且获得了新的信息,统计学家可能将概率重新计算至0.9。但是一般人被

问及新概率时,他们多数将他们的个人概率改变较少。这种保守的行为可能有很多原因,包括不愿意改变自己的思想或以激进的新方式来理解事件。其它的影响因素也影响人们的行为,使他们正确地或是错误地使用信息(见下面“个人概率”)。

个人概率

心理学家们长期以来被个人概率所困扰。他们试图研究一些使个人概率不同于统计学家的概率的因素。在一个调查中,他们拿出大量的睡衣展示给女性看,询问那一件她最可能买下。然后将这些睡衣重新随机排列顺序,使每一件与原来的顺序都有所不同。研究者假设,最漂亮的一件睡衣有最高的概率被选中,而最不吸引人的一件有最小的概率。研究结果表明,一般说来,第一件看到的睡衣得分最高,无论睡衣的样式是什么样的。最后一个看到的睡衣是第二个最被青睐的。这些女性都不知道是睡衣出现的顺序影响了她们的选择。(来源:R. E. Nisbett, E. Borgida, R. Crandall, and H. Reed, "Popular induction: Information is not necessarily informative," in J. S. Carroll and J. W. Payne (eds.), *Cognition and social Behavior*, Hillsdale, NJ: Lawrence Erlbaum, 1976.)

在另一个研究中,传统观念的作用清楚地显示出来。测试对象被告知,心理学家对一个随机地从70个工程师,30个律师中选出的人进行了下述描述:"John是一个39岁的男人,已婚,有两个孩子,对地方政治十分积极。最大的爱好是收集稀有书籍,他好竞争,好辩,表达力强。"然后问测试对象:John是律师而不是工程师的概率有多大?注意,根据样本我们知道他是个工程师的概率是0.7而他是一个律师的概率只有0.3。所以随机选择的人是一个工程师的机会是7/10而他是一个律师的机会只有3/10。但是有95%的测试对象选择了律师而不是工程师。他们倾向于忽略工程师对律师的基本比例而过分强调了看来适合传统观念里律师应该像什么样子的信息。(来源:D. Kahneman and A. Tversky, "On the psychology of prediction," *Psychological Review*, vol. 80(1973), pp. 237-251)

人们依赖于个人概率而不是统计概率还有其它原因:一个事件的具体化——例如,你越来越相信你将染上某种罕见的疾病是因为一个你认识的人已染上它;一个亲密的朋友的强烈的观点——“主修通灵术(mortuary science)是没前途的”;一段个人的经历——“我不在乎统计怎样说,吸烟对我没有伤害”;或者权威的话——“大麻引起衰老。”

5.8 小结

概率是0到1之间的一个数。它告诉了我们事件发生的可能性。

5.1 怎样得到概率

有三种主要的找到概率数值的方法:等可能事件、相对频率和个人评价。当事件是等可能时,把感兴趣的事件结果数除以结果总数的方法来获得感兴趣事件的概率。当一个记录或事件覆盖了很长的时间或有很大的样本,一个事件出现的次数的比例是该事件概率的很好的估计。对于独一无二的事件,在所有可获得的信息的基础上的,个人对事件发生可能性的估计就是其概率。

5.2 概率计算的规则

和数字一样概率也可以进行加减乘除。这种计算可以帮我们简单些的概率来获得更复杂的概率。当两个事件不可能同时发生时,一个事件或另外一个事件发生的概率是这两个事件概率的和。当两个事件是独立事件时,一个事件和另外一个事件同时发生的概率是这两个事件概率的乘积。

5.3 优势:概率的对照物

优势是一个整数,它是事件不出现的次数与事件出现的次数的比率。优势 5 比 1 表示 6 次试验中,事件不出现的次数是 5,出现的次数是 1。

5.4 离散变量的概率分布

二项分布是用来寻找 n 次试验中,两个事件之一出现一定次数的概率是多少的;二项分布只有在样本数量很少时使用才比较方便。对于一个有很少的样本,又有很多种可能性的事件,例无安打比赛的次数,Poisson 分布比二项分布更适合寻找概率。Poisson 分布可以看成是二项分布的特例。

5.5 连续变量的概率分布

四个可以用来寻找概率的理论变量是标准正态 z , t , χ^2 , 和 F 变量。每一种变量都有自己的概率分布并且有自己的特别的分布曲线。

标准正态曲线是钟形的,中点两边各有 50% 的观察值。标准正态分布的均值是 0,标准差是 1。曲线下面有 95% 的面积在 -1.96 和 1.96 之间。样本取值可以转换成特别的 z 得分。从标准正态分布表中,我们可以查出 z 取值大于某一个数或小于某一个数的概率是多少。

t 分布,有时又叫学生分布,其图形与正态分布类似但不完全相同。它们都是最常用的统计分布。自由度的大小,或间接地观察值的个数的多少,决定了哪种 t 分布应该被使用。

χ^2 分布是有偏的,取值从 0 开始的。 F 分布也是有偏的,取值从 0 开始的。 F 分布依赖于一对自由度。

由定义可知, t 变量, χ^2 变量,和 F 变量都是由正态变量派生出来的。如果要使用这些统计变量,数据应服从正态分布。

5.6 使用概率来核对假设

p -值是在有关数据的某些基本假设下,得到所观察的值或更极端的值的概率。

5.7 决策分析:利用概率来作决策

科学家和统计学家常常联合起来开发一些建立在概率基础上的决策分析形式。在获得了新的信息时,人们使个人概率更新的方式通常与数学方法有所不同。

补充读物

Chernoff, Herman. "Decision Theory." In William H. Kruskal and Judith M. Tanur (eds.), *International Encyclopedia of Statistics*. New York: The Free Press, 1978. 简单讨论了有关统计学如何帮助作出决策。

Fairley, William B., and Frederick Mosteller. *Statistics and Public Policy*. Reading, MA, 1977. "People v. Collins. The supreme Court of California" and "A conversation about Collins." 用概率决定有罪还是无罪的有趣用法。

Huff, Darrell, and Irving Geis. *How to Take a Chance*. New York: W. W. Norton, 1959.

Snell, F. Laurie. *Introduction to Probability*. New York: Random House, 1984. vskip 20pt.

公 式

优势和概率

如果给了我们一个事件发生的优势 a 比 b , 那么概率 p 的计算方法为:

$$p = \frac{b}{a + b}$$

这个方程还可以被改写为:

$$p = \frac{b/a}{1 + b/a} = \frac{\text{优势}}{1 + \text{优势}}$$

另一方面我们还可以使用概率来得到优势。如果我们将上面的方程对优势 b/a 进行反解可以得到:

$$\frac{b}{a} = \frac{p}{1 - p}$$

在 5.3 节的奥运会主办地的例子中, 悉尼得到 2000 年夏季运动会主办权的概率是 0.692:

$$\frac{b}{a} = \frac{0.692}{1 - 0.692} = 2.25$$

这就是说优势是 1 比 2.25, 转化为整数是 4 比 9。优势是事件不发生(b)与事件发生(a)的比率。因此事件发生的概率是 $p = b/(a + b)$, 事件不发生的概率是 $p = a/(a + b)$ 。

二项概率

二元变量的两个取值常常被称为成功和失败。用 π 来代表成功的概率用 $1 - \pi$ 来代表失败的概率。在 n 个观察值的样本中, 成功的次数为 x 次, 而失败的次数为 $n - x$ 次。则 x 次

成功和 $n - x$ 次失败的概率, x 次成功的均值, 和成功次数的标准差分别为:

$$\begin{aligned} p(n \text{ 次试验中成功 } x \text{ 次}) &= \binom{n}{x} \pi^x (1 - \pi)^{n-x} \\ &= \frac{n!}{x!(n-x)!} \pi^x (1 - \pi)^{n-x} \end{aligned} \quad (5.4)$$

$$x \text{ 的均值} = \mu = n\pi \quad (5.5)$$

$$x \text{ 的标准差} = \sigma = \sqrt{n\pi(1 - \pi)} \quad (5.6)$$

这里惊叹号表示阶乘:

$$n! = n(n-1)(n-2) \cdots (3)(2)(1)$$

而 $\binom{n}{x}$ 表示二项系数, 即从 n 个中挑选 x 个的方法数。

作为一个例子, 如果 $n = 4$, $x = 3$ 和 $\pi = 0.49$ 。那么有三次成功一次失败的概率是:

$$\begin{aligned} \binom{4}{3} 0.49^3 (1 - 0.49)^{4-3} &= \frac{4!}{3!(4-3)!} 0.49^3 (1 - 0.49)^{4-3} \\ &= \frac{(4)(3)(2)(1)}{(3)(2)(1)(1)} 0.49^3 (0.51) = 4(0.49)^3 (0.51) \\ &= 0.24 \end{aligned}$$

均值和标准差是:

$$\text{均值 } \mu = 4(0.49) = 1.96 \text{ 女孩 / 家庭}$$

$$\text{标准差 } \sigma = \sqrt{4(0.49)(1 - 0.49)} = 1.00 \text{ 女孩 / 家庭}$$

Poisson 分布的概率

让我们用 μ 来表示事件发生次数的均值。那么事件发生 x 次的概率可以用下面公式来计算:

$$p(x) = \frac{e^{-\mu} \mu^x}{x!} \quad (5.7)$$

事件发生次数的均值和标准差则是:

$$\text{均值} = \mu \quad (5.8)$$

$$\text{标准差 } \sigma = \sqrt{\mu} \quad (5.9)$$

如果一个小时内, 电话铃平均响 2.1 次, 那么一小时内电话铃响 5 次的概率是多大? 我们在 Poisson 分布的公式中代入 $\mu = 2.1$ 得到:

$$p(5) = \frac{e^{-2.1} 2.1^5}{5!} = 0.042$$

100 个小时中只有大约 4 个小时会出现电话铃声响过 5 次的情况。呼叫次数的标准差是 $\sigma = \sqrt{2.1} = 1.45$ 次。

超几何概率分布

第 5.6 节中有关委员会成员选择是否公平的例子可以如表 5.3 所示。这十个人中每一个人都属于上述表格中四栏之一。因为我们知道性别变量的分布和需要选人委员会的人数,所以上表中总数一栏是固定的。表中的其它四个栏是随机的,如果形成另一个委员会,这两个表不一定一样。

表 5.3 性别和委员会人员的选取

	女性	男性	总数
选中的	4	0	4
未选中的	1	5	6
总共的	5	5	10

表 5.4 从 b 个第一类对象中选出 x 个,
从 $n-b$ 个第二类对象中选出 $m-x$ 个的概率

	第一类	第二类	总数
选中的	x	$m-x$	m
未选中的	$b-x$	$n-b-m+x$	$n-m$
总数	b	$n-b$	n

在表 5.4 中表示了概括的数据。所有事件的总数是 n , 某一种事件是 b 个, 另一种是 r 个。从 n 个事件中随机的选择 m 个并且属于无放回选择。那么某一种事件出现 x 次的概率是:

$$p(x) = \frac{\binom{b}{x} \binom{n-b}{m-x}}{\binom{n}{m}} \quad (5.10)$$

括号中的两个数是二项系数, 其计算与在二项分布中的解释相同(公式 5.4)。对于委员会成员选取的例子:

$$p(4) = \frac{\binom{5}{4} \binom{5}{0}}{\binom{10}{4}} = \frac{(5)(1)}{\frac{(10)(9)(8)(7)}{(4)(3)(2)(1)}} = \frac{(5)(24)}{5040} = 0.02$$

从概率分布中计算均值和方差

当离散随机变量按相应的概率 $p(x_1), p(x_2), \dots, p(x_k)$ 取值 x_1, x_2, \dots, x_k 时, 均值 μ 和方差 σ^2 可以通过下式来计算:

$$\begin{aligned} \mu &= x_1 p(x_1) + x_2 p(x_2) + \dots + x_n p(x_n) \\ \sigma^2 &= (x_1 - \mu)^2 p(x_1) + (x_2 - \mu)^2 p(x_2) + \dots + (x_n - \mu)^2 p(x_n) \end{aligned}$$

习题

回顾(习题 5.1—5.29)

- 5.1 我们日常生活中所用的概率的同义词是什么?
- 5.2 找到报纸文章中使用概率的例子并说明这个概率是如何被使用的?
- 5.3 根据本书所说,你将如何定义概率?
- 5.4 概率的变化区间是多大?
- 5.5 用基本公式 k/n 来计算概率意味着什么?
- 5.6 叙述得出从一副牌中抽出红心的概率是 0.25 的方法?
- 5.7 如果我们已知一个学生由于欺骗被学校开除的概率是 0.12,你怎样得到其没有被学校开除的概率?
- 5.8 当一个事件的“实际概率”未知时,那么能对其进行估计。说出一种估计事件概率的方法。
- 5.9 a.如果你对一个朋友说,你去参加假日舞会并认识一个新朋友的概率是 0.9,你给出的是哪种概率?
b.你给出的这种概率是用来描述何种事件的?
- 5.10 a.概率与优势之间有什么不同?
b.它们之间有什么关系。
c.为什么在打赌时,更愿意使用优势而不是概率?
- 5.11 一个有两个值的变量,如扔硬币时出现正面和反面,形成了二项分布的基础。二项是什么意思?
- 5.12 在二项分布中,二项变量的不同值的概率之和是什么?
- 5.13 重新访问有四个孩子的家庭。
a.如果用女孩的出生概率乘以 4,那么这个乘积告诉了我们什么?
b.用哪个希腊字母来代表这个乘积?
- 5.14 a.在什么情况下,统计学家使用 Poisson 分布来寻找某个事件的概率?
b.说出你认为什么时候使用 Poisson 分布最为合适。
- 5.15 根据变量的概率分布来计算变量的标准差是可能的,用哪一个希腊字母来表示标准差?
- 5.16 二项分布与 Poisson 分布的主要区别是什么?
- 5.17 a.说出用于统计中的四种主要连续理论变量。
b.这些变量中哪一种是最常用的?
- 5.18 a.说出标准正态分布的三个重要特征。
b.标准正态曲线下的总面积是多少?
c.标准正态分布的 z 变量的均值是多少?
d.标准正态变量的众数是多少?
e.相等地平衡的分布叫作什么?
f.大部分标准正态变量的 z 得分取值在哪两个值之间?

- 5.19 a. 由制酒专家发现的分布叫什么名字?
b. 他使用的假名是什么? 为什么要使用?
- 5.20 我们有一族 t 分布, 一族 χ^2 分布, 和一族 F 分布.
a. 以 t 分布为例, 如何区分一个分布族中的各个分布?
b. 何时一族中的各分布开始看起来互相类似?
- 5.21 你在什么时候使用统计表来寻找 χ^2 分布的概率? 除了 χ^2 变量的取值以外, 我们还需要知道什么?
- 5.22 a. 在两个自由度的 χ^2 分布中, 曲线下的总面积是多少?
b. 对于有三个自由度的 χ^2 分布, 它的面积有变化吗?
- 5.23 a. 如果有人告诉你, 他得到某个特定问题的 χ^2 取值是 -11.11, 你的反应是什么?
b. 如果 χ^2 变量的取值是 11.11, 你会对它的大小印象深刻么? 为什么?
- 5.24 a. F 分布中的 F 代表了什么?
b. 除了很小的自由度以外, F 变量的一般取值范围是什么?
- 5.25 在一个选民的样本中, 有 700 个人说他们会投共和党的票, 在两党选民相同的假设下, 对此数据的 p -值是 0.002. 你是否认为你的关于两党选民一样多的假设是正确的?
- 5.26 如果一个经济学家说她是研究决策分析的, 那么她研究的对象可能是什么呢?
- 5.27 “统计意味着永不必说你确信无疑”可能既有趣又属实; 你能解释为什么吗?
- 5.28 在 5.3 节的优势问题中, 各个城市成为主办地的概率已知。这是那种概率? (是等可能概率, 长期比率概率, 还是个人观点), 请解释?
- 5.29 给出一个基于等可能结果的概率的例子, 一个基于长期比例的概率的例子, 和一个基于个人观点的概率的例子。

解释(习题 5.30—5.52)

- 5.30 当一个政治分析家说: “我相信总统明年又被选上的概率是 0.6,” 这是什么概率? 这个分析家的论述的意思是什么?
- 5.31 在买汽车比赛的彩票时, 你的朋友告诉你, 你准备下注的三个运动员的优势是: Trudi 3 比 2, Andy 8 比 2, Rod 20 比 1. 如果你想使风险最小, 那么你应在谁身上下注? 如果你不在乎损失一些但想获得最大的赔率, 那么你应在谁身上下注? 解释你的战略?
- 5.32 假设 32% 的美国成年男子都拥有至少一只枪, 那么可以表明 300 个成年男人中有 120 个或更多拥有枪的事件出现的概率是 0.0015。
a. 你怎样找到概率 0.0015?
b. 这类概率又叫什么?
- 5.33 a. 给出一个概率几乎为 0 的事件的例子。
b. 给出一个概率几乎为 1 的事件的例子。
- 5.34 一项研究报告表明, 13 至 19 岁的怀孕并人工流产的女子和其双亲谈过此事的概率是 0.61. 这种概率的含义是什么?
- 5.35 解释为什么在知道男孩出生率时, 可以不必收集大量的各种家庭的样本数据而用二项分布来估计一个家庭中的男孩和女孩的期望个数?
- 5.36 a. 以下是计算一个家庭中有 3 个女孩, 1 个男孩的概率的例子, $p(\text{GGGB}) = 0.49 \times$

$0.49 \times 0.49 \times 0.51 = 0.06$ 。向一个不懂统计的朋友解释你这样算的原因。

b. 如何用公式计算一个家庭有四个男孩的概率?

- 5.37 为了估计美国有四个男孩没有女孩的家庭的个数,我们应使用二项分布来计算还是随机的抽取 100 个家庭进行调查。哪一个的结论更精确?
- 5.38 如果你的应答者样本数量很大,如果你不希望自己去计算,如何得到二项概率?
- 5.39 有四个孩子的家庭中女孩个数的均值是 1.96。这个均值的获得是用四个孩子的家庭中的女孩个数的可能取值与取这个值的概率相乘后相加获得的。解释你如何得到有三个孩子的家庭中女孩个数的均值?
- 5.40 在 1992 年的总统竞选中,在任总统布什为了置疑副总统候选人戈尔的亲生态立场,在他最后的竞选日子中说“如果我们不小心,我们将会深深陷人为猫头鹰的工作中而不为人民做任何事。”你怎样用决策论的术语来解释布什的抱怨?
- 5.41 你如何能使一个六年级的班级相信,在以后的六年中,他们有一次骨折的概率是 0.0009?你如何能使一个心理学本科班相信,下一年中他们发生车祸的概率是 0.50?你如何能使一个宝马汽车的拥有者相信一个汽车在世界范围内被挟持的概率是 0.00001?
- 5.42 根据这段文章,“ z 大于 2.5 的概率等于 0.0062。而 10 个自由度的 t 变量大于 2.5 的概率等于 0.0152。”
- a. 对哪一个变量更可能找出大于 2.5 的数?
- b. 为什么 z 变量与 t 变量在这时有区别?
- c. 什么能使这两个统计量之间的差别变得非常小。
- 5.43 如果一个统计学家说,得到某个特定的 χ^2 或者更大的概率是 0.4,你将如何描述这个用 χ^2 来衡量的事件?
- 5.44 a. 以什么方式 χ^2 分布看起来与 t 分布不同? 解释之?
- b. 自由度为 3 的 χ^2 分布不容易取到的值是什么?
- c. 还有什么其它分布与 χ^2 分布类似?
- 5.45 某个社会俱乐部有 52 个老成员 7 个新成员。一个假定随机的选取结果为 5 名新成员但没有老成员被选中在集会后作清扫工作。(统计专家说这种结果出现的概率是 0.000004)。
- a. 你是否认为这次选择有欠公平? 解释原因?
- b. 作为一个新成员,你怎样向俱乐部的头头辩护你的说法。填表来使你的说明更确切。

	老成员	新成员	总数
选中的			
未选中的			
总数			

- 5.46 a. 假设你们校园组织的成员按照是否愿意在每年的献血行动中献血被等分为两组。所以每一个成员参加献血的概率是 0.5,但后来你发现 100 个人中有 10 个人参加了献血行动,你对你假设的原始概率得出了什么结论?
- b. 从你的早期假设的献血支持者的概率中,你可能想得到的两个结论是什么?

- c. 可以帮助你决定哪一个结论是正确的统计决策方法是什么?
- 5.47** 在明尼苏达州夏天打高尔夫球时被闪电击中的概率是 0.00002。
- 用什么方法可以使这个统计数值更容易被人理解?
 - 为什么你这样认为?
- 5.48** 为了度假,你想决定是飞去埃及看金字塔,在尼罗河上畅游还是飞去迈阿密利用租来的车探索佛罗里达群岛。在每一个旅游点,旅游者都不时地会遭遇到不友好的当地人。
- 在你做决定时,你将考虑什么概率?
 - 你怎样算计可能愿意做什么?
 - 在你做决定时,近期的事件报道对你的决策有多重要?
- 5.49** 你今年对于慈善行动很感兴趣。你希望选择一个做可能对牺牲者有最大影响的研究的单位。
- 你将如何在乳腺癌症研究,爱滋病研究基金会,帮助残疾儿童,和心脏病学会中进行选择?
 - 广告在哪方面对捐赠人作出错误的决定产生潜在的影响?
- 5.50** 在 20 个连续的世界级棒球比赛中,一支在常规赛季中有更多偷垒的球队赢得了 13 次冠军,输了 7 次。如果假设在指定的一年中,每个球队赢得比赛的概率是 0.5。那么 20 次中赢 13 次或更多次的概率是 0.13。基于偷垒数据,你对你的输赢可能均等的假设有什么结论?
- 5.51** 在第四章的新娘年龄的例子中,新郎的年龄的均值是 32.3 岁,新娘年龄的均值是 30.0 岁,两者之差是 2.3 岁。将 2.3 年变为一个 t 变量的取值,你发现,如果总体中新娘与新郎的年龄没有差距,那么我们得到相差 2.3 岁的样本的概率是 0.002。从这个 p -值的大小,你对于平均年龄没区别的假设做出的结论是什么?
- 5.52** 美国邮政局声明,83% 邮寄到纽约市的信会第二天送到。一个人为了检验这种说法给自己邮寄了 10 封信,有四封信是在第二天到达的。利用隔夜递送的概率为 0.83 的假设及 10 个样本,应用二项分布来计算发现,有四封或更少封信第二天送到的概率是 0.0027。(来源: Daniel Seligman, "Ask Mr. Statistics," *Fortune Magazine*, July 24, 1995, pp. 170 - 171)
- 我们如何解释 p -值为 0.0027 的含义?
 - 你对这个值有什么保留看法?

分析(习题 5.53—5.79)

- 5.53**
- 滚动一个骰子出现 6 的概率是多少?
 - 滚动一个骰子出现 1 的概率是多少?
 - 滚动一个骰子出现 1 或 6 的概率是多少?
- 5.54** 如果我们知道任何一个学生在心理学期末考试中得 B 的概率为 13%,得 A 的概率为 5%。那么:
- 一个学生在期末考试中得分不是 A 就是 B 的概率是多少?
 - 在一个英国文学课程中,任何一个学生期末得 B 的概率为 20%,得 A 的概率为 10%。那么,一个学生在心理学和英国文学期末考试中都得 B 的概率是多少?
 - 一个学生在两门课中的得分不是 A 就是 B 的概率是多少?

- d. 一个学生在所有的课中得分都是 A 的概率是多少?
- 5.55 你能否想出以下这个习题的实际限制? 四个人打扑克牌, 一个叫做 Chris 的牌手仔细地看了看发在桌面上翻开着的牌, 在这一局已发 16 张牌, 最后 4 张牌是扣着的。桌上有两张 A; 每一只手有一张。
- a. 在最后一轮中, 发出一张 A 的概率是多大?
- b. 在最后一轮中发出两张 A 的概率是多大?
- 5.56 假设你已知当 Loretta Jones 下个月将分娩时 Jones 家会有一对双生子的概率, 你同样知道 Lizzie Smith 生下三胞胎的概率。
- a. Jones 生下双胞胎或 Lizzie 生下三胞胎的概率是多少?
- b. Jones 生下双胞胎同时 Lizzie 生下三胞胎的概率是多少?
- 5.57 Mars 公司宣称(生产 M&Ms 的公司), 他们使用以下的颜色分布:

褐色	红色	黄色	绿色	橘黄色	蓝色	总数
30%	20%	20%	10%	10%	10%	100%

- a. 买一小袋普通的 M&Ms(没有花生那一种的)。
- b. 数一数每一种颜色的 M&Ms 各有多少。
- c. 在表中记录出现的频率。
- d. 用上面的每一种颜色的百分比乘袋中 M&Ms 的总数。在一个表中记录结果。这个表就是期望频率; 你原先的频率是观察频率。算出期望频率精确到一位小数点。
- e. 观察频率不应和期望频率差得太远。一种度量它们差别的方法是对每一种颜色计算一个分数, 分子为观察频率和期望频率的差的平方, 而分母为 M&Ms 的期望频率。再对这六个分数求和。这个和是 χ^2 变量的一个值, 有 5 个自由度, 而且如果你的值位于 1 到 10 的范围之外, 则是很意外的。
- f. 求出你的数据的 p -值, 它是你所观察的和更极端的与期望频率的偏差的概率。你的这一袋是否异常?
- g. 吃掉 M&Ms。
- 5.58 以下数据是自由度为 1 的 χ^2 变量的 50 个观察值:

1.76	1.64	0.38	0.48	0.01	1.90	0.32	0.01	1.92	1.56
0.57	0.73	0.60	0.01	6.86	0.17	1.09	1.01	0.02	0.15
0.09	0.10	0.60	0.38	2.04	0.07	0.95	1.52	0.06	4.21
0.05	0.08	0.25	0.15	0.36	1.84	0.23	0.00	2.19	1.57
1.28	0.30	0.73	0.19	0.07	0.01	0.47	0.91	0.92	0.05

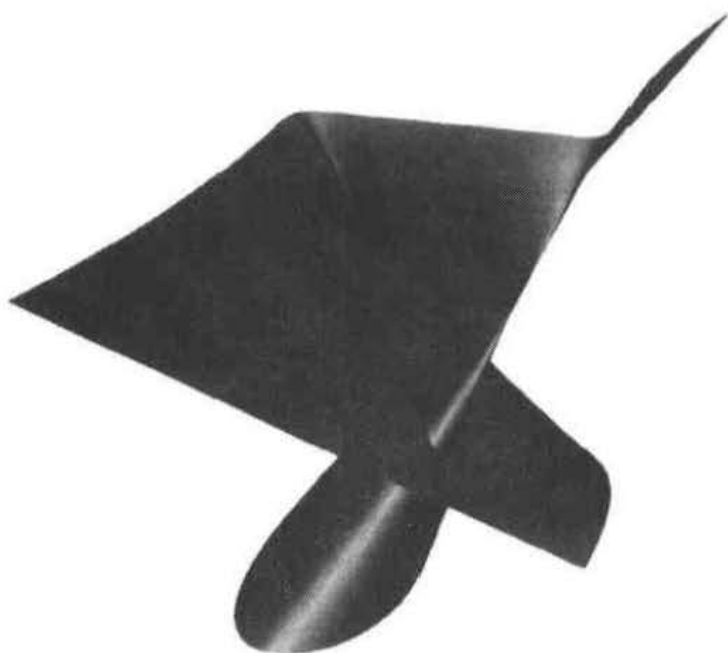
- a. 使用长度为 1.00 的区间画出直方图来显示观察值的分布。
- b. 描述这个分布的形状。(注意: 不同的自由度的 χ^2 分布有不同的形状。)
- c. 在该 χ^2 值的样本中有多少个观察值是大于 3.84 的?
- d. 根据 χ^2 分布的统计表, 自由度为 1 的 χ^2 分布的取值大于 3.84 的百分比是多少?
- e. 将 c 与 d 的答案相比较。
- f. 将上表中第一列的 5 个值相加, 然后将第二列的值相加, 如此进行下去直到将 10 列加完, 这些和为来自有 5 个自由度的 χ^2 分布的十个观察值。

- g. 利用长度 2.00 的区间画出直方图来表示这十个新观察值的分布。
- h. 描述这个分布图的形状并和 a 中的分布比较。
- i. 用 χ^2 分布表来找出自由度为 5 的 χ^2 分布的一个数使得只有 5% 的观察值大于这个数。
- j. 在我们十个值的样本中, 有这样大的值吗?

- 5.59 a. 从另一个学生处收集 M&Ms 的 χ^2 值, 并且为全班的学生画出直方图。
b. 根据理论的自由度为 5 的 χ^2 分布的值, 有一半的值大于 4.35 而大于 9.24 的值只有十分之一。将这些数字和全班观察到的 χ^2 值相比较。
- 5.60 使用二项分布来找到下列概率:
a. 扔十次公平硬币, 出现 8 次正面, 出现 9 次正面和出现 10 次的概率各是多少?
b. 正而出现 8 次或大于 8 次的概率?
- 5.61 心理学家观察了一些小群体并将每群划为或者竞争型或者合作型。她假设任何一个群属于这两类中之一的概率是 0.5, 她把 7 群划为竞争型而 1 群为合作型。
a. 得到大于等于 7 个竞争型群的概率是多少?
b. 是否你认为这个概率太小而竞争型的概率为 0.5 的假设有错误?
- 5.62 使用统计软件程序或统计表来找出下列概率。
a. z 大于等于 2.34 的概率。
b. 自由度为 17 的 t 变量大于等于 2.34 的概率。
c. 自由度为 17 的 t 变量小于等于 -2.34 的概率。
d. 自由度为 17 的 t 变量小于等于 -2.34 的概率或大于等于 2.34 的概率。
- 5.63 使用统计软件程序或统计表来找出下列概率。
a. 自由度为 2 的 χ^2 变量大于等于 6.78 的概率。
b. 自由度为 20 的 χ^2 变量大于等于 27.8 的概率。
- 5.64 使用统计软件程序或统计表来找出自由度为 2 和 46 的 F 变量大于等于 3.45 的概率。
- 5.65 人口普查局报告说 1989 年家庭收入的中位数是 \$35225。
a. 随机选择的一个家庭收入大于 \$35225 的概率是多少?
b. 随机选择十户家庭, 所有十户收入都大于 \$35225 的概率是多少?
c. 为什么在这类问题中, 中位数比均值拥有的信息更多?
- 5.66 人口普查局报告说 1990 年美国年龄的中位数是 32.7 岁。
a. 随机选择一个人, 他的年龄小于 32.7 岁的概率是多少?
b. 随机选择 5 个人, 有 4 个人的年龄小于 32.7 岁的概率是多少?
c. 随机选择 5 个人, 有大于等于 4 个人的年龄小于 32.7 岁的概率是多少?
- 5.67 盖洛普调查公司 1991 年 2 月发现, 在任何给定的一天, 有 33% 的美国人通过读书消遣。(来源: The New York Times, July 26, 1992, p. E5) 在一张桥牌桌上的四个人在前一天都在读书的概率是多少?
- 5.68 如果在一个有四个孩子的家庭中, 有四个女孩的概率是 0.06, 那么少于四个女孩的概率是多少呢?
- 5.69 a. 如果你知道一个事件的概率是 0.25, 另一个独立事件发生的概率是 0.08, 那么如何求出这两个事件同时发生的概率?

- b. 这个答案比任何一个事件单独发生的概率都要小, 你能解释为什么吗?
- c. 你能否将这个方法应用在打雷与闪电同时发生的概率上? 为什么能或为什么不能?
- 5.70 一旦我们得到一个 z 得分(回顾我们为找到标准得分 z 而从原始得分减去分布的均值再除以标准差), 我们能用标准正态分布得出这个得分如何地经常(或不经常)出现。如果我们在几何考试成绩中得到以下 z 得分, 那么我们怎样利用它们如何地经常(或不经常)来解释。
- a. $z = 0.22$
- b. $z = 2.50$
- c. $z = -1.96$
- d. $z = -0.013$
- 5.71 在德国的汽车拥有者中, 有一辆 Porsche 的概率是 0.07。有一辆 Mercedes 的概率是 0.29。
- a. 有一辆 Porsche 或一辆 Mercedes 的概率是多少?
- b. 有一辆 Porsche 和一辆 Mercedes 的概率是多少(假设拥有两辆车是独立事件)?
- c. 这两种车一辆也没有的概率是多少?
- 5.72 Bob 在法学院的第一年中, 一个学生的 GPA 大于等于 3.8 的概率是 0.15。GPA 在 2.5 与 3.8 之间的概率是 0.8。那么一个学生 GPA 不在 2.1 与 3.8 之间的概率是多少?
- 5.73 Robin 想在周四离开学校春游。但是她不想错过周五的数学和物理课。根据过去的记录, 数学老师取消一次课的概率是 0.05, 而物理老师取消课的概率是 0.10。那么, Robin 希望的情况, 这两门课在一天全部取消的概率是多大?
- 5.74 一个连续的概率分布的光滑的曲线勾画出了它下面区域的轮廓。曲线下的面积就是概率, 曲线下的总面积是多少?
- 5.75 在一个正态分布中, 95% 的 z 的取值都在 -1.96 和 1.96 之间。
- a. 值在 -1.96 之下和在 1.96 之上的百分比是多少?
- b. 值仅小于 -1.96 的百分比是多少?
- c. 值在 -1.96 和 0 之间的百分比是多少。
- 5.76 自由度为 10 的 t 分布中, 找出包含 95% 的点的取值区间。
- 5.77 看习题 5.46 中给出的信息, 你能说出只有新成员参与打扫活动出现的概率是 0.000004 的概率值是怎样计算出来的吗?
- 5.78 国家高速公路安全委员会根据过去的数据报告每年在 100000 个有执照的司机中, 大约有 6 个女性涉及与酒精有关的致命的交通事故。
- a. 在有 100000 个有执照的司机的城市中, 使用 Poisson 分布说明如何得出一年中没有女性涉及与酒精有关的致命交通事故的概率是 0.002。
- b. 同一时期, 一个或大于一个女性死于与酒精有关的交通事故的概率是多少?
- 5.79 对一个二项变量, $n = 6$ 及 $\pi = 0.4$ 。
- a. 利用统计表 3(二项分布)画出该概率分布的直方图。
- b. 使用二项概率计算均值 μ 。
- c. 与 $\mu = n\pi$ 比较你算出的数。
- d. 这个概率分布是否看起来在均值处达到平衡?

C H A P T E R 6



6.1 样本统计量和总体参数

6.2 点估计

6.3 区间估计:给结论留一些余地

6.4 小结

作出结论:估计



Gallup(盖洛普公司)就消费者对美国产品质量的看法,对美国、德国、日本的消费者分别进行调查,结果表明:有 55% 的美国人相信美国产品的质量非常好,而持同样看法的德国人和日本人的比例分别是 26% 和 17%。美联社在报导这项调查结果时曾提到“抽样误差在正、负三个百分点之间”。在报导中“正、负三个百分点”这句话有什么作用?(来源: *The Philadelphia Inquirer*, Oct. 2, 1991, p. C-7)

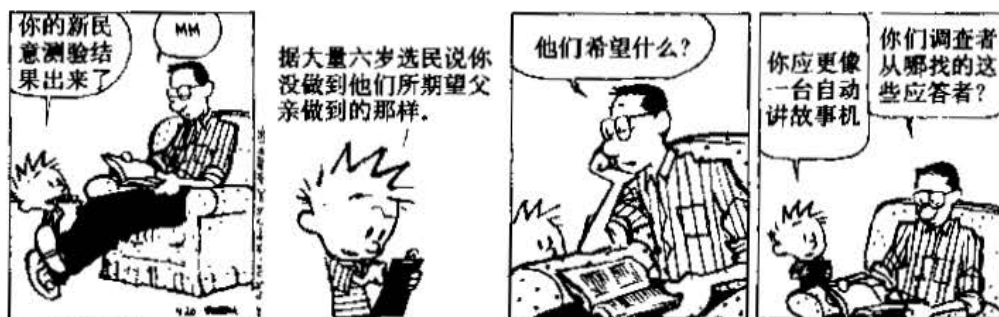
在报纸、杂志和电视新闻上经常可以看到调查结果和其它统计性报导。虽然它们的结论是各式各样的,但都是经过统计研究得出的。在许多结果之中,统计研究表明,在一个样本中有百分之几的非洲裔美国人更喜欢用“非裔美国人”而不是“黑人”作为他们种族的名称(26%; 1989 年 Yankelovich Partners 公司为 Time/CNN 作的电话调查结果);有百分之几的美国白人说他们没有足够的钱来购买食品(13%; 来自 1989 年 Gallup 的调查);女体操运动员的平均年龄是多少(12.3 岁; G. S. Theintz, et al. “Evidence for a reduction of growth potential in adolescent female gymnasts,” *Journal of pediatrics*, vol. 122 (1993), pp. 306-313) 以及人们一生中有百分之几的时间是在睡觉中度过的(30.9%; *New York Times*, Tuesday, September 6, 1995, p. C6)。

当某项研究的结果使人感兴趣时,研究者就会超越样本数据去找到一些样本所属总体的结论。他们想知道如果调查了总体中的全部元素(人口、树木等)将会得到什么样的结果;假如 Gallup 的调查者能够问到所有美国人,则认为美国产品质量好的人占总数的百分之多少? 样本百分数告诉我们的仅仅是在样本中的几百个美国人如何回答询问者的关于质量的问题。在有些情况下没有办法定义被调查的总体,可研究者们还是希望能对产生数据的世界有进一步的了解。

统计推断是一个过程,它能从样本数据得出与总体参数值有关的结论。它由两部分构成:估计和假设检验。

超越实际数据是统计学的一个分支,被称为统计推断(statistical inference)。它由估计(estimation)和假设检验(hypothesis testing)组成。本章讨论的是参数估计,而假设检验则是下一章的内容,这些方法将会在以后的章节中被用到。

研究者之所以用样本替代总体的实用原因是:别说在一般情况下没有办法收集到总体中的全部元素,即使能,所需的时间和金钱也是难以承受的。尽管样本中的信息并不完全,而且来自样本的结果一般都不等于总体真值,研究者们还是满足于样本数据。为了弥补样本结果的不准确,研究者们计算抽样误差——这个数能使 19/20 的抽样结果都位于由总体真值加、减样本误差所得到的区间内。



(“Calvin and Hobbs” copyright 1993 Watterson. Dist by Universal Press Syndicate. Reprinted with permission. All right reserved)

6.1 样本统计量和总体参数

样本统计量是从样本数据中计算出来的数。

总体参数是在原理上可以从整个总体中计算出来的数。

统计量(statistic)是 statistics 的单数形式。最平常的样本统计量(sample statistic)的例子是样本均值 \bar{x} 、样本百分比 P 和样本标准差 s 。人们习惯于用英文中 26 个罗马字母来标记常用的样本统计量。因为统计量是从一个样本数据中计算出来的,所以它们的值是可知的。

在总体中,性质类似的量被称为总体参数(population parameter),一般用希腊字母来标记。例如:总体均值用小写的希腊字母 μ 来标记、总体百分比用大写 Π 以及总体标准差用小写的 σ 等等。

这些概念被表示在图 6.1 中。图 6.1 是一个有关总体、样本、总体参数和样本统计量的图示。最左边的大椭圆代表总体,其中元素可以是人、县、植物、猪、电灯泡或者任何别的东西。从这个总体中随机抽取元素组成的子集就是样本,用右边的小椭圆表示。因为样本是随机抽取的,所以它能很好地代表总体。

值得注意的是总体可以被分成两类:有限总体和无限总体。有限总体是由具体的、大量

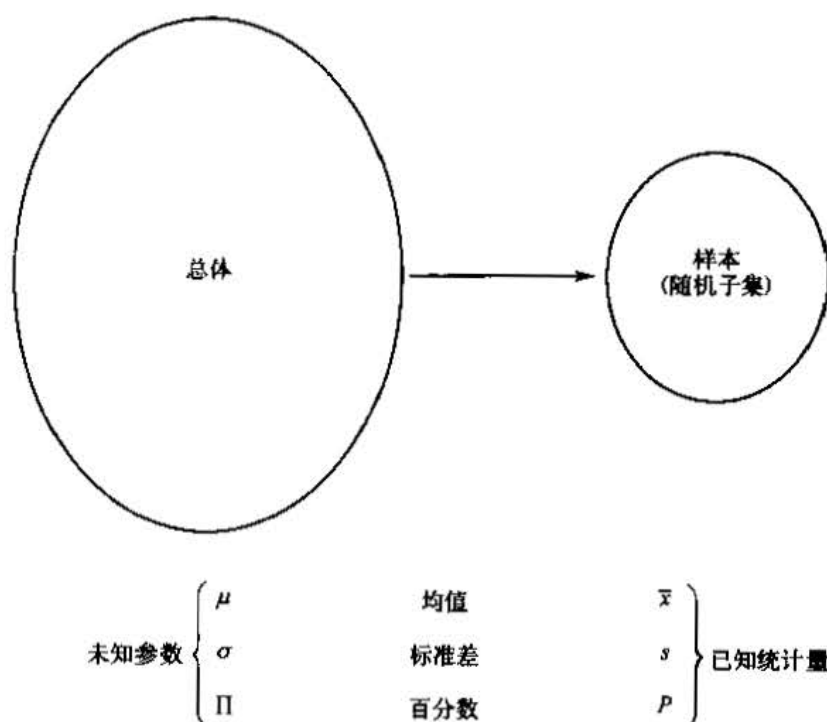


图 6.1 总体和样本, 未知的总体参数和已知的样本统计量。

的、有限的元素组成的集合。比如在最近的总统竞选中, 投票者约有一亿人, 这些人就是投票者的总体。无限总体是元素的一个假设集合, 例如: 所有的已做好的和将要被某些机器做出来的电灯泡集合; 用于检验出现正面和反面概率的某硬币的所有投掷的集合。机器能制造的所有的电灯泡和用一硬币能做的所有的投掷都是不可能真正计数的。

停下来想一想 6.1

请你想出一些出现在现实生活中的总体的例子。在每个例子中什么实际问题可以用抽样来表示? 对调查人员来说这些问题的难点是什么? 它们怎样阻碍了对总体信息的寻找?

6.2 点估计

点估计是一个用来估计总体参数的数。

假设你正在研究平均一个美国人一生中要得到多少交通罚单。报告研究结果的方法有以下两种: “10” 或者 “8 到 12 之间”, 请考虑它们各自的优缺点。这两种结果代表了估计总体参数时所用的两种不同的方式。最简单的是点估计 (point estimation), 像 10 张罚单这个结果就是个点估计。如果在一次抽样调查中有 58% 的回答者赞同提高汽油税, 则这个数就可以作为赞成提税的人在总体中所占比例的点估计。

同样,样本均值也可以作为总体均值的点估计,看来利用样本统计量作为总体参数的点估计是合理的,尽管我们不知道这个估计如何好?

什么是一个好的点估计?

由于一个来自样本的特别的估计值绝不会精确地等于总体参数的真值,所以问某一个值是否是好的估计值是没有意义的。而可以问的是计算估计值的方法是不是一个好方法。

为了确定一个方法的好坏,需要对多次重复同一个研究所得的结果进行比较。下面让我们用一个假想的实验来说明。假设我们做了多次抽样调查,每一次都可以得到一个样本百分比。同时假设这些抽样都是理想抽样,而唯一的误差就是抽样误差。例如,下面的数字代表在每个由 500 个观测值组成的 10 个样本中赞同提高汽油税的人所占的百分比,其结果是:

58.0 57.8 61.0 59.4 55.8 63.2 59.0 60.6 57.4 58.6

尽管所有样本都来自同一个总体而且这个总体只有一个固定的未知百分比,这些样本所得到的百分数仍然互不相同。这一现象正是由抽样的随机性引起的,这也是一个在大选前民意测验中出现的现象。几家不同的调查组织进行抽样,调查人们的投票倾向性,并预测候选人在未来大选中的得票百分率。结果,它们得到的结论各不相同。如果相信它们的抽样方案没有缺陷的话,那么这些结论之所以不同的原因就可以被归结为抽取随机样本所固有的波动和随机性。

一个好的估计方法可以这样被定义:如果在无数个样本上应用该估计方法,得到的估计的均值等于总体参数的真值。上面例子中 10 个估计值的平均是 59.1。这些互不相同的估计值分别来自 10 个样本,而这 10 个样本是由计算机从百分比是 60% 的总体中随机抽取的。如果计算机所抽的样本大大多于 10 组,那这些样本百分比的均值将会等于真值 60%。因此可以说样本百分比是总体百分比的无偏(unbiased)估计。虽然每一次的结果可能不对,但多个重复抽样结果的平均就是对的。如果重复抽样后得到的许多统计量的均值仍不能等于总体的真值,就称这种估计是有偏的(biased)。

一个好的点估计的标准:

1. 如果大量样本的样本统计量的均值等于总体参数的真值,则这种样本统计量是该参数的无偏估计。
2. 许多重复抽样所得的估计值不应离真值太远。

在前面的例子里,10 个样本统计量中的最小值是 55.8,比真值 60.0 小了约 4 个百分点,而最大值比真值大了 3.2 个百分点。这 10 个百分点的标准差近似等于它们的标准误差(本例中是 2.1%),这意味着平均每个样本百分点与这 10 个值的均值相差 2.1%。这样样本百分比是总体百分点的一个好的估计,因为:(1) 多次抽样所得样本百分点的均值等于总体百分点的真值;(2) 许多不同样本中的样本百分点都离总体百分点很近(相近程度通常依赖于样本的观测值数目)。

可以证明,任何别的通过样本来估计总体百分点的方法其效果都因为抽样误差太大而给

出更坏的结果。所以即使上面的结果不太理想,也是能得到的最好的了。同样也可以证明样本均值是总体均值的一个好的估计,比用样本中位数来估计总体均值要好。

战略中使用点估计的例子:德军有多少坦克?

有时不知道应该如何利用样本中的数据来估计总体参数。统计理论可以用来推导点估计的计算公式。一个不寻常的例子是盟军在估计二战期间德军制造的坦克总数时所用的方法。

由于许多战略上的理由,盟军非常想希望知道二战期间德军总共制造了多少辆坦克。德国人在制造坦克时是墨守成规的,他们把坦克从1开始进行了连续编号。在战争进行过程中,盟军缴获一些了敌军坦克,并记录了它们的生产编号。那么怎样用这些号码来估计坦克总数呢?在这个问题中,总体参数是未知的生产出的坦克总数 N ,而缴获坦克的编号则是样本。



一个二战时的德军坦克,有时统计学家们证明他们是值得被攻击的。(来源:Guler Pictures.)

假设我们是盟军手下负责解决这个问题的统计人员。制造出来的坦克总数肯定大于等于记录中的最大编号。为了找到它比最大编号大多少,我们先找到被缴获坦克编号的平均值,并认为这个值是全部编号的中点。因此样本均值乘以2就是总数的一个估计;当然特别要假设缴获的坦克代表了所有坦克的一个随机样本。这种估计 N 的公式的缺点是不能保证均值的2倍一定大于记录中的最大编号。

N 的另一个点估计公式是用观测到的最大编号乘以因子 $1 + 1/n$, 其中 n 是被俘坦克个数。例如你俘获了10辆坦克,其中最大编号是50,那么坦克总数的一个估计是 $(1 + 1/10) \times 50 = 55$ 。此处我们认为坦克的实际数略大于最大编号。

这种方法的各種變形的確用於二戰之中。從戰後發現的德軍記錄來看,盟軍的估計值非常接近所生產的坦克的真實值。記錄仍然表明統計估計比通過其它情報方式做出估計要大大接近於真實數目。統計學家做得比間諜們更漂亮!

停下来想一想 6.2

在上面的例子中,对战争、坦克的部署以及捕获做了哪些假设?

6.3 区间估计:给结论留一些余地

区间估计又称置信区间,是用来估计参数的取值范围的。

第二类估计参数的方法是区间估计(interval estimate)。“取值范围是8到12”是一个区间估计。与此类似,汽油提税问题的答案应该是“52%到64%之间。”这种估计方法提供的信息比“估计值是58%”要多。

对于大多数总体参数,估计区间是用如下方法找到的:首先要找一个样本统计量如均值或者比例;然后从数据中计算出抽样误差;最后用样本统计量加、减抽样误差就得到了估计区间的两个端点。用这三步得到的区间称为置信区间(confidence interval)——一个被统计学家认为能够包含参数真值的区间。一个总体参数的置信区间是用一个样本统计量加、减抽样误差得到的:

$$\text{统计量} - \text{抽样误差} \quad \text{到} \quad \text{统计量} + \text{抽样误差}$$

现在让我们以本章开头的问题为例来作一个说明。在来自美国的样本中,相信美国产品质量好的人占55%,抽样误差是 $\pm 3\%$,则总体百分点的置信区间是:

$$55 - 3 = 52 \quad \text{到} \quad 55 + 3 = 58$$

我们希望未知的总体百分点包含在此区间中。本章末尾的公式6.1可以用来找置信区间。公式6.2是6.1的简化,在许多情况下使用起来比较方便。

让我们再做另一个假想的实验。如果再抽另一个样本,那么它将会产生出一个多少不同的样本百分点和不同的置信区间。即使有许多不同的样本及不同的百分点和不同的置信区间,我们仍能期望这些区间都包含真值参数。

因为统计学家有某种程度的信心认为这个区间会包含真正的固定的参数值,所以给它取名为置信区间。其理由是:如果我们收集了许多不同的样本,并对每个样本都构造了一个置信区间。这些置信区间有足够的宽度使它们中的95%包含了总体百分点的真值,而5%没包含,则95%这个值就被称为置信水平(confidence level)。这个值是比较常用的,当然也能用别的值来做置信水平。

怎样看上面的从52到58这个置信区间?这个区间包含未知的总体百分点的真值吗?对于一个特定的区间,这个问题是无法回答的。如果我们说有95%的把握相信52到58这个区间包含了未知参数,也就等于承认了我们不绝对了解这个特定区间,我们知道的仅仅是——在多次抽样中有95%的样本得到的区间包含真值。

图6.2阐明了这一点。它画出了石油提税问题中10个样本所得的10个置信区间。第一个样本的区间是从54到62,第二个从53.7到61.9,等等。因为这些样本的样本百分点不同,所以这些线的中点不同。

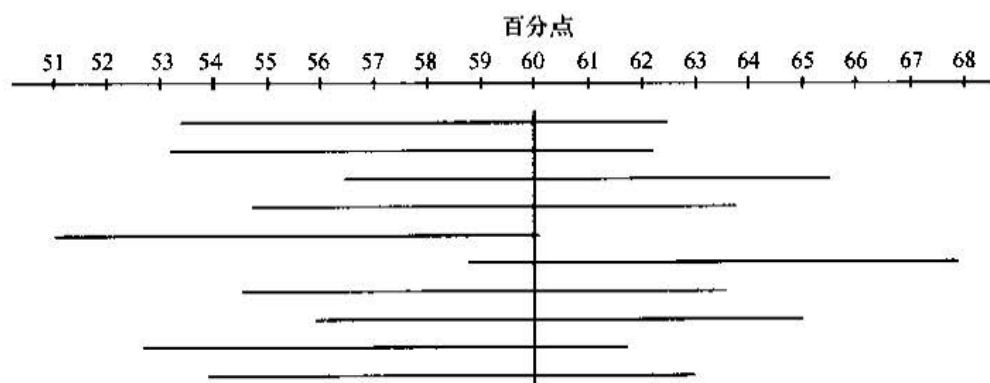


图 6.2 来自 10 个不同样本的置信区间,所有的都包含 60 这个真值。

如果用某种方法构造的所有区间中有 95% 的区间包含真值,5% 的区间不包含真值,那么这些用该方法构造的区间都叫做置信水平为 95% 的置信区间,简称 95% 置信区间。

因为这些数据是计算机产生的,所以真正的总体百分点是已知的,它等于 60,被一条在 60 的竖线标记在了图中。可以看到,所有这 10 个区间都包括真值。如果抽的样本不是 10 组而是 100 组,我们将会期望大约有 5 个置信区间不包含 60,而有 95 个区间包含 60。这意味着:

大多数情况下,研究人员在收集数据时只取一个样本。可是,没人能够知道由这个样本产生的置信区间是否包含均值。所以,他们只能希望这个区间是大量包含真值的区间中的一个,但它也有可能成为少数几个侥幸不包含真值的区间之一。

我们之所以用这种拐弯末角的表达方式其原因在于总体的参数值是固定的、未知的。而我们用样本构造的置信区间是不固定的。如果我们重复这个做法,会得到一个多少不同的区间;在这个意义下,置信区间是一个随机区间。因此置信区间会因样本的不同而不同,而且不是所有的区间都包含参数真值。一个区间就像一个为了捕获未知参数而撒出的网,不是所有撒网的地点都能捕获参数。

警告!

请记住,概率可以告诉我们一件事发生的频繁程度。因此,我们不能说 52 到 58 这个特别的区间以 0.95 的概率包含未知的对美国产品质量持肯定态度的人所占百分比。它真正的意义是如果你做了 100 次抽样,大概有 95 次找到的区间包含真值,有 5 次找到的区间不包含真值。

当然这样说不直观。假设实际参数是 57,则区间 52 到 58 一定包含真值而不是以 95% 的概率包含真值。同样如果假设真值为 50,那么区间 52 到 58 就绝不包含真值,无论你做多少次实验。

这个概率不是用来描述某个特定区间包含未知真值的可能性的,一个特定的区间总是包含或者绝不包含真值,不存在一会包含,一会不包含的问题。用概率可以知道在多次抽样得到的区间中大概有多少个包含了参数的真值。

停下来想一想 6.3

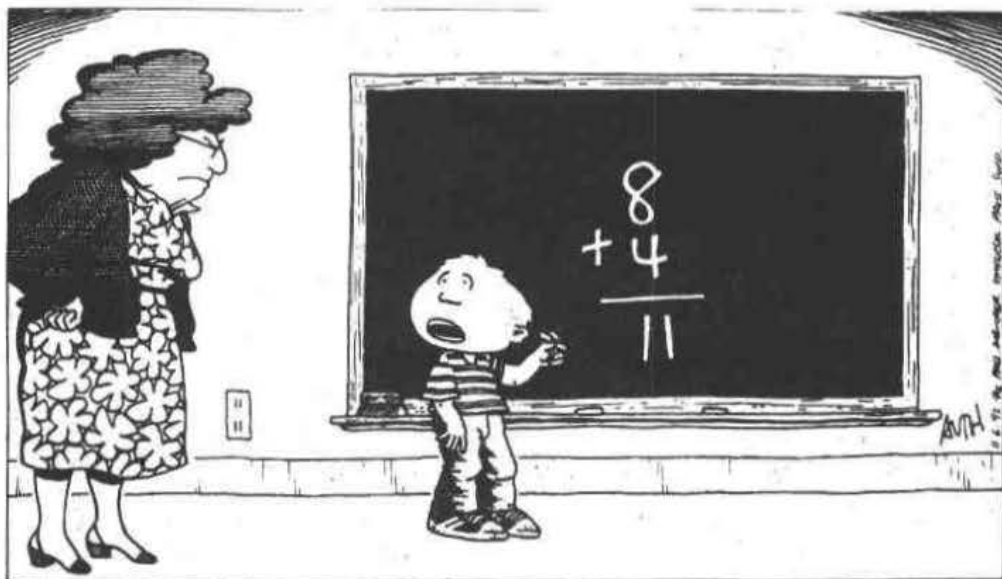
为什么用一个精心挑选的随机样本构造的置信区间还是有 1 比 20 的可能(5%)不包含真的参数?

置信区间的长度**停下来想一想 6.4**

样本中包含的观测个数与置信区间的长度有何关系? 你能否给出一个例子来阐明这种关系。

1. 样本中的观测值个数影响着置信区间的长度。大的样本产生较短的置信区间, 小的样本产生较长的置信区间。
2. 置信区间的长度还受置信水平的影响。低的置信水平(如 90%)产生较短的区间, 高的置信水平(如 99%)产生较长的区间。

短的置信区间能比长的置信区间提供更多的有关总体参数的信息。如果你告诉我一个未知的总体百分比的置信区间是从 0% 到 100%, 那么你等于什么也没说: 因为百分点的值只能在 0 到 100 之间变化。如果你告诉我它的置信区间是从 30% 到 70%, 那么我就对这个未知的参数有了一些了解。如果你告诉我它的置信区间是从 49% 到 52%, 那你几乎已经告诉我这个未知参数的值了。



“我的意见是这个值是 11, 但误差在 ± 2 之间。”

经画家 Carol Cable 许可翻印。

样本容量对置信区间的影响 得到较短置信区间的一个办法是有个包含许多观察值的大样本。因为大样本中包含了较多的信息,而包含信息量大的置信区间都比较短。还有,来自大样本的统计量的值比来自小样本的统计量的值距离真值更近。用数学的观点来说就是在抽样误差的公式中样本观测的个数一般是在分母上,分母越大,分数就越小。

图 6.3 给出了 6 个容量不同的样本计算出来的置信区间。其中假设每个样本的样本百分比 P 都是 60%,这些区间是用本章后面给出的公式计算的。从这个图中可以很清楚地看到样本容量越大区间长度越短。但是置信区间变短的速度不像样本容量增加的速度那么快;如果样本容量加倍,区间长度并不是原来长度的一半。若是要区间长度是原来的一半,样本容量就得是原来的 4 倍。但增加观测的花费可能很贵;所以当样本容量大于某数时就不值得再增加观测值了。这就是为什么在大多数全国性调查对 1,200 个左右的人提问的原因。这个样本容量已经大得足够使抽样误差减小到 3%。

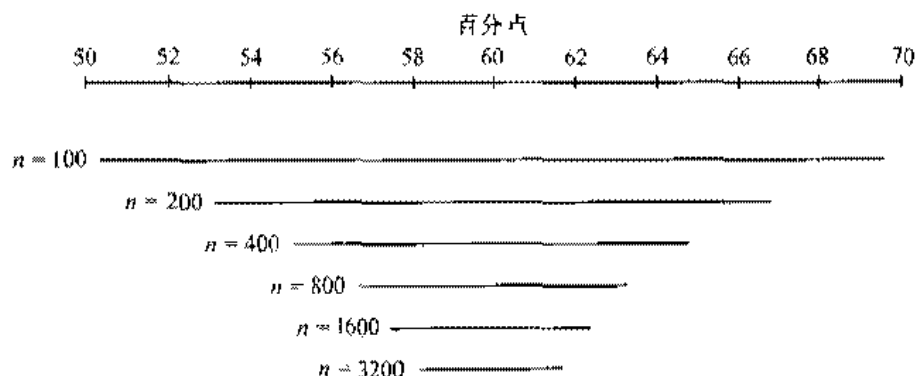


图 6.3 未知的总体百分点的置信区间。它们分别是由容量不同、样本百分比都为 60% 的样本计算而得的。

置信水平对置信区间的影响 得到一个短的置信区间的另一个方法是用一个较低的置信水平。最常用的一个是 95% 置信水平。当然也可以用 90% 置信水平。其置信区间比 95% 的还要短。当然 90% 的置信水平不如 95% 那么让人放心。因为 90% 的置信水平意味着在多次抽样中只有 90% 的置信区间包含真值,而另有 10% 不包含真值。

下面我们来看一看在一个容量为 1,200 的样本上 95% 和 90% 置信区间的差别。假设样本中有 60% 的人赞扬总统的工作。因为抽样误差是 ± 2.8 ,水平为 95% 的置信区间是从 57.2 到 62.8。当然希望这个区间是众多包含总体真值的区间中的一员。类似地,水平为 90% 的置信区间是 57.7 到 62.3,其抽样误差是 ± 2.3 。第一个区间长度是 5.6 而第二个则为 4.6。如果非常希望这个区间包含真值,那么稍长的区间更可能包含它。

置信区间的应用 美联社报导了 Gallup 公司所做的一项关于人们对美国产品质量看法的调查结果;调查了来自三个国家的总共 3,500 人,其中有 55% 的美国人认为美国产品质量好,而只有 26% 的德国人和 17% 日本人持同样看法。报导中还提到了抽样误差是加或减 3,这表明在 100 次抽样中,大约有 95 次得到的样本百分数与总体百分数之差小于 3%。当然真值是未知的(来源: *The Philadelphia Inquirer*, October 2, 1991 p. C-7.)。

用报导中的三个百分点分别加、减 3 就可以得到每个消费群体的 95% 置信区间。这样,

有 52% 到 58% 的美国消费者、23% 到 29% 的德国消费者和 14% 到 20% 的日本消费者认为美国产品质量好。希望这些置信区间的确都包含真值。

这篇新闻报导中没有提到置信水平是 95%，但是按照惯例，在不加特殊说明的情况下所提到的抽样误差都是基于 95% 置信水平的。所有好的调查都应该提供抽样误差，这样读者自己就可以构造置信区间了。

停下来 想一想 6.5

你是否认为美国、德国和日本的消费者在评价美国产品的质量时所持的观点“的确”有所不同？如果调查的 49% 的墨西哥人认为美国产品质量非常好，你是否认为他们比美国人更怀疑美国产品的质量？

差异的置信区间

两个参数的差值的置信区间可以用来研究两个群体是否有所不同或者一个群体是否随时间发生了变化；例如可以用此方法来判断共和党和民主党的参议员对儿童福利法案是否持有不同的立场。

两个比例之间的差异 1989 年 2 月 Time/CNN 委托 YankelovichPartners, Inc. 做了一个电话调查，访问了 503 名非洲裔美国人，询问他们更喜欢用“非洲裔美国人”还是“黑人”来作为他们种族的名称。结果其中有 26% 的人更喜欢前一种称呼。五年后，也就是 1994 年 2 月这个调查被重做，结果是 53% 的人更喜欢“非洲裔美国人”的称呼。这是否能说明与五年前相比人们的看法有所改变？通过检查这两次研究结果的置信区间能得出什么结论？

在 1989 到 1994 年间，这个百分点从 26 变到了 53，相差了 27。这 27 个百分点的差异是不是能够单纯归因于在比较两数时总是存在的抽样误差呢，或是它大到单靠随机性本身所不能解释了？如果我们能够对比 1989 年和 1994 两次百分比的真值，那它们的差异是否和估计的一样大？回答这些问题的一个方法是用样本数据构造两个总体百分比差值的置信区间。

两个样本百分比差异的抽样误差是 $\pm 9\%$ ，这样，二者差值的 95% 置信区间是 $27 - 9 = 18$ 到 $27 + 9 = 36$ 。作为估计差异真值的从 18 到 36 的置信区间的一个显著特点就是它不含有 0。如果差异是 0 即两个百分比相等，就说明人们喜欢被称为什么没有变化。但是这个区间不包含 0，据此我们可以推断真正的差异不是 0；还可以认为有比例大约在 18% 到 36% 之间一部分人变得更喜欢“非洲裔美国人”这个称呼了。从这个样本可以得到结论：在这五年期间，非洲裔美国人所喜欢的称呼已经发生了变化。本章结尾处的公式 6.4 表明如何算出两个总体百分比差值的置信区间。

停下来 想一想 6.6

你能否举出一个类似的你不能肯定它在五年内是否发生了变化的问题？

均值间的差别 在我们的一个班上，我们让学生们数他们一分钟内脉搏的挑动次数。男

生的平均脉跳次数是 71.6 次,女生的是 64 次。他们每分钟平均脉跳相差 7.6 次。这个差值显示了一个性别比另一种性别的心跳快多少。这个差异是应该由任意两个样本均值的随机不同造成的呢,还是已经超过了随机性本身所能解释的范围?

在这两个总体中,均值差异的 95% 置信区间是 -1.8 到 17.0 次。这个区间包含有 0。因此两个总体均值差异有可能是 0,而两个样本均值的不同也就有可能完全是由随机效应引起的。必要的计算公式由公式 6.5 给出。

6.4 小结

统计推断是用来从样本数据中得到关于总体参数值的结论。它由两部分组成:估计和假设检验。

6.1 样本统计量和总体参数

样本统计量是从样本中计算出来的数。例如样本均值 \bar{x} , 样本百分比 P , 以及样本标准差 s 。因为样本统计量是从样本数据中计算出来的,所以它的值是可行的。总体参数是从总体中计算出来的数,如总体均值 μ , 总体百分比 Π 以及总体标准差 σ 。总体参数值几乎总是从来不会被知道的,所以要用样本数据中提供的信息来估计总体参数的值。

一个有限总体是具体的、含有大量元素的,如:在部分选举中的所有投票者。无限总体含有无限个元素。所有用一个硬币可能的投掷为一个无穷总体的例子。

6.2 点估计

点估计是一个用来估计参数值的数。

从一个总体中抽出非常大量不同的随机样本并在每个样本上计算统计量的值;如果这些统计量的均值等于总体参数的真值则这个样本统计量就被称为总体参数的无偏估计。如果大量重复抽样所得的统计量的均值不等于总体的真值,则该估计是有偏的。某个统计量是否无偏通常可以用数学证明。

在估计总体均值时,我们更喜欢用样本均值而不是样本中位数或是其它类型的平均值。这是因为在多数情况下,在多个样本上计算出的样本均值比它们计算出来的其它的样本统计量的值能更紧密地聚集在总体均值周围。

6.3 区间估计:给结论留一些余地

区间估计是用于参数估计值的一个范围。一个区间比一个单值能提供更多的信息,但构造和解释这类区间则更困难些。大多数估计总体参数的区间可以这样被找到:(1)计算样本统计量如均值;(2)计算抽样误差;(3)用样本统计量加、减抽样误差。这样就得了个区间,它被称为总体参数的置信区间。

在来自不同样本的多个置信区间当中包含未知的总体参数的区间所占的百分比称为置信水平。置信水平为 95% 的意思是多次抽样中有 95% 的置信区间包含未知的总体参数值而另

外的 5% 则不包含真值。至于在一次抽样得到的置信区间是包含总体参数的众多区间中的一员呢,还是属于个别不含参数值的区间就永远不得而知了。

短的置信区间包含的信息比长的多。可以通过增加样本容量或降低置信水平这两种方法来获得较短的置信区间。当一个调查有大约 1200 个响应者时抽样误差是 $\pm 3\%$,也就是说在 100 个不同的样本中大约有 95 个样本百分比的值与总体百分比的值相差不到 3%。新闻报导中通常利用给出 95% 置信水平的抽样误差。

也可以对两个参数的差值,比如两个总体均值之差,来构造一个置信区间,用以研究这两个参数是否相等。

补充读物

Burkholder, Donald L. "Point estimation." In William H. Kruskal and Judith M. Tanur (eds.), *International Encyclopedia of STATISTICS*. New York: The Free Press, 1978. 更多地关于如何用一个数估计参数。

Pfanzagl, J. "Confidence intervals and regions." In William H. Kruskal and Judith M. Tanur (eds.), *International Encyclopedia of STATISTICS*. New York: The Free Press, 1978. 更多地关于如何估计参数值。

公 式

总体百分数的置信区间

从一个大的总体中抽取一个由 n 个观测值组成的随机样本,并用 P 来标记样本百分比。我们想要得到总体百分比 π 的一个 95% 置信区间。该区间为

$$P - 1.96 \sqrt{\frac{P(100 - P)}{n}} \quad \text{到} \quad P + 1.96 \sqrt{\frac{P(100 - P)}{n}} \quad (6.1)$$

1.96 这个值来自正态分布。是变量 z 的一个值。它使得 2.5% 的 z 值小于 -1.96 同时 2.5% 的 z 值大于 1.96。也就是说有 95% 的 z 值落在 -1.96 到 1.96 之间,从而构成了一个 95% 的置信区间。对于别的置信水平的置信区间,相应的 z 值可以从正态分布表中查到(统计表 1)。

一个快速计算 95% 置信区间的近似方法是令 $P = 50$,同时四舍五入 1.96 到 2:

$$P - \frac{100}{\sqrt{n}} \quad \text{到} \quad P + \frac{100}{\sqrt{n}} \quad (6.2)$$

在某种意义上,这是一个有些保守的置信区间,因为大多数情况下它比公式 6.1 得到的区间要稍长一些。但是它容易计算,并且与(6.1)的结果差别不太大。由公式 6.2 可知,对于一个有 900 个观测的样本,误差是 $100/30 = 3.3$,而对有 1600 个观测的样本这个误差变成了 2.5,等等。对于误差是 3,则要求样本含有 1111 个观测。一般都要把误差控制在 3 左右。这就是为什么大多数样本要求有 1200 个响应者的原因。

总体均值的置信区间

由 n 个独立的、服从正态分布的观测组成一个样本, 样本均值计为 \bar{x} , 样本标准差记为 s 。则总体均值的置信区间是:

$$\bar{x} - t^* \frac{s}{\sqrt{n}} \quad \text{到} \quad \bar{x} + t^* \frac{s}{\sqrt{n}} \quad \text{d.f.} = n - 1 \quad (6.3)$$

这儿 t^* 是 t 变量的一个值, 它可以从自由度为 $n - 1$ 的 t 分布的统计表 2 中查到。要得到 95% 置信区间, 只需找到 t^* 使 95% 的 t 变量的值都落在 $-t^*$ 到 $+t^*$ 之间即可。

在少数情况下总体的标准差 σ 是已知的, 这时可以用 σ 去替代公式中的样本标准差 s , 同时还要用来自正态分布的值 z^* 去替代来自 t 分布的 t^* 。这时置信区间就变成了:

$$\bar{x} - z^* \frac{\sigma}{\sqrt{n}} \quad \text{到} \quad \bar{x} + z^* \frac{\sigma}{\sqrt{n}}$$

当 $z^* = 1.96$ 时它是一个 95% 置信区间。

两个百分比之差的置信区间

一个样本有 n_1 个观测, 另一个有 n_2 个观测; 相应地, 两个样本的样本百分比分别为 P_1 和 P_2 。则两个总体百分比 Π_1 和 Π_2 之差的 95% 置信区间是:

$$\begin{aligned} (P_1 - P_2) - 1.96 \sqrt{\frac{P_1(100 - P_1)}{n_1} + \frac{P_2(100 - P_2)}{n_2}} \\ \text{到} \quad (P_1 - P_2) + 1.96 \sqrt{\frac{P_1(100 - P_1)}{n_1} + \frac{P_2(100 - P_2)}{n_2}} \end{aligned} \quad (6.4)$$

两个均值之差的置信区间

一个含有 n_1 个观测的样本的样本均值为 \bar{x}_1 , 样本标准差为 s_1 。另一个样本有 n_2 个观测, 样本均值为 \bar{x}_2 , 样本标准差为 s_2 。先由公式:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

给出两个标准差的平均值, 然后就可以得到两个总体均值 μ_1 与 μ_2 之差的置信区间了:

$$(\bar{x}_1 - \bar{x}_2) - t^* s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{到} \quad (\bar{x}_1 - \bar{x}_2) + t^* s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (6.5)$$

我们从自由度是 $n_1 + n_2 - 2$ 的 t 分布表中查得到 t^* 的值, 使 t 变量落入 $-t^*$ 到 t^* 之间的概率是 0.95。

习 题

回顾(习题 6.1—6.21)

- 6.1 统计推断的目的是什么?
- 6.2 如果均值是由样本计算出的,那么这个值称为什么?
- 6.3 如果已知均值是总体的,那么这个值称为什么?
- 6.4 统计学家分别用罗马字母和希腊字母表示哪两种数据?
- 6.5 请分别说出 μ, Π, σ 与 \bar{x}, P 和 s 在定义各种群体的量时的区别。
- 6.6 a. 什么是一个参数的点估计?
b. 什么是一个参数的区间估计?
c. 说出这两类参数估计方法各自的优缺点。
- 6.7 有一个无偏统计估计是什么意思?
- 6.8 请解释总体参数的置信区间。
- 6.9 样本均值是 10, 抽样误差是 ± 2 。
a. 总体均值的置信区间如何计算?
b. 如果你在这种情况下计算了置信区间,你怎样理解这个结果?
- 6.10 假设你从总体中抽取了大量样本,并且用每个样本都构造了一个总体均值的置信区间。如果从这些置信区间中随机选一个,那么它不包含总体参数的区间的机会有多大?
- 6.11 描述缩短置信区间长度的两个办法。
- 6.12 a. 当调查设计者们在全国范围对一个新产品的观点有兴趣时,他们不用总体而用样本来进行研究。请给出他们这样做的三个理由。
b. 请说出用样本替代总体的一个主要缺点。
- 6.13 怎样从样本数据中获取关于总体的信息?
- 6.14 一个参数的点估计与区间估计有什么区别?
- 6.15 如果你从同一个总体中抽取 20 个随机样本,你就可以得到 20 个均值和中位数。于是你就可以计算这 20 个均值的标准差,同样也可以计算中位数的标准差。
a. 是均值的标准差小还是中位数的标准差小?
b. 为什么最好要选标准差小的统计量来估计总体均值?
- 6.16 你认为置信水平这个名字的来由是什么?
- 6.17 “总体参数的值是永远不可知的。”统计学家是否不应试图了解总体参数?请解释。
- 6.18 a. 为什么从总体中抽样时必须强调随机性?
b. 如果样本不是随机抽取的会发生什么现象?
c. 非随机样本会对总体参数的估计产生什么影响?
- 6.19 a. 在盟军估计德军坦克数量的例子中,对俘获的坦克样本做了至少两个假设,请说出它们是什么。
b. 如果这些假设是错误的,你能否想出一种方法,使估计出来的值太小或太大。

- 6.20 为什么在许多全国性调查中包含的响应者的个数是 1200,而不是 500 或者 2000 或其它数目?
- 6.21 在 Dogpacth 举行地方选举之前,市报的编辑做了一个有关市民选取城市迷失家畜管理人员的民意测验。大约 89%的人说他们计划选 Abner Yokum,另外 11%的人说他们将在选票上写上其亲戚的名字。请问这个编辑会用什么置信水平来使他报导的调查结果更加可信?

解释(习题 6.22—6.35)

- 6.22 1960 年,在约翰·肯尼迪和里查德·尼克松竞选总统时做过一项对选民的调查,其结果是 469 名妇女中的 47%和 429 位男士中的 53%投了肯尼迪的票。这个比例相差 6%,而这个差值的置信区间是从 -3%到 15%(来源:这个数据是由 *Inter-university Consortium of Political Research* 提供的。它最初是由 *Angus Campbell*, *Philip Converse*, *Warren Miller* 以及 *Donald Stokes* 为 *the Survey Research Center 1960 American National Election Study* 收集的。原始的数据收集者和 *Consortium* 都不对这儿提供的对该数据的分析和解释负责)。
- a. 请解释“我们以 95%的把握确信差异的真值在 -3%到 15%之间”的含义。
- b. 为什么要研究这个区间?
- 6.23 1990 年 Roper Organization 随机选择了 3000 名妇女,其中 58%的回答者赞同“大多数男人都认为他们自己对世界的看法才是最重要的”这个观点,它的边际误差范围是 $\pm 2\%$ 。(来源: *Contemporary Social Psychology*, 1991 年集 15 卷 53 页,由 *Delwin S. Cahoon* 和 *Edm Edmonds* 所写的文章“*Comments concerning increased female negativism toward males*”)。
- a. 你能说出另外 42%的回答者的一些信息吗?
- b. 在你对这项调查结果下结论之前,你还要对数据的抽样方式做哪些了解?
- c. 你如何用边际误差来解释你的结论?
- 6.24 在一项药物研究中,对照组中携带 HIV 病毒的服用安慰剂后的母亲后发现出生的 154 个婴儿当中有 40 个(26%)婴儿出生时就携带 HIV 病毒。(来源: *The New York Times*, February 21, 1994, P. A1.)可以得到出生时就携带 HIV 病毒的婴儿在总体中所占百分比的置信水平 95%的置信区间为从 19%到 33%。
- a. 你怎样解释这个置信区间?
- b. 在这个研究中的抽样误差有多大?
- 6.25 一项美国人对自己工作热爱程度的调查结果是:在从事同一项工作十年以上的人中非常热爱自己工作的人占的比例在 70%到 75%之间(通过加减 2.5 得到的置信区间)。考查这项结果,说明在真正的这样美国人的总体中喜欢自己工作的人所占的比例是否真的在 70%到 75%之间?
- 6.26 在一项 1989 年 Gallup 调查中表明,13%的美国白人受访者回答说去年他们有时没有足够的钱去购买食物,而非洲裔美国人中有 33%的人是这样回答的。这两个百分比相差了 20%。
- a. 你是否认为这两个百分比差值的置信区间将会足够大,使我们可以认为相应的两个总体百分比没有差异?
- b. 如果要得到两个总体百分比没有差异的结论,这两个样本百分比差值的抽样误差得达到多大?

- 6.27 请说出下列的每项调查是针对样本的还是针对总体的,并解释。
- 从餐厅里随机抽取其中十分之一的学生调查他们对午餐菜单的满意程度。
 - 乐队中所有的单簧管演员对领奏者的满意程度。
 - 所有参加上一次选举的选民来考查总统在选民中的声望。
 - 所有内战中记录的军队的死亡数,判断战争中哪方损失的人员更多。
- 6.28 调查了 22 个女体操运动员,她们的平均年龄是 12.3,标准差是 0.2 年。这就给出了总体年龄均值的 95% 置信区间为从 11.9 岁到 12.7 岁。(来源:the journal of Pediatrics 杂志 1993 年第 122 卷 306 - 313 页 G. E. Theintz, H. Howard, U. Wiess 和 P. C. Sizonenko 的文章“Evidence for a Pediatric growth potential in adolescent female gymnasts.”)你怎样理解这个置信区间?
- 6.29 在来自南方的 205 个人中有 70 个人(34%)称自己是专业人员,而 151 个非南方人中有 62%的人称自己是专家。这两地专业人员百分比之差是 $62\% - 34\% = 28\%$,经过计算,非南方和南方的专业人员百分比之差的 95% 置信区间是从 18 到 38。(来源:American Sociological Review 杂志 1972 年 37 卷第 236 页由 J. C. Mckinney 和 L. B. Rorue 所写的“Further comments on ‘The changing South’: A response to Sly and Weller”)
- 你如何理解这个置信区间?
 - 是否这个国家的两部分的专业人员在各自的总体中的百分比有可能相等?
- 6.30 让我们来看一看橄榄球超级联赛的前 24 场比赛中的赢家与输家。在 24 个获胜的队中有 7 个(29%)第二年仍参加超级联赛,而 24 个失败的队中有 4 个第二年继续留在了联赛中。这两个百分比的差异为 $29\% - 17\% = 12\%$ 。为了判断这个差异是不是偶然发生的,我们找到了它的 95% 置信区间: -12 到 36。
- 你如何解释这个置信区间?
 - 在注意到明年继续打球的队中赢者比负者多这个事实后,你还能下结论说今年是赢是输对明年继续打球没有影响吗?请解释。
- 6.31 在有关秃顶的研究中,秃顶男人组成的样本中有 55%的人患有心脏病,而不秃顶的男人组成的样本中只有 43%的人患心脏病。(来源:1993 年 2 月 14 号的纽约时报第 A1 and C12.)这两个百分点之差是 $55 - 43 = 12$,相应的总体百分数之差的 95% 置信区间是从 6 到 18。
- 你怎样解释这个区间?
 - 从这个区间来看,在两个总体中患病百分比是否应该相同?
- 6.32 一个由大学四年级男生组成的样本中,平均身高是 71 英寸,标准差是 2.1 英寸。(来源:1992 年 7 月 26 日的纽约时报 E5 页。)用这组数据构造的总体平均身高的 95% 置信区间是 70.4 英寸到 71.6 英寸之间。美国男人的身高的均值是 69.1 英寸。
- 你如何理解这个置信区间?
 - 从这个置信区间来看,大学四年级男生的身高和所有男性身高是否有区别?
- 6.33 在研究人造瓣膜中血液流速时抽取了一个样本,其平均流速是 5.96。由此得到总体平均流速的 95% 置信区间是 5.22 到 6.70。
- 你如何理解这个置信区间?
 - 制造商是想要保证这种瓣膜中的血液流速至少为 5.00。你是否觉得制造商能证明这个保证?
- 6.34 在调查了 49 位雇员后得知,他们每年的平均生病天数是 7.0 天。由此得到总体中平均生病天数的 95% 置信区间是 6.3 到 7.7 天。按全国的数字,平均生病天数应该是 5.1 天。
- 你如何解释这个置信区间?

b. 这是否说明, 该样本来自这样一个总体, 其平均生病天数大于等于 5.1 天?

6.35 根据习题 3.46 和 4.66 提供的箭鱼样本得知平均水银浓度是 1.09 ppm, 总体平均浓度的 95% 置信区间是 0.90 ppm 到 1.28 ppm。

a. 如何理解这个置信区间?

b. 总体的平均水银浓度是否看起来满足法定的不超过 1.00 ppm 的限制?

分析(习题 6.36—6.52)

6.36 用下面的抽样结果给出两个相应的总体百分数的点估计: 对于问题“你是否曾有过与妻子/丈夫离婚的想法?”在 682 对被访夫妇中有 30% 的女人和 23% 的男人回答“是”(来源: 纽约 Academic 出版社 1983 年出版的由 Joan Huber 和 Glenna Spitze 所著的书“Sex Stratification, Children, Housework, and Jobs”的第 98 页)。

6.37 一般来讲总体参数的置信区间是如何得到的?

6.38 a. 用计算机软件产生 10 个容量为 50、来自均值为 0 标准差为 1 的正态分布总体的样本, 并记录这 10 个不同的样本均值。

b. 计算这 10 个均值的标准差, 它可以被用作均值的标准误差。

c. 用标准误差乘以 1.96, 并用这 10 个均值的每一个构造一个 95% 置信区间。

d. 像图 6.3 那样画出这样一组置信区间。

e. 有多少个置信区间包含总体均值?

6.39 在 1992 共和党大会后, 有 4 个不同的民意调查报告了有关布什总统的支持率。CNN/USA Today 预测的支持率是 $42\% \pm 4\%$, Newsweek 的是 39 ± 4 , Los Angeles Times 的是 41 ± 3 , Washington Post 的是 40 ± 4 。

a. 为什么 4 个民意测验得到 4 个互不相同的结果并不令人感到意外?

b. 为什么其中一个民意测验的抽样误差是 ± 3 而其它的是 ± 4 ?

c. 用每一个调查结果构造一个未知的总体百分数的 95% 置信区间, 并把它们画到一个图里, 以便比较。

d. 假设这 4 个区间没有不寻常之处而且都包含总体百分数的真值。总体百分数的可能值的范围是什么。

e. 在每个民意测验中大约各有多少个人被访?

f. 合并这 4 个民意测验并表明布什在被访者中的得票率是 40%, 抽样误差是 ± 2 。

g. 请用 f 中的数据构造一个 95% 置信区间。

b. f 中的区间和 d 中所得区间比较起来如何?

6.40 根据人口普查局提供的数据, 在 1990 年时全国平均每套住房有 2.6 间卧室, 标准差是 $\sigma = 0.9$ 。而在芝加哥郊区随机抽取的 100 套住房中平均卧室拥有量是 3.1 间/套。

a. 请给出该社区平均每套住房中含有的卧室数的 95% 置信区间。

b. 是否该郊区的卧室数看来多于全国平均数?

6.41 在一个大商行的 49 个雇员的样本中, 这些雇员一年中平均有 7.0 天在生病, 其标准差为 2.5 天。

a. 请给出整个公司的雇员一年中平均生病天数的 95% 置信区间。

b. 你怎样解释这个区间。

6.42 “如果有两种职业前途供你选择, (1) 你可以自由支配所有的工作时间, 并能很好地照顾

家庭,但职位提升较慢;(2)工作时间非常的严格,较少时间照顾家庭,但提升较快。你愿意选择哪一种?”关于这个问题 1989 年的一项调查表明有 74%的男性和 82%的女性选择第一种(来源:Juliet B. Schar, *The Overworked American: The Unexpected Decline of Leisure*, New York: Basic Books, 1991, p. 148)。假设这个百分比的抽样误差是 ± 3 。

- 请给出男性总体中选择第一种工作的人所占百分比的置信区间。
- 请给出女性总体中选择第一种工作的人所占百分比的置信区间。
- 从这两个区间来看你是否认为这两个总体百分比相等?
- 从结果来看,是否可以说在整个总体中无论男性还是女性选择第一种工作的人实际上占少数?请解释。

6.43 研究 21 个游泳运动员时发现他们的平均年龄是 12.3 岁,均值的标准差是 0.3 岁(来源:G. E. Theintz, H. Boward, U. Weiss, and P. C. Suzonenko, "Evidence for a reduction of growth potential in adolescent female gymnasts," *Journal of Pediatrics*, vol. 122(1993), no. 2, pp. 306 - 313)。

- 请给出游泳运动员总体平均年龄的 95%置信区间。
- 这个区间告诉了你哪些包含该样本的总体的信息?

6.44 几个老病人总是抱怨带状疱疹后神经痛,其中六个男性从确诊到治疗的平均间隔是 30.5 个月,而 12 个女性的平均间隔则是 37.9 个月,这两个等待时间相差了 7.4 个月。男性组的标准差是 17.5 个月而女性组的是 30.8 个月(来源:P. R. Layman, E. Agyras, and C. J. Glynn, "Iontophoresis of vincristine versus saline in post - herpetic neuralgia: A controlled trial," *Pain*, vol. 25(1986), pp. 165 - 170 reported in W. W. Piegorsch, "Complementary log regression for generalized linear models," *The American Statistician*, vol. 46(1992), pp. 94 - 99)。

- 请证明总的标准差是 27.4 个月。
- 请给出这两个总体均值之差的 95%置信区间。
- 是否这两个总体均值可能没有差别?

6.45 在 1989 年,全国健康统计中心对 12 到 18 岁的青年进行了一项买什么牌的烟的调查。41 名非洲裔美国人中的 61%抽 Newport 牌的烟,而 807 名白人中只有 5%的人抽此烟,两者结果相差 $61 - 5 = 56$ 。

- 请证明这两个总体百分比差异的 95%置信区间是 41 到 71。
- 你如何理解这个置信区间?
- 从直观上来看,这两个总体的百分比是相同还是不同?

6.46 用盟军估计二战中德国坦克数的办法来估计一种假想出来的二战新式战斗机的数目。每架飞机都被用一个号码标出它在生产线上的位置。假设被打落飞机的最高数目是 100 架。飞机的样本是 20。请你估计总共生产了多少架飞机?

6.47 Wyatt 公司对 531 家大公司进行调查,发现被调查的商业经理中有 61%的人期望缩小公司规模可以提高对顾客的服务质量,但有 33%的人认为提高服务质量已经由于裁减职员而出现了(来源:R. Reich, "Companies are cutting their hearts out," *The New York Time Magazine*, December 19, 1993, p. 54)。

- 请用快速近似公式 6.2 给出两个相应总体百分比的置信区间。
- 从这些信息中你能对缩小规模得出什么结论?

6.48 用某种准则来为 117 名少年与其父母的亲近程度打分。其中 71 名少年的父亲酗酒,他们的平均接触程度是 78 分,标准差是 25。而不酗酒父亲的孩子的平均亲近程度是 91 分,标准差是 22(来源:Timothy Cavell et al., "Perception of attachment and the adjustment of

adolescents with alcoholic fathers," *Journal of Family Psychology*, vol. 7 (1993), pp. 204-212).

- a. 请分别给出这两个总体均值的置信区间。
 - b. 你是否认为那些酗酒父亲的孩子与父母的亲近程度不如不酗酒父亲的孩子?
- 6 49 在 Gallup 的一项有 502 个回答者的调查中,有 56% 的人说他们是习惯早起的人,44% 的人说他们是“夜猫子”(来源:USA Today, December 13, 1993, p. 1A)。请用快速近似公式 (6.2) 寻找 a, b, c, 中(仔细阅读)的百分比的置信区间。然后继续做 d, e, f。
- a. 在早起者中有 53% 的人认为早起者精力大于大多数人,而持同样观点的夜猫子占 39%。
 - b. 45% 的早起者认为早起者锻炼的比别人多,有 37% 的夜猫子赞同此观点。
 - c. 74% 的早起者认为早起者过着一种更积极的生活,夜猫子们中持此看法的占了 64%。
 - d. 以上二个区间提供的信息好像说明了早起者与夜猫子对于早起的人的理解有所不同。你的观点呢?
 - e. 报告表明是 Quaker 麦片公司发起的这项研究,它的边际误差是 4.4%。这个边际误差应该如何与你的结果进行比较?
 - f. 用 4.4% 这个边际误差代替你自己的计算是否改变对 e 的回答?
- 6 50 选取了 2112 对夫妇,用两种度量来他们对婚姻状况的满意程度进行研究。一种婚姻满意度是描述每个人在婚姻中的快乐程度,另一种理想扭曲度是用来度量每个人评估婚姻的倾向。其结果由表 6.1 给出。为了回答问题,请构造均值之差的量信区间。

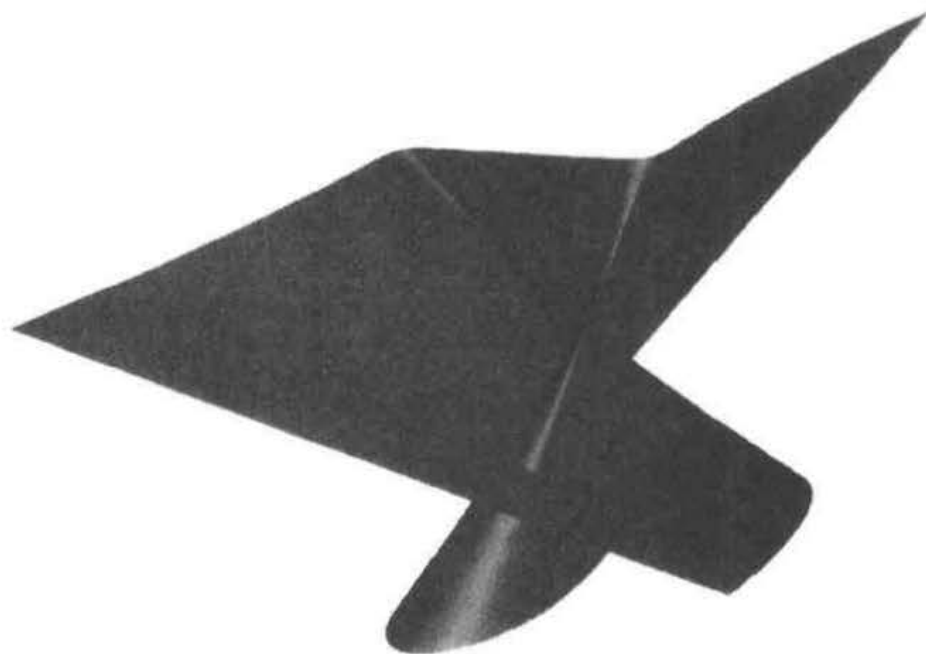
表 6.1 习题 6.50 的数据

准则	男性		女性	
	均值	标准差	均值	标准差
婚姻满意度	31.6	8.7	30.0	9.8
理想扭曲度	16.7	5.1	14.0	5.5

来源: Blaine J. Flowers and David H. Olson, "ENRICH marital satisfaction scale: A brief research and clinical tool," *Journal of Family Psychology*, vol. 7 (1993), pp. 176-185.

- a. 好像男性比女性对婚姻的状况更满意,是不是?
 - b. 你是否认为男性对比女性有更多的扭曲论者?
- 6 51 第五章中我们报告了我们一个初等统计课的 25 个学生的实验结果;该实验为每人将硬币旋转 10 次并记录正面出现的次数。正面出现的平均数为 $\bar{y} = 3.96$, 标准差是 1.74, 则均值的标准误差就是 $1.74/\sqrt{25} = 0.35$ 。
- a. 请用这些数据构造 10 次实验中正面出现次数的总体均值的 95% 置信区间。
 - b. 请用这个置信区间推测:在硬币旋转过程中,正、反面出现的概率是否相等。即在 10 次实验中是否能期望出现 5 次正面 5 次反面?
- 6 52 在习题 6.51 中,25 个学生共旋转硬币 250 次。累计出现正面 97 次,反面 153 次,因而正面出现的比率是 $97/250 = 0.39$ 。请为旋转一枚硬币正面出现的概率构造一个 95% 置信区间。

C H A P T E R 7



- 7.1 作为一个问题的假设
 - 7.2 怎样回答零假设所提出的问题
 - 7.3 显著水平
 - 7.4 总体比例检验
 - 7.5 两个总体比例的差异
 - 7.6 假设检验与构造置信区间
 - 7.7 统计显著与实际显著
 - 7.8 应用:何时拒绝零假设
 - 7.9 小结
-

作出结论：假设检验



是否法国人的地理知识比其它国家的人更丰富？而美国人正好相反？

1988年7月28日的纽约时报上刊登了一篇有关人们地理知识的文章。这篇文章中描述了一个由 The National Geographic Society 委托 Gallup 公司所做的研究结果。研究者们从一些国家抽取许多成年人并请他们鉴别在一个地图上的 16 个地方（包括 13 个国家、中非、波斯湾和太平洋）；然后把每个人答对的个数加起来。

四个国家的样本中答对的个数的均值为

美国 6.9

墨西哥 8.2

大不列颠 9.0

法国 9.2

平均来看，法国的回答者有可能在地图上找到的地方数比其它三个国家的人要多。这篇文章称“从统计显著性方面考虑，得分相差至少应在 0.6 以上才算有差异。”“统计显著性”是什么意思？

法国与大不列颠均值的差异是 0.2，这个小小的不同似乎不能说明任何问题，同时它也小于 0.6 这个统计显著性的界限。另一方面大不列颠和美国均值差异是 2.1，这在统计上已经大到显著了；此因 2.1 大于 0.6。同时墨西哥和大不列颠，墨西哥和法国之间的均值之差都达到了统计显著性的标准。

当对这些样本间的不同感兴趣时，我们不仅想要知道不同国家的样本均值之差，而且还想知道这些国家所有成年人总体均值之间是否有差异。如果我们敢下结论说两个国家的总体均值不同，则它们样本均值之间的不同就是统计显著的。

估计的主要任务是找参数值等于几;假设检验的兴趣主要是看参数的值是否等于某个特别感兴趣的值。

这个问题的一种解法要用到第六章中的思想,即估计总体的均值并计算它们的差异。在这一章我们将用另一种方法来解决这个问题,它叫做假设检验。这两种解法感兴趣的都是总体参数,如百分比、百分比间的差异、均值、均值间的差异等等。只不过,在假设检验中我们注意的焦点是某个特殊值,并想知道参数是否等于这个特殊值。在那个有关地理的问题中,参数是总体均值。为了说明本章的主旨,我们探讨墨西哥的总体均值是否等于美国的总体均值。即使我们观察到了样本均值之间的 1.3 的差别,我们仍想要知道墨西哥与美国的总体均值的差是否等于零。样本均值的不同有可能仅仅归于随机性。

7.1 作为一个问题的假设

我们以一个问题来开始假设检验:来自墨西哥和美国的总体均值差异是否为零?在两个样本中,均值差为 $8.2 - 6.9 = 1.3$ 。即平均起来,每个墨西哥人能比美国人在地图上多找出 1.3 个地方。当然,即使两国的总体均值没有差异我们也不能指望两个样本均值相同。因为两个随机样本都会受抽样变化的影响。但是这个变化所能造成的差异也许不足以大到可以解释 1.3 这样的差距。

零假设

要判断 1.3 这个值是否超出了样本变化所能造成的差异的范围,我们先要问一问在总体均值相等的情况下,样本均值会发生什么情况,即是否两个均值的差等于零。在统计中这样的问题被称为**零假设**(null hypothesis)。零假设总是通过一个或多个参数来表示的;而且它设定这(些)参数等于某个特殊值。在目前这个例子中,零假设问这两个总体均值之差是否等于 0。

在更正式的统计语言中是用一个等式来表示零假设的。如果用 μ_M 来标记墨西哥的总体均值,用 μ_{US} 来标记美国的总体均值(记住希腊字母 μ 代表总体均值),那么在这个有关地理的问题中零假设可以表示成统计等式:

$$H_0: \mu_M - \mu_{US} = 0$$

字母 H 代表假设,下标 0 表明是零假设。之所以用**零**这个字来修饰假设,其原因是假设的内容总是没有差异或没有改变,或变量间没有关系等等。

一个统计上的**零假设**提出一个参数是否等于某个特殊值的问题,形式上零假设则被写成:

$$H_0: \text{参数} = \text{值}$$

尽管零假设陈述的是两个总体的均值之差等于 0,却并不等于这一定是事实。它仅仅是个假设而已。它仅仅是提出两个总体的均值之差是否等于 0 的问题的简单方式。大多数情况

下,收集数据是为了表明两个均值不相等。如果零假设是错的,那就意味着两国的总体均值是有差异的。零假设所提问题的答案是以样本数据为基础的。

值得强调的是零假设总是一个与总体参数有关的问题,所以总包含希腊字母。关于样本统计量如样本均值 \bar{x} 或样本均值之差的零假设是没有意义的,因为样本统计量是已知的,当然能说出它们是否相等。例如:上面例子中的样本均值之差是 1.3,它显然不等于 0。但是这并不等于总体均值之间就一定不同,所以我们要问总体均值的差是多少?

备择假设

对于内容是“无区别”的零假设,其逻辑上的反面假设是“两个参数间有区别”。这种反面假设称为**备择假设**(alternative hypothesis)。前面例子的备择假设应表述为“两个总体均值之间的差不等于 0”:

$$H_a: \mu_M - \mu_{US} \neq 0$$

因此,当零假设所提问题的答案被否定时,备择假设的答案就是肯定的。如果两个均值不相等,则它们必有差别;即在零假设中的对差别的否定和对零假设的回答中的否定互相抵消了。如果样本数据能证明对于零假设提出的问题应该否定,那么我们就**拒绝(reject)**零假设而倾向于备择假设。

回答问题时的错误

零假设提出的问题是“是”或“不是”就能回答的问题。墨西哥的总体均值与美国的总体均值的差异是否等于 0? 或者“是”或者“不是”。这个问题的答案是由样本数据所提供的信息决定的。既然样本所携带的信息来自样本而不是总体,其信息量就会受到限制,就有可能提供错误答案。

假设检验问题有些像陪审团判断被告有罪还是无罪的问题。如果被告的确是清白的但我们判他有罪,那我们就犯了一个错误。反过来,如果被告有罪而被判无罪的话,我们又犯了另一类错误。

对于零假设检验所提的问题有两种可能:

1. 两个总体间均值差异为 0, 正确答案为“是”。
2. 两个总体间均值差异不为 0, 正确答案为“不是”。

第一类错误
(α 错误): 在假设检验中拒绝了本来是正确的零假设。

如果零假设是对的 如果零假设是正确的即两个总体均值的差异的确是 0, 那么对零假设的正确答案为“是”。如果我们回答“是”, 那就答对了; 但如果我们回答“不是”, 那就犯了错误。我们管这类错误叫做 α 错误或是**第一类错误**(type I error), 即当零假设正确时我们确认为它错了。

第二类错误 (β 错误): 在假设检验中没有拒绝本来是错误的零假设。

如果零假设是错的 如果零假设是错的,两国的均值差异不是 0,那对零假设的正确答案为“不是”。如果我们回答“是”并认为两个均值差异就是 0,那我们就犯了另一类错误,它被称为 β 错误或第二类错误 (type II error)。

停下来想一想 7.1

一个人因为杀妻而受审。他实际上是有罪的,但陪审团确认为他无罪。这里的零假设是:一个人是无罪的除非你能在一些怀疑之外证明他有罪。则在此案中陪审团犯的是第一类错误还是第二类错误?你能够设想出一个陪审团犯了与以上错误不同的另一类错误的案例吗?这两种错误中的哪一种是我们法律系统更情愿容忍的?

7.2 怎样回答零假设所提出的问题

零假设所提问题的答案是以样本数据为基础做出的。如果数据表现出支持零假设,它就不会被拒绝;如果数据和零假设不一致,那它就会被拒绝。例如:如果零假设说两个总体均值差异是 0,而实际上样本均值差和零相距甚远,零假设就会被拒绝。

我们确定样本数据是否和零假设不一致的办法是先问一问:在零假设正确的情况下,是否能期望得到现在所得的这组数,如果美国和墨西哥的总体均值是相等的,我们能否期望得到 1.3 那么大的样本均值差?换句话说就是:因为零假设说总体均值差异是 0,如果零假设是对的,样本均值的差异也应该接近于 0。

如果样本真的来自均值相等的两个总体,那样本均值差异达到 1.3 的可能性有多大?如果总体均值的差为零,那么样本均值差为 1.3 是否属于样本均值差中的一个不寻常的集合?换句话说就是在总体均值差异是零的情况下,样本均值差大于等于 1.3 的概率有多大?

概率: p -值

p -值是当零假设正确时,得到所观测的数据或更极端的数据的概率。

为了确定像 1.3 这么大的差异是否属于一类不常见的数据集合,我们计算当总体差别为零时,得到一个大于等于 1.3 的样本均值差的概率。这个概率称为 **p -值** (p -value)。当 p -值如此之小,以至于几乎不可能在零假设正确时出现目前的观测数据时,我们就拒绝零假设。 p -值越小,拒绝零假设的理由就越充分。但什么才算“小”呢?概率是 0 到 1 之间的一个数,因此小概率就应该是接近 0 的一个数。著名的英国统计学家 Ronald Fisher 把 20 分之 1 作为标准,这也就是 0.05,从此 0.05 或者比 0.05 小的概率都被认为是小的。Fisher 没有任何深奥的理由解释他为什么选择 0.05,只是说是他忽然想起来的。

我们应该在 p -值多小时才拒绝零假设的问题原则上来说应该由错误拒绝零假设的后果来决定,但是这个后果通常又是难以确定的。

当 Fisher 设定 0.05 时,他考虑的是观测到的样本统计量大于或小于零假设中参数值的概率。这也就是说显著水平概率的一半(0.025)在参数值的左边,另一半(0.025)在参数值的右边。

如果 p -值是观测样本或者比它更极端的数据的概率,则这个 p -值是单边的,其为小概率的标准就减小到 0.025 了。当 p -值是由 t 变量和 z 变量计算得到的时候,会出现这种情况。有时人们用双边 p -值,有时他们又用单边的。

警告! 请注意 p -值!

人们有时错误以为 p -值与零假设的对错的概率有关,但这是不可能的。当我们抛硬币时,有时正面向上,有时反面向上,因此我们可以讨论硬币落地时某面向上的概率。这个概率告诉我们在多次抛硬币的过程中某一面出现的经常程度。

但是我们不能讨论一个特别的零假设对还是错的概率,一个零假设不是有时候对有时候错。一个零假设或者是对的或者是错的,它永远只是这二者之一。

假设你的名字是 David,我们总不能说你的名字是 David 的概率是 0.04 吧!如果是那样就意味着我们问你 100 次“你叫什么?”,而你只有 4 次回答自己叫 David,另外 96 次都回答的是别的名。这与零假设是否正确是一类问题。

实际上, p -值指的是关于数据的概率。 p -值告诉我们,在某总体的许多样本中,某一类数据出现的经常程度。假设你们班有 10 个学生,其中两个名字是 David。从中抽样,每次抽一个学生作为样本并用该生的名字给这个样本命名。则名为 David 的样本的 p -值是 0.2,因为如果抽了很多样本,其中叫这个名字的样本占了 20%。

当数据导致拒绝零假设时,这个经验结果是统计显著的。换句话说就是:如果 p -值很小,经验结果就是统计显著的。

如果零假设被拒绝,就可以说样本结果是统计显著的。

在地理的例子中,零假设是墨西哥和美国人正确识别的地方数相等,而观测到的两个样本均值差异是 1.3。因为在两个总体均值相等的前提下两个样本均值差大于等于 1.3 的概率小于 0.025。所以可以认为 1.3 这个差异是统计显著的。实际上,对于 0.6 的样本均值差, p -值恰好是 0.025。

假设检验的机制

为了求 p -值,统计理论指出要把观测到的 1.3 这个样本均值之差变换成标准得分。两样本均值之差的标准得分(standard score)是统计 t 变量的一个值。对于当前的数据,样本均值之差 $\bar{x}_M - \bar{x}_{US} = 1.3$ 变成 $t = 4.25$ 。又因为美国的样本中包含了 1600 个观测,墨西哥的样本有 1200 个观测,这使得 t 变量的自由度非常大,以至于可以认为它等于标准正态分布—— z 变量。(t 变量及自由度在第五章讨论。)

把变量 $(\bar{x}_M - \bar{x}_{US})$ 变换成另一个变量(t)类似于把温度从华氏转换成摄氏。 t 的值可以通

过本章末的公式(7.2),用样本均值、标准差以及样本观测个数计算出来。具体计算可以用纸和笔或者用计算机统计软件。

把两样本均值之差变换成 t 变量的值的过程由图 7.1 给出。这个图表示了两种尺度的关联:观察的差为 0 时 $t = 0.00$, 观察的差为 1.3 时 $t = 4.25$ 。 t 值大于 4.25 的概率也就是观察均值之差大于 1.3 的概率。

任何超出 +2 或 -2 以外的标准得分都是不寻常的大,所以 4.25 为是一个极不寻常的大 t 值。得到样本均值差大于等于 1.3 的概率是不能被直接计算的,但 t 大于等于 4.25 的概率是能够被算出的。许多统计计算机软件包能直接计算 t 的概率,或者也能用统计表来找到变量大于等于任何值的 t 的概率。

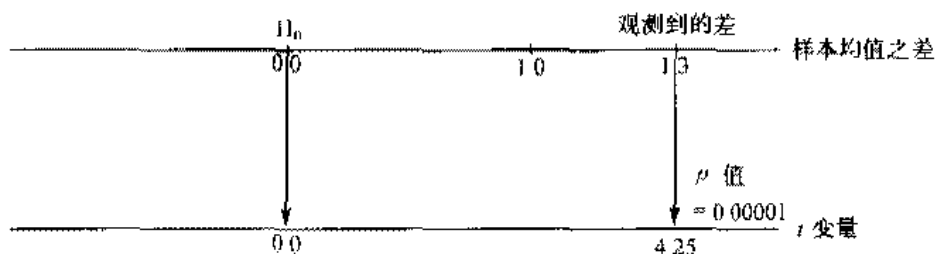


图 7.1 两个样本均值之差与 t 变量值的对应关系。

对于这个例子,对于一个观测数大于 2000 的样本来说, t 值大于等于 4.25 的概率是 0.00001。因此样本均值的差大于等于 1.3 的概率也是 0.00001。换句话说就是如果两总体均值之差等于零,从均值相等的总体中抽取 10000 个样本才有可能碰到一次样本均值相差在 1.3 以上的情况;即在总体均值相等的情况下样本均值差异有这么大是一件很少见的事情。

第五章曾经讨论过, t 变量会受样本容量的影响,而且,当样本中的观测个数不同时,即使 t 值相同得到的概率也不一样。统计学家们把观测个数的概念抽象为自由度,并把它记为 df 或 df 。

在考虑两样本均值之差时, t 的自由度等于两个样本中总的观测个数减去 2。例如在上面的例子中,总观测个数是 2800 个,这个数是如此之大,以至可以用标准正态 z 变量来代替 t 值。

拒绝或不拒绝零假设

基于两个总体均值相等的假设,得到两个样本均值之差大于等于 1.3 的概率是 0.00001 或者是十万分之一。于是如果总体均值相等,观测到的两样本均值差 1.3 属于非常不可能的均值的集合,有一个非常小的发生概率。

对此情况有两种解释:一种是零假设是正确的,观测到的数据恰好是不常发生的那一类;另一种是数据倒是常见的那一类,只是零假设是错的。(这和第五章中研究概率时的讨论是类似的,那也是有关假设检验的讨论,只不过这里用词不同罢了。)因为当总体均值相等时样本均值有这么大的差的概率是 0.00001,所以我们选择第二种可能性,即认为导致这个小概率出现的假设——两总体均值相等是错的。我们拒绝了两总体均值相等的零假设,而认为两个总

体均值差异不是零。这样我们就有可能以样本数据为基础得到了总体参数与某个特殊值之间关系的结论,也因此了解了总体的概貌。

上面例子的分析过程能用如下术语表示:为了检验 $H_0: \mu_M - \mu_{US} = 0$, 从墨西哥抽了一个容量是 $n_1 = 1200$ 的样本, 从美国抽了一个容量是 $n_2 = 1600$ 的样本, 经计算有

$$\bar{y}_M - \bar{y}_{US} = 1.3 \quad (t = 4.25, \quad df = 2798, \quad p = 0.00001)$$

用平常的话来说:为了检验墨西哥和美国两国的总体均值差等于零的零假设, 从墨西哥抽取了一个含有 1200 个观测的样本, 从美国抽取了一个含有 1600 个观测的样本。经计算两样本均值差异是 1.3, 换算成 t 值是 4.25 并有相当大的自由度, 由此可知两样本均值差异大于等于 1.3 的概率是十万分之一。

停下来想一想 7.2

把以下用术语描述的过程转用平常的话来叙述: 检验 $H_0: \mu = 0.50$, 收集了有 22 个观测的一个样本, 发现 $\bar{x} = 3.5$ ($t = -2.50$, $df = 21$, $p = 0.01$)。

因果关系: 过犹不及

在没有进一步的证据时, 我们不能再多走一步去断言国家之间的文化差别是在世界地图上查找地方能力差别的原因。这是一个比统计显著更强的命题, 而我们又没有足够的证据来支持这一结论。也许对于其它诸如教育大纲等因素的了解可以帮助我们进一步解释美国与墨西哥在观测中的不同。

一些统计理论和计算游戏

墨西哥和美国均值差的 p -值是 0.00001。即在总体均值相等的情况下, 在 100000 次抽样中只有一次得到的样本均值差异大于等于 1.3。但该研究不能真的有 100000 个样本; 实际只有一个样本。而抽 100000 个样本的想法似乎太遥远了。

一些合适的软件可以从均值相同的两个总体中产生许多不同的样本。我们实际上已经这样作的: 我们另外选了 99 对样本, 并计算每一对的样本均值之差。这样我们在原先的观察的样本均值差之外从均值相等的总体中得到了 99 个差异值。用这些样本差 (共 100 个) 做直方图, 得到图 7.2 (从总体中选的样本服从均值是 8.6 方差是 6.0 的正态分布)。

图中显示, 那一个观测到的样本均值差在图右边很偏的地方, 而其它差值都集中在 -0.5 到 0.5 之间。它们都聚集在 0 周围并不奇怪, 因为它们本来就是从已知均值相等的总体中抽取的。这个直方图的一个显著特点就是实际观测到的 1.3 这个样本差异自己孤独地躲在一旁; 而计算机产生的样本差异没有一个接近 1.3 的。如果我们抽了 100000 个样本而不是 100 个, 那可能会有多几个比较极端的差值。但即使只有 100 个样本我们也可以清楚地看到 1.3 这个观测到的差值有多么不寻常, 换句话说就是它的 p -值非常小。

把图 7.2 那样的直方图直观化有助于理解小的 p -值。图 7.2 明显地提示什么叫做小的

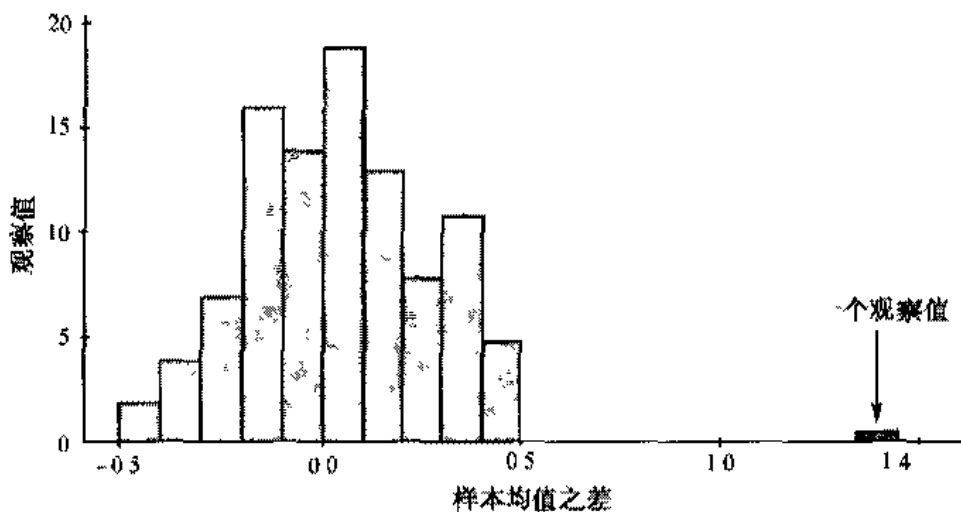


图 7.2 99 对由计算机生成均样本均值之差和 1 个观察值之差组成的直方图。

p -值,并显示出“不寻常的样本”的意思。在这个直方图中,差异 1.3 不属于其它样本差值的聚集范围。在这种情况下,我们应该下结论说:观测到的样本不像其它样本那样来自均值相等的两个总体。这也就是说来自墨西哥的总体与来自美国的总体具有不同的均值。根据样本提供的信息我们拒绝了零假设。图 7.3 阐明了关于地理的这个例子中所涵盖的统计理论。

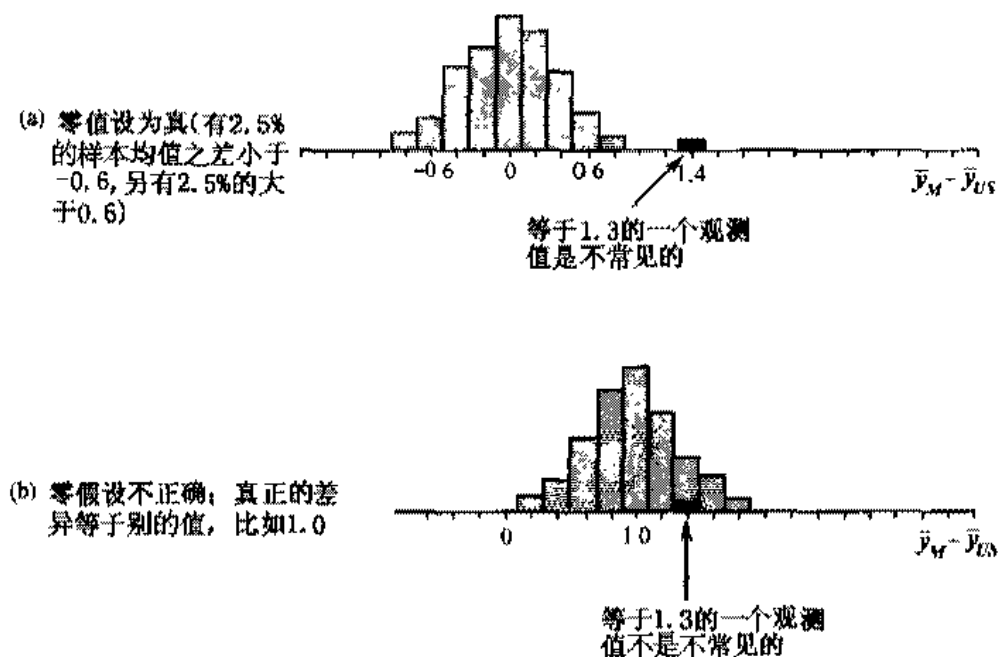


图 7.3 来自许多不同样本的样本统计量的分布。零假设为真时是(a),为假时是(b)。

图 7.3 显示了在地理例子后面的统计理论。图 7.3a 是零假设正确时的情况。在假设中,参数(总体均值、总体百分数、总体均值间的差异——这里指的是两个总体间的差异)等于某个在零假设(两个总体均值相等)中给定的值。如果我们现在抽取许多不同的样本,则相应的样

本统计量(样本均值、百分比等等)应聚集在总体参数值的周围。图 7.3a 显示了观测的样本统计量在参数周围的聚集情况。因为样本是互不相同的,所以它们的结果也不尽相同。但请注意图 7.3a 中观测的样本统计量远离与其它样本的结果。

图 7.3b 表明了其它可能性,即当总体参数等于别的值而不是零假设中的值。同样,来自许多不同样本的结果聚集在真值的附近,而观测到的样本统计量的值与其它样本的结果混在了一起。

在任何实际调查中我们只有一个样本,而且我们也不知道零假设是否正确。问题是我们的样本结果是属于图 7.3a 所示的那一类呢?还是属于图 7.3b 所示的那一类。如果我们能说出它属于哪一类,我们就能断定零假设是否正确了。我们要做的就是假设它是图 7.3a 那一类,然后计算样本结果的 p -值。任意属于图 7.3b 的样本结果都会有很小的 p -值,而基于这个小的 p -值我们就能正确地拒绝零假设。如果观测真的属于 7.3a,那它就是一个很不常见的值而且 p -值非常小。可是如果它属于 7.3b,则它就不那么不寻常了。这样,假设检验就是在回答一个问题:一个观测值属于哪一群样本统计量?其答案是根据 p -值大小确定的。

停下来想一想 7.3

一家报纸称澳大利亚人比美国人喝更多的啤酒。它的论据是下面的事实:在 12 月中澳大利亚平均每人喝掉了 2 升啤酒而美国人平均只喝了 1.7 升。在你同意澳大利亚人比美国佬喝更多的啤酒之前,你应该先弄清哪些问题?

7.3 显著水平

一个检验的显著水平 α 是抽样所得的数据拒绝了本来是正确的零假设的概率。

在统计软件能方便地计算出 p -值以前,假设检验是用另一种稍有不同的方法来实现的。不是在数据收集完毕之后计算 p -值,而是在收集数据以前就已经确定好的小概率来构造一个区间。当样本数据落入这个区间时就拒绝零假设。这个小概率称为检验的显著水平(significant level),通常选 0.05。显著水平是 0.05 的意思是:在零假设正确的情况下进行 100 次抽样,会有 5 次错误地拒绝了零假设。显著水平 0.05 通常认为是一个合理的风险。

图 7.4 显示了显著水平在地理知识的那个问题中是如何被应用的。那个问题中的零假设是:两个总体均值差异为 0.00。这个值标在了水平线的正中。为了检验这个假设,计算两个样本均值的之差,并在轴标为“样本统计量”的水平线上的某处标出。

如果零假设是正确的 当零假设正确时,样本统计量的值常落在零假设提到的那个值的附近。如果总体均值差异是零,则两个样本均值的差异应该接近于零。但随机抽取样本偶尔也会产生远离总体值的统计量。用恰当的公式可以找到两个样本统计量的值使得有 2.5% 的可能的样本统计量大于其中一个值;另有 2.5% 的可能的样本统计量小于另一个值。这些都是样本统计量的极端的和很少出现的值。(注意:图中的显著水平 0.05 被平分为两个部分。

一部分在中点以左,一部分在中点以右。)在地理的例子中,按照对该例的计算,任何比 ± 0.6 更极端的值都会导致拒绝零假设。如果零假设正确,则样本均值之差将会有 2.5% 的机会次小于等于 -0.6 , 2.5% 的机会大于等于 $+0.6$ 。这两个数 -0.6 和 $+0.6$ 被叫做临界值(critical values)。

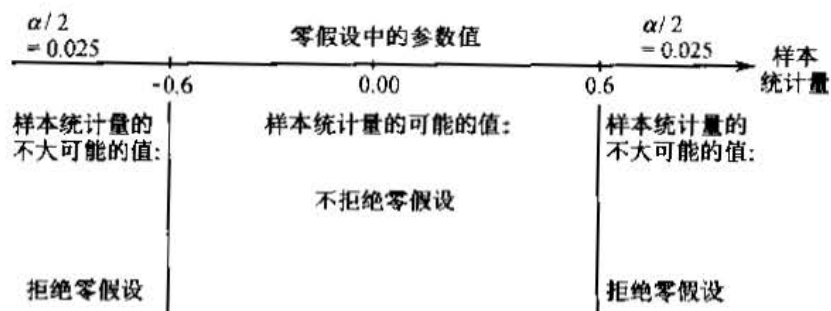


图 7.4 显著水平是 0.05 的关于两样本均值差异的假设检验。

对于双边检验,样本统计量的临界值是两个值。当假设正确时,来自不同样本的样本统计量中只有 5% 的值超越了这两个值之一。

如果零假设是错误的 如果来自墨西哥和美国的两个总体差异不是零,则零假设就是错误的。这时所谓的样本统计量的极端值并不是少见的,而是经常发生的。因此当这些所谓的极端值之一被观测到了,我们就认为零假设是错误的并拒绝它。既然观测到墨西哥和美国的样本均值之差为 1.3,大于 0.6 这个上限;就可以说观测到的这个差值是统计显著的,并拒绝均值相等的零假设。于是,这个问题的结论是:这次测验表明墨西哥人在这个测验上的地理知识比美国人要好。

较小的显著水平 如果你用了一个比 0.05 还要小的显著水平,比如 0.01,则这两个临界值应该分别向相反的两个方向移动,其绝对值应该比 0.6 要大。在显著水平是 0.01 时,临界值是 ± 0.79 。而因为 1.3 这个观测值仍然在该区域之外,所以在这个显著水平上我们还是拒绝零假设。事实上对于任何小于 1.3 的值域,我们都会拒绝零假设。这意味着我们可以在大大小于 0.05 的显著水平拒绝零假设。能拒绝零假设的最小显著水平就是该数据的(双边) p -值。

双边与单边的假设检验 备择假设是:来自墨西哥和美国的总体均值差异不是零。不是零的意思是可以大于零也可以小于零。正因为这个原因,它称为双边备择假设。

有时单边假设检验也是有用的。如果我们对于两国人如何学习地理和在地图上找地点的测验中的得分情况有所了解,同时已经知道墨西哥的人均水平不低于美国的人均水平,则墨西哥人在测验中表现或者一样好或者更好。有了这些附加知识,备择假设可以被改为

$$H_0: \mu_M - \mu_{US} > 0$$

既然我们已知差值不会是负的,那唯一的对于差值是零(零假设)的备选就是它是正值。

这种备择假设为零假设定义了一个新的拒绝域:样本均值差为较大的正数。任何负的差值都被认为是抽样误差,可以被忽略。剩下的问题是样本均值差要多大才能拒绝零假设?

因为整个显著水平 0.05 都集中在了 0 的右边,所以当计算出来的 t 值大于等于 1.64 时我们就可以拒绝零假设了。与这个 t 得分相对的两样本均值之差为 0.5,所以任何大于 0.5 的差异都是统计显著的。对于双边假设检验,均值的差则需要超过 ± 0.6 之外就拒绝零假设。

图 7.5 中用图示说明了双边和单边假设检验。图中分别显示了何时具有双边备择假设的零假设被拒绝;何时具有单边备择假设的零假设被拒绝。这两种情况的显著性水平都等于 0.05。

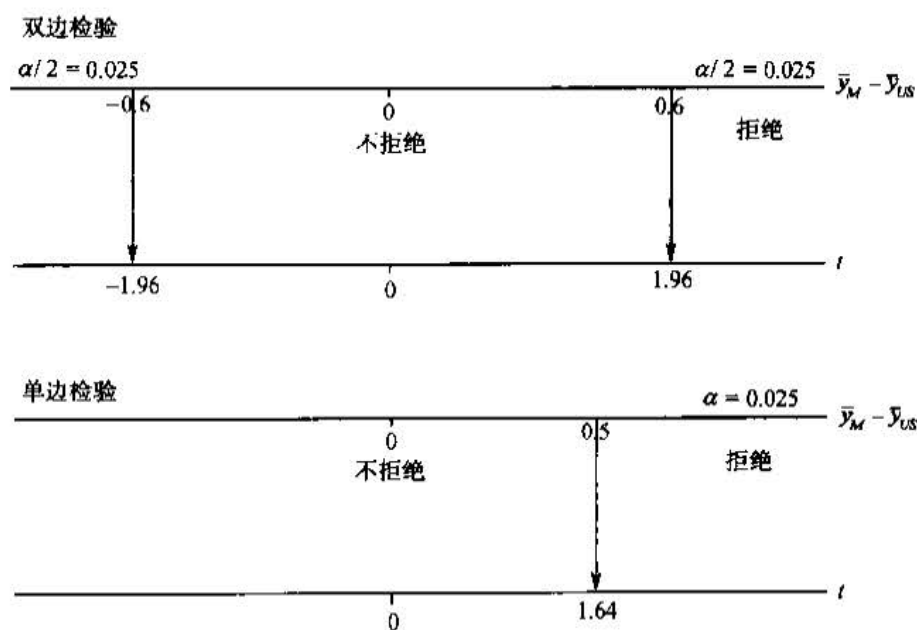


图 7.5 双边假设检验与单边假设检验的拒绝域。

7.4 总体比例检验

调查或民意测验常常是为了弄清楚:我们中有多少人持这种或那种观点;我们中有多少人做这种或那种事;我们中有多少人知道这种或那种事等。报纸报道过:1995 年有 22% 的经理担心被解雇,而 1993 年时这个数字是 6%;61% 的青少年堕胎者告诉了父母双亲;60% 的大学四年级学生不知道哥伦布何时在美洲登陆等等。选举前的民意测验经常提供一些细节,诸如候选人在全体选民中的地位,某位候选人的支持率比另一位高几个百分点等等。

对于一位政治候选人来说,在两人竞选时 50% 是一个相当重要的标志。得票率超过 50% 的将在选举中获胜,而低于 50% 的将被淘汰出局。所以选举前的民意测验可以给候选人如何与对手竞争一个重要的概念。为此,候选人的统计专家能检验零假设:支持率等于 50%。形式上,这能写成

$$H_0: \pi = 0.50$$

这里的 π 是指这位候选人的民众支持率(我们都习惯于用 π 代表 3.14..., 所以把它用在这里

代表一个 0 到 1 之间的数时是有点儿容易混淆,但是我们前面已经说过,习惯上总体参数都用希腊字母来标记,没有别的希腊字母来标记一个概率或总体比例)。

零假设也可以被写成相应的百分数,即用这个比例乘以 100%。如果我们用 Π 来代表某位候选人的总体支持率,则零假设可以写成:

$$H_0: \Pi = 50\%$$

计算百分比比计算比例要容易,而且在研究分类型变量时总是用百分比来报告结果。

检验零假设的第一步是要确定样本容量。样本容量越大,样本百分比就越接近总体百分比。包含大约 1200 个响应者的样本经常用在新闻媒体对选举和其它事件所做的民意测验中。

让我们来看一看如果双方选民势均力敌的零假设正确,会发生什么现象?如果零假设正确,我们用含有 1200 个响应的样本算出的百分比中将会有 95% 落在区间 47% 到 53% 之间。这两个数是用来检验零假设的临界值,而且只要样本容量不变,这个临界值也不变。也许你已经在电视或报纸上注意到:在许多重要的民意测验中常用的抽样误差是 $\pm 3\%$ 。

如果观测到的样本值落在 47 到 53 之外,我们就以 0.05 的显著水平拒绝零假设。比如,如果在样本中有 55% 的选民称他们支持某位候选人,我们就有理由相信选票将不会被平均分配,而是对某位候选人有利,因为观测值大于 53% 这个上限。

我们当然也可以取实际的样本百分比 55% 并计算其 p -值。如果 p -值很小,我们就拒绝零假设。当零假设等于 50 时,对于一个含有 1200 个观测的样本,得到 55 这个百分比的 p -值是 0.0008。这样,如果总体百分比真的是 50,则在 10000 个样本中只有 8 个样本的样本百分比大于等于 55。计算 p -值时先要计算 z 值(公式 7.4),然后才能用统计软件或查统计表 1 来确定 p -值。

请注意“因为 p -值等于 0.0008 而拒绝零假设”这种说法比“以 0.05 的显著水平拒绝零假设”能提供更多的信息。现在 p -值的方法比显著性水平的方法要普遍的多,因为现代统计软件能常规地计算出精确的 p -值。

停下来想一想 7.4

看一看报上政治候选人或政府官员在选举或声望的民意测验的报道。注意样本中的响应者是否有 1000 个之多?如果不是,是多少?报上是以什么方式提供民意测验结果的?这些报导正确吗?平衡吗?是否反映了你所理解的统计不确定性?如果不是,问题又是什么?

7.5 两个总体比例的差异

来自两个不同总体的比例相等的零假设记为

$$H_0: \pi_1 - \pi_2 = 0$$

这里 π_1 是第一个总体的比例, π_2 是第二个总体的相应的比例。当然事实中这两个比例不一定相等。如果我们能用数据表明应该拒绝零假设,就等于已经证明了这两个比例的不同。

为了检验零假设,我们从每个总体中抽一个样本,并分别计算的这两个样本的比例 p_1 和 p_2 。如果这两个样本比例相差很多,我们就拒绝总体比例相等的零假设。到底样本比例要差到多大才能拒绝零假设是要由两个样本的容量来决定的。

再看第三章研究罪犯上没上过文学课的例子。在 32 个上过文学课的犯人中有 6 个人被判犯了新罪,比例为 $6/32 = 0.19$;而 40 个没上过文学课的人中被判犯了新罪的比例是 $18/40 = 0.45$ 。观测到的这两个样本比例的差异是 $0.45 - 0.19 = 0.26$ 。这个差异是否已经大到可以拒绝零假设了? 还是观测到的差异只是样本随机性引起的?

检验零假设

有了从样本得到的比例值。我们首先把观测到的这两个比例之差(用公式 7.5)变换成统计中的 z 变量。我们然后用 z 分布表即正态分布表(统计表 1)或者统计软件找到相应的 p -值。如果 p -值很小,比如说小于 0.025,我们就拒绝总体比例相等的零假设。

观测到的两个样本比例的差异是 0.26,变换成 z 值就是 $z = 2.35$ 。 z 值大于等于 2.35 的概率是 0.009 即千分之九。这意味着在零假设正确即总体比例相等的情况下观测到两样本比例之差大于等于 0.26 的概率也是 0.009。

这个概率是这样小,以至于它是拒绝零假设的一个强有力的证据。我们拒绝零假设,并且可以得出这样的结论:在上过和没上过文学课的犯人中,重新犯罪的人所占比例的差异是统计显著的。

估计差异值

既然拒绝了零假设,就说明两个比例之差不是零。那到底差多少呢? 观测到的差 0.26 是总体比例差异的最好估计。

我们当然可以为差异的真值构造一个置信区间。对于观测到的差 0.26,则 95% 的置信区间是 0.06 到 0.46。因为这两个样本中的观测个数少,所以区间才会这么长。作为研究者,我们希望这个区间是众多包含真值的区间之一,而不是少数几个不含真值的区间中的一员。正如我们已经期待的,因为我们已经拒绝了真值差别是 0 的零假设,置信区间不包含 0。

z 还是 t ?

在前面的章节中,变量 z 和 t 已经被多次用作代替原始得分的标准得分。初学统计者的一个难题是:什么时候,用哪一个作为原始得分的替换? 到目前为止,我们已经在均值的假设检验中使用过了 t 变量,在比例检验中用到了 z 变量。

在含有一个或两个均值的假设检验中要用 t 变量,在包含一到两个比例的统计检验中要用 z 变量。

在后面的章节中还会出现 χ^2 变量和 F 变量。它们都是用来寻找不同背景下观测数据的 p -值的,只是应用的地方各有不同。

从数学上也可以导出应该用哪一个变量来作为原始得分的替换,但这已超出了本书的范围。因而,我们这些作者被推上了一个不舒服的位置:必须要简练地说出哪种变量适用于哪种数据。许多时候统计软件给出了正确的变量,而使用者却不以为然。我们能否正确地解释从分析中计算的 p -值是一个很重要的问题。

7.6 假设检验与构造置信区间

假设检验与构造置信区间两者都是为作出关于参数值的结论,并继而认识现实世界的方法。它们都是以样本数据为基础的。在假设检验中,我们的焦点是一个参数的一个特别的值,并且问是否该参数有可能等于该值。例如,智商测验的总平均值是否有可能等于 100。我们用置信区间来估计参数的真值。例如我们为总体均值找到 102 到 107 的置信区间。

对我们在 7.5 节中提到的总体中比例差的问题,置信区间的范围是从 L 到 U ,我们希望这个区间包含参数的真值。如果零假设中的相关的参数值在 L 与 U 之间,我们就不拒绝零假设,如该值在这个区间之外的某个地方,则拒绝零假设。



"shoe" reprinted by permission of Tribune Media Services.

在许多方面置信区间比假设检验提供的信息要多。置信区间给了我们一个参数值的可能范围,而假设检验只考虑到一个可能值。例如在假设检验中如果总体参数不是 100,我们就不清楚它是多少了。有时这一个值是非常重要或有意义的,像检验两均值之差是否等于零。可是即使当我们拒绝了零假设,并得到均值之差不为零的结论之后,紧接着的一个问题就是该差异是多大。这个问题是由置信区间来负责回答的。

尽管人们可能更希望得到置信区间,假设检验还是被广泛地应用于大多数领域中。主要原因就是统计软件包并不自动计算置信区间。

7.7 统计显著和实际显著

本章已经强调了统计显著的重要性,但有时一个统计显著的结论在实际中却并不重要。

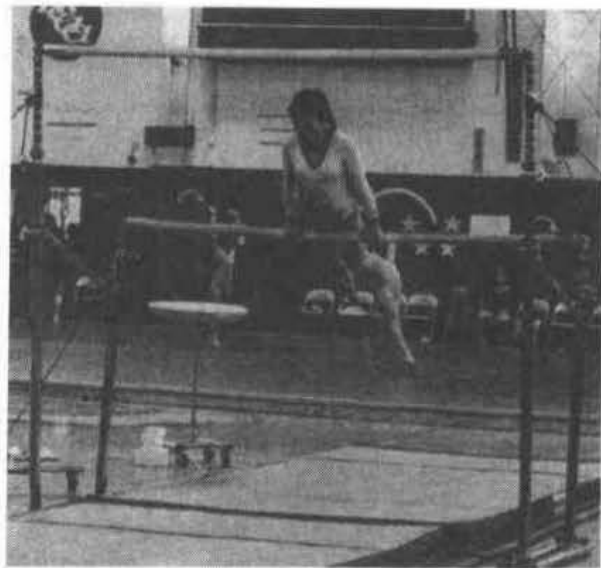
这里所要强调的是:一个统计显著的结果在实际中不一定真是一个显著结果。在大样本中,大多结果都是统计显著的。一个结果在实际中显著与否只有在研究清楚了来龙去脉后才能下结论。

停下来想一想 7.5

你下面将要读到的内容来自一个研究体育心理学的刊物：

在一个包含 75 个得奖者的样本中，发现奥林匹克队的成员在高低杠项目上的得分与受训年龄有关。那些 7 岁前就受到正式训练的运动员在此项目上的得分与那些 7 岁后才接受正式训练的运动员的得分差异是统计显著的($p = 0.017$)。

有关体操运动员的得分情况，这段话告诉了你什么信息？没告诉你什么？对于那些不甚了解统计的读者，怎样用“统计显著”这个词来夸大两类之间微弱的差别？



(来源:Gudmund Iversen)

零假设注意参数是否等于一个特殊值。在地理例子中，我们问的是：在地图上，墨西哥人均找到的地方数与美国人均找到的地方数是否相等？如果事实中他们相差 0.1，则零假设就是错的，因为差别不是 0。而这个零假设所提问题的正确答案应该是“否”，并且应该拒绝它。在这种情况下，当样本很大时，零假设就会被拒绝。

然而 0.1 这个差异其实很小，简直和 0 差不多。一个地理学家可能会说 0.1 这个差异很小，以至于它在这个地理问题中没有什么显著意义。这也就是说，尽管 0.1 这个差异是统计显著的，但不是实际显著的。所以如果我们在这个例子上一味地只是看统计上的显著性而不注意实际显著性，我们对数据的分析就是不完整的。

7.8 应用：何时拒绝零假设

为了在总结所有这些想法上获得一些经验，让我们先来看两个在研究中应用假设检验原理的简单例子。

设计这些例子的目的是为了表明如何将本章的内容应用到其它“现实世界”的情况中去。所有的分析也是一种警告：任何据说是来自一个样本，或是出自对两个或更多样本的比較的

息都需要认真检查。

关于合作性与竞争性的心理测试

一个心理学家正在研究对一个课题如何能有效地使一小群人在他们的工作策略上进行合作。该心理学家通过一个单面镜观察每一群人在该课题的工作状况;并在工作结束时把他们划分成合作类或竞争类。在观测了 8 组之后,有 7 组人被划为合作类。这一现象是偶然的吗?还是这个课题的设计使他们更容易合作?

让 π 代表一群人能合作的概率。如果仅有随机性决定结果,这个概率应该是 0.5(该研究结果与扔 8 次硬币出现 7 次反面的结果相同)。心理学家想知道的是:合作的情况是否比不合作的情况更可能出现?现在建立零假设:

$$H_0: \pi = 0.5$$

为了确定是否应该拒绝零假设,这位心理学家需要找到这种观测数据的 p -值。观测数据中包含了 7 个合作的组,所以计算 p -值就相当于计算合作组数大于等于 7 的概率;即已知 $\pi = 0.5$,有 7 个或 8 个组合作的概率是多少?

这是一个小样本,因此任意数目的合作组出现的概率都可以用有 $n = 8$ 个实验和 $\pi = 0.5$ 的二项分布得出。这个二项分布能利用二项分布的统计软件或表找到(公式 5.4)。有 7 组或 8 组合作的概率由图 7.6 给出。该数据的 p -值是 0.035。如果合作与竞争发生的概率相同,观测 1000 个样本(每个样本包含 8 个组),其中只有 35 个样本仅因偶然性使其中合作的组数达到了 7 或 8。这个概率大于 0.025 这个双边假设的检验标准,所以不是一个反对零假设的有利证据。每一组合作与否可能完全是由运气决定的。

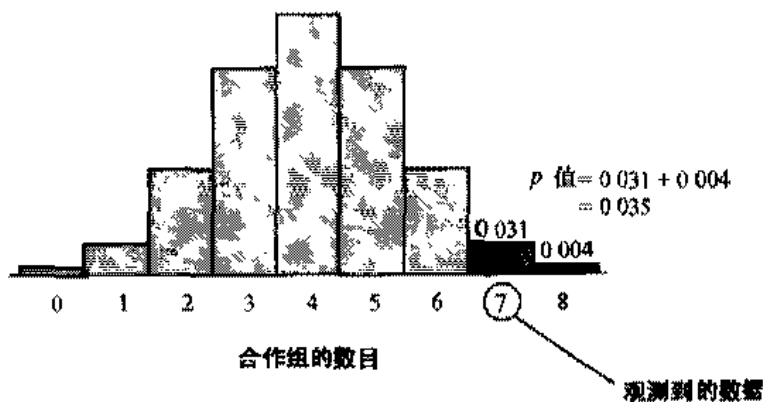


图 7.6 假设检验中二项分布的例子($n = 8$ 个观测, 概率 $\pi = 0.5$)。

如果这位心理学家把观测组数加倍, 变成 16 组而发现其中有 14 组合作, 则 p -值就变成了 0.002, 也就是在 1000 个样本(每个样本包含 16 个组)中只有 2 个样本中的合作组数能在 14 以上。这是一个拒绝“机会而不是课题决定了是否一组人合作”的零假设的很好的证据。

对社区的蓝领工人的研究

假设 1990 年时,在一个社区蓝领工人占工薪阶层的 60%,随着美国制造业工作的衰退,城市劳工委员会想知道是否该城市中的蓝领工人比例发生了变化。

用 Π 来代表蓝领工人所占的总体百分比,零假设为:

$$H_0: \Pi = 60\%$$

这个零假设认为蓝领工人的百分比没有发生变化。

为了检验这个零假设,该市在工薪阶层选了一个由 400 人组成的随机样本,发现其中有 215 个蓝领。因此,蓝领工人的样本百分比是 $215/400 = 53.75\%$,该样本的百分比与零假设的比例要小一些,但该市劳工委员会还不足以确定 53.75% 与 60% 的差距是否已经大到足以拒绝零假设的程度。有可能真正的比例仍然是 60%,而 53.75 只不过代表了在 60 附近的随机性波动。

为了寻找观测到的这个百分比的 p -值,先要把它按照公式 7.4 变换成正态分布 z 变量的值,得到 $z = -2.55$ 。正态分布表(统计表 1)显示 z 小于等于 -2.55 的概率是 0.005。也就是说 1000 个 z 值当中只有 5 个小于等于 -2.55 。所以在总体百分数为 60 时,观测到 53.75 这个百分数是一种很少见的情况。只有想一想,在 1000 次不同样本中只有 5 次的样本百分数小于等于 53.75,就会知道现在的情况有多么不常见了。该市官员认为该样本并不是不寻常的,因此拒绝零假设,并下结论说蓝领工人的比例小于 60%。

7.9 小结

当我们进行参数估计时,我们试图在找它的真实值。而当我们进行假设检验时,注意力主要集中在参数是否有可能等于和我们感兴趣的问题有关的某个值。

7.1 作为一个问题的假设

零假设(null hypothesis)说参数等于某个值。它描述的内容的含意通常是“参数值没有改变”,“两个参数没有不同”或者是“两个变量无关”等。备择假设(alternative hypothesis)是零假设逻辑上的反面假设。因此备择假设通常都描述的是“参数值改变了”、“两个参数间存在差别”以及“两个变量相关”等。

在检验零假设时,我们先假设它是真的,然后再分析数据,看它是否支持零假设。如果样本数据能提出比较有说服力的证据来挑战零假设,我们就拒绝(reject)零假设。

作出关于零假设的错误结论能导致两类错误: α 错误(第一类错误)是零假设为真时被误认为是错的; β 错误(第二类错误)是零假设为假时没能被拒绝。

7.2 怎样回答零假设所提出的问题?

p -值是在零假设为真(即参数等于某个值)时得到观测到的或比它更极端的数据的概率。

它给出了在多次抽样中能得到某种数据的机会的大小。它不是零假设为真的概率,零假设只能是真的或者假的之一。如果 p -值非常小(一般小于 0.05 或 0.025)就拒绝零假设。

备择假设可以是双边的,也可以是单边的。如果一个备择假设是双边的,则相对于零假设中设定的总体值,如样本统计量的值太大或太小时我们都应该拒绝零假设。如果一个单边备择假设的内容是“参数值大于零假设中的值”,则只有当样本统计量的值比较大时才拒绝零假设;反之若单边备择假设的内容是“参数值小于零假设中的值”我们就在样本统计量较小时拒绝零假设。

当一个零假设被拒绝时,我们可以说样本结果是统计显著的。

样本得分应该被转换成标准得分(如 t 得分、 z 得分),为的是寻找 p -值。标准得分大于 +2 或小于 -2,都是不寻常的大而且伴随着非常小的 p -值。

7.3 显著水平

显著水平一般用希腊字母 α 来表示。样本统计量的临界值是这样被选出来的:当零假设正确时落在临界值以外的样本统计量只占全部的 5%。

7.4 总体比例检验

在检验总体比例时,零假设叙述:总体比例 Π 等于诸如 0.5 那样的某个特殊值。对于一个样本(1200 个观测),有 95% 的样本比例落在真的百分比加、减 3% 之间。如果观测到的样本值落在临界值范围(+3 到 -3)之外,就拒绝零假设。

7.5 两个总体比例的差异

这类问题的零假设一般是“没有差异”。如果样本统计量间的差是统计显著,就拒绝零假设。

一旦零假设被拒绝,下一个任务就是估计这个差异有多大,这能通过构造 95% 的置信区间的办法来解决。

7.6 假设检验与构造置信区间

假设检验与构造置信区间都是对总体参数作出结论的有用工具。假设检验提出“一个参数是否等于某个特定的值”这一类的问题,而置信区间则给出了一个可能包含参数值的范围。因为多数统计软件的设计,假设检验应用得较多。

7.7 统计显著与实际显著

一个大样本的结果可能是统计显著的,但对于现实问题的目的它并不一定都是实际显著的。

7.8 应用:何时拒绝零假设

通过两个例子说明了假设检验方法的用处。这两个例子之一是用小样本来研究心理测验

的结果,它利用了二项分布表(表5);另一个是用大样本来研究在一个城市人口中蓝领工人所占的比例。

补充读物

Henkel, Ramon E. *Tests of Significance* (Sage University Paper Series on Quantitative Applications in Social Sciences, series no. 07-004). Beverly Hills, CA: Sage, 1976.

Mohr, Lawrence B. *Understanding Significance Testing* (Sage University Paper Series on Quantitative Applications in Social Sciences, series no. 07-073). Newbury Park, CA: Sage, 1990.

公 式

单个均值的检验

它的零假设要问的是“总体均值 μ 是否等于一个特殊数值 μ_0 ”:

$$H_0: \mu = \mu_0$$

一个样本包括 n 个观测,均值是 \bar{x} ,标准差是 s 。为检验零假设先要把样本均值做一个变换,变换成 t 变量的一个值,公式如下:

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \quad \text{d.f.} = n - 1 \quad (7.1)$$

用 t 的这个观测值,通过统计软件或统计表可以计算出相应的 p -值。如果 p -值很小就拒绝零假设。

考虑零假设 $H_0: \mu = 4.0$ 。如果样本容量 $n = 12$,均值 $\bar{x} = 2.0$,标准差 $s = 1.54$,则有

$$t = \frac{2.0 - 4.0}{1.54 / \sqrt{12}} = -4.50 \quad \text{d.f.} = 12 - 1 = 11$$

t 小于等于 -4.50 的概率是 0.0005 ,即 p -值是 0.0005 。如果用 5% 的显著水平或 t 分布表(统计表2)来判断,对于自由度是 11 的 t 分布来说 $t < -2.20$ 或 $t > 2.20$ 的概率是 5% 。现在 t 值为更极端的 -4.50 ,所以拒绝零假设。

检验两均值的不同

设两个总体的均值分别是 μ_1 和 μ_2 ,零假设是:两个均值相等:

$$H_0: \mu_1 - \mu_2 = 0$$

为检验零假设,从两个总体中抽取样本数据,第一个样本有 n_1 个观测,均值是 \bar{y}_1 ,标准差是 s_1 ;第二个样本有 n_2 个观测,均值是 \bar{y}_2 ,标准差是 s_2 。

通过这些可以计算出下面的 t 统计量:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s \sqrt{1/n_1 + 1/n_2}} \quad \text{d.f.} = n_1 + n_2 - 2 \quad (7.2)$$

如果计算出的 t 值超过了从 t 分布表中查到的临界值,就拒绝零假设。(用 p -值方法可以得到更多的信息, p -值是出现观测到的 t 值和比这个值更极端的值的概率。)

要计算 t ,我们首先要找到公式(7.2)中的标准差 s 。首先要计算的是 s^2 而不是 s 。它是两个样本方差的加权平均。它被称为来自两个样本的联合方差,计算公式如下:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad (7.3)$$

注意公式中用的是两个样本的方差而不是标准差。使用这个公式的前提是样本所属的两个总体的方差相等。因为若总体方差相等,则这两个样本方差就是对同一个参数的估计,两个估计也可以联合起来。联合方差是估计的联合。如果两个总体的方差不相等,那就有必要修改计算方法了。

联合方差的平方根就是联合标准差 s ,用它、两个样本均值和它们的样本容量就可以计算 t 变量的观测值了。

在地理的例子中,我们虽然知道分子上的任何两国样本的两个均值但由于并不知道联合的 s 或它们的样本容量,所以无法计算 t 值。但是我们知道大不列颠和法国的 t 值一定很小,因为在地图上找对的地方数之差仅为 0.2,从统计上看是不显著的。同时墨西哥和美国的均值之差却达到 1.3,这个差异应该是统计显著的,它们的 t 值很可能会比较大。一般来说 t 值大于 2.00 就被认为是相当大的,而且是统计显著的。反之,若算得的 t 值小于 2.00 就不能算大,也就不能说这个结果是统计显著的。

在总体方差相等的情况下,有一个计算 t 值的简化公式,它可以避开联合标准差的计算。与上面类似,用下标 1 和 2 来区分统计量来自哪个样本,一个计算 t 的近似值的公式为:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \quad \text{d.f.} = n_1 + n_2 - 2$$

注意在分母的根号中是用第一个样本的观测个数除第二个样本的方差,同时用第二个样本的观测个数除第一个样本的方差。在绝大多数情况下这个公式的效果和公式 7.2 一样好。

比较成对数据均值的 t 检验将会在第十二章给出。

总体比例检验

零假设为总体比例 π 等于一个特殊值 π_0 :

$$H_0: \pi = \pi_0$$

我们把样本比例 p 的值变换成 z 变量的一个值,公式如下:

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} \quad (7.4)$$

此处 n 是样本尺寸。

作为一个例子,假设有 $n = 1000$ 个观测的样本的样本比例是 $p = 0.60$ 。我们想检验零假设

$$H_0: \pi = 0.50$$

首先把 p -值变换成 z 值

$$z = \frac{0.60 - 0.50}{\sqrt{\frac{0.50(1 - 0.50)}{1000}}} = 6.32$$

而 z 值大于等于 6.32 的概率小于 0.0001。这样,在总体比例是 0.50 的情况下,含有 1000 个观测的样本之比例大于等于 0.60 的概率是如此之小,于是我们只好拒绝零假设,并下结论说样本不可能来自这样的总体,真正的总体比例应大于 0.50。

比例之差的检验

零假设为两个总体比例相等:

$$H_0: \pi_1 - \pi_2 = 0$$

为了对零假设进行检验,我们把两个样本的观察比例之差 $p_1 - p_2$ 用下面公式变换成为 z 得分:

$$z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}} \quad (7.5)$$

其中 n_1 和 n_2 分别是两个样本中观测的个数。

在罪犯的那个例子中一个样本中的 32 个人中有 6 个被判犯新罪,而另一个样本的 40 个人中有 18 个被判犯新罪。这就给出了 $p_1 = 6/32 = 0.19$, $p_2 = 18/40 = 0.45$ 。有

$$z = \frac{(0.45 - 0.19) - 0}{\sqrt{\frac{0.45(1.00 - 0.45)}{40} + \frac{0.19(1.00 - 0.19)}{32}}} = 2.35$$

而 $z = 2.35$ 的 p -值是 0.0094,因此我们拒绝两个总体比例相等的零假设。

这个方法也可以作少许变动。当零假设为真而且两个总体比例确实相等时,两个样本就可以联合在一起来估计这个公共的值。重新犯罪的总人数所占的比例为

$$p = \frac{6 + 18}{32 + 40} = \frac{24}{72} = 0.33$$

现在可以用这个公共的 p -值来代替两个分别的值去计算 z 值:

$$z = \frac{(0.45 - 0.19) - 0}{\sqrt{\frac{0.33(1.00 - 0.33)}{40} + \frac{0.33(1.00 - 0.33)}{32}}} = \frac{0.45 - 0.19}{\sqrt{0.33(1.00 - 0.33)}\sqrt{\frac{1}{32} + \frac{1}{40}}} = 2.33$$

其 p -值为 0.0099, 得到同样的结论: 拒绝总体有相等比例的零假设。

这个公式看上去很像在检验两个均值差异时求 t 值的公式; 分子为两个均值之差, 分母的 s 用 $p(1-p)$ 和两样本容量的倒数之和的平方根来代替了。

也可以通过比较两个比例来研究 2×2 列联表中两个分类型变量间的关系。这将在第九章中讨论。

习 题

回顾(习题 7.1—7.24)

- 7.1 统计显著是什么意思?
- 7.2 a. 当我们对某组数据做假设检验时我们想知道某个参数的什么情况?
b. 请给出一个零假设的例子。
- 7.3 a. 什么是零假设?
b. 零假设与备择假设有什么不同?
c. 请写出它们各自的符号。
- 7.4 a. α 错误与 β 错误有何不同
b. 想一种记忆的方法来分清楚谁是谁。(记忆方法用来帮助记忆, 可以是一聪明的话、故事、歌曲、视觉想象、联想或任何能帮你正确地记忆某事的方法) 例如, α (α) 误差用于零假设为精确(α ccurate)时; β (β) 误差用于零假设糟糕(β ad)时。你一定能想出更好的!①
- 7.5 一般来说, 如果样本均值与零假设中所设的总体均值相差很大, 是否应该拒绝零假设?
- 7.6 a. 什么叫做标准得分?
b. 当零假设正确时, t 变量的值主要集中在什么范围之内?
c. 如何使用标准得分来进行假设检验?
d. 如果已知样本均值对应的 t 值, 在零假设的情况下你如何找到等于或比该值更极端的值的概率? 请给出两种方法。
- 7.7 a. p -值能告诉我们什么信息?
b. 当相应的 p -值较小时为什么要拒绝零假设?
c. 显著水平与 p -值有何区别?
- 7.8 如果“总体均值等于 4”的零假设在研究过程中被错误地拒绝了, 请问这是犯了第几类错误?
- 7.9 a. 最常用的显著水平是多大?
b. 请用日常用语来解释它的意思。
- 7.10 若零假设正确, 则大多数样本统计量的值将会[接近; 远离]零假设中的那个特殊的参

① 这里的例子是为说英语的人举的, 它利用英语单词的第一个字母同来联系 α 和 α ccurate, β 和 β ad——译者注。

数值。

- 7.11 a. 显著水平经常用哪个希腊字母来定义?
b. 这个字母在英语中如何拼写?
- 7.12 a. “这项研究的临界值是 ± 2.3 ”的意思是什么?
b. 如果在你的结果中的标准得分是 -3.0 , 你应该做出什么样的结论?
c. 如果显著水平被改成了小于 0.05 的其它值, 比如 0.01 , 临界值将会发生什么样的变化?
d. 你是否认为 b 中的结论应该随显著水平的变化而变化?
- 7.13 a. 用正式的叙述来表达下面的零假设: “在总人口中, 支持总统的比例等于 0.50 ”的检验。
b. 用正式的叙述来表达下面的零假设: “在总人口中, 支持总统的百分比等于 50% ”的检验。
- 7.14 用正式的叙述来表达下面的零假设: “民主党和共和党中支持最近税务改革议案的人在各自党派中所占的比例没有区别”。
- 7.15 当两个样本比例之差达到某个值时, 我们就拒绝“两个总体比例相等”这个零假设。现在请问有哪些因素会影响这个值的大小?
- 7.16 你正在努力倡议一种在职培训工人的新方法, 因为你相信它将会使他们对他们的工作更满意。做一个新旧方法的对比研究。你有什么希望: 数据支持零假设呢还是备择假设? 解释你的感觉。
- 7.17 让一年级的一半学生在语文和算术课之间休息 20 分钟, 同时让另一半学生在这 20 分钟里做作业。我们做这个实验的目的是想说明户外运动有助于提高学生的算术成绩。
a. 这个研究的零假设是什么?
b. 这个研究的备择假设又是什么?
- 7.18 你相信在改进汽车设计、革新汽油添加剂、控制泄漏法案等一系列措施的影响下, 芝加哥今年的污染状况比去年有所好转。但你的汽车修理师不赞同你的观点, 他认为尽管有这些措施, 空气污染并未减轻。于是你同意就你们的对立的观点做一个检验。你应该如何叙述零假设?
- 7.19 村镇教育系统中工资级别最高的教师们在 1985 年的平均工资是 \$43000, 而 1995 年时在这个水平的教师的一个样本表明该工资均值为 \$53000。
a. 如果你想知道在这十年内该系统的教师最高工资待遇是否有所提高, 你应做一个什么样的零假设?
b. 你是否认为在过去的十年内该变量的总体均值有了一些变化? 解释你的回答。
- 7.20 根据粗略的统计原则, 下列 p -值中哪些能导致拒绝零假设? 哪些不能? 哪些难以决定?

$$p = 0.50 \quad p = 0.25 \quad p = 0.001 \quad p = 0.10 \quad p = 0.05 \quad p = 0.025$$

- 7.21 在日常生活的语言中 p -值等于 0.50 是什么意思?

- 7.22 统计研究中,小样本和大样本哪个更容易获得统计显著的结果?
- 7.23 单、双边假设检验的主要区别是什么?
- 7.24 如果你没能拒绝零假设,是否等于已经证明了零假设是对的?

解释(习题 7.25—7.43)

- 7.25 一罐草莓酱上写到“净重 18 盎司。”
- 解释这句话的意思是不是当你仔细称这罐酱时,它将恰好为 18 盎司。
 - 解释是否其它罐中酱的份量都与第一罐的相等。
 - 你应该怎样设计一个实验及分析结果数据来判断制造商是否有权声称酱的净量是 18 盎司?
- 7.26 从 1976 年到 1980 年全国健康统计中心分析美国成年男子身高的报告中可知:美国成年男子平均身高 69.1 英寸。(来源:1992 年 7 月 26 号的纽约时报 pE5。)在一个含有 50 名大学四年级男生的样本中平均身高是 71 英寸,标准差是 2.1 英寸。从这两个数据中可以算出 $t = 6.40$, 自由度为 49 ($p < 0.0001$)。
- 如果我们想发现是否高校学生的平均身高与总人口的平均身高有区别,应该对高校四年级学生做什么样的零假设?
 - 我们能对样本所属的总体的平均身高做什么结论?(我们拒绝还是不拒绝零假设?)
- 7.27 Gallup 的一项调查表明全国有 53% 的人拥有枪支(来源:1992 年 7 月 26 日的纽约时报 pE5)。为了了解枪拥有率是否与中西部一个中等城市的等价,做了一个包含 300 个响应的调查,其中 45% 的人拥有自己的枪支。在检验“该城市中持枪人数所占百分比等于全国枪拥有者的百分比”这个零假设时得到 $z = 2.78$, 相应的 p -值是 0.003。
- 为什么我们可以由此下结论说该城的枪拥有率不等于全国的平均水平?
 - 该调查的百分比和全国的百分比之差是统计显著的,可它是否能大到能产生实际的兴趣?
- 7.28 在一次有关剃须习惯的电话调查中,有 200 人原意和提问者在电话上花费 20 分钟回答问题。但是研究者们的报告中忽略了有 70% 的人挂断电话或是拒绝完成受访的事实。由于样本有偏差,这个研究的结果有缺陷。为什么对于假设检验来说有一个正确的抽样方案是很重要的?
- 7.29 为什么零假设用“零”作定语?
- 7.30 当统计学家拒绝零假设时,他们能有绝对确定的把握认为他们的结论是正确的吗? 解释你的回答。
- 7.31
- 统计报告中提到的“ p -值是 0.025”是什么意思?
 - 什么是拒绝零假设所导致的可能实际后果?
- 7.32 在本章提到的地理知识的那个例子中,为了得到在总体均值相等的假设下得到观察的样本均值差或更大的概率,要把样本均值差异变换成 t 值,为什么这一步是必要的?
- 7.33 在一项口味偏好的研究中,随机抽取了 200 个可乐饮料消费者,调查他们对两种主要

可乐的满意程度。零假设是消费者对两种可乐没有偏好。如果满意程度是用 7 分制来打分, 可乐 A 的平均得分是 5.0 分而可乐 B 的平均得分是 4.6 分。

a. p -值是 0.001; 为什么拒绝零假设?

b. 在拒绝零假设时, 我们出错的可能性有多大?

c. 你认为可乐公司将会从中找到他们实际感兴趣的结果吗?

- 7.34 政府调查表明 1986 年制药业的利润是从处方的总费用中获利 62.8%, 到了 1992 年这个数字变成了 69%。如果报告认为改变 3% 就算统计显著, 那你是否认为这六年中药业利润发生了变化? (来源: *Adapted from a report by Stephen Schondelmeyer, economist, Health Care Financing Administration, Office of Technology Assessment, U. S. Government, 1992*)

- 7.35 著名统计学家 Ronald Fisher 给出了最适当的(最大的)水平, 在该水平能判断是否 p -值足够小以至于拒绝零假设。请用平常的话来重说这句话, 并就 Fisher 为什么选 0.05 而不是 0.20 或者 0.001 或别的什么值提出你的看法。

- 7.36 a. 如果一个样本相对较小(例如 50 个观察值), 而且因为 p -值是 0.1 而没能拒绝零假设, 下一次再进行类似实验时, 你应该怎样做才能提高拒绝零假设的机会?

b. 我们能否认为经济上宽裕的研究者比拮据的研究者更容易拒绝零假设?

c. 人们是否总是希望能够拒绝零假设?

- 7.37 拒绝了零假设(有一个统计显著的结果)是不是就意味着你已经发现了现实世界中的一个令人激动的新的事实? 能否结果是统计显著的, 但在现实生活中只是微不足道或没有意义的? 讨论诸如估计广告和市场营销的结果。

- 7.38 1987 年 Diane 和 George Weiss 允诺为从费城的 Belmont 小学毕业的 112 名学生中考上大学的人提供大学学费。6 年后, 这批人的 45% 高中毕业而与之相应的是在 1986 年 Belmont 的小学毕业生中只有 28% 的人在 6 年后高中毕业。新闻报导由此下结论说这两个总体百分比的差异是统计显著的。(来源: 1993 年 6 月 25 日的费城问讯报 pp A1, A18)

a. 在这个研究中, 零假设是什么?

b. 用的是哪种统计检验来得到统计显著的?

c. 报导中说结论是统计显著的, 这给我们提供了什么信息?

- 7.39 对血液在人造瓣膜中的流速问题:

$$H_0: \mu = 0.50, \bar{x} = 5.96, t = 2.59(47 \text{ d.f.}), p = 0.0049$$

a. 请用文字把上面的过程描述一遍。

b. 我们能从这个结果中得到什么有关这种人造瓣膜的信息?

- 7.40 FBI 报导说全国有 55% 的凶杀是枪伤的结果。在一个最近的来自一个社区的样本中 66% 的凶杀是枪伤的结果。从该社区的百分比并比较全国的百分比, 你能作出哪三种结论? 要从中选择一个你还需要什么附加信息?

- 7.41 这儿有一项对加州新婚墨西哥裔妇女抑郁症的研究结果。下面是患此病的妇女与未患此病的妇女的对比情况:

感觉到的受歧视程度 ($\bar{x} = 17.3$ 对 $\bar{x} = 11.4$), $t(df 136) = -3.7, p < 0.001, \dots$ 和对在

美国成家的关心程度($\bar{x} = 16.3$ 对 $\bar{x} = 12.0$), $t(df\ 117) = -2.5$, $p < 0.05$, 被识别为使这一群移民妇女有发展忧郁症危险的重要因素。(来源: Adapted from Salgado de Snyder and V. Nelly, "Factors associated with acculturativestress and depressive symptomatology among married Mexican immigrant women," Psychology of Women Quarterly, vol. 11(1987), pp 475 - 488)

a. 用文字描述上述内容。

b. 有关这群妇女患抑郁症的问题, 以上那些内容能为你提供什么信息?

- 7.42 一家商店最近在收款台添加了自动从银行提款这种新的结帐方式。把采取这项措施以前的6个月中人均消费量与此后一个月中人均消费量进行对比, 发现购买的增加量是统计显著的。能否告诉商店经理购买力的增加应该归功于新添加的付款方式? 解释你的回答。
- 7.43 一项全国调查表明在900个投票者中有47%的人支持总统最近的对外政策。请用按照一般抽样调查的粗糙的统计准则, 是否这个结果说明总统的行为不被大多数选民支持? 为什么?

分析(习题 7.44—7.59)

- 7.44 我们可以用计算标准得分的方法来判断某个样本来自几个可能总体中的哪一个。在这道题中我们提供了两个不同的总体, 并从中各抽取10个样本。当然在实际中不可能抽这么多的样本。这个习题证明了统计得分是很有用的。这些样本的均值如下:

总体 A: 61.2 62.6 40.1 51.7 38.0 59.8 47.6 47.7 56.3 35.0

总体 B: 83.0 93.7 82.1 72.4 92.3 68.7 76.5 88.4 79.6 63.3

总体 A 中均值的平均是 50.0, 均值的标准差是 10.0, 而在 B 中相应的两个数是 80.0 和 10.0。

- a. 把 A 中所有的样本均值变换成标准得分。
- b. 把 B 中所有的样本均值变换成标准得分。
- c. 若第 21 个样本均值是 75.0, 则当它的确是来自总体 A 的样本时, 其标准得分是多少? 若是来自总体 B 的, 标准得分又是多少?
- d. 基于你对 a - c 的答案, 你认为这个新样本属于来自总体 A 的可能的样本集合或是来自总体 B 的可能的样本集合? 为什么你认为该新样本来自总体 A 或 B?
- 7.45 在包含了12个家庭的样本中, 平均每家有2.0个孩子。在某国这个均值是每家1.4个孩子。你想知道是否这个样本来自该国。
- a. 假如你从该国抽取了许多不同的样本, 请解释为何这些样本的均值不都等于1.4, 但所有这些样本均值的均值等于1.4?
- b. 如果这些样本均值的标准差(标准误差)是0.5, 请给出你的包含这12个家庭的样本的标准得分。
- c. 这个样本的标准得分是不是特别大, 或者你的样本有没有可能来自这个国家? 请解释。
- 7.46 一个含有50个响应的样本中, 支持某位政治候选人的选民比例是38%。你希望知道总体中相应的比例是不是50%。如果总体比例是50%, 就意味着大量样本比例的均

- 值等于 50%。若这些样本比例的标准差(标准误差)是 0.071,则:
- 求这个样本比例的标准得分。
 - 这个样本标准得分是否很大,或者此样本有无可能来自比例是 50%的总体?请解释。
- 7.47 据全国健康统计研究中心的研究表明工人和学生每年人均病假 5.1 天。而在一个包含 49 位商业雇员的小公司的样本中每年人均病假天数是 7.0 天,标准差是 2.5 天。公司领导希望知道他们的雇员是否比平常人容易生病?
- 顾主感兴趣的零假设是什么?
 - 利用报道的数据作为来自一个随机样本的数据并对这个观察的均值找出 p -值。
 - 顾主能得到什么关于检验零假设的结论?
- 7.48 据调查,有 64% 的美国成年男子喝酒。在一个包含了 25 名大学生的样本中有 19 名学生说他们喝某种酒。学监想知道高校学生中喝酒的人所占的百分比是否高于全体成年人中的这个百分比。
- 学监要用的零假设是什么?
 - 请在这个零假设下计算样本中观察数的 p -值。
 - 从这个 p -值可以得到什么结论?
 - 如果 p -值较大而且学监不能拒绝零假设。这是否等于已经证明了学生喝酒的百分比实际上等于全国的百分比?
 - 用样本数据构造总体喝酒百分比的 95% 的置信区间,并解释它是否包含 64%。
- 7.49 据人口普查局的调查表明:有 73.2% 的 16 岁或 16 岁以上的工人独自开车去上班。在实行共同开汽车上班的办法以后,一个城市调查发现 300 名工人中有 67% 的人独自开车上班。城市管理者想知道该城市中独自开车上班的人所占比例是否比全国的这个比例要低?
- 城市管理者的零假设是什么?
 - 用调查到的数据检验这个零假设。独自开车者的比例减少是否统计显著?
 - 独自开车者的比例缩减量在实际中显著吗?
- 7.50 参看本章开头所提供的地理知识的那个例子中各国样本的平均得分表。像本章中对比墨西哥和美国均值差异时所做的那样,比较表中每一对国家的均值。用本章学到的对比均值的方法来判断哪几对国家的人们所具有的地理知识有显著差异。这些对比为:美国和墨西哥、美国和大不列颠、美国和法国、墨西哥和大不列颠、墨西哥和法国、大不列颠和法国。
- 7.51 一个来自 Detroit 的随机样本包含了 103 名浸礼教信徒和 87 名卫理公教信徒(来源: H. Schuman, "The religious factor in Detroit: Review, replication and reanalysis," *American Sociological Review*, vol. 36(1971), pp 30-48)。很明显样本中浸礼教的信徒比卫理公教的信徒要多。你能就此做出结论说:此项研究中,包含样本的总体中浸礼教信徒在调查的时候比卫理公教信徒要多吗?请建立一个适当的零假设,并用这组数据来检验它。
- 7.52 在习题 4.66 中 28 条箭鱼中平均水银含量是 1.09 ppm,标准差是 0.48 ppm。既然当箭鱼中水银含量超过 1.00 ppm 时就不宜被食用,你当然想知道所有箭鱼中水银含量的

平均值是否等于 1.00。

a. 请检验这个零假设。

b. 对箭鱼总体中水银含量的均值你能做出什么结论?

- 7.53 教授停车厂总是挤满了各种各样的汽车。毋庸置疑每一辆车的牌子在一定程度上反映了其主人对一些事物的看法。在一次调查中,询问车主两个问题。一个是他们开的是什么车,另一个是在刚刚结束的选举中选的是谁。结果表明在开 Saabs 车的教授中有 98% 的人选民主党候选人。类似地,在开 Volvos 的教授中有 80% 的人选民主党候选人。(来源: *The Laddn - Ipsos Survey*, *The Chronicle of Higher Education*, April 5, 1976, p. 18) 报告中没有提到样本中开 Saabs 和 Volvos 的人各是多少。假设有 50 人开 Saabs, 有 200 人开 Volvos。如果零假设是: 在所有开 Saabs 的人和开 Volvos 的人中民主党的支持率相等, 请用以上的数据检验这个零假设。
- 7.54 1980 年中期 NCAA^① 收集了一些数据, 目的是研究 Division I 中的运动员的毕业率。在 2332 名男性中有 1343 人没有大学毕业, 而在 959 名女性中有 441 人没有毕业。用这个数据检验零假设: 男性毕业率与女性毕业率相等。
- 7.55 在一个当地的娱乐场中乘坐 Mile High Terrifying Trojans Roller Coaster^② 中的前 200 人中有 134 名男性。是否一般来说乘游戏车的男性多于女性, 还是该区别完全是随机性引起的?
- 7.56 进口丝质头巾子样本之间的瑕疵数之差一定要超过 2.1 才会认为是统计显著的。表 7.1 中列出了进口丝质头巾各子样本中发现的瑕疵数

表 7.1 习题 7.5 的数据

质量控制 群号	被拒 领头 巾数
1	17
2	14
3	12
4	18
5	21
6	20

a. 假设头巾有瑕疵是随机的, 在检查瑕疵时是否有些子样本与别的子样本相比查得更精确?

b. 哪一些子样本互相有显著差异?

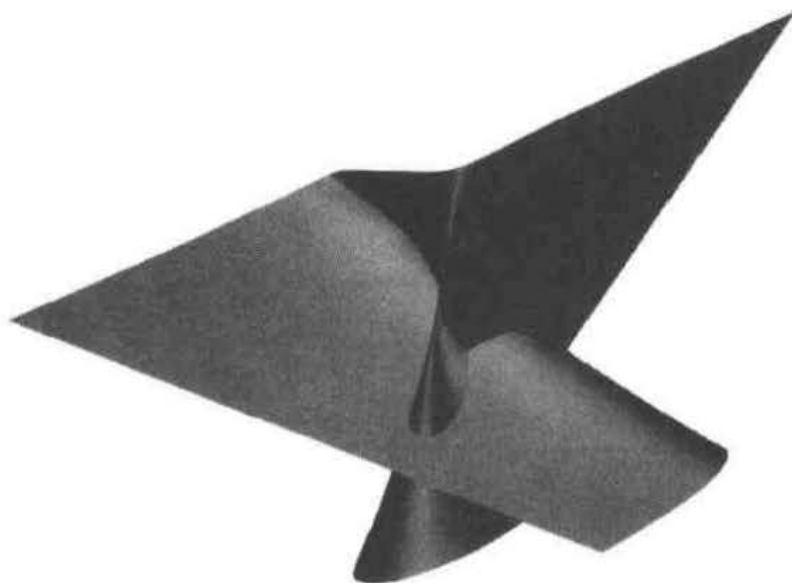
- 7.57 习题 4.59 中的表列出了数学与统计系教师拥有的孩子数。在 10 个孩子中有 3 个女孩 7 个男孩。做一个统计检验来看是否数学家和统计学家男孩和女孩数目一般来说相等?

① National Collegiate Athletic Association, 全国大学生体育协会——译者注。

② 一种游戏车, 亦称过山车——译者注。

- 7.58 在我们一个初等统计课上,学生们共旋转硬币 250 次。他们观测到 97 次正面和 153 次反面。请检验零假设:硬币出现正面的概率是 0.5。
- 7.59 在高校学生的随机样本中,36 名男生中有 16 人说他们买食品时读了营养说明,而 36 名女生中则有 28 人说他们读了说明(来源:Data used by permission of Jasa Porciello, Swarthmore College.)。检验零假设:在大学男生中读营养说明的人占的比例与女生中读此说明的人占的比例相等。

C H A P T E R 8



8.1 关于两个变量的 4 个问题以及它们之间的关系

8.2 预测：从一个变量到另一个变量

8.3 自变量和因变量

8.4 不同类型的变量：分类型变量、顺序型变量和数量型变量

8.5 回到因果关系的问题

8.6 小结

变量间的关系



大学的一个生态俱乐部计划增加开会时的饮料和娱乐经费以吸引缴纳会费的成员。作为俱乐部的一个成员,由你负责判断增加娱乐和饮料经费实际上是否能增加会费收入。这笔支出和会费收入这两个变量之间有关系吗?

一个本地的商店被请求购买学校年鉴上的广告。他们想知道在年鉴上刊登广告是否能增加和学生有关的生意。作为商业经理,你需要回答这个问题。

一个橄榄球队面临着在人造草地还是在草地上打球两种选择。他们的回答依赖于在人造草地上和在草地上打球的受伤率。如果要你去给这个队提一个建议,你应该如何做?

现在让我们转到一些更普遍的话题上来。在最近 10 年间,上班族中妇女所占的比例是否发生了变化?是不是不同种族的人吸不同牌子的烟?患爱滋病的孕妇服用 AZT 是增加还是减少传染病毒给她们的孩子的机会?在年平均气温不同的地方乳腺癌发病率是否也不同?死刑的数目会影响凶杀数量吗?

在前面的章节中我们已经了解了在每次有一个变量时如何收集数据、汇总数据以及从数据中得到结论等统计方法。现在我们要在同时有两个或两个以上的变量时分析有关的数据。也就是说我们要研究变量间的关系。大多数科学都会遇到寻找两个变量之间关系的问题,在这种工作中统计学扮演着重要的角色。在这一章中我们提到了统计分析所涉及的范畴。本书剩下的部分就是用各种方法研究变量之间的关系。

在我们研究两个变量时,我们研究一个变量的某些值是否对应于另一个变量的某些值。例如:增加用于娱乐的预算是否会带来更多的会费?学校年鉴上的广告是否能增加商业收入?橄榄球比赛中受伤次数与受伤程度

是否与场地类型有关?当我们确实从数据中发现了这样一些规律时,我们就称变量间存在统计关系(statistical relationship)。

我们在第二章中介绍过,用于研究两个变量的数据按其来源分主要有两种:实验型数据和观测型数据。从这两类数据中获得有关统计关系本质的结论常常是非常不同的。

假设生物学家关心的是采光量如何影响某种作物的生长。为了回答这个问题,她设计了一个实验:尽可能地选择相似的幼苗,并且除了采光量以外为它们提供完全相同的生长条件。当不同光照下的作物成熟以后,分别测量它们的生长情况。这里,每个植株有两个变量的观察值:采光量和植株的生长量。这位生物学家把这两个变量的数据在一个数据文件中写成两列数(从上往下)。其中一列记录的是每个作物期望得到的采光量,另一列记录的是每个作物的生长量。每一行(横着读)中包含的两个数据为某一株作物的采光量和生长量。



为什么估计光线对这些植物的影响是很困难的?(来源:Hans Reinhard/Okapia, Photo Researcher)

一位政治学家想知道一个人的投票倾向性与其年龄是否有关。为了回答这个问题,该政治学家利用观察数据。他实施一个民意调查,让调查者记录被访者的年龄和最近一次选举中投了谁的票。像作物的那个例子一样,他所记录的数据文件也由两列组成;一列是年龄变量的数据,另一列是选举变量的数据,而且每一行由一位被访者的年龄和投票组成。

现在上面两个例子中的数据已经被收集好,可以进行分析了。

8.1 关于两个变量的4个问题以及它们之间的关系

分析由两个变量控制的数据其主要目的就是回答以下4个重要问题。这些问题为我们进行统计关系的分析提供了框架。分析来自两个变量以上的数据通常也应该回答以下4个问题。

问题一:从数据来看变量间有关系吗? 首先我们要试图确定观测到的数据中是否含有某种关系模式。如果我们发现变量间确实有关系则继续回答以下问题。

问题二:如果变量间有关系,这个关系有多强? 如果数据间存在某种关系,我们就应该试着去确定这个关系有多强。变量间的关系可能强,也可能弱。

问题三:是否不仅在样本中,而且在总体中也有这种关系? 换句话说,我们如何能从观



对投票倾向性的关注已经激发了调查研究的极大的兴趣。(来源: Rob crandall, stock Boston)

测值推广到现实世界? 我们对观测型数据或实验型数据中两个变量间的关系感兴趣,但通常我们更感兴趣的是在包含这些数据的更大的总体中,变量间是否也有同样的关系。在一次政治民意调查中,在含有几百个人的样本中两个变量相关是一回事,在整个选举过程中这两个变量相关又是另一回事。

有时第三个问题还可以换成另一种说法:这个结果是完全由偶然因素引起的,还是受某种系统影响而产生的? 如果我们发现这个结果只是偶然发生的,于是我们通常就可以作出在包含该样本的总体中变量之间无关的结论。

问题四:这个关系是不是因果关系? 当然,这是最难回答的一个问题。但它却常常也是最重要的一个问题,尤其是处理数据观测型时,这是统计学最没有办法解决的问题。对于观察数据,问题常常得不到回答,此因我们不知道观测到的这两个变量间的关系是否由根本就没被考虑进来的一些变量引起的,就好像下面将要提到的冰淇淋和孩子受伤的例子中发生的那样。对于实验得到的数据,情况往往会不同。在一个按照适当的统计原则设计的实验中,我们常常能控制其它变量来消除其影响,这样使得我们自己能确定因果关系。

如果观测到的两个变量间的关系可以通过引入第三个变量来解释,这种关系就被称为伪关系。

许多变量可能有统计关系但没有因果关系。例如我们有一年中每个月的冰淇淋销量和儿童在交通事故中受伤次数的数据。在冰淇淋销量高的月份,许多儿童在交通事故中受伤;而在冰淇淋销量少的月份中事故次数也少。基于这样的规律,我们可以说冰淇淋销量和儿童事故次数这两个变量是统计上相关的,但这绝不等于说该关系是因果关系。冰淇淋销量高的月份受伤的孩子也多这一事实并不意味着冰淇淋的销量能导致孩子在交通事故中受伤或儿童由于受伤而喜欢吃冰淇淋。这种关系被称为伪关系(spurious relationship),它可以通过加入某一个新变量比如温度而诡辩过去。在天较热时,孩子们吃冰淇淋就多一些,而



孩子们的冰淇淋消费量与受伤率之间的关系是伪关系。(来源: First Imagewest, Inc.)

这种月份也正是他们放假和到处乱跑的月份,而更多的人在路上开车。在这个例子中,两个变量之间很明显是没有因果关系的,可是在许多情况下两个变量之间是有因果关系还是仅为统计关系是很难分清的。

鸕能带来孩子吗?



一个典型的伪关系的例子就是鸕的数量与丹麦乡间婴儿出生率的关系:在鸕数量多的地方婴儿出生率高,鸕数量少的地方婴儿的出生率低。尽管在统计上这两个变量是相关的,但是它们之间并无因果关系。不过这个统计相关到也许可以解释为什么“鸕能带来孩子”的迷信说法会这么流行。(来源: Stock Montage, Inc.)

表 8.1 两个变量在 10 个个体上的观测数矩阵

人	性别	活动偏好
1	M	W
2	F	W
3	M	S
4	F	S
5	F	S
6	M	W
7	F	W
8	M	S
9	M	W
10	F	S

表 8.1 给出了习题 9.41 中的一小部分数据以说明我们如何试图回答前面的 4 个问题。这个表中的数据描述了男性或女性喜欢工作还是社会生活。每一行代表一个人,每一列是一个变量。每个人都用编号代替姓名作为标识,性别列中 M 代表男性, F 代表女性,活动偏好列中 W 代表工作, S 代表社会生活。现在我们试着用表中的数据回答本章开头的 4 个问题。

问题 1. 变量间有关系么?

表 8.1 的数矩阵中,性别和活动偏好这两个变量间有关系吗? 性别变量的某些值是不是对应于活动偏好变量的某些值? 只需观察一下我们就能发现一种模式: F 倾向于与 S 对应, 而 M 倾向于与 W 对应。这也就是说女性可能偏好社会生活而男性则更喜欢工作。

像处理大多数数据那样,在寻找变量关系之前我们先要简化和初步处理这组性别/活动偏好数据。像第三章中讨论的那样,可以做图表或计算统计量。这里把数据安排在一个 2×2 的表中(表 8.2)是很方便的。从表中我们可以清楚地看到一个变量的某个值与另一个变量的某个值之间的对应情况。女性倾向社会生活而不是工作,男性则倾向于工作而不是社会生活。以这个模式看来,可以说这个数据集中的两个变量是有关系的。

表 8.2 在活动偏好和性别两个变量上 10 个人的分布情况

		性别		总计
		女性	男性	
活动偏好	工作	2	3	5
	社会生活	3	2	5
	总计	5	5	10

表 8.3 表 8.2 中数据的不同强度

(a) 强度 = 0.20			
2	3	5	
3	2	5	
5	5	10	
(b) 强度 = 0.60			
1	4	5	
4	1	5	
5	5	10	
(c) 强度 = 1.00			
0	5	5	
5	0	5	
5	5	10	

问题 2. 关系的强弱程度?

性别与活动偏好间的关系是强还是弱? 这个关系的强度显然应该用某种统计量来标识。从(表 8.2 或 8.3a 的)左下到右上的对角线上的人数比别的对角线上的多。当然如果 5 位男性都喜欢工作而 5 位女性都喜欢社会生活,即女性那一列是 0 和 5 而男性那列是 5 和 0(见表 8.3c),则变量间的关系就会强得多。如果女性列是 1 和 4 而男性列是 4 和 1(见表 8.3b),那么这个关系就没有上一种情况中那么强了。

后面的几章将要讨论到如何计算表示变量之间关系强度的各种系数。这样一个系数是个统计量,它的取值范围是从 0 到 1。当系数为 0 时,两个变量间没有关系;当系数为 1 时,关系达到最强。就拿表 8.2 中的数据来说,它的关系强度系数是 0.20。类似地,表 8.3b 中的强度系数是 0.60,而表 8.3c 中的强度系数是 1.00。按照从 0 到 1 的尺度,目前该统计量的观测值是 0.20,这表明两个变量间的关系是比较弱的。

问题 3. 变量在总体中的关系如何?

数据中所反映的关系是不是只是一种偶然? 如果它不是偶然现象, 那么我们可以从样本推广到总体, 并作出总体也存在关系的结论。

想像表 8.2 放大成一个飞镖的靶子, 它有两行两列。为了简单起见, 我们保持同样的男性和女性个数, 和同样的工作和社会生活倾向数。然后我们随机朝该表投掷 10 枚飞镖, 我们是否单凭这种随机扔就能得到与表 8.2 类似的模式? 还是表 8.2 中的数据中含有一些非随机的东西? 如果我们从数据中能找到一个非随机因素, 我们就可以由此下结论说两个变量间的关系不仅存在于样本中, 而且仍然存在于产生样本的总体中。

为了弄清性别/活动偏好的数据分布状况是否只是由偶然因素引起的, 先让我们建立一个认为两个变量无关的零假设, 然后来看看数据能否导致拒绝这个零假设。我们不必相信两个变量间一定不相关, 事实上, 我们想要表明它们的确相关。

如果变量没有关系的零假设被拒绝, 我们就认为变量间确有关系。

我们通过计算观测数据的 p 值来检验没有关系的零假设, 如果 p 值很小就拒绝零假设。但是表 8.2 中数据的 p 值太大了, 无法用它来拒绝零假设; 因此出现这种数据可能是偶然的。在这个例子中我们无法拒绝该例总体中无关的零假设——这并不奇怪, 因为样本太小了。如果样本大一些(1000 或甚至 100 也行), 我们拒绝零假设的可能性就会更大。

问题 4. 是因果关系吗?

一个人的性别是否以某种方式决定着喜爱的活动? 仅从获得的有关这两个变量的数据来看, 我们无法回答这个问题。我们还需要其它相关变量的数据——即使这样我仍有可能无法判断在性别和活动偏好之间的关系是因果关系还是其它变量的副产品。

停下来想一想 8.1

给出了下面的统计信息:

- 在抽烟与否和患肺癌与否两个变量之间的关系强度是 0.53。
- 怀孕期妇女的饮酒量与婴儿出生体重之间的关系强度是 0.34(喝酒越多的妇女所生孩子的体重越轻)。
- 纳税者年龄和他们缴纳税款的数量之间的关系强度是 0.32(年龄越大的纳税者缴纳的税款越多)。

你如何根据每一句话提供的信息, 回答有关因果关系的问题 4? 若要更自信地回答问题, 你还希望知道什么信息?

8.2 预测: 从一个变量到另一个变量

数据中两个变量间是否存在关系是与预测(prediction)紧密相连的。假设我们知道样本中

的一个观测是位妇女。在表 8.2 提供的信息的基础上我们能否预测这位妇女喜欢如何度过时间? 知道那个人是位妇女就意味着我们只需要用女性那一列的 5 个人的信息而不是表上所有 10 个人的信息。既然有三分之二的女性喜欢社会生活, 我们就预测这位妇女也喜欢社会生活。如果我们对样本中所有 5 名女性做同样的预测, 我们不会每次都正确, 但是正确的次数肯定比不正确的次数多。的确, 在这 5 次预测中我们将对 3 次错 2 次。预测样本中男性的活动也是同样的道理。对男性做“喜欢工作”的预测, 我们将在 5 次预测中对 3 次。

用一个变量的值预测另一个变量的值时它们之间不必非得有因果关系。

如果个体的两个变量之间存在某种关系, 我们就可以用个体中一个变量的值所提供的信息来预测的另一个变量的值。在性别/活动偏好的例子中, 如果我们知道一个人的性别, 就可以预测他(她)的活动偏好。我们的预测不可能总是正确的, 但是两个变量间的关系越强, 预测正确的可能性就越大。所以关系的强度就代表了我们用一个变量去预测另一个的可行程度。

停下来想一想 8.2

为什么你可以用两个相关变量之一的值去预测另一个的值, 即使它们之间并无因果关系? 为什么可以用冰淇淋的销量来预测儿童交通事故的发生率? 你认为可以用郊区滑雪机的销量来预测城市单元房火灾的发生情况吗?

8.3 自变量和因变量

对于性别/活动那个例子中的两个变量来说有一些东西是不对称的, 如果变量间有因果关系, 我们将会说性别变量影响活动偏好变量。类似地在植物的例子中, 我们只能说是光线影响产量。我们不能认为一个人的活动偏好好歹会影响他的性别, 也不能认为植物的生长情况影响它在实验中得到的光照量。

以上这两个例子中都是一个变量在另一个变量之前就确定下来了。人们生来就有性别而活动偏好要多年后才能表示出来, 一株植物先要有光照才能生长。在我们研究具有类似关系的两个变量时要注意: 通常有一个变量发生得早, 而且会影响另一个变量。我们把先发生的变量叫做自变量(independent variable), 而受其影响的那个变量则称为因变量(dependent variable)。自变量和因变量有时仍叫做解释变量(explanatory variable)和响应变量(response variable)。我们可以用一个从自变量指向因变量的箭头来表示这种关系:

自变量 \longrightarrow 因变量

有时, 为了简单起见, 用字母 X 来表示自变量或解释变量, 字母 Y 表示因变量或响应变量:

$X \longrightarrow Y$

一般的自变量又被称为 X -变量,而一般的因变量又被称为 Y -变量。(为了便于记忆可以把它联想成:因为有两腿的 X 比一条腿的 Y 强健,所以是 X 影响 Y 。)

停下来想一想 8.3

在下面的几对变量中,哪一个自变量(或 x)哪一个因变量(或 y)?

- 闪电和雷。
- 销售税的总量和商品总成本。
- 电影院里爆米花的销售率和垃圾袋的使用率。
- 发电量和热天的天数。
- 节目的广告时间和城中水的消费量。

8.4 不同类型的变量:分类型变量、顺序型变量和数量型变量

我们用哪种数据分析方法是由两个变量的属性决定的。现在已经发展了多种统计方法以适应不同类型的变量。

因变量和自变量都可以是下面三种类型之一:

- 分类型变量(categorical):** 它的值是非数量的范畴;例如对于性别变量,它的值就是男和女。
- 顺序型变量(rank):** 它的值是有序的;例如对态度变量,它的值就是反对、中立和赞同;对比赛名次变量,它的值是第一、第二和第三。
- 数量型变量(metric):** 它的值是可以作数学计算(加、乘)的有意义的数值;比如收入,重量,年龄等。

自变量和因变量不一定是同一类变量。所以自变量和因变量就有 9 种(3×3)可能的组合关系。表 8.4 列出了两个变量之间可能出现的关系。在这个表中自变量 X 被标在水平方向上,因变量 Y 则在竖直方向上被标出。自变量类型从左到右的顺序是从简单的分类型变量到复杂的数量型变量。因变量从下到上则逐渐变复杂,即从分类型变量到数量型变量。

表 8.4 变量类型可能的组合对

		自变量 x		
		分类型变量	顺序型变量	数量型变量
因变量 y	数量型变量	D(第 12 章)		B(第 10 章)
	顺序型变量		E(第 11 章)	
	分类型变量	A(第 9 章)		C(第 10 章)

性别/活动偏好的例子属于该表左下角的情况(A),此因两个变量都是分类型变量;它们

的关系将在第9章被讨论。光和植物生长的例子属于右上角的情况(B),此因两个变量都是数量型变量;这种关系将在第10章被提及。研究一个国家的识字率和其政府类型之间的关系属于右下角的情况(C),此因因变量是识字率—数量型变量而自变量是政府类型—分类型变量;它也将第10章讨论。研究三种不同的教学方法(分类型自变量)和学校成绩(数量型因变量)之间的关系,它的类型是左上角的(D)—相当常见的一种情况—将在第12章被讨论。

一般情况下很少出现顺序型变量,本书仅考虑一种两个变量都是顺序型变量的情况(E)。例如在两年中篮球队的排名情况;这个关系将会在第11章出现。如果在某一研究中,一个顺序型变量和一个数量型变量的组合出现,通常把它们作为两个顺序型变量来处理。类似地,当出现的组合是分类型变量与顺序型变量时,通常把它们看作两个分类型变量,不过这时我们已经损失了部分信息。(注意:总可以把数量型变量改换成顺序型变量或分类型变量,但反过来不行。例如,收入水平就可以变成贫穷、一般和富裕的顺序型变量。但是贫穷、一般和富裕确不能变成数量型的收入水平,除非有精确的收入信息作为转换的基础。)

研究两个变量之间的关系时必须先确定它们的组合属于表中哪一类,因为对不同的组合有不同的统计分析方法。两个数量型变量含有的信息量比两个分类型变量所含的要多得多,许多统计方法虽然不能处理其它类型的组合但却很适合处理两个数量型变量的组合。当不可避免地遇到分类型变量和顺序型变量时,有一些适当的方法可以从中提取尽可能多的信息。

停下来想一想 8.4

下面是一些变量对,请判断它们是表8.4中的哪一种情况?

- 平均温度和欧洲各主要城市公共场所的清洁程度。
- 少数民族的划分和吉隆坡的社会阶层。
- 日本的神户与京都之间的火车平均车速和每公里车票的平均价格。

8.5 回到因果关系的问题

本章第一节中的问题4问的是自变量是否能影响因变量。因果关系是一个困难的概念,哲学家们已经为了它的确切意义争论了几百年。我们不在这里解决因果关系问题,统计学家们自己也不能确定一个变量是不是另一个变量的原因。但统计方法却可以为这个复杂的问题提供一种观点。我们将要在第14章就一些判断变量之间是否有因果关系的途径做详细的讨论。

我们已经注意到了,在统计分析中发现两个变量之间有关系并不等于证明它们之间有因果关系。我们可以猜测基因中的某种染色体在多大程度上决定了一个人的性别,但是同样用一个人的生物组成来判断他喜欢如何打发时间就行不通了。甚至我们根本不能相信冰淇淋的月销量与孩子当月受伤人数之间会有因果关系。我们可以猜测文化与政府类型之间超出了统计的关系,还可以断定采光量和植物生长之间存在着因果关系。

别的变量的角色

我们在这章例子中所做的推测好像没有基于任何统计方法；它们仅仅是建立在我们平时对变量的了解的基础上的常识判断。当两个变量间好像没有因果关系时，我们就假定观测到的相关是由另一个变量引起的。这样在我们判断两个变量是否有因果关系之前，应该先看一看是否有其它变量能引起观测到的这个关系。例如在性别/活动偏好的例子中我们应该先问一问诸如怎样进行社交等问题是否也会有影响？再比如，在冰淇淋和孩子受伤的例子中气温似乎可以解释观测到的关系；夏天温度高，孩子们放假，因此可能容易卷入车祸中，同时他们也因为气温高而吃更多的冰淇淋。

图 8.1 是冰淇淋和车祸例子的图示。从温度指向两个变量的箭头代表了温度是这个关系的起因。冰淇淋销量和车祸次数的数据显示了它们是统计相关的，但是这个关系是伪关系，因为我们知道它们之间并无因果关系。如果例子中提供了温度的数据，我们就可以用多元统计方法来证明冰淇淋销量和车祸次数之间的关系确是伪关系。这个方法将在第 13 章中介绍。

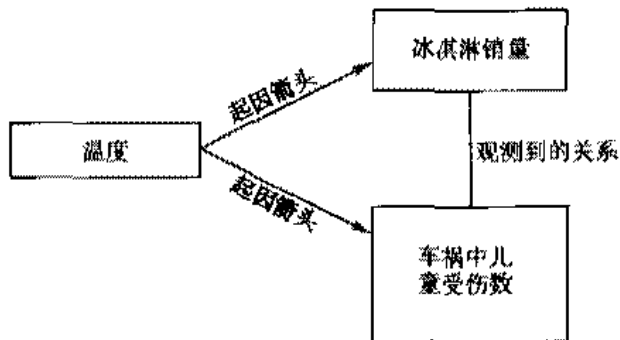


图 8.1 观测到的两个变量间的伪关系是由背后的第三个变量引起的。

时间的角色

两个变量之间有因果关系的一个前提是自变量发生在因变量之前。光线先照到植物上，然后植物才能生长。另外一个前提是自变量是因变量的起因，因变量会随着自变量的改变而改变。如果我们改变光照时间的长度，则植物生长情况也随之发生了变化，我们就有理由说光线影响植物的生长。反之就是伪关系即无因果关系；我们不能期望一个变量的改变会引起另一个变量的变化。我们认识到“冰淇淋销量的变化能引起车祸的变化”是荒谬的。

有无因果关系的不同更突出了实验型数据与观测型数据的差别。实验型数据可以抵消掉其它变量的影响，同时还可以通过控制自变量来看它是否对因变量产生影响。而对于观测型数据我们则没有办法去操纵它。所以基于试验型数据比基于观测型数据容易确定变量之间的因果关系。处理观测型数据时，我们总是要对付其它变量的影响。

一个成功地从观测型数据中确立了因果关系的著名例子就是你经常在香烟盒或广告上看到的“吸烟有害健康！”。这个结论以各种形式表达出来，警告吸烟可能引起各种健康问题。如果研究者可以随机抽取新生婴儿，并自由地安排他们成为吸烟者或不吸烟者，那么多年以后再

来研究他们的健康状况时就可以得到试验型数据,并以它为基础进行分析。因为这种试验是不可能的,于是医疗总监出色地做了这项工作;他请求统计学家研究观测的有关吸烟者和不吸烟者的健康数据。基于能够得到的证据,大多数人已经接受了吸烟与健康问题之间有因果关系这一结论。



尽管已发表了大量的关于吸烟的危害研究,但迄今为止仍然没有精确地试验分离出诸如吸烟是否能引起肺癌这类因果因素。

多元因果关系

在日常生活中,几个变量间的因果关系往往是复杂的。常常是多个变量而不是一个来决定某个因变量的。例如,一个人的薪水会受到工作性质、所需的训练、需要的工作经验、本人的才能、当时情况、甚至性别和种族等现实因素的影响。而当我们仅考虑一个变量提供的影响时我们不能期望它是影响因变量的唯一的原因。

最后,即使两个变量之间有因果关系,这个关系也不一定在每一次事实中都发生。例如不是所有的吸烟者都有健康问题,同时吸烟者的健康问题也不都是由吸烟引起的。这个关系只是对某些人成立。为了更加全面地了解吸烟者的健康问题,还应该考虑除吸烟之外的其它因变量。

8.6 小结

这一章讨论的是两个变量之间的关系。数据的来源主要有两个：试验和观测。

8.1 两个变量的4个问题

得到两个变量的数据之后需要回答4个问题：

问题1. 从数据看来两个变量之间有关系吗？

问题2. 如果有，有多强？

问题3. 总体中是否有和样本中一样的关系？

问题4. 观测到的这个关系是因果关系吗？

要回答问题1，就先要寻找样本数据中的关系模式。如果样本中确有关系，我们再来回答问题2。在回答问题2时我们需要计算一个系数来测量这个关系的强度。系数接近1表示关系强，而系数接近0表示关系弱。回答问题3时，我们应先建立一个“两个变量之间没有关系”的零假设，看能否拒绝该假设。我们计算观测数据的 p 值，如果 p 值很小就拒绝零假设。

要回答关于因果关系的问题4往往是很困难的。没有因果关系的两个变量之间也会有非常强的统计关系。试验型数据往往比观测型数据更容易确定是否有因果关系，因为在试验中可以控制其它的变量对结果的影响。

8.2 预测：从一个变量到另一个变量

即使两个变量之间并无因果关系，我们仍然可以用一个变量的某一个体的观察值去预测另一个变量的相应个体的观察值。关系的强度表明我们用一个变量去预测另一个变量的可行性有多大。

8.3 自变量和因变量

在研究变量之间的关系时，通常把变量分成两种：自变量和因变量。如果变量间有因果关系，那么原因变量就叫做自变量，而受自变量影响的变量就称为因变量。自变量通常发生在因变量之前。（不是所有先发生的变量都是自变量。）一个一般的自变量记为 X -变量，而一个一般的因变量记为 Y -变量。

8.4 不同类型的变量：分类型变量、顺序型变量和数量型变量

自变量和因变量都可以三种类型之一。分类型变量是那些值为两个或多个类型的变量，例如性别变量的取值就是男性和女性。顺序变量是那些按照从低到高的顺序取值的变量，例如比赛结果。凡是其值可以进行数学运算（加、乘等）的变量都称为数量型变量。收入、体重和年龄都是数量型变量的例子。对不同类型的变量组合有不同的统计分析方法。

8.5 回到因果关系的问题

为了判断自变量与因变量之间的关系是否为因果关系(一旦两个变量之间的关系建立在总体中),我们应该:(1)用常识来判断这种关系是否在我们所知道的世界上有意义;(2)注意自变量是否发生在因变量之前;(3)如果可能,适当更改自变量,并观察因变量的值是否会受影响(也就是做个试验);(4)即使自变量是决定因变量的一个原因,也要认识到其它没被考虑进计划的其它重要变量可能影响因变量。

补充读物

Davis, James A. *The Logic of Causal Order* (Sage University Paper Series on Quantitative Application in the Social Sciences, series no.07-055). Beverly Hills, CA: Sage, 1985. 一本小的关于因果关系的平装书。

Liebetrau, Albert M. *Measures of Association* (Sage University Paper Series on Quantitative Application in the Social Sciences, series no.07-032). Beverly Hills, CA: Sage, 1983. 用不同方法度量两个变量关系强度的入门。

Simon, Herbert A. "Causation." In William H. Kruskal and Judith M. Tanur (eds), *International Encyclopedia of Statistics*. New York: The Free Press, 1978. 一个权威在因果关系和统计之间联系上的观点。

习题

回顾(习题 8.1—8.23)

- 8.1 用平常的活来解释“两个变量的统计关系”的含意。
- 8.2 在研究两个变量时,关于收集的数据,研究者提出哪 4 个问题?
- 8.3 在回答问题 1 时,如果在样本数据中没找到两个变量之间的关系,你会对产生样本的总体做出什么判断?
- 8.4 为什么判断样本中是否有关系以及关系的强弱是很重要的?
- 8.5 请解释问题 3“它们在总体中也有同样的关系吗?”的意思。
- 8.6 为什么用实验型数据比用观测型数据更容易判断出变量之间是否存在着因果关系?
- 8.7 “两个变量之间的伪关系”是什么意思?
- 8.8 历史上,妇女裙子的长度与经济的好坏有关系:裙子越短,经济越景气。请问这个关系是因果关系还是伪关系?
- 8.9 统计学家们试图通过分析可能与问题中的变量有因果关系的其它变量来解决伪关系问题。
 - a. 有统计知识的人愿意你认为两个变量之间的一个关系是伪关系,主要原因是什么?

- b. 你为什么认为冰淇淋的销量与儿童出事故的次数之间是伪关系?
 - c. 在判断变量之间的关系是否是伪关系时如何会犯错误?
 - d. 请想出一个历史环境,在其中本来是因果关系却被误认为是伪关系或者反之?
- 8.10 讨论下面的关系是因果关系还是伪关系?如果不能决定,为什么?
- a. Summons 和 Blyth(1987)年发现早熟的女孩会有以下问题:身体形态失调、学术成就低以及在学校制造麻烦。
 - b. 街上警察数量增加使得犯罪数量也增加。
 - c. 高中生中,有规律吸毒者比那些不吸的人的成绩要低。
- 8.11 研究者们通过计算一个系数来判断两个变量之间关系的强弱程度。作为例子请给出一个表示关系强的值和一个表示关系弱的值。
- 8.12 如果你被告知两种不同形式的测验之间的关系系数是 0.96,你会怎样看待这两种测验?
- 8.13 如果一个研究者能够拒绝零假设,那么他可以就两个变量之间的关系做出什么结论?
- 8.14 当 p 值是多大时拒绝零假设?
- 8.15 “如果你不能证明数据间有因果关系,你就不能用自变量的值来预测因变量的值”这种看法对吗?为什么?
- 8.16
- a. 什么叫自变量?
 - b. 什么叫因变量?
 - c. 它们分别用哪个字母表示?
 - d. 你打算如何记住这些?
- 8.17 在研究两个变量之间的关系时,你是依靠什么因素来判断谁是自变量,谁是因变量的?
- 8.18 为什么用观测型数据来确定是否两个变量之间的关系为因果关系是很困难的?
- 8.19 为什么在研究两个变量之间的关系时先弄清楚它们属于哪一类变量(分类型变量、顺序型变量和数量型变量)是很重要的?
- 8.20 判断下列变量是什么类型的变量(分类型变量、顺序型变量和数量型变量)。
- a. 宗教
 - b. 全国橄榄球联盟(NFL)名次表
 - c. 高度
 - d. 马在肯塔基赛马会上的定位
 - e. 雪橇比赛中的铜牌、银牌和金牌得主
 - f. 参加奥林匹克队的资格
 - g. 一周中的天数
 - h. 年龄
- 8.21
- a. 请给出自变量是分类型变量,因变量是数量型变量的例子。
 - b. 请给出所研究的两个变量都是数量型变量的例子。
 - c. 请给出所研究的两个变量都是顺序型变量的例子。
- 8.22 如果要研究的两个变量一个是数量型变量而另一个是顺序型变量,你应该如何分析它们之间的关系?
- 8.23 用于分析多于两个变量的方法的一般名字是什么?

解释(习题 8.24—8.33)

- 8.24 全国民意调查中心于 1972 年和 1991 年所做的调查表明一个人对其工作的满意程度和他从事该工作的年限有关系。工作时间越长越喜欢。表 8.5 给出了两组共 270 人对工作的喜爱程度和干该工作的年限之间的关系。
- 表的哪个部分的统计相关性最强?为什么?
 - 找到仅应用于样本的关系和找到应用于产生样本的总体中的关系有何不同?
 - 有时你会发现两个变量在样本中表现出来的关系在总体中并不存在,这是怎么回事?

表 8.5 习题 8.24 的数据

(a)	工作年数			(b)	工作年数		
	10 年以下	10 年以上	总计		10 年以下	10 年以上	总计
高兴	50	100	150	高兴	70	80	150
不高兴	100	20	120	不高兴	80	40	120
总计	150	120	270	总计	150	120	270

- 8.25 请解释下面的话:“在统计分析中发现两个变量之间有关系并不等于证明了它们之间存在着因果关系。”
- 8.26 在研究两个变量之间的关系时,为什么常常把其中一个称作自变量,而另一个则被称为因变量?
- 8.27 在什么背景下烟草公司雇佣的统计学家可以声称目前还没有足够证据能证明吸烟会危害健康?
- 8.28 怎样才能证明嚼烟会使嚼烟的人得各种癌症?为什么不能真的去做这个实验?
- 8.29 一旦知道两个变量之间存在关系,哪几种主要方法可以确定这个关系是不是因果关系?
- 8.30 为什么一个数量型变量可以被转换成顺序型变量而反之却不行?请用例子来说明。
- 8.31 观测表明吸烟的孕妇所生的婴儿比不吸烟的要小。你能就此下结论说吸烟会降低婴儿体重吗?讨论之?
- 8.32 在鹤与婴儿出生率的图中,人们可以用一个地区鹤的数量来相当准确地预能出该地区的婴儿出生率。说明为什么在预测一个变量值的时候不必要理解原因变量?
- 8.33 a. 调查研究者认为在北京有 74% 的妇女说她们对自己的身材满意,而在东京却有 84% 的妇女说她们不满意自己的身材。请解释其作出结论的步骤。(来源: *Newsweek*, February 12 1996, p. 41)
- b. 这些研究者认为,由于日本引进了“挑逗性”的西方内衣广告而导致了这个不同。你认为他们所说的原因与观测的结果是因果关系还是伪关系?

分析(习题 8.34—8.40)

- 8.34 本题中,一个变量是参加一个特别教育计划的少年犯人数,另一个是与警察的接触。如果我们知道一个人参加了该计划,我们能不能由此推测出这个人是否与警察有更多的接触?表 8.6 给出了 100 个少年犯的数据。

表 8.6 习题 8.34 的数据

		少年犯参加特别计划的状况		总计
		在其中	不在其中	
与警察的更 进一步接触	无	37	20	57
	有	13	30	43
	总计	50	50	100

来源: T. Hirschi and M. J. Hindelang, "Intelligence and delinquency: A revisionist review" *American Sociological Review*, vol. 42(1977), p. 575

- 假设你经介绍认识其中一个少年犯,但你不知道他是否已参加该特别计划。看一看表 8.5,为什么你的最好的预测是认为这个少年犯与警察没有更进一步的接触?
- 如果你预测这 100 个人与警察都没有更进一步的接触,你将预测错多少个?
- 如果你被告知一个人已参加该特别计划,为什么你的最好预测为此人与警察没有更进一步的接触?
- 如果你打算预测参加该特别计划的 50 个人与警察都没有更进一步的接触,你将会有几次错误预测?
- 如果你被告知一个少年犯没参加该特别计划,你的最好预测为此人与警察有还是没有更进一步的接触?
- 如果你打算预测没有参加该特别计划的 50 个人与警察有更进一步的接触,你将会有几次错误预测?
- 为什么在 d 和 f 的答案的总和等于当是否参加计划是已知时作出的错误预测的总和?
- 为什么在了解了参加计划状况之后再进行预测就会少犯 10 次错误?
- 为什么 $10/43 = 0.23$ 这个比率可以作为一个在 0 和 1 之间变动的标准来判断预测精度的提高程度?

8.35 在因首次犯心脏病而住进新英格兰医院的 665 名男性病人中有 214 位病人秃顶。在另外 772 名不是因为心脏病而住院者的病人中有 175 人有类似的秃顶。(来源:1993 年 2 月 14 日的纽约时报 pp. A1, C12.)

- 在一个表中表示这些数据。
- 在那些因首次犯心脏病而住院的人中秃顶症患者所占的百分比是多少?
- 那些非因心脏病住院的病人中患秃顶的人的比例是多少?
- 这两个比例之间的不同给你提供了什么信息?
- 比较心脏病患者中患秃顶的人和没患秃顶的人的比例差异,以及比较心脏病患者和未患心脏病的人中秃顶者的比率差异。为什么我们对前者更有兴趣?换句话说就是我们对表中两行的比例差异感兴趣呢,还是对两列的比例差异感兴趣?
- 尽管 d 中的百分比可以通过数据计算出来,为什么它没有多大的意义?
- 这个数据是否表明秃顶能引起心脏病?

8.36 在底特律市区随机采访了 190 人并收集对“工作是很重要的而且给人们以成就感”这个看法的意见。用它的来研究宗教派别(浸礼教和卫理公教)与人们回答“赞同”或“不赞同”的关系。

表 8.7 习题 8.36 的数据

对问题的回答		宗教派别		总计
		浸礼教	卫理公教	
	是	36	51	87
	否	67	36	103
	总计	103	87	190

来源: H. Schuman, "The religious factor in Detroit: Review, replication and reanalysis," *American Sociological Review*, vol. 36 (1971), pp. 30-48

- 该项研究的两个变量中,谁是自变量,谁是因变量?
 - 从数据来看,这两个变量之间有关系吗?
 - 这两个变量的关系是强还是弱?
 - 有关底特律这两个教派的不同,你能从这个表中获得什么信息?
- 8.37 在习题 6.45 中你看到了年轻吸烟者中谁抽 Newport 牌烟。在这里我们还给出了其它几种年轻人喜欢的香烟。表 8.8 包括了据说在两组吸烟者中占的比例超过 10% 的三种牌子的香烟。

表 8.8 习题 8.37 的数据

香烟的牌子		种族		总计
		黑人	白人	
	Marlboro	4	576	580
	Newport	25	45	70
	Kool	4	5	9
	其它	12	181	193
	总计	45	807	852

来源: *Teenage Attitudes and Practices Survey*, 1989, by the National Center for Health Statistics, as reported in *Chance*, vol. 5 (1992), nos. 1-2, p. 27

- 在进行计算之前,你能从这个表中看出什么?
 - 把每一列换成百分数,你能从中看出黑人和白人之间有什么不同吗?
 - 表中体现的两个变量之间的关系是强还是弱?
 - 对于这三种香烟如此吸引年轻烟民的事实你是怎样看的?
- 8.38 表 8.9 给出了婴幼儿看护收费情况的数据矩阵,单位是美元/小时。
- 应该如何安排这些数据才能更好地体现出不同看护类型之间收费的差异?
 - 在看护类型和收费高低之间有关系吗?
 - 如果你认为有关系,你认为这个关系强吗?
 - 你认为这个结果纯粹是个偶然呢? 还是总体中可能存在着类似的关系?
 - 通过看护类型来预测收费数目的效果如何?
 - 它们之中哪一个是自变量,哪一个是因变量?

表 8.9 习题 8.38 的数据

婴儿	看护类型	收费标准(美元/小时)
1	亲属照管	4.90
2	雇佣保姆	7.00
3	亲属照管	5.00
4	全日制托儿所	6.60
5	私人家庭照管	5.35
6	雇佣保姆	7.50
7	私人家庭照管	5.50
8	全日制托儿所	6.75
9	亲属看护	5.25
10	私人家庭照管	5.15
11	雇佣保姆	7.55
12	全日制托儿所	6.67
13	亲属照管	5.10
14	私人家庭照管	5.35
15	雇佣保姆	7.40
16	全日制托儿所	6.75

来源: Sandra L. Hofferth, Urban Institute

8.39 本习题中一个变量是年(从 1960 年到 1995 年),另一个变量是 6 岁以下孩子的母亲中上班者的百分比。你能用年代的数据来预测一个 6 岁以下孩子的母亲是否在上班吗?

表 8.10 习题 8.39 的数据

	1960	1965	1970	1975	1980	1985	1990	1995
上班	20	25	32	38	47	52	58	58
不上班	80	75	68	62	53	48	42	42

来源: Bureau of Labor Statistics

- 假定你被引见了这些母亲之一,并且你不知道她是否上班。观察该表,然后解释为什么你的最好的预测是她没有上班?
- 假设每 5 年中,这类妇女上班的比例是不变的。如果你认为所有 6 岁以下孩子的母亲都不上班,那么你的判断出错的百分比是多少?
- 如果你已知一位妇女在 1960 年时有一个 6 岁以下的孩子,为什么你的最好的预测是她没有上班?
- 如果你预测所有在 1970 年样本中的母亲都不上班,那你会预测错几回?
- 现在你已知一位妇女在 1995 年时有一个 6 岁以下的孩子,根据这个信息,你是应该认为她上班呢? 还是应该认为她不上班?
- 如果你认为 1995 年时所有 6 岁以下孩子的母亲都上班,你出错的次数是多少?
- 当你已知了有 6 岁以下孩子的母亲所在的年代,你可以用什么方法来提高推断的准确性,从而少犯错误?
- 是否有迹象表明在有小孩子的母亲们中间,出去工作的人的比例是否在发生变化?

你的证据是什么?

1. 年代与工作百分比这两个变量之间有因果关系吗?

表 8.11 习题 8.40 的数据

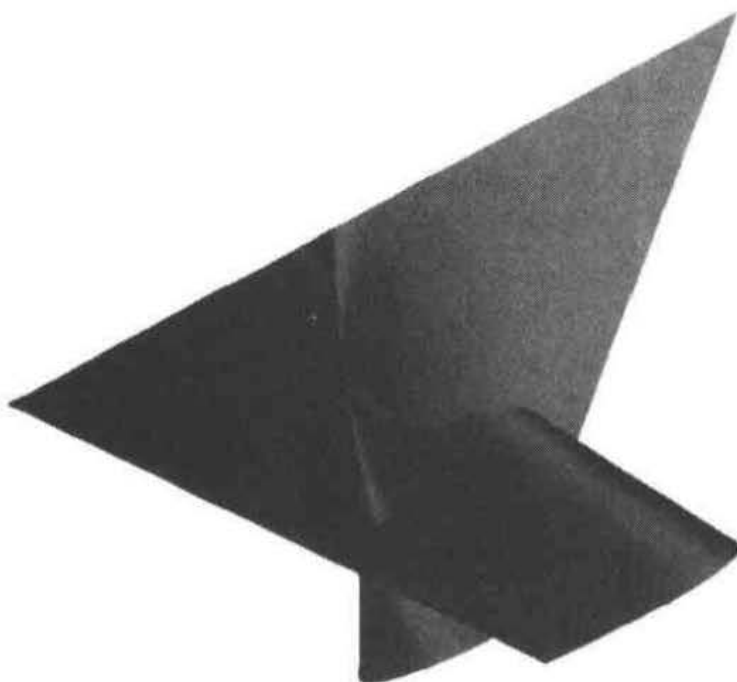
		说话者的态度		总计
		否定	肯定	
回答者 的态度	否定	444	181	625
	肯定	435	679	1114
	总计	879	860	1739

来源: V. L. Walsh et al., "Impact of message valence, focus, expressive style, and gender on communication patterns among maritally distressed couples," *Journal of Family Psychology*, vol. 7(1993), pp. 163-175

8.40 表 8.11 的数据被用来研究婚姻不幸福的夫妻之间的交流模式。问题是这样的:当配偶之一以肯定或否定的语调来叙述一件事时,另一位反应是肯定还是否定?所以一个变量就是说话者的态度,其值为肯定或否定。另一个变量是回答者的态度,它的值也是肯定或否定。这个数据来自 52 对夫妻总共 1739 个评论。

- a. 假设你选了一个说话者的评论但是你不知道他的态度,那么你最好认为回答者的态度是哪一类?
- b. 如果你断定每位回答者的态度都是否定的,你的判断出错的次数将会是多少?
- c. 如果你被告知说话者的态度是否定的,为什么对回答者态度最好的推断是否定?
- d. 如果你打算预测回答者对所有否定语调的评论也以否定态度回答,你的推断会出多少次错?
- e. 现在你被告知说话者的态度是肯定的,你认为回答者的态度应是什么样的?
- f. 如果你打算预测回答者对所有肯定语调的评论也以肯定态度回答,你的推断会出多少次错?
- g. 为什么 d 和 f 答案的总数等于当你知道了说话者的态度后作出错误预测的总数?
- h. 从数据上来看怎样才能使谈话愉快和不愉快?
- i. 从这组数据来看,你认为在归纳这些婚姻不幸福的夫妻之间的谈话模式时会有什么困难?
- j. 你认为在选修统计课的高校学生中总结这一关于对话的结果时会有什么潜在的困难?

C H A P T E R 9



9.1 数据分析：在态度上有可靠的差异吗？

9.2 问题 1. 变量间的关系？

9.3 问题 2. 关系的强度？

9.4 问题 3. 总体中的关系？

9.5 问题 4. 是因果关系吗？

9.6 更大的表：更多的可能性

9.7 小结



两个分类变量

的 χ^2 分析



不同的国家的人们用同样的眼光来看待陌生人吗？在欧洲的几个国家的抽样调查过程中，人们经常被问到下面的问题：“一般说来，你是同意大多数人都是可信赖的呢，还是认为和人们相处时再怎么小心也不过分？”

列联表是一个描述两个分类变量分布的频率表。

让我们来比较两个国家，丹麦和法国（表 9.1）。表 9.1 称为列联表(contingency table)。这个表描述了这次民意测验中人们所在国家和对他人的态度这两个分类变量上的分布情况。因为这两个变量都有两个取值，所以这个列联表有两行和两列（还有一行一列是表示总和的）。国家这个变量有量个值：丹麦和法国，态度变量也有两个取值：信任和怀疑。当然，分类变量也有可能多于两类或者两个取值。如果有四个国家，这个表将含有四列、两行。

一个列联表描述的是频率(frequency)，也就是说不同类别中元素的个数。分析列联表的数据需要分几步。首先分别考虑各个变量。对于表 9.1 中国家这个变量，我们看到有 985 个丹麦人和 969 个法国人。对于态度变量，有 831 个人认为大多数人是可信赖的，有 1123 个人认为无论你多么小心都不为过。其次，要断定这两个国家之间是否有差异，我们同时考虑这两个变量。我们发现共有 625 个丹麦人认为大多数人是可信赖的，而有略多于这个数目一半的人则认为无论多么小心都不为过。对于法国人，有 206 个人认为大多数人是可信赖的，有 763 个人则认为无论多么小心都不为过。

所以,大多数丹麦人持信赖的态度,而大多数法国人则持怀疑态度。这样粗略地看一下这张表,我们就获得了被调查者态度的一些信息。也许丹麦人和法国人之间确实存在着某些差异。

表 9.1 国家和人们对他人的态度

		国家		总计
		丹麦	法国	
对他人 的态度	信任	625	206	831
	怀疑	360	763	1123
	总计	985	969	1954

来源: Jacques-Ren'e Rabier, Helen Riffault, and Ronald Inglehart, *Euro-barometer 25: Holiday Travel and Environmental Problems*, April 1986, Ann Arbor, MI: Inter-University Consortium for Political and Social Research, 1988. Codebook p.10.

停下来想一想 9.1

举出一个有两个可能相关的分类变量的例子并为这两个变量构造一列联表。你所举的变量为什么是分类变量?

9.1 数据分析:在态度上有可靠的差异吗?

在这个分析中,哪个是自变量哪个是因变量呢?在这个样本中,至少对于本国出生的人来说,很明显国籍在先,而后才会形成关于信赖的态度。所以,我们选择国籍作为自变量,而态度作为因变量。

注意在表 9.1 中,国籍这个自变量是水平排列的,而因变量(态度)是垂直排列的。水平安排自变量和垂直安排因变量是构造列联表的常用方法。这种方法和我们在第三章中和以后几章中安排其他类型变量数据时所采用的方法是一致的。

停下来想一想 9.2

Phi Beta Kappa 学生分会准备请一位演说者到学校来。委员会缩小了选择的范围,即在两个演说者中选一个。一个是一位浪漫体裁语言教授,他的主题是优雅爱情的赞美诗;另一个学者的主题是 Lacanian 心理分析及痛苦的政治。该委员会拿不定主意到底请哪一位,所以就征求整个社团成员的意见。在社团的高级成员中,14 个人投票赞成赞美诗的讲座,有 7 个人赞成另一个。在初级成员中,有 10 个赞成讲 Lacanian 讲座,5 个则选择优雅爱情的赞美诗的演讲。

构造一个描述这些投票结果的列联表。这个表告诉你两个变量间什么样的关系呢?

当然这并非一个一成不变的规则。如果自变量只有几类而因变量有许多类,这两个变量则可以倒过来安排,也就是说,将自变量垂直排列而将因变量水平排列。由于行数比列数少,

这样安排可以节约一些空间。

条形图

三维 在第三章中看到,用图要比表格更容易描述数据的总体趋势。表 9.1 中的数据可以用各种各样的图来描述。一个可能的图就是基于条形图的思想得到的。因为要处理两个变量,我们用一个矩形棱柱,而不是一个条,来描述两个变量每个值的观测个数(图 9.1)。

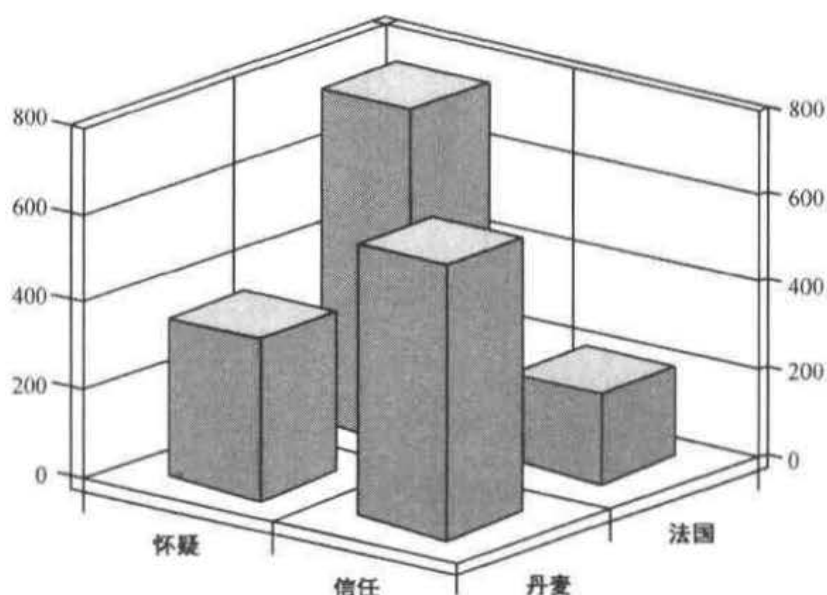


图 9.1 表 9.1 中数据的三维条形图。

作这样一个图要比作一个变量的条形图困难而且不容易看懂。频率最大的一个通常放在图的后而以防它“挡住”其他几个频率。在这里最大的频率代表 736 个认为别人不可信任的法国人。从这个图上我们可以知道,在丹麦人中有更多的人持信任的态度,而在法国人当中有更多的人持怀疑态度。我们还看到,最小的一组是持信任态度的法国人,第二小的是持怀疑态度的丹麦人,第三小的是持信任态度的丹麦人。在一个三维的图上,很难看清楚每个频率等于多少,然而相对大小却很清晰。

相同宽度、不同高度的条 这类的数据也可以用普通的条形图来描述。我们可以把每个国家的两个条迭在一起,也可以并排放在一起(图 9.2)。在图 9.2 的两个图形中,每个条有相同的底,但高度不同。我们一眼就可以看出丹麦和法国被调查的人数是不相等的。在丹麦人中,有更多的人持信任的态度。而在法国人当中,更多的人则持怀疑的态度。在每个图的左边都有一个刻度,但却很难准确地读出每一类有多少人。对图 9.2a 由于持信任态度的人的条形的底不是从 0 开始的,所以很难看出有多少人持信任的态度。而在图 9.2b 中,由于同一个国家的两个条形都是从 0 开始的,所以很难知道每个国家总共有多少个被调查者。正如第三章所注明的,在这些图上我们发现不同的图各有优缺点。

不同宽度、相同高度的条 相同的数据也可以用具有相同高度、不同宽度的条来描述(图 9.3)。每个条形面积表示每组中观测的百分比,当然也可以用频率。我们再次从图上看

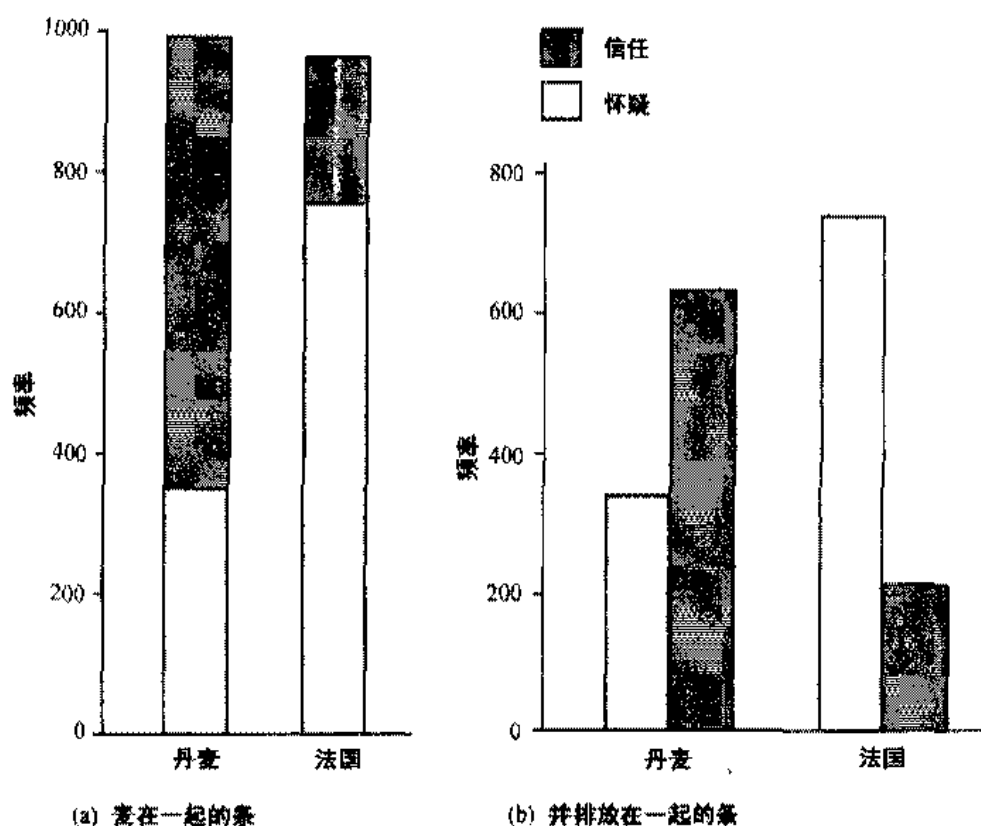


图 9.2 表 9.1 中数据的二维条形图·相同宽度、不同高度的条。

到,在被调查者中丹麦人比法国人稍多一些。在丹麦人当中,大约三分之二的人认为多数人是可信赖的。在法国人当中,大多数人则认为和别人相处应该谨慎小心。条形图的构造将在本章末公式一节中讨论。

到底选择三个图中的哪一个并不是显而易见的事情。图 9.3 的简单明了对于有经验的人来说是很有吸引力的,而对于不习惯于这种描述形式的人来说就并非如此了。

停下来想一想 9.3

图 9.1、9.2 和 9.3 是对于国籍和信任态度数据的三种不同的描述。如果你打算选择其中的一个用于教科书或商业报告,你最喜欢哪一个? 为什么作这样的选择呢?

分类变量的汇总计算

列联表是用来研究分类变量间关系的。对于列联表中的这些数据,我们要问在第八章中谈到的四个问题。(1) 我们想知道在这些数据中,被调查者的国籍和他们对于他人的信任态度间是否有一定的关系。(2) 我们想了解这两个变量间的关系强度如何。一个国家的人们更易于信赖他人这只是一中轻微的偏向呢,还是一种很强烈的偏向?(3) 我们想知道,这次民意调查的结果是否不仅适用于样本而且适用于两个国家的所有人。如果我们发现对于整个总体

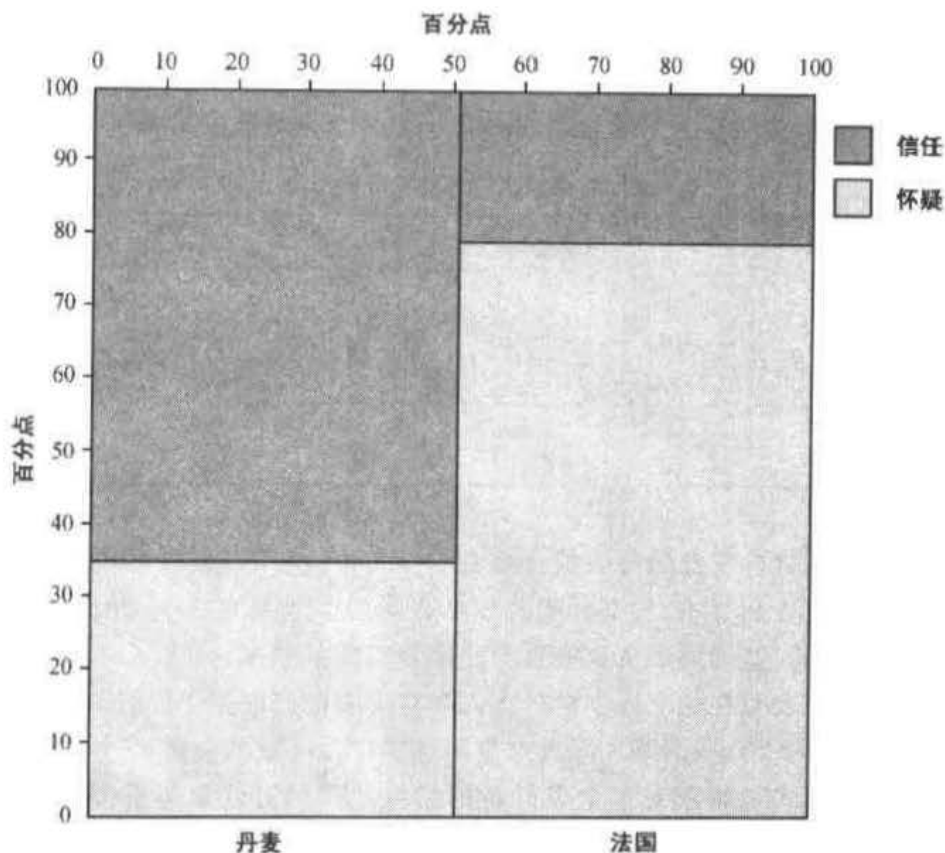


图 9.3 表 9.1 中数据的条形图：相同高度、不同宽度的条。

这两个变量间存在某种关系,那么我们就更一般地了解一些人们思考的方式和人类的行为。

(4) 最后,这种关系是因果关系吗?

让我们看一下表 9.1。表格下面附带的三个数是这些数据的统计分析的一部分。它们描述了两个变量间关系的各个方面。我们在以后几页将讨论怎样解释这些数字。在本章末,我们将解释这些数是怎样得来的并给出它们的计算公式。

9.2 问题 1. 变量间的关系?

我们问的第一个问题是国籍和对人态度之间是否存在一定的关系,也就是说,给定一个变量的某些取值,另一个变量的某些取值是否更容易发生呢?换句话说,丹麦人和法国人对他人的态度是否不一样呢?

快速的看一下表 9.1,在被调查的丹麦人当中,大多数人认为人们是可信赖的。而在法国人中大多数人则认为应当谨慎小心。注意在这张表中,自左上角至右下角的对角线上有更多的观测。这种频率结构表明在这些数据中这两个变量是相关的。图 9.1, 9.2 和 9.3 也表明这两个变量是相关的。

判定两个变量是否相关的一个常用方法是将频率变为百分比,然后比较这些百分比。比

较 64% 和 21% 要比较 $\frac{625}{985}$ 和 $\frac{206}{969}$ 容易的多。

我们通常对自变量的值计算百分比。所以我们首先计算丹麦人中那些持信任态度和怀疑态度的人的百分比, 然后对法国人计算这些百分比。表 9.2 是计算的结果。

表 9.2 两个国家的态度百分比分布

		国家	
		丹麦	法国
对他人的态度	信任	64	21
	怀疑	36	79
	总计 (n)	100 (985)	100 (969)

一个百分比表应该对自变量的每一组注明总计 100%, 这样可以告诉读者百分比是在哪一个方向上计算的。在表 9.2 中, 底部标明的 100 表明列加起来为 100。另外, 每组中观测的总数也列在了括号中(n), 这样可以从此些百分比恢复实际的频率。

这两列中的百分比表明在这个样本中丹麦人和法国人的态度是不同的。从两列间有差异的事实我们可以得出以下结论: 国籍和对人态度间确实存在一定的关系。

我们可以用两种方式来解释对这个表分析的结果。我们可以聚焦于国籍这个自变量的值, 认为两个国家间的确存在差异。也可以着重注意这两个变量本身, 认为国籍和对人态度间存在一定的统计关系。有时把这个结果描述为由自变量所定义的不同组间的差异是令人感兴趣的, 但强调变量间的关系会和以后的几章保持一致。

9.3 问题 2. 关系的强度?

我们问的第二个问题是国籍和信任态度间的关系的强度。统计上的关系是由一个取值于 0 到 1 之间的一个系数来度量的。当这个系数等于 0 时, 认为变量间没有关系; 如果两列中的百分比相等, 就表明在态度上没有差异。当这个系数等于 1 时, 认为变量间的关系最强; 如果所有的丹麦人都认为他人是可信的, 而所有的法国人都认为应当谨慎小心, 这时两个国家的人们的态度之间差异达到了最大。一个系数如果在 0.00 到 0.30 左右之间, 可以认为这个关系是比较弱的。如果在 0.30 到 0.70 之间, 表明这个关系是适中的。如果在 0.70 左右到 1.00 之间, 表明这个关系很强。

停下来想一想 9.4

仅看表 9.1 的数据, 你猜猜这个强度系数的值是多少?

样本中的 ϕ

对于一个像表 9.1 那样由两行和两列观测数据构成的列联表,我们计算出来的系数叫做 ϕ (参见本章末公式 9.1)。在表 9.2 中 $\phi = 0.43$ 。由于两列的百分比不同,所以这些数据中确实存在着一定的关系。对于 $\phi = 0.43$,我们可以认为这个关系是中度的。一个中度的关系意味着国籍对于人们是否信赖他人的态度具有一定的影响,但除此之外,还有其他的因素影响着一个人的态度。

若 $\phi = 0.00$,列联表看上去将像表 9.3a 那样。若 $\phi = 0.43$,625 个丹麦人认为大多数人是可信赖的;若 $\phi = 0.00$,将有 419 个丹麦人持这种观点。所以与 $\phi = 0.00$ 时的数据比较, $\phi = 0.43$ 时将有更多的 ($625 - 419 = 206$ 个) 丹麦人实际上认为他人是可信赖的。对于法国人,则有类似的结果:和 $\phi = 0.00$ 时相比,实际数据有多于 206 ($763 - 557 = 206$) 个的法国人认为应该持谨慎的态度。这 412 ($206 + 206 = 412$) 个人决定了观测到的 $\phi = 0.43$ 。

表 9.3 不同 ϕ 值的假想表

(a) $\phi = 0.00$

		国家		
		丹麦	法国	总计
对他人的 态度	信任	419	412	831
	怀疑	566	557	1123
	总计	985	969	1954

(b) $\phi = 1.00$

		国家		
		丹麦	法国	总计
对他人的 态度	信任	985	0	831
	怀疑	0	969	1123
	总计	985	969	1954

(c) 观测和 ϕ 的对照情况

持信任态度的丹麦人数

419

625

985

0.00

0.43

1.00

ϕ 的值

表 9.3b 是 $\phi = 1.00$ 时的数据。所有的丹麦人认为人们是可信赖的,而所有的法国人则认为必须谨慎小心。对于 $\phi = 1.00$,将会有 985 个丹麦人认为他人是可信赖的,也就是说,比实际 625 个持这种观点的多 360 个人。

表 9.3c 描述了丹麦人中认为他人是可信赖的人的数目与 ϕ 的联系。随着我们将丹麦人中持这种观点的人的数目从 419 增加到 625 再增加到 985, ϕ 值也将随之从 0.00 增加到 0.43 再增加到 1.00。就像 625 这个数大约在最小值 419 和最大值 985 之间的 $\frac{4}{10}$ 处一样, 观测到的 $\phi = 0.43$ 大约处在 ϕ 的最小值 0.00 和最大值 1.00 之间的 $\frac{4}{10}$ 处。

当每个变量只有两类的时候, 还有其他许多系数来衡量两变量间关系的强度, 但 ϕ 是最常用的。但对于多于两行和(或)两列的更大的表, 就不能再用 ϕ 了, 这时可从许多其它的系数中选一个。

总体中的 ϕ

因为我们没有丹麦和法国所有人的态度的观测数据, 我们就不知道对于这两个国家的所有人来说, 这个关系到底有多强, 但是我们可以利用样本值 $\phi = 0.43$ 作为两个国家中国籍和对人态度的相关程度的一个估计。

9.4 问题 3: 总体中的关系?

我们问的第三个问题是对于这两个国家的所有人来说, 国籍和对他人的信赖态度是否有一定的联系。这两个变量是否不仅对这 1954 个样本来说是相关的, 而且对于所有的丹麦人和法国人来说也是相关的? 这大约有 6 千万人。

提出零假设

如果我们知道这两个国家中每一个人是如何想的, 我们就可以把所有这 6 千万人填入到一个两行和两列的列联表中。然后为了发现在那张表中两个变量是否有某种关系, 我们只需要简单地比较一下百分比分布就可以了。如果这两列百分比存在差异, 那么这张表将表明对于所有的丹麦人和法国人来说, 这两个变量间存在一定的联系。我们还可以计算 ϕ 来看一看这个关系有多强。

但这只是统计上的美梦。我们不可能搜集到要构造一张总体的列联表所要求的所有数据。因为有太多的限制, 我们不可能去问每一个人到底是怎样想的, 因为这样花费太高。另外, 我们没有理由相信每个人都会回答这个问题, 或者即使回答了, 也没有理由相信他们一定会说实话。然而, 我们却可以利用样本中得到的信息来外推到现实世界。

对于丹麦人和法国人从样本到总体的外推是通过假设检验来完成的。回忆第七章, 我们从并非显然的一步开始。我们首先提出零假设, 认为对于所有的丹麦人和法国人这个总体来说, 这两个变量不相关, 然后看一看这些数据是否提供了拒绝这个零假设的证据。如果我们可以拒绝零假设, 我们就找到了两个变量相关的证据。这是做事情时许多办法中向回倒推的一种, 但却是我们能有的最好的一种办法。

我们是否拒绝零假设依赖于两个因素: (1) 样本的关系强度(ϕ)和(2)样本中观测的个数(n)。对于任何样本的数据, 两个因素作用一样。如果我们有一个很大的样本, 那么即使 ϕ 很

小也足以拒绝零假设。而对于一个较小的样本,我们需要 ϕ 很大才可以拒绝零假设。对于很小的样本而且 ϕ 也很小时,我们也许就不可以拒绝零假设。然而,这却并不一定意味着零假设是正确的。事实也许是零假设是错误的,只不过我们找不到足够的证据来证实它罢了。

国籍和信赖态度这两个变量的关系强度为 $\phi = 0.43$;而且在这个样本中有 1954 个观测数据。这样大小的相关程度和样本数联合起来就有足够多的证据来拒绝零假设。所以我们可以下结论:对于所有的丹麦人和法国人构成的总体来说,国籍和信赖态度间确实存在着某种关系。

检验零假设

我们需要前面几章所述的方法来做出拒绝零假设的结论。首先假定零假设是正确的;我们假设对于两个国家的所有人(这个总体)所构造的假想表中这两个变量是不相关的,且 $\phi = 0$ 。然后我们问是否单凭偶然就可以得到一个 ϕ 大于或等于 0.43 的样本。

考虑同样问题的另一个方法是问:如果我们从所有的丹麦人和法国人中抽取许多样本,并假定两个变量是不相关的, ϕ 大于或等于 0.43 的可能有多少呢?更正式地,我们从两个变量不相关的总体中抽取一个样本,然后计算得到一个 ϕ 大于或等于 0.43 的概率。这个概率暗示了 $\phi = 0.43$ 是否属于一个不正常的 ϕ 的集合,对于我们的数据来说,这就是 p -值。对于这个样本,表 9.1 括号中给出的 p -值小于 0.0001 或者说小于 $\frac{1}{10000}$ ①。

这样大小的 p -值是很小的,至少可以这样说,它告诉我们,从没有相关性的总体中随机抽取一个 1954 人的样本并得到 ϕ 大于或等于 0.43 几乎是不可能的。从两个变量毫无关系的总体中抽取大量样本,得到 ϕ 大于或等于 0.43 的样本的可能性将不超过 $\frac{1}{10000}$ 。

我们可以用两种方法来解释这个 p -值,或者说零假设是对的而且得到一个 $\phi = 0.43$ 的随机样本是极其不可能的;或者说总体中存在一定的关系, $\phi = 0.43$ 并不奇怪。因为 p -值太小,我们有足够的证据来拒绝零假设。我们认为我们的样本并非很特别,从而可以认为它来自两个变量相关的总体而拒绝认为这两个变量不相关的零假设。

从 χ^2 到 p -假

怎样得到 p -值呢?很不幸的是,没有一张统计表使我们能够寻找样本为 1954、 $\phi = 0.43$ 的样本的 p -值。我们可以构造一个总体,在这个总体中我们已知变量是无关的。然后我们可以从这个总体中抽取大量不同的样本,并看一看这些样本中有多少个的 ϕ 大于或等于 0.43。这就会告诉我们得到大于或等于 0.43 ϕ 的概率。可以用计算机程序来帮助我们来做这件事情,但这样仍然很烦琐。我们可以通过把 ϕ 变换为第五章中介绍的理论统计量的值来寻找 p -值。

回头看一看表 9.1 的数据。表下方表明 1 个自由度的 $\chi^2 = 355.78$ ②。计算这个 χ^2 -变量的值时用到了样本的大小和 ϕ 的值。如前所述,这个计算只是从一个变量到另一个变量的简单变换,就像把华氏度变为摄氏温度那样。在其他章我们已把数据变换为了 z -或 t -变量;在这里我们要用 χ^2 变量。这样一个变换不改变我们的结果的值;一个变量用起来比另一个更方

① 原书表 9.1 并未给出 p -值——译者注

② 原书表 9.1 并未给出这些值——译者注

便。本章末附有由样本数和 ϕ 值计算 χ^2 值的公式(公式 9.3);图 9.4 显示了由 ϕ 到 χ^2 的变换。这个图说明了 1954 个人的样本的 ϕ 大于或等于 0.43 的概率为什么等于 1 个自由度的 χ^2 变量大于或等于 355.78 的概率。

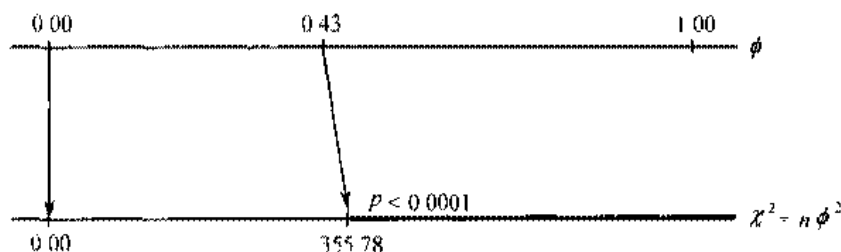


图 9.4 为计算 ϕ 的 p -值把 ϕ 变换为 χ^2 值。

我们的一个学生, Maura McDermott, 在她的一篇文章中描述她对 χ^2 的看法时说:“ χ^2 是一个很神秘的东西, 有点像业余厨师发粉一样; 我们不太知道它做了什么, 但我们知道我们需要它!” χ^2 实际上并不那么神秘, 它很像我们观测的其他变量。就像我们走上浴室的称上看看我们有多重一样, 我们把我们的数据放在 χ^2 这个称上来看看我们的数据有多重。如果我们花点时间看看它的数学推导, χ^2 也许就不那么神秘了, 不过这个推导并不能把我们变成更出色的统计学家。

得到 χ^2 的值后, 我们就可以利用计算机上的统计软件来计算 p -值了, 也可利用 χ^2 分布的统计表查找相应的 p -值。但是 χ^2 分布表中并没有 355.78 这样大的数, 所以我们利用统计软件来计算这个 p -值。

当 p -值小于 $\frac{1}{10000}$ 时, 从一个两个变量不相关的总体中抽取一个样本, 得到 ϕ 大于或等于 0.43 几乎是不可能的。对于这样大的 ϕ 的样本, 唯一可能的解释是这个样本来自于两个变量相关的总体。由于 p -值很小, 我们拒绝零假设, 并可断言, 对于整个总体来说, 国籍和对人态度是相关的。

χ^2 分析的自由度

为了计算 p -值, 我们不仅仅需要 χ^2 值, 这个概率不仅依赖于 χ^2 这个值有多大, 而且还依赖于列联表中行和列的数目。表的大小用一个称作自由度(记作 d.f. 或 df)的数来度量。表 9.1 有两行和两列, 自由度就为 1。

为了搞明白为什么一个两行和两列的表只有一个自由度, 可以想像从表 9.1 中去掉四个频率的格子, 只保留行总和和列总和。这时这个表看上去像表 9.4 那样。对于丢失的四个频率, 我们必须知道几个失去的频率才能补齐这张表呢? 这个表丢失了四个数据, 但是我们只需要知道其中的一个就可以了, 这是因为我们可以从总和中减去这个来得到另外的三个。比如, 如果我们知道这个样本中有 625 个丹麦人持信赖的态度, 就可以把它填在表的左上角。要得到丹麦人中持怀疑态度的人的频率, 我们可以从总数中减去持信赖态度的丹麦人的数目就得到 $985 - 625 = 360$ 个持怀疑态度的丹麦人。利用类似的减法可以得到法国人的频率来完成这张表。

表 9.4 没有格子中元素的表 9.1

		国家		
对他人 的态度		丹麦	法国	总计
	信任			831
	怀疑			1123
	总计	985	969	1954

既然我们只需要知道丢失的频率中的一个就可以得到其他的几个,从而就说这个表和它的 χ^2 具有一个自由度。一般地,对于一个特别的列联表,其自由度等于行数减 1 和列数减 1 的乘积:

$$\text{自由度} = (\text{行数} - 1) \times (\text{列数} - 1)$$

对于两行和两列的列联表,

$$\text{d.f.} = (2 - 1) \times (2 - 1) = 1$$

9.5 问题 4. 是因果关系吗?

我们感兴趣的第四个也是最后一个问题是国籍和对人态度间的关系是否为因果关系。国籍是一个影响人们是否信赖他人的变量吗?更具体的说,是否因为是丹麦人才导致一个人信赖别人,而是法国人就多疑呢?据我们所知,回答这个问题一般来说要比回答其他三个问题困难得多。在国籍/态度这个例子中,要回答这个问题,在统计上我们很少能做什么。尽管我们已经发现了一个统计上的关系,我们还没有证据说这个关系就是因果关系。特别地,对于民意测验或抽样调查的数据,也许其他变量影响着观测的结果。所以,也许有其他变量导致或有助于导致丹麦人比法国人更信赖别人。

停下来想一想 9.5

你认为其他哪些变量会有助于解释丹麦人和法国人在态度上的差异?

9.6 更大的表:更多的可能性

当一个或两个分类变量的取值多于两类时,列联表就会多于两行或两列。一个表明四个国家的人们对于他人的态度的列联表将含有两行和四列(表 9.5)。

表 9.5 国家和人们对他人的态度

		国家				
对他人 的态度	信任	丹麦	法国	荷兰	西德	总计
	怀疑	625	206	468	393	1692
	总计	360	763	463	513	2099
		985	969	931	906	3791

($V = 0.32$, $\chi^2 = 367.94$, $df = 3$, $p\text{-值} < 0.0001$)

来源 Jacques-René Rabier, Helen Riffault, and Ronald Inglehart, Euro-barometer 25: Holiday Travel and Environmental Problems, April 1986, Ann Arbor, MI: Inter University Consortium for Political and Social Research, 1988. Codebook p. 10

对于一个更大的表,我们仍然可以列的百分比以确定数据中两个变量间是否有一定的关系。但当表扩大时,分析的其他方法发生了变化。其中一个,不再能用 ϕ 来检验相关的程度; ϕ 对两行和两列的表才有定义。对于较大的表一个常用的系数是 Cramer 的 V 。 V 是 ϕ 的推广,如果我们把计算 V 的公式用在两行和两列的列联表上,将会得到与用 ϕ 计算的同样的结果。公式 9.2 给出了计算 V 的公式。像 ϕ 一样,它的值从 0 变到 1。

对于一个更大的表,仍然需要寻找一个 χ^2 变量来看看对于更大的总体,我们是否能够拒绝零假设,即两个变量间没有关系。公式 9.4 给出了对于任何列联表如何寻求 χ^2 。最后,由于这个表多于两列,所以自由度大于 1。为了计算自由度,我们采用用于两行和两列的列联表的计算自由度的公式: $df = (\text{行数} - 1) \times (\text{列数} - 1)$ 。对于一个三行和四列的表, $df = (3 - 1) \times (4 - 1) = 2 \times 3 = 6$;我们除了知道总和之外,还要知道格子中的 6 个数才能填进其他的 6 个数。

χ^2 是基于一个列联表中行和列的总和是固定的这一思想来计算的。对于固定的总和,要完成这张表,除了最后一行和最后一列外我们还要知道其他所有的频率(图 9.5)。如果我们知道除了最后一行和最后一列的所有频率,那么,我们就可以从相应的总和中减去已知的这些频率来得到最后一行和最后一列。

				总
				数
	总	数		

图 9.5 寻找自由度。

有了 V 、 χ^2 和自由度以后,为了得出对于总体两个变量是否有关系的结论,我们还要计算 p -值。

表 9.5 是除了丹麦和法国之外还有荷兰和德国的调查数据。自变量共有四列,因变量共有两行,构成了一个 2×4 的列联表。让我们利用关于两个变量关系的四个问题来分析这些数据。

问题 1. 两变量间的关系?

为了更好地了解国籍和信赖态度间的关系和更好地看清楚国家间的差异,像对于 2×2 列联表那样我们把频率变为百分比。因为国籍是自变量,对每个国家我们把频率变为百分比,如表 9.6 所示。从百分比可以很明显地看出,从样本来说这四个国家人们的态度是有差异的。持认为大多数人是可信赖的这种观点人的百分比丹麦最高,而法国则最低。

表 9.6 四个国家的百分比分布

		国家			
		丹麦	法国	荷兰	西德
对他人的态度	信任	64	21	50	43
	怀疑	36	79	50	57
	总计	100	100	100	100
	(n)	(985)	(969)	(931)	(906)

问题 2. 关系的强度?

为了寻找国籍和人们的态度间关系的强弱,我们计算 Cramer 的 V 系数。对于这些数据 $V = 0.32$,在 0 到 1 之间这是一个相对较弱的关系。

停下来想一想 9.6

假定现在你对到了上学年龄的学生在选择公共学校还是私立学校与学生家长的宗教信仰间的相关程度感兴趣。家长分为天主教、犹太教、新教或其他。

对这个问题你怎样来构造一张列联表呢? 哪一个系数将是一个很好的表示它们的关系强度的指标呢? 如果宗教信仰只有天主教和其他你将用哪个系数呢?

问题 3 总体中的关系?

要看我们是否能把 3791 个被调查者这个样本的关系推广到四个国家所有成年人的总体上去,我们还要计算 χ^2 和相应的 p -值。如果 p -值很小,那么我们就可以拒绝认为这些数据仅是由于偶然性才得到的零假设。在这里, $\chi^2 = 367.94$, 自由度 $d.f. = 3$ 。现在我们用计算机软件或查 χ^2 分布表得到 p -值小于 0.0001,这就是说,如果这四个国家中两个变量间没有关系,那么得到 V 大于或等于 0.32 的概率将小于 $1/10000$ 。所以,如果我们从两个变量没有关系的总体中抽取大量不同的样本,得到 V 大于或等于 0.32 的机会将不超过 $1/10000$ 。

由于 p -值如此之小,我们有足够的证据认为这些数据并非偶然产生的。所以,我们拒绝总体两变量间没有关系的零假设,即对于四个国家成年人这个更大的总体来说,变量间确实存在着一定的关系。我们并不知道是否对所有的国家都彼此各不相同,或者是否只是其中几个国家,和其他的不一样。要知道这一点还需要进一步的分析。

问题 4. 是因果关系吗?

最后一个问题是,住在一个特定的国家里是否会导致人们更加信赖别人或者相反?正如我们反复强调的,如果我们只有基于这两个变量的数据,就不能用统计的方法来回答这个问题。然而,我们可以预测,持信任态度的丹麦人将比法国人多。我们也许可以推测在人口较少、风俗习惯更加相似的地方的人们较容易信赖他人,但我们却不能在统计上支持这个推测。

9.7 小 结

一个列联表是一张频率表,它描述对于两个分类变量所有值的组合数据是如何分布的。

9.1 数据分析:在态度上有可靠的差异吗?

在一张列联表中,自变量通常在表中水平排列,因变量垂直排列。用一个图来描述这些数据通常是很有用的,因为用一个图比用一个列联表更容易看出数据的整体趋势。列联表对于研究分类变量的关系有很大用处。

9.2 问题 1. 变量间的关系?

要寻求观测数据中变量是否相关,我们要比较列联表中列的百分比。如果百分比分布不同,我们就可以认为数据中变量间存在一定的关系。

9.3 问题 2. 关系的强度?

关系的强度由一个用数据计算出来的系数来度量。这个系数在 0 到 1 之间取值。一个接近于 0 的系数表示关系较弱,而系数若接近于 1 则表明关系很强。

对于一个两行和两列的列联表,一个常用的表示关系强度的系数称作 ϕ 。对于多于两行和(或)两列的表常用的系数是 V 。

9.4 问题 3. 总体中的关系?

要了解产生样本的总体的变量间是否存在一种关系,我们通过统计假设检验利用对样本的知识来推广到总体。第一步,提出零假设,即认为总体的两个变量不相关。是否拒绝零假设依赖于两个因素:(1)样本的关系有多强(ϕ 或 V),和(2)样本有多少个观测(n)。如果样很大,即使 ϕ 或 V 很小也足以拒绝零假设。如果样本只有少数几个观测,则需要 ϕ 或 V 很大才能拒绝零假设。

要对于一个特定的样本计算 p -值,需要将 ϕ 或 V 变换为一个 χ^2 变量的值。要计算这个值需要利用样本数和 ϕ 或 V 的值。

要计算与 χ^2 变量相应的 p -值,需要知道列联表的自由度。自由度 $df = (\text{行数} - 1) \times (\text{列数} - 1)$ 。

9.5 问题 4. 是因果关系吗?

对于只有两个变量的观测数据,不可能回答两变量间的关系是否为因果关系。统计的关系并不能说明是因果关系。

补充读物

Reynolds, H. T. *Analysis of Nominal Data*, 2nd ed. (Sage University PaperSeries on Quantitative Applications in the Social Sciences, series no. 07-007). Beverly Hills, CA: Sage, 1984, χ^2 检验和分类变量关系的度量。

公 式

用不同宽度的条作成的条形图

图 9.3 是一个 2×2 的列联表的条形图,在这个图中,条的宽度不同,每个矩形的面积和那一类的频率成正比。对于任何两个变量的频率表都可以作这样的图。图 9.6 说明对于一个 2×2 的列联表这样的条形图是怎样画出来的。

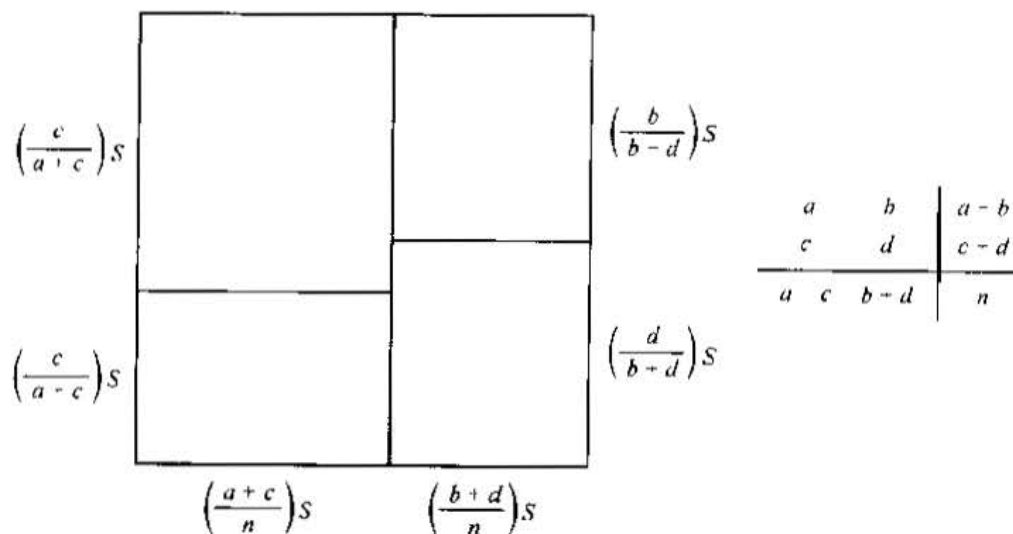


图 9.6 在一个边长为 S 的正方形中一个由高度相同、宽度不同的条所构成的条形图。

这个图是基于它右边的列联表的频率。字母 a, b, c 和 d 代表每个格子中观测的个数。整个表的总和记作 n 。从这个表我们要做一个包括一个正方形的图,这个正方形的面积与总频率 n 成正比。正方形的边长为 S 。正方形边长可以用英寸、厘米或其他任何单位来度量。

下一步我们将这个正方形分为两个竖直的条,每个条代表表中的一列。第一列共有 $a+c$

个观测,这一列观测的比例为 $\frac{a+c}{n}$ 。我们用同样的比例把正方形的底分为两半。左边的条宽度为 $\frac{a+c}{n}S$, 右边的宽度为 $\frac{b+d}{n}S$ 。这两个宽度的和自然就是 S 。

现在我们利用相应列中的数据来划分图中的每一个竖条。表中左边一列上边那类占这一列观测的比例为 $\frac{a}{a+c}$, 所以上边矩形的高度为 $\frac{a}{a+c}S$, 下边矩形的高度为 $\frac{c}{a+c}S$ 。类似地, 可以把第二列划分为相应的比例, 如图所示。

边长为 S 的正方形面积为 S^2 。这四个矩形每个的面积都是这个总和的一部分。这个部分相应于表中每个格子中观测的比例。例如, 右下角矩形面积为 $\frac{d}{n}S^2$, 相应格子中观测数的比例为 $\frac{d}{n}$ 。

总的区域不必是一个正方形。它可以是一个底长为 B 高为 H 的矩形。可以用对 S 同样的比例来划分底和高。

当每个变量只有两个取值(类)时, 常用 ϕ 在 0 到 1 范围内的尺度来度量两个分类变量的关系强度。如果我们用字母 a, b, c, d 和 n 来代替观测的频率, 列联表看上去如表 9.7 所示。

表 9.7 用字母代替频率的列联表

	a	b	总计
	c	d	$a+b$
总计	$a+c$	$b+d$	$c+d$
			n

要看看两个变量间是否存在某种关系, 我们将频率转化为百分比然后比较这两列百分比。在国家/态度这个例子中, 丹麦和法国人如果持信赖态度的人的比例相等, 则认为这两个变量间没有相关性。这可以记作:

$$\begin{aligned}\frac{a}{a+c} &= \frac{b}{b+d} \\ a(b+d) &= b(a+c) \\ ad - bc &= 0\end{aligned}$$

乘积 ad 是表中一个对角线上两个频率的乘积, bc 则是另一个对角线上两个频率的乘积。当两个变量之间没有关系时, 这两个乘积相等。这两个乘积的差异越大, 关系越强。 ϕ 可以由对角线上两个乘积的差异计算出来; 公式如下:

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \quad (9.1)$$

加进分母一项是为了保证 ϕ 不会大于其最大值 1.00。由表 9.1 中数据可以得到

$$\phi = \frac{625 \times 763 - 206 \times 360}{\sqrt{831 \times 1123 \times 985 \times 969}} = 0.43$$

如果 $\phi < 0$, 我们通常忽略负号, 只取其正值。因为我们只需要交换表中的两列(或两行)就可以改变 ϕ 的符号。因为变量是分类变量, 我们可以这样交换而不改变表的意义。

Cramer 的 V

如果分类变量有两个或多个, 而至少其中一个分类变量取值多于两个, 我们常用取值于 0 到 1 之间的 V 来度量变量间的相关程度。计算 V 的公式是:

$$V = \sqrt{\frac{\chi^2}{n(L-1)}} \quad (9.2)$$

其中 n 是观测的个数, L 是行数和列数中较小的那个数。对于表 9.5 的数据, 共有两行和四列, L 是 2 和 4 中较小的一个, 即 $L = 2$ 。

$$V = \sqrt{\frac{367.94}{3791 \times (2-1)}} = 0.31$$

V 是 ϕ 的推广。对于两行和两列的列联表, V 和 ϕ 等价。

χ^2 变量

2×2 表 对于一个两行和两列的列联表, 可由下面的公式计算 χ^2 :

$$\begin{aligned} \chi^2 &= n\phi^2 \\ &= \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \\ &= \frac{1954(625 \times 763 - 206 \times 360)^2}{831 \times 1123 \times 985 \times 969} \\ &= 355.78 \quad \text{d.f.} = 1 \end{aligned} \quad (9.3)$$

更大的表 对于更大的表, 我们用另一种方法也就是基于所谓的期望(expected)频率来计算 χ^2 。

可以用一个 2×2 表来说明这种方法。一个有期望频率的表与初始的表含有相同的行总和和列总和, 但对于这个新表, $\phi = 0.00$ 。 χ^2 度量了观测到的频率与期望频率间有多大的差异。

表中一个特定格子中的期望频率由下面的表达式计算:

$$\text{期望频率} = \frac{\text{行总和} \times \text{列总和}}{\text{表的总和}}$$

对于表 9.1 中数据, 我们得到如表 9.8 所示的期望频率。这些期望频率和表 9.3a 中频率相同。

注意由于行总和和列总和与初始表相同, 所以只需要计算这些期望频率中的一个就可以了, 这可以通过行总和乘以列总和再除以表的总和来计算。其他三个可以由总和减去这个数来得到。这就是为什么相应的 χ^2 的自由度为 1 的理由。

表 9.8 期望频率

		国家		
		丹麦	法国	总计
对他人的 态度	信任	$\frac{(831)(985)}{1954} = 418.90$	$\frac{(831)(969)}{1954} = 412.10$	831
	怀疑	$\frac{(1123)(985)}{1954} = 566.10$	$\frac{(1123)(969)}{1954} = 556.90$	1123
	总计	985	969	1954

下一步可以通过比较观测频率和期望频率的差异来计算 χ^2 , 由此可以知道观测的表格与没有任何关系的表格有多大的差别。我们可以根据下式来计算 χ^2 :

$$\begin{aligned}
 \chi^2 &= \frac{\text{观测频率} - \text{期望频率}}{\text{期望频率}} \\
 &= \frac{(625 - 418.9)^2}{418.9} + \frac{(206 - 412.1)^2}{412.1} \\
 &\quad + \frac{(360 - 566.1)^2}{566.1} + \frac{(763 - 556.9)^2}{556.9} \\
 &= 355.78 \quad \text{d.f.} = 1
 \end{aligned} \tag{9.4}$$

对于更多行和列的表计算过程相同。首先计算期望频率, 然后由公式 9.4 计算 χ^2 。当然和式中将多于四项。

作为近似的 χ^2 变量 对于我们的数据可以利用 χ^2 来计算 p -值, 但这种方法只是给我们提供了一个近似真实的 p -值。特别地, 如果 p -值仅是在显著的边沿, 这种近似将是令人担忧的。表中的观测数据越多, 近似得越好。我们是否够用 χ^2 来计算 p -值的一个方法是看一看期望频率的大小, 在一个 2×2 的列联表中, 所有的期望频率都应大于 5。对于行和列比较多的表, 这个要求则并不那么重要。

习 题

回顾(习题 9.1—9.8)

- 9.1 在一张报纸、一份新闻杂志或者其它地方找一个抽样调查的报告, 构造一个列联表并说明这个调查中两个分类变量的关系。利用这个表讨论两个变量间的关系。
- 9.2
 - a. 什么时候可以用 ϕ 来度量两个变量间的关系? 什么时候不可以用 ϕ ?
 - b. 举一个 ϕ 值较小的例子。一个较小的 ϕ 值说明了两个分类变量间的怎样的关系?
 - c. 一个小的 p -值说明两变量间有怎样的关系?
 - d. 仔细讨论为什么可能有一个小的 ϕ 值并同时 p -值仍然可以很小。
- 9.3 描述一个列联表。
- 9.4
 - a. 自变量用哪一个轴表示(水平或垂直)?
 - b. 因变量用哪一个轴表示?

- c. 这种安排自变量和因变量的方法是一成不变的统计法则吗? 解释你的答案。
- 9.5 对于一个有两个变量, 每个变量有两个取值的列联表, 我们用来衡量两变量间的关系强度而计算的统计量的名字是什么?
- 9.6 a. 计算 χ^2 时, 我们用来度量表的大小(行数和列数)的统计术语叫做什么? 这个术语通常如何缩写?
b. 怎样计算这个度量?
- 9.7 如果你从一个很大的 χ^2 值知道两变量间有一定的统计关系, 你总能断定这个关系是因果关系吗? 解释你的结论。
- 9.8 一个列联表包括每个变量的行的总和与列的总和。
a. 如果把行或列加起来, 你能得到什么结论?
b. 这些和为什么能够检验列联表的结构?

解释(习题 9.9—9.27)

- 9.9 1986 年秋, 众议院投票表决是否资助尼加拉瓜反政府武装(里根总统赞成的一项建议)。他们还对一项一般拨款提案进行了投票。否决票被视作支持里根总统的政策, 赞成票被认为反对该项政策。除去那些弃权的人, 表 9.9 是投票的结果。

表 9.9 习题 9.9 的数据

		拨款提案		
		是	否	总计
支持反政 府武装	是	42	167	209
	否	156	33	189
总计		198	200	398

$$(\phi = 0.62, \chi^2 = 156.75, df = 1, p = 0.000)$$

来源: Kenneth Janda and Philip A. Schrodt, *Crosstabs: Student Workbook for American Government*, Boston: Houghton Mifflin, 1987. Data disk

- a. 两个投票间有一定的关系吗?
b. 相关程度如何?
c. 样本中存在的这个关系仅仅是由偶然产生的吗?
d. 这个关系是因果关系吗?
- 9.10 在 1950 年一项研究居住地(包括国家的南部和非南部地区)和职业(专业人员或农民)之间的关系, 抽样调查所得样本的频率如表 9.10 所示。

表 9.10 习题 9.10 的数据

		居住地		
		南部	非南部	总计
职 业	专业人员	70	93	163
	农民	135	58	193
总计		205	151	356

$$(\phi = 0.27, \chi^2 = 26.38, df = 1, p = 0.000)$$

来源: Adapted from J. C. McKinney and L. B. Borque, "Further comments on 'The changing south': 4

response to Sly and Weller," American Sociological Review, vol. 37 (1972), p. 236.

- a. 对于这些数据,居住地和职业间有关系吗?
- b. 相关程度如何?
- c. 产生样本的较大总体中变量间有关系吗?
- d. 这个关系是因果关系吗?

- 9 11 在前三十次橄榄球决赛中,有 60 个队参加。有一半队伍获胜,另一半失败。在获胜者和失败者中分别有 8 支和 7 支队伍参加了第二年的决赛。这些数据可以像表 9.11 那样安排。说明这两个变量间的关系。

表 9 11 习题 9 11 的数据

		结果		总计
		胜者	负者	
第二年	参加	8	7	15
	不参加	22	23	45
	总计	30	30	60

($\phi = 0.04, \chi^2 = 0.089, df = 1, p = 0.76$)

- 9 12 表 9.12 是 1880 年美国被划分为 6 个地区的数据。列表示人的出生地,行表示现在的居住地。这些数据来自 1880 年的人口普查。我们用调查到的数据除以 10000,也就是把这 4361 个人的数据看作从大约 4300 万个人的总体中抽取的样本。最后得到的 6×6 的列联表社会学家们也称之为“迁移表”。它表明有多少个人从他们的出生地迁移到了现在的居住地。这些数据可以读作:“306 个在新英格兰出生的人现在仍然住在新英格兰,18 个新英格兰出生的人现在住在中大西洋州”,等等。

- a. 说明两个变量之间的关系。
- b. 19 世纪末期人口地理流动的主要趋势是什么(比较主对角线上方的频率和主对角线下方的频率)?

表 9 12 习题 9 12 的数据

		出生地						总计
		新英 格兰	中大西 洋州	中 北部	南大西 洋州	中 南部	山区和太 平洋州	
居 住 地	新英格兰	306	11	2	2	0	0	321
	中大西洋州	18	806	8	14	1	0	847
	中北部	30	127	1180	39	55	2	1433
	南大西洋州	2	11	5	717	8	0	743
	中南部	1	5	23	76	790	0	895
	山区和太平洋州	8	13	24	3	2	72	122
总计		865	973	1242	851	856	74	4361

($V = 0.85, \chi^2 = 15733, df = 25, p < 0.0001$)

来源: U. S. Bureau of the Census, Historical Statistics of the United States: Colonial Times to 1957, Washington, DC, 1960, Series C 15-24, pp. 42, 44.

9.13 受教育程度不同的组间有什么不同吗? 表 9.13 是一张关于 988 个亚洲裔、西班牙裔和白人受教育程度的随机样本。

- 该表中两变量间有关系吗?
- 这个关系有多强?
- 三组间在受教育程度在统计上有显著的差异吗?
- 这个关系是因果关系吗?

表 9.13 习题 9.13 的数据

		组			
教 育	中学或更低	亚洲裔	西班牙裔	白人	总计
	上过大学的	24	98	419	514
	专业人员或研究生	27	34	310	371
	总计	9	6	61	76
		60	138	790	988

$$(V = 0.11, \chi^2 = 23.26, df = 4, p = 0.0001)$$

来源: Column percentages equal those found by the U. S. Bureau of the Census, as reported in The Chronicle of Higher Education, vol. XXIV, no. 1, August 26, 1992, p. 12

9.14 在关于家庭和家庭的一次全国调查中, 其中一个问题是问妇女们当她们 14 岁时的家庭情况。表 9.14 是对于三组妇女的部分数据。对于这些数据, $V = 0.16$, $\chi^2 = 255.29$, 自由度为 6, p -值小于 0.0001。

- 从这些数据你可以得出两变量间的关系如何?
- 6 个自由度的 χ^2 变量典型的值从 0 变到 15。在这里你如何解释对于这样大的 χ^2 值, 而 V 的值却很小。

表 9.14 习题 9.14 的数据

		组			
状 况	完整家庭	白人	黑人	西班牙裔	总计
	单亲母亲	2583	526	292	3401
	继父母家庭	297	239	75	611
	其他	317	107	25	449
	总计	175	106	34	315
		3372	978	426	4776

来源: L. L. Wu and B. G. Martinson, "Family structure and the risk of opremarital birth," American Sociological Review, vol. 58 (1993), p. 217

9.15 表 9.15 是习题 8.35 中研究秃头和心脏病关系的数据。

表 9.15 习题 9.15 的数据

		秃头		
心脏病 发作	是	是	否	总计
	否	214	431	665
	总计	175	597	772
		389	1048	1437

$$(\phi = 0.11, \chi^2 = 16.37, df = 1, p = 0.0001)$$

来源: The New York Times, February 14, 1993, pp. A1, C12

- a. 哪个是自变量? 解释你的选择
- b. 在这些结果中你得到变量间的何种关系?
- 9.16 习题 8.36 中的表 8.6 是研究宗教信仰和关于工作的重要性的观点的选择的数据。在这里, $\phi = 0.24$, $\chi^2 = 10.64$ (自由度 = 1), $p = 0.0011$ 。基于这些计算出的数据, 你能对两变量间的关系下什么结论?
- 9.17 在习题 9.16 同样的研究中, 社会学家还注意到天主教徒和新教徒的区别。这项研究要求被调查者对一些观点进行排序, 其中一项是“工作很重要”。表 9.16 是两种宗教信仰的人把这种观点排为第一位的人的百分比。
- a. 宗教信仰和对某些观点的排序间是否有一定的关系?
- b. 相关程度看上去很强吗? (把百分比转化为频率并补齐这张 2×2 的表也许会有所帮助。)
- c. 对于这些数据既可以用两个百分比间差异的检验来分析, 也可以用两行和两列的表的 χ^2 来分析。对于两种方法, p -值均为 0.013。在这个研究中, 新教徒和天主教徒之间在底特律有什么差异?

表 9.16 习题 9.17 的数据

观 点		宗教信仰	
		新教	天主教
	排第一	62%	50%
	不排第一	38%	50%
	总计	100%	100%
	(n)	(165)	(145)

- 9.18 美国很早就是一个有许多志愿者组织的国家, 两个社会学家对于从 20 世纪 50 年代中期到 60 年代早期人们属于多少个志愿者组织的数目是否有变化感兴趣。表 9.17 是在两个不同时间在调查中人们报告的他们属于的志愿者组织的数目。人们所属的志愿者组织的平均数目在 1955 年是 0.64 个, 在 1962 年为 0.80 个, 即人们所属的志愿者组织的数目稍有增长。严格地说, 所属组织的数目是一个度量变量, 关于这样两个变量的研究属于 12 章的内容。在这里, 作为一个开始, 我们把它作为一个五行和两列的列联表来研究。从表中的数据可以得到下面的结果: $\chi^2 = 25.44$ (自由度 = 1), $p = 0.00004$, $V = 0.08$ 。关于两变量间的关系你能得出什么样的结论?

表 9.17 习题 9.18 的数据

组 织 的 个 数		年份		总计
		1955	1962	
	0	1523	1012	2535
	1	476	390	866
	2	214	195	419
	3	95	106	201
	4 +	71	71	142
	总计	2379	1774	4163

来源: H. H. Hyman and C. R. Wright, "Trends in voluntary association membership of American adults: Replication based on second analysis of national samples surveys," American Sociological Review, vol. 36 (1971).

pp 191 - 206.

9.19 1970 年左右在 Baltimore 的一项研究中询问不同年级的小孩,对于美国社会分化的观点。在回答这样的问题:“在美国所有的孩子都有同样的成长和享受人生的机会吗?”时,社会学家们报告了如表 9.18 所示的数据。

- a. 关于孩子们是如何回答这个问题时的一些令人吃惊的事情是什么?(在回答这个问题以前你可以先将频率转化为百分比。)
- b. 统计上计算的结果告诉你,孩子们所受教育的水平和对于社会分化的观点间有何关系?

表 9.18 习题 9.19 的数据

回答		教育			总计
		小学	初中	高中	
是	是	207	110	67	384
	否	496	327	234	1057
	不知道	330	79	34	443
总计		1033	516	335	1884

$$(\chi^2 = 99.94, df = 4, p\text{-值} < 0.0001, V = 0.16)$$

来源: R. C. Simmons and M. Rosenberg, "Functions of children's perceptions of the stratification system," American Sociological Review, vol. 36 (1971), pp 235 - 249

9.20 习题 8.36 的数据是关于种族和抽烟品牌两个变量的表。对于这些数据, $\chi^2 = 181.93$, $df = 3$, $p\text{-值} < 0.0001$ 和 $V = 0.46$ 。从这些数据你可以知道两变量间的什么关系?

9.21 在马萨诸塞大学 Dartmouth 校园的 Robert P. Waxler 教授的鼓励下,马萨诸塞的 New Bedford 联邦地方法院的法官 Robert Kane 给了那些在其法院被判有罪的人一个或者蹲监狱或者听 Waxler 教授的一门文学课的选择。Indiana 大学的 G. Roger Jarjoura 教授跟踪了上这个课的 32 个男子,发现有 6 个人又判犯新罪。而另外 40 个有同样背景的人中,有 18 个又判犯新罪。(来源: The New York Times, October 6, 1993, p. B10.)

- a. 用一个两行和两列的列联表来描述这些数据。
- b. 对于这些数据, $\phi = 0.28$, $\chi^2 = 5.51$, $df = 1$, $p = 0.019$ 。这些数据告诉你,参加这个文学课与重新犯罪有什么关系?
- c. 在 b 部分的统计方法用于这种类型的数据看上去适当吗?

9.22 在每年美国大约有 100000 病人适宜于接受心脏搭桥手术或者所谓的血管清障术 (angioplasty) 也就是将气球插入动脉来清除障碍物。对 392 个病人进行了为期 3 年的跟踪研究以比较这两种方法。(来源: "Study finds angioplasty as good as heart bypass," The New York Times, November 11, 1993, p. A19.) 假设有一半的人接受血管清障术,另一半接受心脏搭桥手术,我们可以用报纸上文章中的百分比构造出表 9.19a、b 和 c。表 a 表明两组中有多少个病人需要进一步手术,表 b 说明病人是否还发过心脏病,表 c 说明血流是否完全恢复。

- a. 为什么保留 ϕ 的符号^①,即第一、三个表中 ϕ 为正的,而第二个表中 ϕ 为负的仍然有意义?

① 原书没有保留符号——译者注。

b. 对于每个表, 两变量间有什么关系?

c. 基于这些数据, 对这两种方法你能得出什么样的结论?

表 9.19 习题 9.22 的数据

(a)		处理		
进一步 手术		血管清障	搭桥	总计
	是	122	24	146
	否	74	172	246
总计		196	196	392
$(\phi = 0.52, \chi^2 = 104.82, p\text{-值} < 0.0001)$				
(b)		处理		
进一步 发作心脏病		血管清障	搭桥	总计
	是	29	39	68
	否	167	157	324
总计		196	196	392
$(\phi = 0.07, \chi^2 = 1.78, p = 0.18)$				
(c)		处理		
血流 恢复		血管清障	搭桥	总计
	是	110	67	177
	否	86	129	215
总计		196	196	392
$(\phi = 0.22, \chi^2 = 19.05, p\text{-值} < 0.00001)$				

9.23 在一组视力差的男孩中, 有些戴眼镜, 有些不戴。有些犯过错误, 其他则没有。表 9.20 是每一类男孩的数目。对于这些数据, $\phi = 0.62, \chi^2 = 6.11$ (d.f. = 1), $p = 0.013$ 。(这些频率太小, 所以利用 χ^2 是否合适是值得怀疑的, 但所谓的 Fisher 精确检验给出了大概相同的结果。所以对于这些数据仍用 χ^2 。)这些数据告诉你, 少年犯错误和是否戴眼镜有何关系?

表 9.20 习题 9.22 的数据

	犯错		总计
	是	否	
是	1	5	6
否	8	2	10
总计	9	7	16

来源: A. M. Weindling, F. N. Bamford, and R. A. Whitall, "Health of juvenile delinquents," British Medical Journal, vol. 292 (1986), p. 447.

9.24 蔓越橘汁可以减少老年妇女的尿感染吗? 在波士顿妇女医院进行的一项研究中, 153 个妇女中有一半在六个月中每天都喝一杯蔓越橘汁, 而另一半每天喝一杯相同颜色和口味的安慰剂。六个月后, 15% 的喝蔓越橘汁的人和 28% 的喝安慰剂的尿中发现了导致感染的细菌。对于这些数据, $\phi = 0.16, \chi^2 = 3.96$ (d.f. = 1), $p = 0.047$ 。

a. 用一张 2×2 的表描述这些数据。

b. 讨论这个结果。

- 9.25 美国统计协会每年通过选举其一些成员为特别会员来给他们荣誉。1994 年在 77 个被提名的男子中有 36 个被选为特别会员,而被提名的 20 个妇女中有 13 个被选为特别会员。(来源: Daniel L. Solomon, "Turning women into Fellows - Continued," Newsletter, Caucus for Women in Statistics, vol. 4(1997), no. 4, p. 11.)

a. 用一张 2×2 的表描述这些数据。

b. 对于这些数据, $\phi = 0.17$, $\chi^2 = 2.12$ (d.f. = 1), p -值 = 0.15。这些结果告诉你性别和选举间的何种关系?

- 9.26 1952 年, Adlai Stevenson (史蒂文森) 是民主党的总统候选人, 而 Dwight Eisenhower (艾森豪威尔) 是共和党总统的候选人。在大选前进行的一次民意调查中, 被调查者被问到的一个问题是他们是否认为两党有一定的差别及谁将获得选举的胜利。对于那些认为两党之间有很重要区别或没有区别的数据如表 9.21 所示。对于这些数据, $\phi = 0.10$, $\chi^2 = 6.73$ (d.f. = 1), $p = 0.01$ 。对于人们是如何看待两党的及人们期望的选举结果如何, 你能作出什么结论?

表 9.21 习题 9.26 的数据

		政党间的差异		
		非常明显	没有	总计
谁将 获胜	Stevenson	86	234	320
	Eisenhower	79	340	419
	总计	165	574	739

来源: Data utilized in this exercise made available by the Inter - University Consortium for Political and Social Research, data originally collected by Angus Campbell, Gerald Gurin, and Warren Miller. Neither the original collectors of the data nor the Consortium bear any responsibility for the analysis or interpretations presented here.

- 9.27 通过询问 72 个大学生在购买食物时是否看营养说明得到了一个随机样本。调查者还记录了每个学生的性别(表 9.22)。对于这些数据, $\phi = 0.48$, $\chi^2 = 8.48$, $df = 1$ 。性别和读营养说明间有何关系?

表 9.22 习题 9.27 的数据

		性别		
		女	男	总计
阅读 商标	是	16	28	44
	否	20	8	28
	总计	36	36	72

来源: Data used by permission of Jaso Porciello, Swarthmore College.

分析(习题 9.28—9.59)

- 9.28 过去认为服用药物 AZT 会减少 HIV 呈阳性的孕妇将这种病毒传给她们孩子的可能性。在美国和法国许多医疗中心进行的一项研究中, 一些 HIV 呈阳性的孕妇服用了 AZT 而另一部分人则服用安慰剂。表 9.23 是 364 个新生儿的检测结果。

a. 在数据中处理和结果有关系吗?

- b. 相关程度如何?
- c. 表明这些数据的 $\chi^2 = 27.95$, (d.f. = 1, p -值 < 0.0001)。
- d. 对于处理和结果间的关系,你能得出什么样的结论?
- e. 这些实际上都是初步的结果。你是在得到这些数据后就停止实验呢,还是继续实验直到更多的婴儿出生呢?

表 9.23 习题 9.28 的数据

		对母亲的处理		
		AZT	安慰	总计
新生儿 的状况	HIV 阳性	13	40	53
	HIV 阴性	197	114	311
	总计	210	154	364

来源: The New York Times, February 21, 1994, p. A1

- 9.29 在 1969 年,美国重新恢复了用抽签的方法来决定谁服兵役的法规。抽签的目的是用一个随机机制来决定一个年轻人是否服兵役的可能性。一年中的每一天被指定对应于一个 1 到 366 之间的所谓的随机整数作为征兵号码。比如在那一年抽签得到的 9 月 14 日为号码 1。每个人被指定对应于他的生日的征兵号码。从 1 开始,按照征兵号码顺序让适龄人入伍。抽到一个比较小的号码的概率不应当依赖于一个人是一年中什么出生的,但在 1969 年那次抽签时却并非这样。第一次抽签的随机性就产生了很严重的问题,因为在对于比较小的号码(1-183),生日在前半年的有 73 个,而生日在后半年的有 110 个。对于比较大的号码,这个趋势恰好相反。表 9.24 表示生日和抽签号码的关系。
- a. 对于一个完全随机的抽签,列联表将是什么样子?
- b. 从这个表看来,1969 年的抽签看来是随机的吗?
- c. 列出计算这些数据的 χ^2 的表达式。
- d. 我们发现 $\chi^2 = 14.16$ (d.f. = 1, p -值 < 0.0001)。对于这次抽签你能得到什么结论? 这次抽签可能是随机的吗?

表 9.24 习题 9.29 的数据

		出生的月份		
		1-6 月	7-12 月	总计
征兵 号码	1-183	73	110	183
	184-366	109	74	183
	总计	182	184	366

- 9.30 1976 年在 Kansas 市召开的共和党大会上,当时的总统 Gerald R. Ford 在下面的几个州获得了代表团选票的大多数: AL, CO, DE, FL, HI, IL, IO, KS, KY, MD, MA, MI, MN, MS, NH, NJ, NY, ND, OH, OR, PA, RI, VT, WV 和 WI。当时的州长 Ronald Reagan 在其他州赢得了大多数选票。
- a. 构造一张列联表来说明密西西比河东部和西部选举 Ford 和 Reagan 的情况。
- b. 表中哪个是自变量? 哪个是因变量?
- c. 这个数据中的两变量间是否存在某种关系?

- d. 两变量的关系有多强?
- e. 计算期望频率。
- f. 包含期望频率的表中两变量的关系有多强?
- g. 计算 χ^2 。它是否足够大使你认为密密西比河东西部选举的差异并非由于偶然得到的?
- h. 评价你的关于 1970 年代末期共和党 a - g^① 部分的回答。
- 9.31 智商与犯罪是否有关曾经有一场持久的争论。一项研究表明在 486 个低智商的白人男子中有 24.3% 的人有过两次或更多次的犯罪。类似地, 1053 个高智商的白人男子中有 9.4%, 702 个低智商的黑人男子中有 37.6%, 266 个高智商的黑人男子中有 23.3% 的人有过两次或更多次的犯罪。(来源: T. Hirschi and M. J. Hindelang, "Intelligence and delinquency: A Revisionist review," American Sociological Review, vol. 42(1997), p. 575.)
- a. 把这些数据安排在一张 2×2 的表中。
- b. 两变量间有关吗?
- c. 两变量的关系有多强?
- d. 假设这些数据来自一随机样本, 产生这些数据的总体中存在某种关系吗?
- e. 评价智商是否会影响犯罪。
- 9.32 通常认为少年犯应当受到特别地、个别的对待以防止他们再犯罪。在 100 个少年犯中, 随机地选一半接受特殊对待而另一半则作为对照组没有接受特殊对待。四年后, 得到了表 9.25^② 中的数据。
- a. 两变量的关系有多强?
- b. 有理由相信这种特殊的对待会对总体有什么作用吗?

表 9.25 习题 9.32 的数据

		特别对待		总计
		1 - 6 月	7 - 12 月	
重新 犯罪	否	37	20	57
	是	13	30	43
	总计	50	50	100

(来源: T. Hirschi and M. J. Hindelang, "Intelligence and delinquency: A Revisionist review," American Sociological Review, vol. 42(1997), p. 575.)

- 9.33 北美州和南美州的人的宗教信仰有什么差异吗? 一个来自北美州的 500 个被调查者的样本包括 190 个信仰天主教的、10 个信仰犹太教的、120 个信仰新教的和 180 个信仰其他教的。一个来自南美州的 450 个被调查者的样本包括 310 个信仰天主教的、10 个信仰犹太教的、30 个信仰新教的和 100 个信仰其他教的。
- a. 对这些数据构造一张列联表。
- b. 把每一个地区宗教信仰的分布化为百分比。
- c. 这些数据中的两变量间有关吗?
- d. 两变量的关系有多强?

① 原文误为 0 - 9——译者注。

② 原表中的数 57 误为 56——译者注。

- e. 这个关系比你认为的强、相同、还是弱?
- f. 基于这些数据, 你能认为产生这些数据的总体中存在一定的关系吗?
- g. 数据中犹太人总数很少, 去掉犹太人这一类, 重新作 χ^2 分析。
- h. 比较两个 χ^2 , 并说明他们有什么不同。

9.34 要对两组人在一个社会经济变量上的得分进行比较。在第一组, 有 80 个人的得分小于全部数据的中位数, 而 40 个人的得分大于中位数。在第二组, 有 10 个人的得分小于中位数, 50 个人的得分大于中位数。

- a. 构造一个 2×2 的列联表, 把组作为变量, 小于或大于中位数作为另一个变量。
- b. 为什么这两行有相同的观测并没有什么奇怪的?
- c. 两组之间在社会经济变量上的值有什么不同吗?
- d. 两变量的关系有多强?
- e. 你能下结论说产生这两组的总体之间存在不同吗?

9.35 关于注册的选民的宗教信仰和所属党派的一个随机样本数据如表 9.26 所示。

表 9.26 习题 9.35 的数据

		宗教信仰				总计
		天主教	犹太教	新教	其他	
政 党	民主党	575	75	275	90	1015
	共和党	325	25	325	110	785
	无	150	10	120	50	330
	总计	1150	110	720	250	2130

- a. 基于观察这张表并比较这些列, 数据的两个变量间有一定的关系吗?
- b. 这个关系看上去是强还是弱?
- c. 计算 χ^2 值, 并讨论总体中变量间是否有一定的关系?
- d. 计算 V 。

9.36 多年前进行了一次关于服用阿斯匹林和心脏病发作的研究。有 22071 个医生参与了这项研究。每天服用一片阿斯匹林的 10037 个医生中, 104 个人在研究期间有过一次心脏病发作。在服用安慰剂的 10034 个医生中, 有 189 人在研究期间有过一次心脏病发作。

- a. 在一个表中有效地安排这些数据。
- b. 两变量间的相关程度如何?
- c. 基于你所了解的由收集数据所产生的争论, 这些结果是否意味着每人每天都应服用一片阿斯匹林?

9.37 在宾夕法尼亚东部的一个小文科学院里, 教师的性别和职称分布如表 9.27 所示。

- a. 性别和职称间有关吗?
- b. 相关程度如何?
- c. 产生这些数据能仅归于偶然因素吗?
- d. 你怎样解释数据中的模式?

表 9.27 习题 9.37 的数据

		性别		
职 称		女	男	总计
	教授	9	63	72
	副教授	27	20	47
	助教	32	30	62
总计		68	113	181

来源: Swarthmore College Bulletin, 1995-1996.

- 9.38 关于医疗渎职的投诉越来越普遍了。表 9.28 表示对 567 个妇产科的投诉在下面两个变量上的分布情况: 即医生是否有行医的执照及医生是否接受过外国、讲非英语的医学院的培训。关于医疗渎职的投诉这个表告诉你什么?

表 9.28 习题 9.38 的数据

		医学院		
执 照		讲英语	不讲英语	总计
	是	443	42	485
	否	50	32	82
总计		439	74	567

来源: Bruce Cool, "Using medical malpractice data to predict the frequency of claims: A study of Poisson process models with random effects," Journal of the American Statistical Association, vol. 86 (1991), p. 286.

- 9.39 欧洲共产主义政权失败以前的最后一次奥运会是在 1988 年举行的。关于这些共产主义国家是如何强调其体育运动及其妇女在体育中的角色。表 9.29 是那年获金牌最多的三个国家中不同性别的人获金牌的数目。
- 这些数据中国家和性别间有关系吗?
 - 该关系的强度如何?
 - 这种相关性能仅仅归于偶然性吗?
 - 这个表给了我们当时这些国家中体育和性别的关系的什么信息?

表 9.29 习题 9.39 的数据

		国家			
性 别		苏联	东德	美国	总计
	男	41	15	23	79
	女	14	22	13	49
总计		55	37	36	128

来源: Information Please Almanac, 1992.

- 9.40 在关于美国人是怎样消磨时间的一次调查中, 人们被问到他们最喜欢什么活动。表 9.30 是受过中学或以下教育的在职男子和妇女认为工作还是社会活动是他们最喜欢的活动的人数。对于这些数据, 存在明显的性别差异吗?

表 9.30 习题 9.40 的数据

		性别		
		男	女	总计
喜欢的 活动	工作	64	37	101
	社会生活	27	33	60
	总计	91	70	161

($\phi = 0.18$)

来源: John P. Robinson, *How Americans Use Time*, New York: Praeger, 1972, p. 122.

- 9.41 在习题 9.40 中所述的关于美国人是怎样消磨时间的研究中, 表 9.31 是受过中学或以下教育的在职男子和妇女认为看电视还是或参加社会活动是他们最喜欢的活动的人数。这些数据告诉你性别和最喜欢的活动间有何差异?

表 9.31 习题 9.41 的数据

		性别		
		男	女	总计
喜欢的 活动	看电视	43	21	64
	社会生活	27	33	60
	总计	70	54	124

- 9.42 NCAA 收集了 1980 年代中期运动员毕业率的数据。在 2332 个男子中, 有 1343 个未能从大学毕业; 在 959 个妇女中, 有 441 个没有毕业。(来源: *The Chronicle of Higher Education*, July 10, 1991, p. A30.)
- a. 将性别作为自变量, 是否毕业作为因变量构造一张 2×2 的列联表。
- b. 分析两变量间的关系。
- 9.43 对于一个三行、四列的列联表自由度为多少?
- 9.44 对于某一样本 $\phi = 0.69$ 而 p -值 < 0.0001 。你怎样向不上这个课的人解释这句话的含义? (试图用两种完全不同的方法来解释。)
- 9.45 用习题 7.53 中数据构造一 2×2 的列联表, 并分析汽车的主人和总统选举的关系。(如果你用收集到的所有的民主党人的百分比, $209/250 = 83.6\%$ ^①, 这个习题中 $\sqrt{\chi^2}$ 等于习题 7.53 中的 z , 而且对于 z 的双边 p -值, 这两个 p -值是相等的。)
- 9.46 到 1992 年春天, 在过去的 15 年中有 34 个州对罪犯执行过死刑。对于北部和南部各州在过去 15 年执行过死刑情况的比较如表 9.32 所示。

表 9.32 习题 9.46 的数据

		州		
		北部	南部	总计
执行 死刑	0 次	14	2	16
	1 次或更多	7	11	18
	总计	21	13	34

来源: NAACP *Legal Defense and Educational Fund*, as reprinted in *The New York Times*, April 21, 1992, p. A14.

① 原文为 83.6——译者注。

a. 分析两变量间的关系。

b. 在得出这种关系是因果关系前, 你认为应该考虑其他什么变量?

9.47 分析习题 8.37 中数据。

9.48 分析习题 8.40 中数据。

9.49 妇女怀孕期间也许会由于高血压或者蛋白尿或者两者的原因出现妊娠毒血症。在一次对英国妇女的调查中, 下面数据是四类中每一类的数目。

高血压和蛋白尿	仅蛋白尿	仅高血压	其它	总计
28	82	21	286	417

来源: P. J. Brown, J. Stone, and C. Ord - Smith "Toxaemic signs during pregnancy," *Applied Statistics*, vol. 32(1983), pp. 69 - 72

a. 用一个两行和两列的列联表来描述这些数据。第一行作为高血压的妇女的数据, 第二行为没有高血压的妇女。类似地, 第一列为呈现蛋白尿的妇女, 第二列为没有蛋白尿的妇女。

b. 分析高血压和蛋白尿的关系。

9.50 打鼾不仅听起来令人不悦, 而且对打鼾的人也没有好处。表 9.33 是一次调查所得的数据。分析这些数据。

表 9.33 习题 9.50 的数据

	从不打鼾	每一晚都打鼾	总计
有心脏病	24	30	54
没有心脏病	1355	224	1579
总计	1379	254	1633

来源: P. G. Norton and E. V. Dunn, "Snoring as a risk factor for disease: An epidemiological survey," *British Medical Journal*, vol. 291 (1985), pp. 630 - 632

9.51 何杰金氏病(Hodgkin's disease)是一种淋巴结癌。表 9.34 是根据组织学分类和按对治疗的反应进行的分类。分析这些数据。

表 9.34 习题 9.51 的数据

		组织学分类				
		淋巴细胞显著	结硬化	混合蜂窝状	淋巴细胞枯竭	
对处	有反应	74	68	154	18	314
理的	有部分反应	18	16	54	10	98
反应	无反应	12	12	58	44	126
总计		104	96	266	72	538

来源: B. W. Hancock et al., *Clinical Oncology*, vol. 5(1979), pp. 283 - 297

9.52 经过长期的争论后, 1994 年英国国教终于设置了它的第一批女牧师。表 9.35 是 27 年前, 即 1967 年, 关于这种事情投票的结果。关于 1967 年选举, 这些数据告诉你什么?

9.53 在抽样调查中, 那些未回答者和回答者有什么不同吗? 一项在费城进行的对 293 个非洲裔美国妇女跟踪调查中, 研究者采访了其中的 95 个。表 9.36 表明这些妇女第一次被采访时她们可能会说期望得到多少教育。

- a. 比较两组人百分比并说明它们是否不同。
b. 把两组数据看成为随机样本, 并检验两者没有区别的零假设。

表 9.35 习题 9.52 的数据

选 举		投票组			总计
		主教院	教士院	俗人院	
	同意	1	14	45	60
	同意	8	96	207	311
	弃权	8	20	52	80
	总计	17	130	304	451

来源: The Daily Telegraph, July 4, 1967.

表 9.36 习题 9.53 的数据

		是否前次调查的回答者		总计
		是	否	
所 期 望 的 教 育	部分完成中学	2	4	6
	中学	17	32	49
	大学	32	60	92
	专业学位	8	15	23
	技术/商业学校	34	87	121
	不知道	2	0	2
	总计	95	198	293

来源: Roberta R. Iversen, Income and Employment Consequences for African-American Participants of a Family Planning Clinic: A Seven-Year Follow-up, Doctoral dissertation, Bryn Mawr College, 1991. Dissertation Abstract International 52: 1522A.

- 9.54 有些学生在一门课开始时选定用“通过/不及格”的计分方法, 而不是用字母制。我们根据学生的表现把我们的一个初等统计的班分为前后两半。然后数出每组喜欢用“通过/不及格”的学生的个数(表 9.37)。分析表现和计分方法的关系。

表 9.37 习题 9.54 的数据

记 分		表现		总计
		前一半	后一半	
	字母	37	29	66
	通过/不及格	5	13	18
	总计	42	42	84

- 9.55 对习题 9.54 中同一个班, 我们根据学生的姓氏分为两半(表 9.38)。

表 9.38 习题 9.55 的数据

记 分		名字		总计
		A-K	L-Z	
	字母	35	31	66
	及格/不及格	7	11	18
	总计	42	42	84

a. 分析姓氏和计分方法的关系。

b. 评价这里和习题 9.54 的数据和结果的差异。

9.56 1995 年秋,在 Swarthmore 学院随机对一组学生进行了调查,学生被问到他们的专业是哪一类。调查者还记录了学生的性别。数据如表 9.39^① 所示。分析两变量间的关系。

9.57 一项关于在 School Sisters of Norte Dame 的 Milwaukee 女修道院的修女的阿尔茨海默氏^②的研究中,研究者研究了修女们作为年轻姑娘刚进入修道院时的写作风格——这表现在她们的自述中。在死后进行的大脑阿尔茨海默氏病检查中得到的修女的结果列在表 9.40 中。

表 9.39 习题 9.56 的数据

		性别		总计
		男	女	
分 支	自然科学/工程	13	4	17
	人文学	8	7	15
	社会科学	7	8	15
总计		28	19	47

来源: Data used by permission of Heather Repenning, Swarthmore College

表 9.40 习题 9.57 的数据

		写作风格		总计
		语言能 力较弱	语言能 力较强	
阿尔茨海 默氏病	有	9	1	10
	无	2	13	15
	总计	11	14	25

来源: The Philadelphia Inquirer, February 21, 1995, p. A1.

a. 文章报导尸体解剖后发现患阿尔茨海默氏病的 10 个修女中有 90% 的语言能力较低,而没有患这种疾病的修女中有 13% 的修女语言能力较低。这篇文章选自变量和因变量的方法意味着什么? 在这项研究中,你选什么作为自变量和因变量?

b. 分析两变量间的关系。

9.58 统计分析的理想情况是对于那些适当收集的随机样本进行分析。然而,要使收集数据简化,对两个分类变量(女/男、高/矮、瘦/不瘦,等等)观测 20 个左右的学生。

a. 对于这项数据构造一张列联表。

b. 作一个像图 9.1 那样的条形图。

c. 通过把频率变为百分比创建一个新表。注:可以用不同的方法计算百分比,所以谨慎遵循本章的线索(并参考表 9.2)。你能否说你的数据的百分比很大?

d. 两变量间关系的强度如何?

e. 两变量间有统计上显著的关系吗?

9.59 a. 作为班上的一个计划,找出班上所有学生身高的中位数。

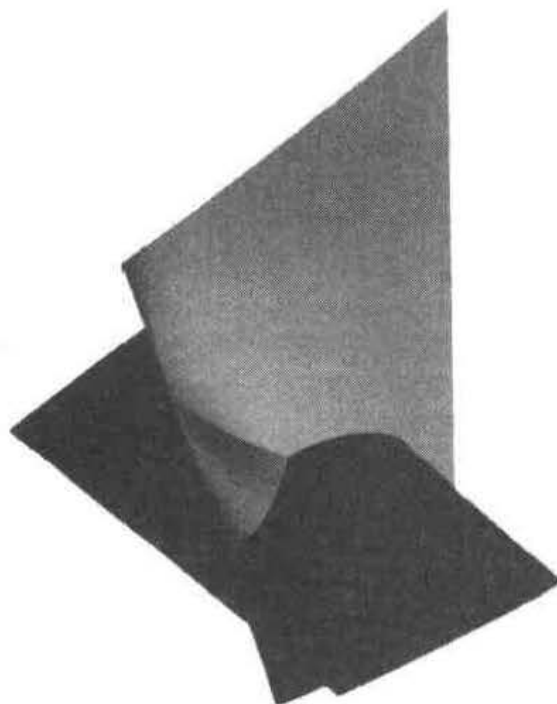
b. 构造一个关于每一个学生的性别和其身高是否高于(或低于)中位数的列联表。

c. 性别和身高关系的强度如何?

d. 这个关系统计上显著吗?

① 原表中最后一行三个数目误为 22,28 和 50——译者注。

② Alzheimer's disease, 早老性痴呆——译者注。



10.1 问题 1. 两个变量间的关系?

10.2 问题 2a. 关系的强度?

10.3 问题 2b. 关系的形式?

10.4 问题 3. 总体中的关系?

10.5 警告: 所测即所得

10.6 用虚拟变量时怎样变得聪明些

10.7 问题 4. 是因果关系吗?

10.8 小结

两个数值型变量的

回归分析和相关分析



高脂肪的食物比低脂肪的食物含有更多的热量吗？汽车的重量和它消耗一加仑汽油所行的平均英里间有什么关系？在不同的州，犯罪率与偷窃率有何关系？在不同的国家，香烟消费与患癌症率有关系吗？在不同的州，低收入人口的百分比与低教育水平的人的百分比有何关系？父母的身高与他们的孩子的身高有关吗？在过去的几十年中，恶性黑色素瘤有什么变化吗？自从 1954 年 Roger Bannister 突破 4 分钟大关后，男子一英里赛跑的世界纪录有什么变化？

回归分析描述的是一个或多个自变量的变化是如何影响因变量的一种方法。相关分析描述的是两个数值变量间关系的强度。

回归分析 (regression analysis) 和相关分析 (correlation analysis) 这两种统计方法可以回答一些明确定义的度量单位的数值变量之间的关系的问題；这些数值变量包括诸如食物的热量和脂肪的含量的关系，汽车一加仑汽油所行的平均里程和其重量间的关系等等。回归分析和相关分析代表了分析两个数值变量间关系主要和互补的两种方法。

统计上，回归是比日常用语更加专业化的术语。通常，我们认为回归是能力或表现的向回走。统计的回归名称来自于早期研究父母和他们的孩子的身高时所采用的方法（见下面的方框中）。这个研究发现一个趋势就是孩子比身材很矮或很高的父母更加趋于平均值。这种向中间值的趋势称作回归效应。

趋向中间高度的回归

回归这个术语是由英国著名统计学家 Francis Galton 在 19 世纪末期研究孩子及他们的父母的身高时提出来的。正如我们所预料的, Galton 发现身材较高的父母, 他们的孩子也较高。但是这些孩子平均起来并不像他们的父母那样高。对于比较矮的父母情形也类似: 正如我们所预料的那样, 他们的孩子比较矮, 但这些孩子的平均身高要比他们的父母的平均身高高。(这也是很自然的情况, 因为如果身材高的父母他们的孩子比他们更高, 而身材矮的父母, 他们的孩子比他们更矮, 那么许多代以后, 我们将高矮相差越来越大。) Galton 把这种孩子们的身高向中间值靠近的趋势称之为一种回归效应, 而他发展的研究两个数值变量的方法称为回归分析。

作为老师, 我们在我们的班上也看到了同样的回归现象。期中考试时成绩比较高的好学生在期末考试时成绩也好, 但平均不像期中考试那么好。类似地, 期中考试时成绩比较差的学生期末考试时平均要好一些。在体育上, 回归效应也是一种很突出的现象: 第一年成绩很突出的新手, 第二年成绩往往就不是那么好。然而, 在所有用到回归分析的情形中, “回归”的特征并非那么明显。

相关分析度量了数值变量间的关系强度。两个变量间可以有高的相关性, 也可以有低的相关性。这依赖于它们的关系有多强。相关这个词在统计上和日常用语中的意义很类似。

停下来想一想 10.1

举出两个可能相关的数值变量的一个例子, 并列出让变量的一些值。你所举的变量为何是数值变量?

在这一章, 我们研究两个变量的回归分析, 这种类型的分析称为简单回归分析 (simple regression analysis)。你也许会发现这个名字有点讽刺意味, 因为乍一看上去, 事情并非那么简单。但是这里“简单”是表示两个变量而不是多个变量的一种方法。简单回归分析是回归分析最简单的情形。从 Galton 时代到现在, 这种方法已被推广到多个变量的情形, 我们在 13 章将



用作回归分析的食物。(来源:1992. Comstock.)

对此再作研究。

让我们从节食者的困境开始。你试图节食却又非常馋。站在一台自动售货机前,看着这诱人的选择:炸薯片、椒盐卷饼、爆玉米花、糖果条。当你犹豫的时候,脂肪和碳水化合物看上去在向你打招呼呢。哪份小吃对你的节食危害最小呢?哪种含有最少的热量呢?是高脂肪的,还是低脂肪的?有什么区别吗?你现在需要知道的是随着食物中脂肪的含量的增加,热量是增加还是减少。回归分析将给你一个寻找答案的办法。要处理脂肪含量和热量的关系的问题,我们先看看已有的数据(表 10.1)。要做一个简单的回归分析(和相关分析),数据文件中必须含有两列数,每个变量一列。另外,数据文件中还应该包括一列,以便能够识别每一行。表 10.1 中的每一行包含了一特定小吃的数据。第一列表示小吃的名字,第二列是热量,第三列是脂肪的含量。

表 10.1 点心食物中的热量和脂肪

食物	热量(卡 ^①)	脂肪(克)
玉米饼(15)	110	4
炸薯片(18)	120	6
奶酪味小吃(34)	120	6
炸面饼圈(1)	164	8
苹果馅饼(1/6 个 8 英寸大的饼)	430	19
爆玉米花(3 杯)	192	11
冰激凌(1/2 杯)	175	12
巧克力条饼干(1 大号)	236	12
奶酪饼干(2 盎司,10 个薄的)	429	26
鸡翅膀(2)	318	21
奶酪面包圈	249	11
花生酱杯(2)	281	16
干烤花生(1oz)	160	14
巧克力条(1oz.)	147	9
奶酪或花生酱饼干(6)	210	9
麦片条(1)	120	5

来源: ASDA data and manufacturer's data shown as an advertisement in The New York Times Magazine, April 20, 1990, p. 20.

停下来想一想 10.2

看看表 10.1 中的数据。为什么所有这些食品所含的热量都不同?是因为它们含有不同量的脂肪吗?如果是这样的,我们感兴趣的是热量值的变化。你是否能由数据得出来这些观察值互相之间如何不同吗?

从表 10.1 出发,我们能在多大程度上回答脂肪和热量有怎样的关系这个问题呢?大致地

^① 原表上把卡误为千卡(kcal)——译者注。

看一下数据,我们发现高脂肪的食物含有的热量也较高,而低脂肪的食物含有的热量则较低,这两个变量看上去是相关的。但要得到数据包含的详细信息——例如,一种食物如果含有两倍于另一种食物的脂肪,其热量是否也为另一食物的两倍呢——我们要利用回归分析和相关分析。

10.1 问题 1. 两个变量间的关系?

当我们注意到在脂肪变量值比较小相应的热量变量值也较小,和当脂肪变量值比较大相应的热量变量值也较大时,我们已经回答了这个问题,这时两变量间存在一定的关系。为了解这一关系的细节,我们需要分析这些数据。和以前一样,从这些数据出发,我们可以作一个图或者一张表,还可以计算一、两个数。

作这些数据的散点图

散点图是以横轴为自变量,纵轴为因变量的一个图。图上每一个点代表一对观测值。

对于两个数值变量,我们总是用一个图来开始分析这些数据。作图是为了使我们对于两个变量间的关系有一个直观上的印象。这个图还告诉我们,对于这些数据,我们是否可以利用统计上的回归分析和相关分析的方法。并非关于任意两个数值变量的数据都可以用这两种方法。一个图通常会告诉我们是否可以这样作。

我们作的图称为散点图(scatterplot)。水平的 x 轴为自变量,垂直的 y 轴为因变量。对于表 10.1 的数据,由于我们假设脂肪含量(自变量)影响热量(因变量),所以脂肪含量作为 x 轴,而热量作为 y 轴。看一下数据,玉米饼的 x 值为 4, y 值为 110。我们在坐标(4,110)处用一个点标出这两个数。我们同样可以把其他小吃的值在图中用点描出来,最后作出这个散点图如图 10.1 所示。

停下来想一想 10.3

对下面的几例,说明:在一个散点图上,哪个变量应在 y 轴,哪个在 x 轴? 什么时候不清楚?

- a. 投掷的曲线球数和本垒打得分。
- b. 头盔的大小和棒球手套的大小。
- c. 打棒球时三击不中出局次数和年薪。
- d. 偷垒的次数和打棒球的次数。

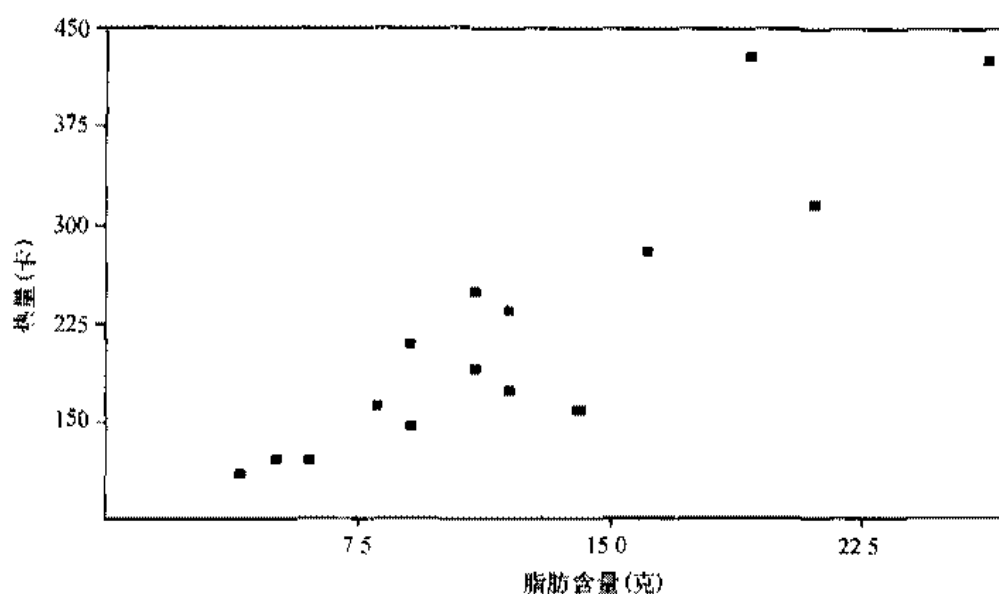


图 10.1 表 10.1 的数据的散点图。

停下来想一想 10.4

下面的数据是 1983 年到 1993 年间,来自赞助和商业(非联邦政府和其他来源)用于公共广播的基金。单位是百万美元。

年份	来自赞助商	来自商业
1983	180	110
1984	195	135
1985	225	175
1986	240	175
1987	275	195
1988	300	210
1989	320	240
1990	340	260
1991	370	290
1992	390	300
1993	395	300

来源: *Foundation for Public Broadcasting* .

- 对于这些数据作一个散点图。
- 数据中两变量间看上去有一定的关系吗?
- 是负相关还是正相关?
- 如果变量间有很强的关系,你认为散点图会是什么样的? 这个关系看来很强吗?
- 在这一阶段末,两个变量间的关系发生了什么变化? 这是否意味着由于 1995 年由于共和党国会提议削减公共广播的基金,来自大商业之间的赞助高涨了?

对于散点图总是自变量作横轴,因变量作纵轴。这和我们构造两个变量的列联表时的安

排是一样的。当然,有时并不清楚哪一个是自变量,哪一个是因变量。例如,对班上学生的身高和体重,到底是哪一个在影响另一个并不明显。这种自变量和因变量的选择只影响回归分析,但并不影响相关分析。对这两种选择,相关分析的结果是一样的。

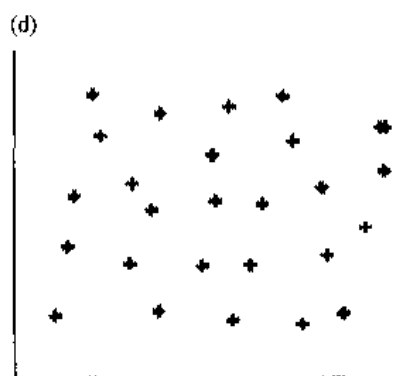
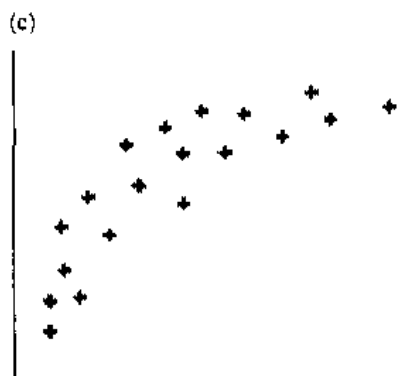
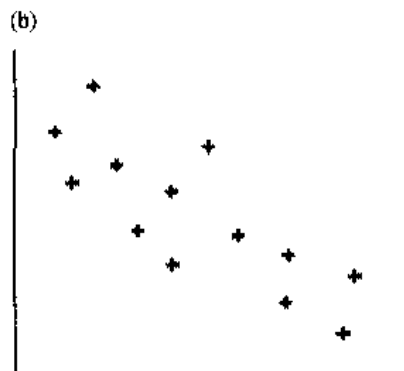
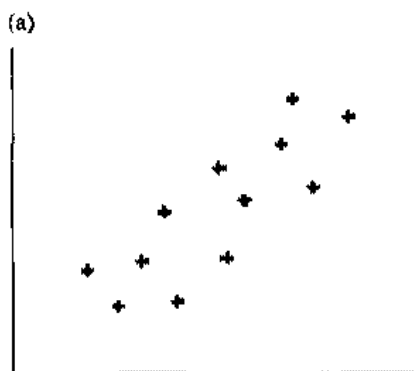
了解散点图

散点图 10.1 表明,一份小吃中脂肪含量越高,热含量亦越高。图中点的趋势说明两变量间确实存在一定的关系。这个图支持了我们仅仅从数据表所得出的结论。从这个图我们确信这两个变量是相关的。

另外,由于这些点散布在从左下角到右上角的区域,说明这两个变量是正相关的,也就是说,小吃中脂肪越高,热量亦越多。一些变量,比如汽车的重量和汽车消耗一加仑的汽油所行的平均里程是负相关的;汽车越重,消耗一加仑的汽油所行的平均里程就越短。在一个两变量负相关的散点图中,这些点散布在从左上角到右下角的区域。

停下来想一想 10.5

下边哪一个散点图表明可以用回归分析和相关分析? 用一句话来描述每一个散点图表达了什么样的关系?



许多统计软件都可以作散点图。图 10.1 就是把数据输入一个计算机文件后用计算机作出来的。如果数据中观测个数不是很多,我们也可以用手工来作散点图。

线性关系

再考虑另一个问题:当 x 值(脂肪)增加或减少时, y 值(热量)以什么方式不同?要知道两个分类变量间是否有一定的关系,我们取自变量的每个值来看看因变量相应的百分比分布。这里因变量是一个数值变量,所以我们用其平均值而不是百分比。

我们处理数值变量的方法和处理分类变量的方法差不多。我们取自变量脂肪的一些值,看看因变量(热量)的相应的值。例如,对脂肪为大约 7.5 g 时,相应的热量平均值为 150,脂肪为大约 15.0 g 时,相应的热量平均值大约为 250。回归分析就是基于对于自变量的不同取值,因变量相应的平均值也不同这一事实。如果数据足够多——对于脂肪含量的每一个值,热量变量都有许多值——我们就可以对脂肪变量的每一个值来计算热量的实际平均值了。

如果对脂肪含量不同的值,热量平均值也彼此不同,那么我们可以认为这两个变量是相关的。另外,在一个散点图中代表平均值的那些点分布在通过散点图中心的一条直线旁,我们就可以对这些数据用回归分析和相关分析。在这里我们没有足够的数据来计算平均值,但这些数据点多少分布在一条直线旁边,我们可以继续作下去。

如果散点图中的点的分布看上去像一条曲线,我们就不能用这些分析了。如果这些点像云一样,没有任何模式,这些数据也许是随机的,而变量间没有任何关系。

10.2 问题 2a. 关系的强度?

相关系数 r 在 -1 到 0 到 1 这个范围里刻画了两个数值变量间的关系强度。

当这些数据沿一条直线排列时,我们可以计算一个系数来衡量两个变量间的关系。对于两个数值变量,计算出来的系数记作 r ,尽管它有许多名字,我们简称其为**相关系数**(correlation coefficient)。还可以把它称为**线性相关系数**(linear correlation coefficient),**Pearson 相关系数**(Pearson's correlation coefficient, 这是为了纪念英国统计学家 Karl Pearson, 他在这方面作了许多重要的工作),或者**乘积相关系数**(product-moment correlation coefficient),它反映了相关系数的计算方法。

r 是正的还是负的? 大还是小?

对于脂肪/热量数据的相关系数 $r = 0.91$ (由公式 10.1 计算所得)。比计算这个系数更重要的是这个系数的意义。首先注意到对这些数据 r 是正的。这意味着,如果一个变量的值比较小,那么另一个变量的值也较小,而如果一个变量的值比较大,那么另一个变量的值也较大。所以,对于脂肪含量较低的玉米饼,热含量也较低。而对于脂肪含量较高的奶酪饼干热含量也较高。正值 r 证实了散点图的趋势。

我们注意到的关于 r 的第二件事情是它的大小。很明显,0.91 几乎等于最大的可能值 1,这意味着两变量间很强的相关性。根据最常用的准则,介于 -0.75 和 -1.00 之间的任何 r 值

代表了一个很强的负相关性；介于0.75和1.00之间的任何 r 值代表了一个很强的正相关性。类似地，介于-0.70和-0.30之间和0.30到0.70之间的 r 值代表了一个适中的相关性。而-0.25和0.25之间则表示相关程度比较弱。

这些准则仅仅是比较粗糙的方法。对于在不同领域工作的人们，寻找 r 的范围亦不同，高和低仅仅是看作相对于某一领域的 r 的普通值来说的。一个社会学家通常会认为 $r = 0.50$ 就很高了，而一个经济学家也许会认为 $r = 0.50$ 比较低。然而，0.91对于几乎所有的人们来说都代表了一种很强的关系。

四种不同的散点图：关系从强到弱

让我们通过几个散点图来看看为什么散点图的不同趋势会导致不同的 r 。图10.2是四个不同的散点图，每个有100组观测值。这些数据是由计算机生成的，并没有实际意义，所以两个轴没有标识或刻度。

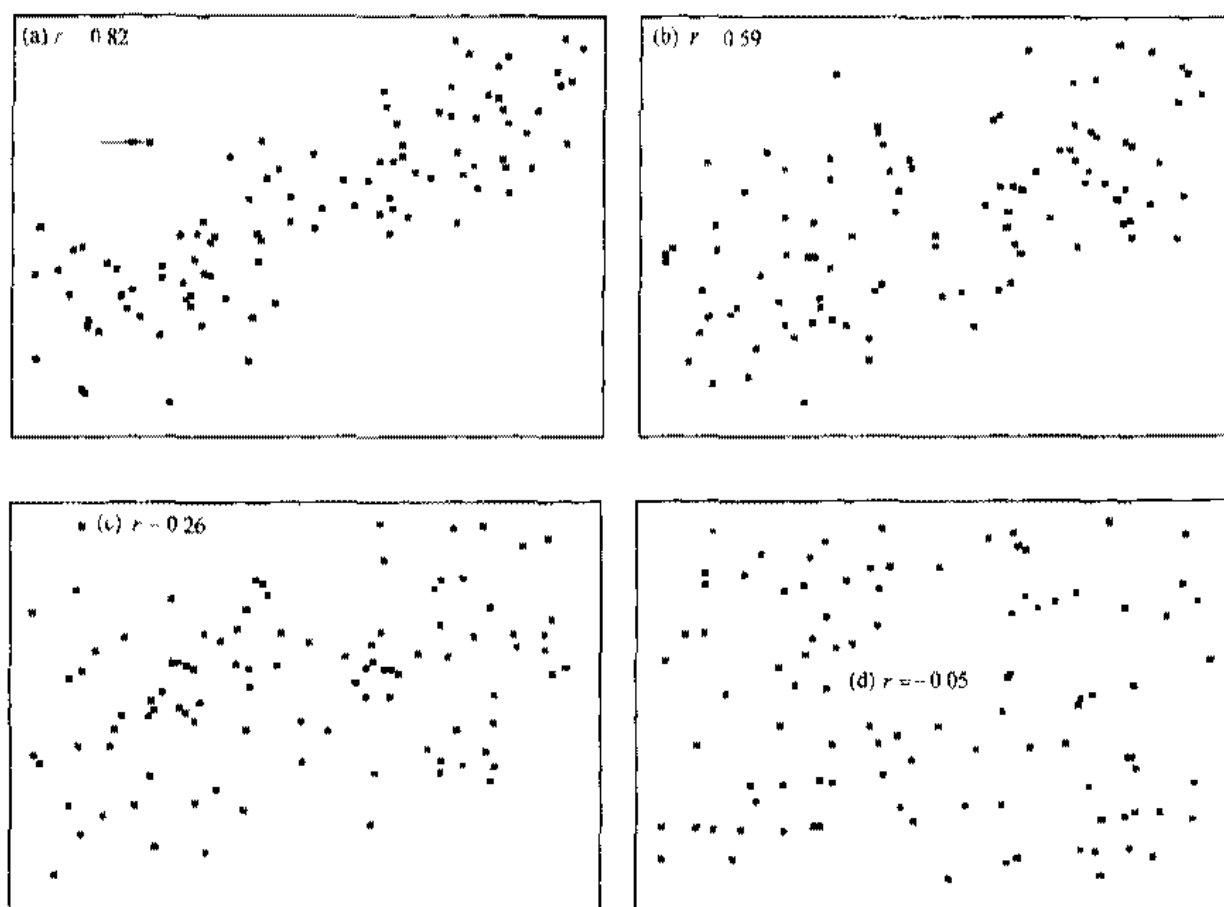


图 10.2 不同相关性的散点图。

在图10.2a中，这些点排成一束，彼此很接近。我们在图上可以看到一种从左下角到右上角的明显的直线趋势。这些点沿对角线呈一种规则的分布。该两变量间的关系应该很强，相关系数证实了这一点，在这里 $r = 0.82$ 。

在图10.2b中，这些点不像10.2a中那样，明显的排成一束。但从这些数据中我们仍然可

以看到一种确定的正相关性,而由散点图得到的相关系数 $r = 0.59$ 。在图 10.2c 中,相关系数减小为 0.26,这意味着一个较弱的关系。对于这样大小的 r ,几乎不可能从这些点中看出任何模式,也看不出这两个变量是否相关。对图 10.2d,这些点是随机散布的,两变量间几乎没有关系。

一个异常的观测

在图 10.2a 中,标有箭头的那个点离其他点较远。它离其他点有较大的距离。一个散点图中离其他点较远的那些点会对相关系数产生不成比例的较大的影响。在分析时如果包含了图 10.2 中的这个孤立点, $r = 0.818$ (精确到三位小数)。而在计算 r 时如果不包含这个点, $r = 0.835$ (这个数通常记为 0.82 和 0.84)。如果 100 个数据中有 5 个这样的离群点,计算 r 时包含和不包含这些点将有显著的差异。这些点如果离其他点越远,对于计算相关系数的影响就越大。

这些相关系数的差异表明了相关系数对远离数据主要趋势的点如何敏感。这就是为什么我们计算 r 前要看一看散点图。通过看散点图,我们可以了解到是否有那些值得怀疑的数据点。

图 10.2 表明了两变量间的正相关性。也可以生成 r 为负值的数据。在这样的散点图中,点会分布在另一条对角线,即从左上角到右下角。在这种情形, x 变量越大, y 变量则越小; x 变量越小, y 变量则越大。例如,若 x 代表汽车的价钱, y 代表汽车的销售量, r 则为负值:贵的汽车相对于便宜的汽车销售量要小。

这里有一些其他的 r 值来说明相关的程度。在一个汽车的例子中,车重和消耗一加仑的汽油所行的平均里程的 $r = -0.90$,对马力和消耗一加仑的汽油所行的平均里程的 $r = -0.87$ 。许多核电站的耗费和发电量间的相关系数竟令人吃惊的小,仅为 $r = 0.47$ 。表 10.1 中的两个变量加上另外两个变量的相关系数 r 如表 10.2 所示。

表 10.2 表 10.1 中的两个变量加上胆固醇和钠的相关系数表

变量	热量	脂肪	胆固醇	钠
热量	1.00			
脂肪	0.91	1.00		
胆固醇	0.62	0.69	1.00	
钠	0.73	0.59	0.41	1.00

在表 10.2 中对角线上相关系数为 1.00。每个 1.00 代表了一个变量和它自己的相关系数。一个变量和它自己的相关系数始终为 1.00。设想一个散点图,横轴和纵轴均为同一个变量。所有的观测将落在一条 45 度的直线上。在一条直线上的观测值之间的相关程度达到最大,且 $r = 1.00$ 。

我们已有了关于脂肪和热量的散点图(图 10.1),对于其他 5 个相关系数我们可以通过已知的 r 来作近似的散点图。最小的 r ——钠和胆固醇的相关系数——为 0.41。因为这个关系并非那么强,所以在这些数据的散点图上,我们将看到这些点有一个一般的向上趋势,但相对比较松散。



核电站的耗费和它们的发电量的相关系数小得令人吃惊： $r = 0.47$ 。(来源：1992, Comstock.)

停下来想一想 10.6

对于下面两变量间的相关关系,是正相关还是负相关?

- a. 激光唱片是否畅销及其价格
- b. 办公室的大小和任职的工资
- c. 汉堡包的价钱和销量
- d. 室外的温度(在 40 到 80 华氏度之间)与游泳馆卖的票数

r 的解释: 不那么严谨

我们已经说过相关系数 r 度量两个变量的关系强度; r 的取值为从 -1 到 $+1$ 。然而,很难想出一个对 r 的确切解释。我们知道, $r = 0.91$ 意味着两变量间有一个很强的关系,而 $r = 0.41$ 则代表一个适中的相关关系。但除了强和适中这些词以外, r 到底意味着什么?

要了解两个数值变量间相关程度的确切解释,我们来看一看 r^2 而不是 r 本身。对脂肪和热量的例子, $r^2 = 0.91^2 = 0.83$;对钠和胆固醇, $r^2 = 0.41^2 = 0.17$;对于汽车的马力和消耗一加仑汽油所行驶的平均里程间的相关系数为 $r^2 = (-0.87)^2 = 0.76$ 。0.83, 0.17, 0.76 这些数有一个很具体的解释,我们在后面讨论。

10.3 问题 2b. 关系的形式?

回归分析是对两个数量变量进行分析的另一部分。相关分析和回归分析同等重要,对两个变量间的关系进行完整的分析应该包含两者。

统计的一个基本思想是更好地理解数据。我们可以用从这些数据计算出来的一个或几个

数来代替它们。让我们先用一个变量来说明这一点。如果我们看一看表 10.1 中的热量那一列, 可以看到它们的值从较低的 110 卡(玉米饼)变到较高的 430 卡(苹果馅饼)。因为同时理解这些热量值是很困难的, 我们用这个变量的平均值来代替这些数据。这个平均值等于 216.3 卡, 出于许多目的, 我们可以用这个数据来代替初始的数据。

为了找到两变量间的关系, 我们可以计算什么样的一个数据来代替观测到的数据呢? 关于散点图 10.1 的讨论说明通过这些点的中心的一条直线可以代表所有的点。我们可以用这条直线而不是所有的数据来讨论两变量间的关系。

停下来想一想 10.7

为下面的讨论作准备, 什么是一条直线的方程? 代表这样的方程所必须的两个数叫做什么? 一条直线的倾斜度是由纵坐标和横坐标之比来度量的是什么意思?

一条通过点的中心的直线

图 10.3 是在图 10.1 的基础上, 又加了一条穿过这些点的中心的直线。这样的一条直线就是回归直线(regression line)。如果我们擦去这些点而只保留这条直线, 我们仍然可以很清楚地了解脂肪含量和热量的相关性。就像平均值很好地代表了一个变量的数据一样, 这条直线很好地代表了两个变量的数据。

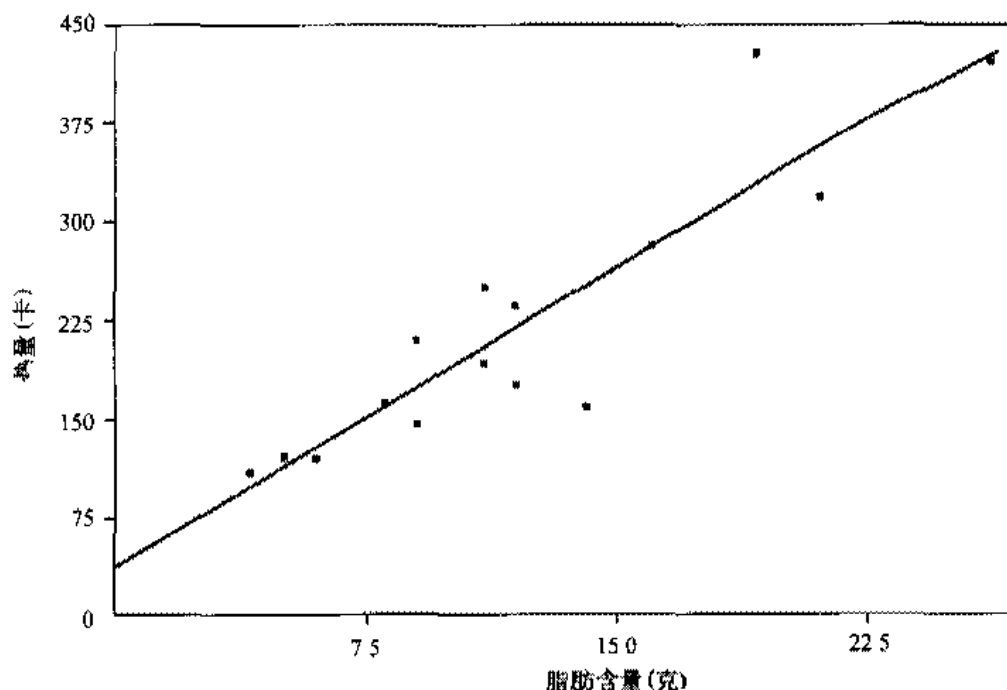


图 10.3 脂肪含量和热量的包括回归直线的散点图。

和这些点间有一个正相关性一样, 这条直线从图的左下角到右上角有一个正的斜率; 也就

是说,低脂肪的食物热量也低,高脂肪的食物热量也高。直线越陡,脂肪含量的单位变化所导致的热量差异就越大。一条回归直线的倾斜度由它的斜率来度量,如果我们知道了图 10.3 中直线的斜率,我们就可以清楚地知道脂肪含量相差一个单位,热量相差多少。

我们可以用一把尺子来测量回归直线的斜率的近似值:测量相应于横向移动的纵向的升起高度。我们还可以由观测到的数据来计算斜率(公式 10.2 和 10.3)。这条直线的斜率等于 15.3 卡/克。所以,两份小吃如果脂肪含量相差 1g,热量将平均相差 15.3 卡。

这一点如图 10.4 所示。这个图表示两种食物,A 和 B。B 比 A 的脂肪含量多 1 克。因为直线的斜率等于 15.3,所以平均起来 B 的热量比 A 多 15.3 卡。我们说“平均起来”是因为两种食物所观测的数据点并不恰好落在直线上。对于两种具体的食物,热量的差异将多于或少于 15.3 卡。但是对于许多对食物,脂肪含量相差 1 克,热量的平均差异将为 15.3 卡。

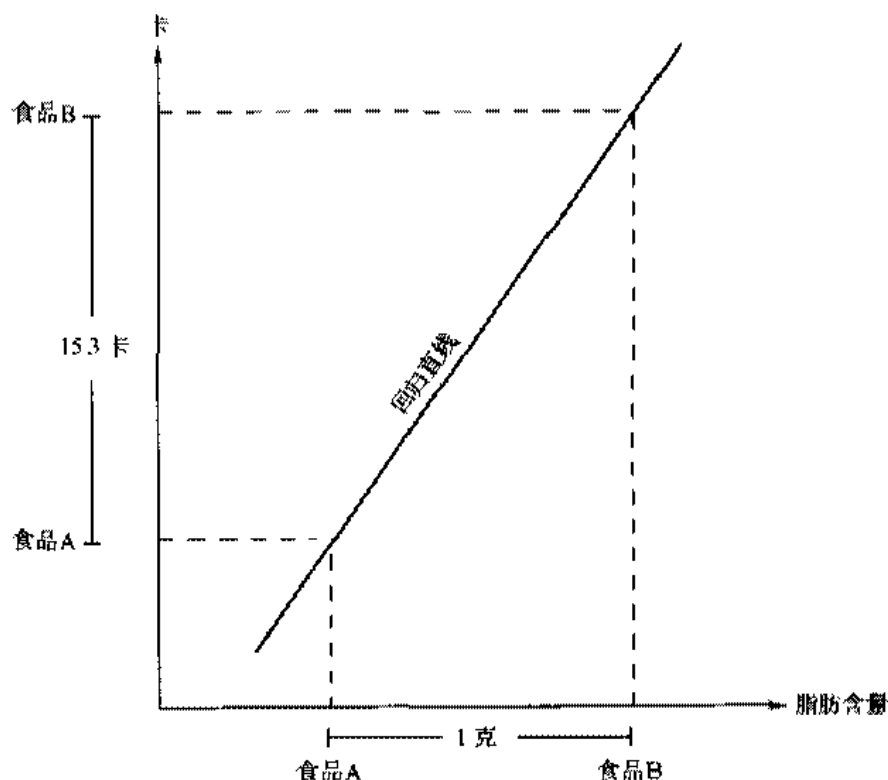


图 10.4 对回归系数 15.3 卡/克的解释。

图 10.3 中的直线在 y 轴上有一个截距,也就是说,当 $x=0$ 时,这条直线与 y 轴相交的那一点。图 10.3 所示的直线向左延伸超过了观测点,这条直线与纵轴相交,交点处脂肪含量的值为 0。这条直线的截距为 36.1 卡。所以这条回归直线的方程为:

$$\text{热量} = 36.1 + 15.3 \text{ 脂肪含量}$$

回归直线的方程称为回归方程(regression equation)。一条回归直线的方程可以写作:

$$\text{因变量} = \text{截距} + \text{斜率} \times \text{自变量}$$

用符号表示为

$$y = a + bx$$

这里 a 为直线的截距, b 为直线的斜率。在图 10.4 中, 斜率 15.3 是自变量(脂肪含量)的回归系数(regression coefficient)。这个回归方程比表 10.1 中的两列数据告诉了我们更多的信息。它用一种很简洁的形式总结了两变量间的关系。

停下来想一想 10.8

学生们总是被鼓励继续他们的教育, 这样他们毕业后就可以挣到更多的钱。分析一个人在学校多呆几年能够多挣多少钱的一个方法是建立一条回归直线来度量受教育年限和年薪的关系。

把受教育年限作为自变量, 也就是横轴 x 轴, 工资作为因变量也就是纵轴 y 轴。如果你有了教育和收入的数据, 你可以作一条回归直线吗? 自变量从 8 年的教育开始到 20 年结束, 年薪可以从一个你认为比较低的数变到一个你认为较高的数。

怎样计算回归直线: 最小二乘原理

回归直线由它的斜率和截距所决定, 这两个数可由本章末的公式计算出。公式 10.2 说明怎样计算斜率 b , 公式 10.4 说明怎样计算截距 a 。推导这些数学公式需要更多的我们不想涉及的数学知识。但我们可以解释得出这些公式的原理。

如果我们每个人都拿一把尺子对图 10.1 中的散点图作一条通过这些点的中心的直线。我们每个人都会作出一条稍微不同的直线来。但每条直线都会和图 10.3 中由作此分析的计算机软件所作的直线差不多。计算机所作的这条直线有一个特点使得它很特别。这条直线是基于每个数据点离这条直线有多远而作出来的。散点图上每个点到回归直线的垂直距离可以用一把尺子量出来, 也可以由数值计算得到, 这样结果更精确一些。得到这些差异后, 我们把每个距离平方然后加起来。对于小吃食物的数据, 距离的平方和等于 27182。这个数给了我们一个这些点到这条直线有多远的一个总体的度量。

如果我们对其他任何直线实施同样的过程, 最后得到的数总会大于 27182。没有其他直线的距离平方和能比它小。所以, 计算机所作的这条直线使这些点到这条直线的距离的平方和最小。从最小平方(least squares)^① 的意义上说, 这条直线就是距所有的点最近的那条直线, 而且在这个意义上, 它比任何其他直线都更好地代表了这些数据。

作为一个说明, 让我们计算几个距离和他们的关系。例如, 对于麦片条, 含有 5 克脂肪。如果我们把这个数代入回归方程中, 就可以得到由麦片条的脂肪含量预测的热量的值:

$$\text{预测热量} = 36.1 + 15.3(5) = 112.4$$

这就是我们在回归直线上得到的热量的值, 也就是说点(5, 112.4)在直线上。(你自己可以通过看图 10.3 的散点图和那条直线来验证这一点。)为了实际计算时更精确一些, 我们对斜率和截距用了更多的小数。因为麦片条实际上含有 120 卡热量, 从观测点到直线的垂直距离为

^① “最小平方”在中国古文中称“最小二乘”, 因此这两个名称目前都在用——译者注。

$120.0 - 112.4 = 7.6$ 。麦片条的实际热量值在回归直线估计值上方 7.6 卡处。其他的垂直距离也可由类似的方法得出。直线上方的点距离为正,直线下方的点距离为负。将这些距离平方并加起来,就得到了和 27182。

麦片条对这个和贡献了 $7.6^2 = 57.7$ 。这不是一个很大的数,因为麦片条这个点离直线很近。其他一些点离直线较远,这些距离的平方将很大。

如果我们大家都同意用最小二乘的方法来寻求这条直线,则我们得到的直线也将一样。如果我们用其他的什么方法,得到的直线将会有所不同。例如,从这些点到这条直线的距离的绝对值的和达到最小的直线就是一条不同的直线。这是由于统计方法而不仅仅是数据决定分析结果的又一例子。



"Cathy" copyright 1995 cathy Guisewite. Reprinted with permission of Universal Press Syndicate. All right reserved.

用回归分析进行预测:从脂肪到热量

把自变量的值
代入回归直线的方
程就得到了因变量
的预测值。

你已经看到了回归直线可以用来进行预测。如果我们知道了食物中含有多少脂肪,我们就可以用回归直线来预测食物中的热量。(我们总是由自变量来预测因变量。)

由于回归分析有这种预测的特性,这个回归方程有时可以写成如下的形式:

$$\text{预测热量} = 36.1 + 15.3 \text{ 脂肪含量}$$

公式左边的“预测”这个词强调了左边这个值仅仅是一个预测值,而不是实际的观测值。另一种表示方法是在方程左边这一项上加上一个“帽子”:

$$\widehat{\text{热量}} = 36.1 + 15.3 \text{ 脂肪含量}$$

“帽子”符号($\widehat{}$)意味着“预测”。另外,有时不用文字而用 y 来代表因变量, x 代表自变量。这样预测值 y 作为 x 的一个函数的方程可以写作:

$$\hat{y} = 36.1 + 15.3x$$

在小吃的例子中,对于含有 5 克热量的麦片条其热量预测值为 112.4 卡。对于其他食物的预测值为:

玉米饼	$36.1 + 15.3(4) = 97.1$
炸薯片	$36.1 + 15.3(6) = 127.6$
奶酪味小吃	$36.1 + 15.3(6) = 127.6$
炸面饼圈	$36.1 + 15.3(8) = 158.1$
⋮	⋮
麦片条	$36.1 + 15.3(5) = 112.4$

另外,我们可以用回归方程来预测一个已知脂肪含量的新食物的热量。我们可以将脂肪含量的值代入回归方程并计算预测的值。

停下来想一想 10.9

你在市场上随手拿起一个糖果条。商标说明这个糖果条含有 3g 脂肪,利用这个例子的结果估计其热量的值。

效果的度量: r^2 的解释

残差变量包含了除自变量外其他所有变量对因变量的效应。

不同小吃中的热量除了受脂肪含量的影响外,还受许多其他变量的影响。其他这些变量称为残差变量(residual variable),这些变量和脂肪共同决定了食物中热量的含量(图 10.5)。图 10.5 中两个箭头表示我们如何认为脂肪含量和残差变量对热量的影响,而箭头上方的问号表示我们并不知道脂肪含量和残差变量对热量的影响程度。我们可以度量脂肪含量和残差变量对热量的影响吗?

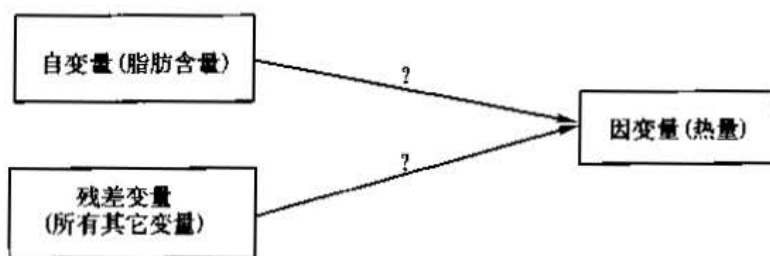


图 10.5 脂肪含量和残差变量对热量的效应。

首先,设想没有变量影响小吃食品中的热量。那么所有的食物将含有相同的热量。脂肪含量的不同不会对热量造成任何不同,其他变量也一样。对于这个想象中的实验,假定这个相同的热量是所有观测的热量值的平均,也就是 216.3 卡。那么所有的热量值都应为 216.3 卡,如表 10.3 所示。

表 10.3 如果无任何变量影响热量变量,脂肪和热量的观测值

食物	脂肪(克)	共同的热量(卡)
玉米饼(15)	4	216.3
炸薯片(18)	6	216.3
奶酪味的小吃(34)	6	216.3
炸面饼圈(1)	8	216.3
苹果馅饼(1/6个8英寸大的饼)	19	216.3
爆玉米花(3杯)	11	216.3
冰激凌(1/2杯)	12	216.3
巧克力条饼干(1大号)	12	216.3
奶酪饼干(2盎司,10个薄的)	26	216.3
鸡翅膀(2)	21	216.3
奶酪面包圈	11	216.3
花生酱杯(2)	16	216.3
干烤花生(1盎司)	14	216.3
巧克力条(1盎司)	9	216.3
奶油或花生油的饼干(6)	9	216.3
麦片条(1)	5	216.3

在一个由脂肪含量和表 10.3 中最后一列的值作出的散点图上,所有的点将落在同一条水平的直线(图 10.6)。但是观测到的数据并没有落在一条水平的直线上。它们的散布情况如图 10.6 所示。这意味着,因变量(热量)的值的变化程度表明热量值受其他变量的影响。我们

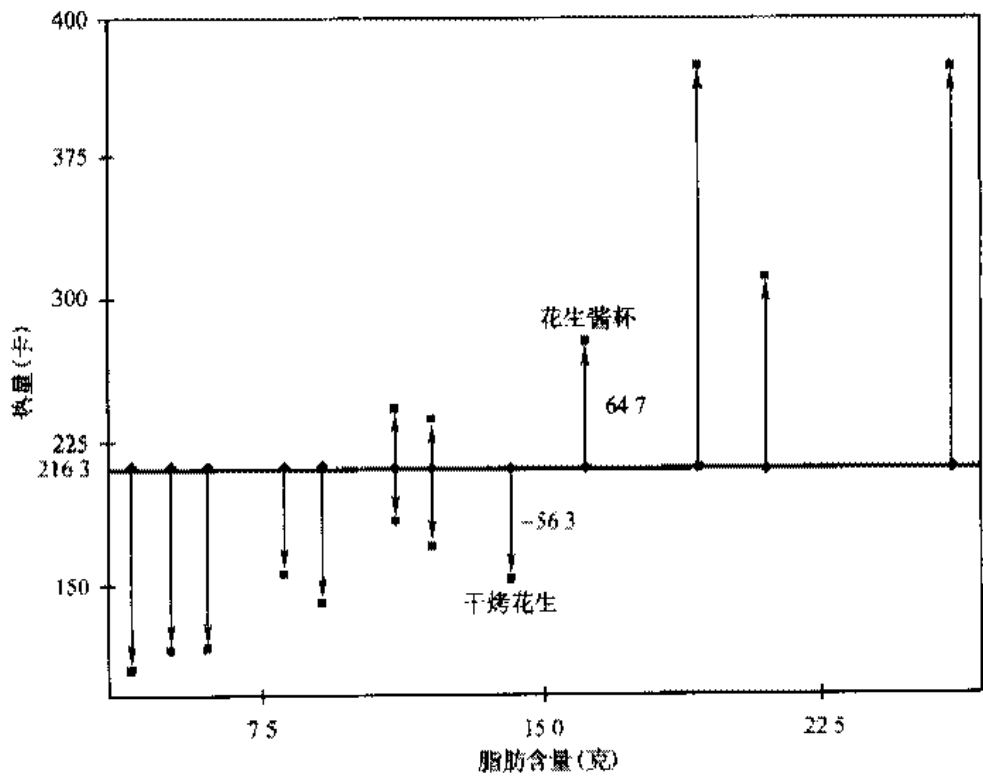


图 10.6 因变量和残差变量的效应。

怎样度量这里有多大的变化呢? 这个变化多大程度上与自变量脂肪含量有关, 多大程度上与残差变量有关?

例如, 对花生酱杯观测到的热量值并没有落在水平直线上, 而是 281。自变量和残差变量共同把花生油的值从 216.3 推到了 281 处, 相差 64.7 卡。所以, 64.7 是脂肪含量和残差变量的组合效应。类似地, 干烤花生的组合效应为 -56.3, 因为观测到的热量值的点在水平线的下方。我们可以用这种方法对所有不同的食物计算组合效应。

下一步, 我们想要把所有的这些效应总结在一个数中。由于许多历史和数学上的原因, 我们把每个效应(观测减去总平均)平方再加起来。对于小吃这个例子, 这个和为 159060, 称为总平方和(total sum of squares)。

总平方和度量了自变量和残差变量在因变量上的效应, 它等于

$$(\text{观测} - \text{平均})^2 \text{ 之和}$$

现在我们想知道这个总平方和有多少来自于自变量(脂肪), 有多少来自于残差变量(所有其他变量)。假设残差变量对热量没有影响, 也就是说, 热量仅受脂肪含量的影响。那么散点图中所有的点将完全落在回归直线上。但是, 由图 10.3, 数据点并没有全落在直线上。这些点散布在回归直线附近, 所以一定是残差变量把这些点从回归直线上推开了。

残差变量的效应是由一个观测点到回归直线的垂直距离。

观测数据点和它相应地在回归直线上的位置的差异是残差变量的效应。我们可以计算这个差异。为了把它们的大小汇总起来, 我们把它们每一个都平方然后再加起来。这个和称作残差平方和(residual sum of squares), 有时也称作误差平方和。它代表了残差变量的效应。对于小吃的例子, 残差平方和等于 27182。当点落在直线附近时, 残差平方和相对于总平方和较小。类似地, 如果至少有一些点离回归直线较远, 残差平方和则较大。

由于自变量和残差变量的组合效应等于 159060, 而残差变量的效应为 27182, 自变量的效应则为差: $159060 - 27182 = 131878$ 。这个平方和也称作回归平方和(regression sum of squares)。本章末的公式 10.5 说明怎样计算不同的平方和。

表 10.4 小吃数据的平方和及比例

来源	平方和	比例
脂肪含量	131878	0.83
残差	27182	0.17
总计	159060	1.00

平方和通常如表 10.4 所示。表中有一行表示每一个变量, 还有一行表示列总和。第一列是变量的名字, 第二列是每个变量的效应的大小, 它由适当的平方和度量。为了更容易地看出这两个效应相比到底有多大, 第三列表示每一个平方和占总平方和的比例。为了计算这个比

例,我们用每一个平方和除以总平方和。这里,自变量对总效应贡献了 0.829,或者说 83%,残差变量贡献了剩余的 0.171 或者说 17%。所以,脂肪含量对于热量这个因变量的效应比残差变量的效应大得多。

自变量的效应的比例在这里等于 0.83,它总是等于自变量和因变量的相关系数的平方(如公式 10.6 所示)。在 10.2 节,我们发现脂肪含量和热量的相关系数为 0.91,所以 $r^2 = 0.91^2 = 0.83$,这恰好是表 10.4 中由脂肪含量贡献的效应的比例!利用回归分析,我们已经能够说明食物中的脂肪含量相对于其它变量来说对于热量含量的影响有多重要。而且还说明了这些点离回归直线越近,残差平方和就越小,相关系数也就越大。

所有这些意味着我们遇到相关系数时,就应当立即计算它的平方。这个平方告诉我们自变量对于因变量的效应占总效应的比例。 $1 - r^2$ 是残差变量占总效应的比例。有时,结果会很令人失望。假设我们得到了一个适中的值 $r = 0.5$ 。此时, $r^2 = 0.25$ 。这意味着自变量对于因变量的效应仅占 1/4。还不到一半。

在一份小吃的分析报告中,我们可以写上这样一句话“脂肪含量解释了 83% 的热量变化,”或者说“83% 的热量变量的变化可以归于脂肪变量。”除了用“解释”或“归于”这些词外,还可以用“是...的原因”这样的词。所有这些广泛应用的词都表达了同样的意思。

相关和/或回归? 多多益善

如果我们对于回归直线一无所知,仅从两变量间的高度相关系数可以得出什么结论呢? 我们容易注意到比较大的相关系数,并认为已经了解了两变量间的主要关系。类似地,如果我们仅知道回归直线而不知道相关系数等于多少,我们将怎样想呢? 斜率越大从而直线越陡,自变量看上去越重要。但是,仅仅知道相关系数或者回归直线并不能很好地分析两个变量。我们两个都应该知道。

在脂肪含量和热量的例子中,相关系数为 0.91,这个相关系数很大。解释这个相关系数的一个困难是,对于具有完全不同的回归直线的许多不同数据集都可以有同一个相关系数。假设回归直线有下面的方程:

$$\text{热量} = 248 + 1.1 \text{ 脂肪含量}$$

而不是我们得到的方程(热量 = 36 + 15.3 脂肪含量)。对于新的回归直线,0.91 这样大的相关系数并不能给人以深刻的印象。比较大的 r 只不过告诉我们这些点很近地分布在直线的周围,但是由于斜率较小,仅为 1.1,直线几乎是水平的。

让我们随便拿两种食物来说,一个的脂肪含量较低,含有 7.5 g 脂肪。而另一个脂肪含量较高,含有 25.5 g 脂肪。把这些数代入直线的方程,我们发现第一个食物的预测热量为 256,第二个食物的预测热量为 273。256 和 273 间仅相差 17 卡,这个数很小,不值得引起我们多大的注意。即使一种食物含有 3 倍于另一种食物的脂肪,它们的热量含量也没有太大的差异。尽管由相关系数(图 10.7)知道两变量间有很强的相关性,从一个营养学的观点来看,这个结果是不合逻辑的。这个例子帮助我们说明了仅仅知道相关系数是不够的,我们还需要知道回归直线。

图 10.7 是两个数据集的散点图,这两个数据集有相同的较高的相关系数。图 a 中的直线

较陡,图 b 中的直线则几乎是水平的。如果我们仅知道相关系数,尽管两个数据集有很明显的差异,我们却不能区别它们。

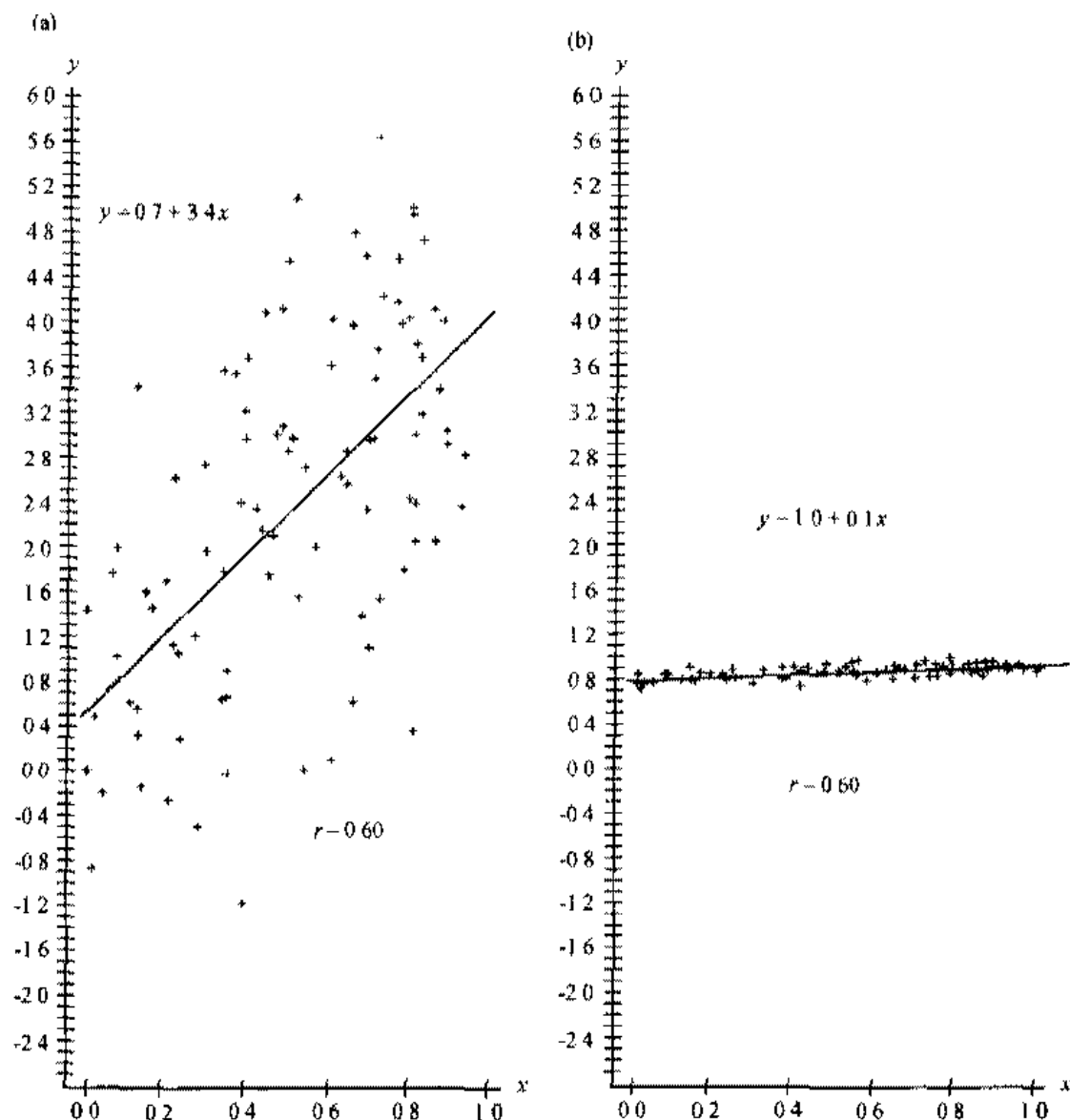


图 10.7 相关系数相同、而回归直线不同的两个数据集的散点图。

当我们仅知道回归直线而不知道相关系数时,就会发生相反的情况。我们仅知道直线的倾斜度而不知道数据点离直线有多远。如果这些点很宽地散布在直线附近,相关系数就比较小,直线所包含的信息要比数据点离直线很近时要少。图 10.8 是两个不同的数据集的散点图。这两个数据集的回归直线相同,而回归系数却不同。

许多统计的计算机程序都在进行回归分析时给出了相关系数或它的平方。但做相关分析时它们却并不给出回归直线。在一个特别的情形,研究报告给出了相关系数而没有回归直线。在不清楚哪个变量是自变量、哪个变量是因变量的时候就会出现这种情形。例如,在研究语文和数学的考试成绩时,选择哪个变量作为自变量和因变量并不是很显然的事。所以,报告时就只有相关系数。

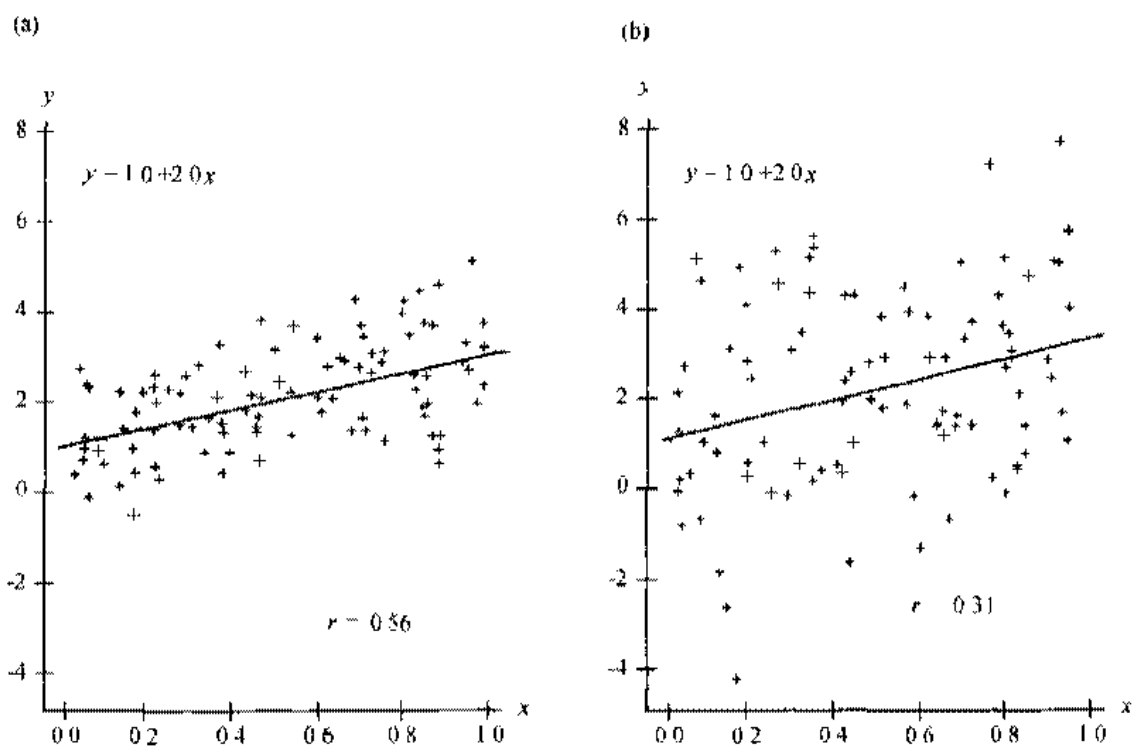


图 10.8 相关系数不同、而回归直线相同的两个数据集的散点图。

变化数据的回归分析

至此,我们在解释回归系数(斜率) b 时一直是很小心的,我们说如果两个观测在自变量上相差一个单位,它们在因变量上将相差 b 个单位。在小吃的例子中, $b = 15.3$,如果两份食物的脂肪含量相差 1 g,它们的热量将平均相差 15.3 卡。我们有对于不同食物的数据,解释 b 时就要依据不同的食物。我们不能从这些数据得出这样的结论:一份小吃如果脂肪含量增加 1 g,其热量将平均增加 15.3 卡。要作出这样的解释,我们需要对这份特别的食物的脂肪含量和热量的变化值。

如果我们有了变化的数据,我们就可以用变化的术语来解释回归系数 b 了。例如,如果产科医生测量到了某一个每次来看门诊的婴儿重量,我们就可以对这些数据进行回归分析并认为这个婴儿的体重每月增长了多少盎司。

在另一个例子中,*The Philadelphia Inquirer* (1943 年 4 月 7 日, p. A2)报导了 *Journal of the American Medical Association* 上的一件事情。这件事情是关于减少孩子们血液中的铅含量和“认知指数”得分增加的联系。铅的含量以 mg/dl 来度量。认知指数来自标准化智力测验。从文章中的数据可以得出这两个变量的回归方程为:

$$\text{认知指数} = 90 - 0.33 \text{ 铅含量}$$

回归系数为 -0.33 表明如果铅含量增加,认知指数就会下降;如果铅含量减少,认知指数就会上升。特别地,这个系数说明如果血液中的铅含量减少 1mg/dl,那么认知指数将平均上升 0.33 点,或者像报纸上所说的那样,如果铅含量下降了 3mg,认知指数将上升 1 点。如果公共政策用的是为社会变化服务,这种类型的回归分析常常是适当的。

10.4 问题3. 总体中的关系?

在小吃的样本中,发现脂肪含量和热量间有一定的关系是一回事,而知道对于所有食物的总体,这两个变量间也有一定的关系则是另一回事。因为我们没有对总体的数据,我们可用样本数据来推广到总体。可以用两种方法达到这个目的:构造一个总体的回归系数 β 的置信区间或者检验认为总体中没有相关性的零假设。为了说明这个问题,我们把小吃样本作为一个从所有小吃食物总体中随机抽取的样本。

置信区间的方法

在第6章我们用置信区间来估计总体中未知参数的值。在这里,我们首先计算观测的回归系数然后加上或减去一个样本误差项。在小吃的例子中,样本回归系数是 15.3 卡/克。我们计算的样本误差为 4.0。从而总体的斜率 β 的置信区间为 $15.3 - 4.0 = 11.3$ 卡/克到 $15.3 + 4.0 = 19.3$ 卡/克。我们希望,区间 $[11.3, 19.3]$ 很可能就是所有 95% 地包含总体的回归系数 β 的区间中的一个,而不是那些很少的不包含 β 的区间中的一个。公式 10.7 说明怎样计算置信区间。

这个置信区间最值得注意的一点是它不包含 0。我们提及这一点是说明 0 不是总体回归系数的个可能的值。既然 β 不可能等于 0,我们可以认为 β 一定与 0 有差别。如果总体的直线的斜率不等于 0,对于包含所有食物的总体,而不仅仅是样本,两个变量脂肪含量和热量间一定存在一定的关系。

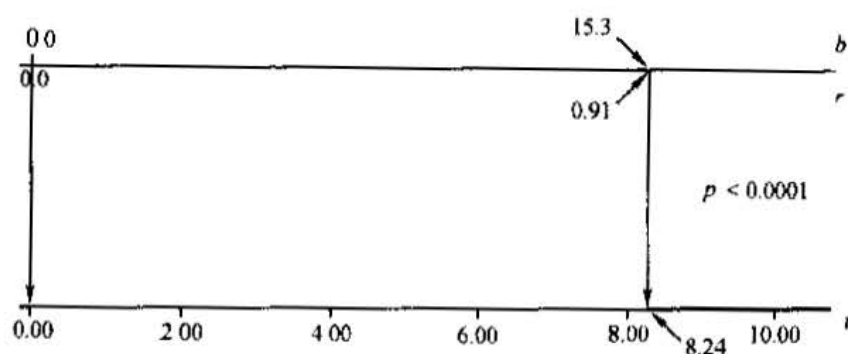
用 t 进行假设检验

第7章中的假设检验方法是基于这样的零假设:即认为两变量间没有关系。为了检验这个零假设,我们可以用观测的样本回归系数 b ,也可以用观测的样本相关系数 r 。它们都得到相同的 t 统计量值。从得到的 t 统计量的值,我们可以计算 p -值并对零假设下结论。 p -值是从变量间没有关系的总体中抽取样本点或得到更加极端的数据的概率。

图 10.9 说明了这个过程。这里, $b = 15.3$ 卡/克,这个 b 值相应于 $t = 8.24$, ($df = n - 2 = 14$) (公式 10.8)。类似地, $r = 0.91$, 也得到相应的 $t = 8.24$ (公式 10.9)。从计算机输出结果或从一个 t -分布表,可以得到 $t \geq 8.24$ 的概率小于 0.0001。所以,如果总体中的两变量无关,在 10000 个不同样本中得到 $t \geq 8.24$ 的样本将小于一个。这意味着,仅仅由于偶然而出现观测到的或更强的样本关系几乎不可能。由于 p -值很小,我们拒绝认为两变量间无关的零假设。这么小的 p -值说明,如果对于所有数据的总体两变量间没有关系,来自这个总体的样本的斜率就几乎不可能总大于或等于 15.3,或者说它们的相关系数就不可能大于或等于 0.91。

利用 F 进行假设检验

除了可以从斜率 b 或相关系数 r 得到 t 统计量外,还有第三个方法可以进行假设检验。两个不同的方法已经比我们需要的多一个了,不过第三个方法可以进行推广。这里引进它是

图 10.9 把回归系数 b 和相关系数 r 变为它们的 t -值。

因为我们在第 12 章和第 13 章还要用到它。

第三种方法是基于表 10.4。我们把那张表再加上三列(表 10.5)。这样我们可以计算一个 F 变量,我们在第 5 章中介绍 z -、 t -和 χ^2 统计量时曾介绍过这个统计量。 F 的值可以用来检验零假设。

标有“均方”的那一列是由每个平方和除以它们相应的自由度得到的。由于回归平方和只有一个自由度,所以其均方还是 131878。残差均方(RMS)是 $27182/14 = 1941.6$ 。然后用回归的均方除以残差的均方就得到 $F = 131878/1941.6 = 67.90$ 。它有两个自由度,分别为 1 和 14。最后,得到 $F \geq 67.90$ 的概率小于 0.0001,所以拒绝零假设。

p -值是从变量间没有关系的总体中抽取样本而得到一个 F 变量的值大于或等于 67.90 的样本的概率。我们已经知道 $t = 8.24$,我们现在发现 F 的 p -值等于 $t < -8.24$ 的概率加上 $t > 8.24$ 的概率。所以, F 统计量的 p -值等于 t 统计量的双边 p -值。

表 10.5 用 F 检验零假设

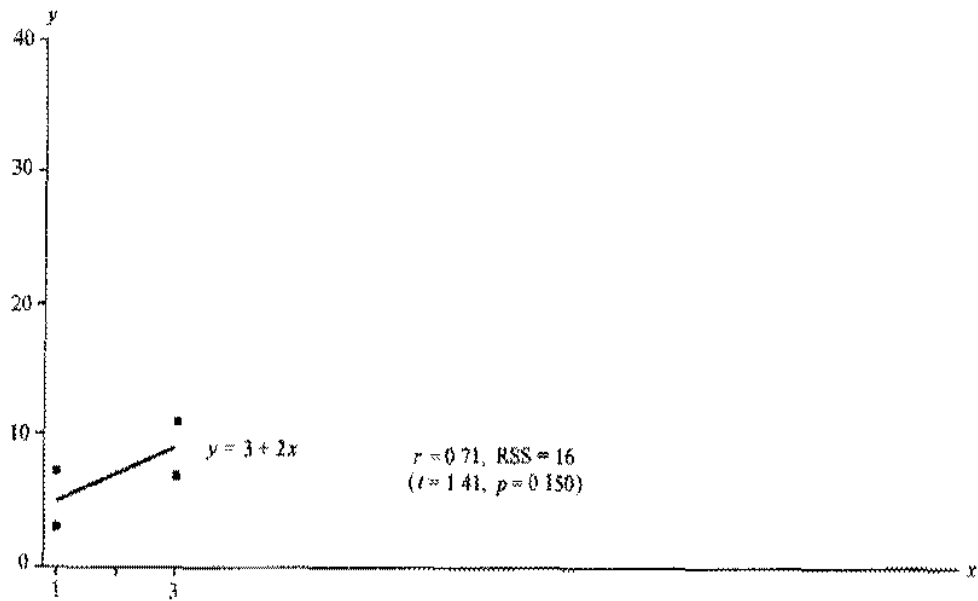
来源	平方和	比例	自由度	均方	F -比	p -值
脂肪含量	131878	0.829	1	131878.0	67.90	0.0000
残差	27182	0.171	14	1941.6		
总计	159060	1.00	15			

10.5 警告：所测即所得

对于观测的数据,我们通常不能选择我们观测的变量的取值范围。我们问人们的年龄并记录下来他们的回答。然而对于实验,我们经常可以选择自变量的取值。有时这种选择会影响分析的结果。

我们用下面的例子中的两个数据集来说明这一点。我们想研究 Y_1 和 X_1 及 Y_2 和 X_2 的关系。我们首先做一个散点图,然后对每一组数据拟合一条回归直线(图 10.10)。这两条直线有相同的斜率和截距,但第二个数据集的直线更长一些。两个图中的点到直线的距离都相同,都在直线上方或下方一个单位。

(a) 数据集 1



(b) 数据集 2

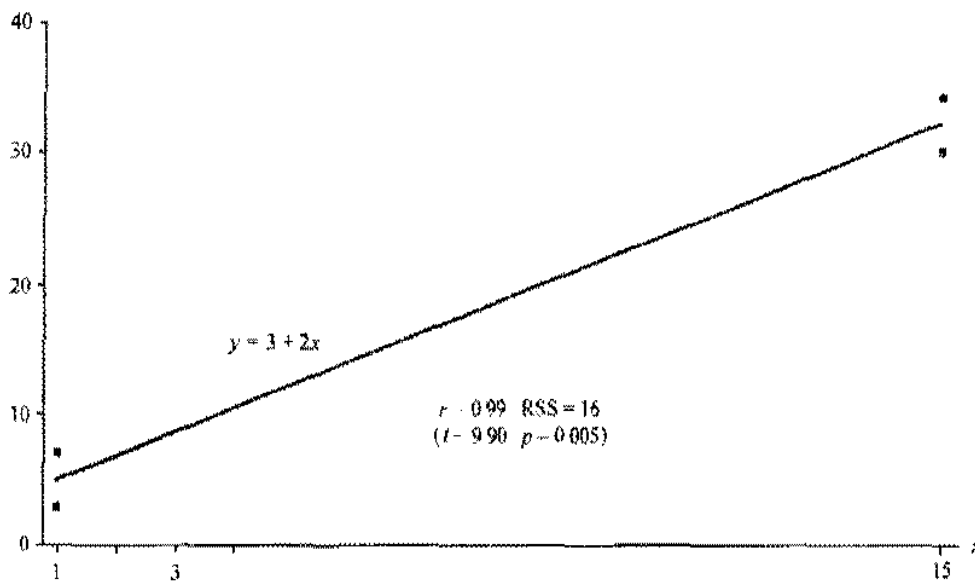


图 10.10 对于相同的斜率,相关系数和显著性的差异。

数据集 1		数据集 2	
X1	Y1	X2	Y2
1	3	1	3
1	7	1	7
3	7	15	31
3	11	15	35

当我们看这两条直线时,图 b 中的点距这条长直线看上去比图 a 中的点到那条短直线的距离更近一些。和在视觉实验中一样,有些事情尽管相同但看上去却并不一样。这个看上去的区别可以由两个相关系数反映出来。在图 b 中 $r = 0.99$,而在图 a 中 $r = 0.71$ 。这意味着相关系数不仅度量了点到回归直线有多近,而且还度量了 x 和 y 的值有多分散。

这两个数据集的 p -值也不同。对于短的直线,因为 $t = 1.41$, $p = 0.15$,这个 p -值很大,所以其斜率和 0 并没有显著的区别。而对于长的直线, $t = 9.90$, $p = 0.005$,所以其斜率显著地不为 0。

这意味着我们不得不怀疑统计上的显著和两变量相关程度的意义。如果控制自变量 x 的值,并选一些比较分散的值就经常会得到显著和比较大的 r 。但这通常发生在我们自由选择自变量的值的实验中。在观测研究中,我们不得不用样本提供的自变量的值。

10.6 用虚拟变量时怎样变得聪明些

至此,我们已对数值变量用了回归分析和相关分析的方法。然而,在有些情况下,对其他类型的变量利用相关分析和回归分析也是方便的。在这一节,我们用它们来处理包含一个分类变量和一个数值变量的问题。

自变量是有两个取值的分类变量和因变量是数值变量

如果你在寻找一个新的居所,气候将对你的选择很重要。然而如果你仅依赖于年平均温度便想选择一个地方会是很困难的,因为不同的地方在夏天和冬天都有不同的温度变化范围。这个问题可以用相关分析和回归分析来研究。

假定你想靠近海,不管是在美国的东海岸还是西海岸,而且你还喜欢比较温和的气候。分析之前,我们把 7 月和 1 月份的平均温度之差作为温度的变化范围。比如,费城 7 月份的平均温度为 76 华氏度,而 1 月份的平均温度为 32 华氏度。你确实可以感觉到季节的变化,温度的变化范围是 $76 - 32 = 44$ 华氏度。在圣迭戈,7 月和 1 月的平均温度分别为 70 度和 55 华氏度,变化范围仅为 15 华氏度。你想知道对于东海岸和西海岸的城市温度变化范围普遍的不同。

作为自变量的地域有两个取值:东海岸和西海岸。这是一个分类变量。因变量是温度的变化范围,它是一个数值变量。由于地域变量仅有两个取值,可以定义一个虚拟变量(dummy variable)来进行分析。对于一个虚拟变量的两个数值可以取任何数来表示,但通常取作 0 和 1。这个方法仅在初始的分类变量只有两类时才可以用。如果分类变量多于两类,我们就必须用其他方法。

一个虚拟变量是一个只有两个数值的变量,它经常用来表示一个有两类的分类变量。分类变量中,第一类的所有观测都取虚拟变量的一个值,第二类的所有观测都取虚拟变量的另一个值。

虚拟变量所做的一切实际上就是辨别这两类。为了计算的目的,利用虚拟变量可以把区域

分类变量变为只有两个数值 0 和 1 的一个变量。在这里,我们把西海岸的城市赋值为 0,东海岸的城市赋值为 1。关于地区的虚拟变量和温度的变化范围的散点图如图 10.11 所示。如果我们现在用计算机软件对这些变量进行回归分析,得到的回归直线如图中所示。计算机软件并不知道虚拟变量是一个分类变量。它只不过简单地计算进行回归分析所需要的所有的数量。

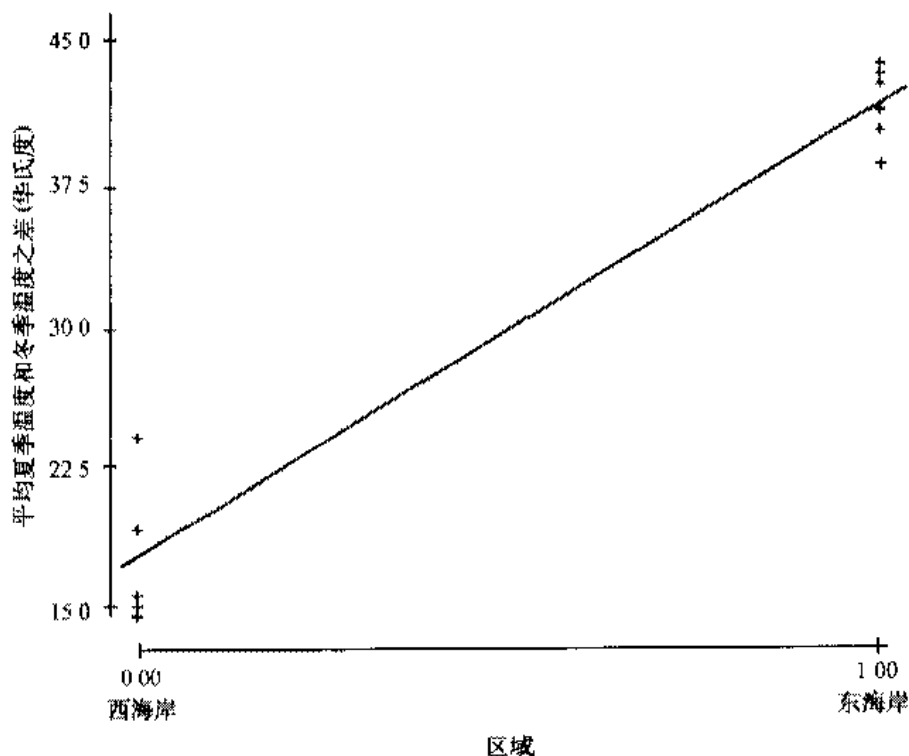


图 10.11 自变量是分类变量和因变量是数值变量。(来源: Data Desk SMSA file)

在散点图上,东海岸的六个城市的温度变化范围是从大约 39 华氏度到 44 华氏度,而西海岸的五个城市温度变化范围是从大约 15 华氏度到 24 华氏度。既然西海岸的温差范围比东海岸小,所以回归直线有一个正的斜率。如果我们把东海岸的城市赋值为 0,西海岸的城市赋值为 1,散点图上的那些点将逆转过来,回归直线的斜率将为负的。

这些数据的回归方程是:

$$\text{范围} = 17.6 + 24.7 \text{ 地区}$$

因为赋予西海岸的城市虚拟变量的取值为 0,所以截距就是这些城市温差范围的平均值,它等于 17.6。当地区取值为 1 时,我们把这个值代入回归直线的方程,得到一个预测值 42.3,它是东海岸城市的温差范围的平均值。斜率 24.7 是东海岸和西海岸城市温差范围的平均值的差异。

回归直线的斜率不等于 0 这个事实表明对于这些数据地区和温差范围间有一定的关系。关系的强弱由 $r = 0.98$ 来度量。为了检验零假设,即认为由所有的城市所组成的总体中两变量间没有关系,我们由相关系数 r 或者回归系数 b 来计算 t -值。对于这些数据, $t = 13.30$, $df = 9$ 。利用统计软件,我们发现 t 是显著的, p -值小于 0.0001。 t -分布的统计表 2 对 9 个自由

度的 t -值只列出了到 $t = 4.30$ 的值,这时相应的 p -值为 0.001。由于观测到的 t -值远大于 4.30,所以这些数据的 p -值一定远远小于 0.001。

这样小的 p -值充分说明了在东海岸城市夏天和冬天的温差范围要比西海岸城市的大。如果两个海岸城市间没有差异,在 10000 个样本中,产生这样的或更极端的数据的样本的可能性将小于 1 个。所以,东海岸的季节变化比西海岸明显这一事实是有证据的。如果我们用第 7 章和第 11 章的方法来研究两个均值的差异这个问题,结果将会是一样的,所得到的 t -值和 p -值也一样。

因变量是有两个取值的分类变量和自变量是数值变量

这个问题是温差范围/地区问题的逆。在这个问题中,因变量(汽车的产地)是一个具有两类的分类变量,自变量(驱动比)是一个数值变量。如果我们用一个虚拟变量来代表因变量,这些数据的散点图看上去就像图 10.12 那样。汽车产地这个因变量有两个值:外国和国内。

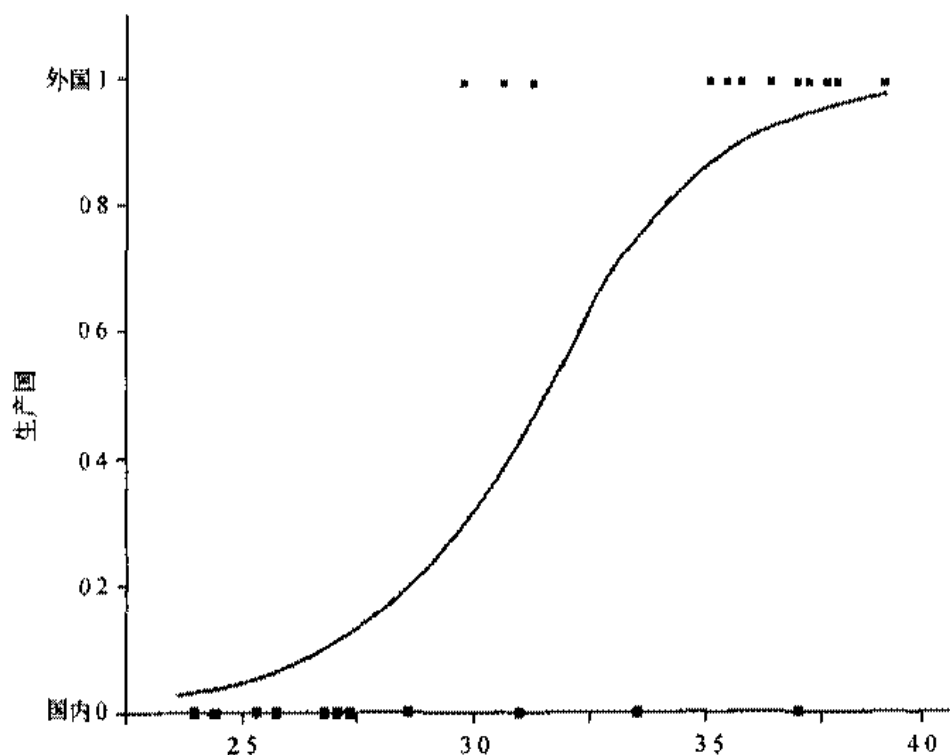


图 10.12 因变量是分类变量和自变量是数值变量。

对这种情形,这些数据的散点图是非线性的,我们不可能拟合一条穿过这些点的直线,因为散点图中所有的点都落在两条水平的直线上。一条是 $y = 0$ 另一条是 $y = 1$ 。对于这些数据,我们拟合一条 S 形的曲线,而不是一条直线,如图所示。这种类型的分析称作 **logistic 回归** (logistic regression)。这条曲线从图的右上角的点开始沿这些点,然后很快的下降到图左下角的那些点。这条曲线比任何一条通过这些点的直线的残差都小。

这里虚拟变量代表汽车的产地,1 代表国外,0 代表国内。0 和 1 之间的任何数可以解释

为汽车是由国外生产的概率。所以,如果我们知道驱动比为 3.5,我们可以在水平轴上确定那个值,然后向上与曲线在一点相交,并确定在纵轴的值 0.87。这个数是驱动比为 3.5 的汽车是一个外国汽车的估计的概率。

10.7 问题 4. 是因果关系吗?

帮助我们回答这个问题的统计方法确实存在;我们在第 13 章再考虑这个方法。这些方法基于引进其他的变量看看是否能够解释初始两个变量间的关系。目前,直观上认为脂肪含量会影响热量是有意义的,但在统计上我们却不能这么讲。尽管这个关系可能不是因果关系,如果我们知道了脂肪含量我们仍然可以用这些分析的结果来预测热量的值。

10.8 小结

相关分析和回归分析是分析两个数值变量关系的两个相互补充的方法。相关分析描述了两个变量的相关程度。回归分析则描述了因变量是怎样受一个或多个自变量的影响的。简单回归分析是指只有一个自变量的回归分析。

10.1 问题 1. 两变量间的关系?

为得到两变量间是否有一定的关系的一个直观印象,可以用一个散点图来描述这些数据。散点图可以用来了解这些数据是否适合用相关分析和回归分析。散点图上,横轴代表自变量,纵轴代表因变量。如果在图上数据看上去散布在一条直线的左下角到右上角的直线附近,两变量间就有一定的正相关性。在两个变量负相关的散点图上,这些点散布在从左上角到右下角的一条直线附近。如果散点图中的点看上去是随机散布的,两变量间的关系就很弱或者没有关系。

10.2 问题 2a. 关系的强度?

如果 x 变量和 y 变量间存在一定的关系,我们就可以知道随着 x 值增加或减小, y 值有何变化。度量两个数值变量的关系强度的统计量是相关系数 r 。 r 的值总是在 -1 和 1 之间。如果 r 的值接近于 1 或 -1 ,两变量间就有很强的相关性。如 $r = 0$,两变量间没有相关性。如 r 为正的,两个变量的值同时增加或减少,如 r 为负的,一个变量的值会随着另一个变量的减小而增加。

10.3 问题 2b. 关系的形式?

在回归分析中,一条回归直线代表了两个变量间的关系。这条直线通过了散点图中数据点的中心。直线的斜率度量了直线的倾斜度。直线的斜率越大,自变量的单位变化所引起的因变量的差异就越大。回归直线的截距是自变量等于 0 时在纵轴上的那一点,也就是回归直

线和纵轴相交的地方。回归直线由斜率和截距所决定。因变量的估计值等于截距加上斜率乘以自变量的值。回归直线可以通过最小二乘方法得到。

回归方程可以由一个自变量的值来预测因变量的值。因变量的预测值是真实值的估计。

除了被选择的自变量,其他的自变量的组合效应称作残差变量。总的平方和度量了所有的变量对因变量的效应。这是对于所有因变量的观测求下面的和: $\sum(\text{观测值} - \text{均值})^2$ 。残差平方和度量了除了自变量外其他的变量对于因变量的效应,它等于对所有因变量的观测求下面的和: $\sum(\text{观测值} - \text{估计值})^2$ 。回归平方和度量了自变量对于因变量的效应。它等于对所有因变量值求下面的和: $\sum(\text{估计值} - \text{均值})^2$ 。总平方和等于回归平方和加上残差平方和。因变量取值变化的效应中可归于自变量的比例总是等于自变量和因变量的相关系数的平方。

要了解两个变量间的关系,知道相关分析和回归分析的结果是很重要的。一个相关系数可能很大,然而回归直线却几乎是水平的,对此我们通常并不感兴趣。一条回归直线可能很陡,然而如果这些点离直线很远,相关系数也会很小。

为了预测一个变量的变化,回归系数 b 一定是基于随时间变化的数据计算出来的。如果不能得到随时间变化的数据,就很难预测变化。

10.4 问题 3. 总体中的关系?

要知道产生样本的总体中两个变量间是否有一定的关系,我们可以通过构造总体的回归系数 β 的置信区间或者检验认为两变量间无关的零假设从而把样本的性质外推。要计算这些数据的 p -值,可以把回归系数 b 或相关系数 r 变为一个 t 变量的值。也可以利用这些数据计算一个 F -统计量的值然后计算 p -值。

10.5 警告:所测即所得

在解释统计上的显著和两变量间的相关系数的大小时,我们必须小心。统计上的显著可以很容易地从大的样本中得到,而如果自变量的值比较分散时,就更容易得到比较大的 r 值。

10.6 用虚拟变量时怎样变得聪明些

对于一个数值变量和一个有两类的分类变量,可以用简单的相关分析和回归分析。虚拟变量代表了分类变量的值。虚拟变量的值通常取作 0 和 1。如果因变量是分类变量,我们可以用一个 S-形的曲线来拟合这些数据。这种分析方法称作 logistic 回归。

10.7 问题 4. 是因果关系吗?

从只有两个变量的数据我们不能知道这个关系是否是因果关系,但是如果我们知道了自变量的值,仍然可以预测因变量的值。

补充读物

Draper, N. R., and H. Smith. *Applied Regression Analysis*, 2nd ed. New York: John Wiley &

Sons, 1981. This well-known book has an introductory chapter on simple regression.

Kleinbaum, David G., Lawrence L. Kupper, and Keith E. Muller. *Applied Regression Analysis and Other Multivariable Methods*, 2nd ed. Boston: PWSKENT, 1988. More applied than Draper and Smith with a longer introduction to simple regression.

Lewis - Beck, Michael S. *Applied Regression* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07 - 022). Beverly Hills, CA: Sage, 1980. Short introduction to regression analysis.

Tufte, Edward R. *Data Analysis for Politics and Policy*. Englewood Cliffs, NJ: Prentice - Hall, 1974. Chapter 3 includes a good discussion of issues that arise in simple regression analysis.

公 式

两个变量 x 和 y 的 n 个观测数据可以用下面的符号表示:

x	y
x_1	y_1
x_2	y_2
x_3	y_3
\vdots	\vdots
x_i	y_i
\vdots	\vdots
x_n	y_n

相关系数和回归系数(斜率)

相关系数 r 可以由下式计算:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n \sum x_i^2 - (\sum x_i)^2][n \sum y_i^2 - (\sum y_i)^2]}} \quad (10.1)$$

回归直线的斜率 b 可由下面的表达式来计算:

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} \quad (10.2)$$

关于 r 和 b 左边的公式有时是用来定义 r 和 b 的。因为每一个观测都要减去一个均值,所以它们不易计算。当没有计算机时,右边的公式计算起来更快一些。

注意 r 和 b 的分子相同, 而分母也几乎相同。从相关系数 r 和回归系数 b 的公式可以得出 r 和 b 有以下关系:

$$b = r \frac{s_y}{s_x} \quad (10.3)$$

这里, 两个 s 是两个变量 x 和 y 的标准差。

截 距

回归直线的截距 a 可由下式计算:

$$a = \bar{y} - b\bar{x} \quad (10.4)$$

作为回归分析的一个例子, 考虑下面的关于 x 和 y 的数据。根据计算相关系数和回归系数的公式 10.1 和 10.2, 我们需要计算 x 和 y 的乘积然后再把这些乘积加起来, 还要计算 x 的平方和及 y 的平方和, 以及所有的 x 的和及所有的 y 的和。这些计算最好列在一张表中:

	x	y	x^2	xy	y^2
	1	3	1	3	9
	2	2	4	4	4
	3	5	9	15	25
	4	6	16	24	36
总计	10	16	30	46	74

对于这些数据可以得到

$$b = \frac{(4)(46) - (10)(16)}{(4)(30) - 10^2} = \frac{184 - 160}{120 - 100} = \frac{24}{20} = 1.20$$

$$a = \frac{16}{4} - 1.20\left(\frac{10}{4}\right) = 4.0 - 1.20(2.5) = 4.0 - 3.0 = 1.0$$

$$r = \frac{(4)(46) - (10)(16)}{\sqrt{[(4)(30) - 10^2][(4)(74) - 16^2]}} = \frac{184 - 160}{\sqrt{[120 - 100][296 - 256]}}$$

$$= \frac{24}{\sqrt{[20][40]}} = \frac{24}{28.28} = 0.85$$

不同的平方和可由下式计算:

$$\begin{aligned} \text{总平方和(TSS)} &= \sum (y_i - \bar{y})^2 \\ \text{回归平方和(RegrSS)} &= \sum (a + bx_i - \bar{y})^2 \end{aligned} \quad (10.5)$$

$$\text{残差平方和(RSS)} = \sum (y_i - a - bx_i)^2$$

$$r^2 = \frac{\text{RegrSS}}{\text{TSS}} \quad (10.6)$$

总体的回归系数 β 的置信区间

β 的置信区间是:

$$b - t^* s_b \quad \text{到} \quad b + t^* s_b \quad (10.7)$$

这里 b 是观测的回归系数, t^* 是从一个从 t 分布表得到的 $n - 2$ 个自由度的 t 统计量的 $1 - \alpha/2$ 分位点的值, s_b 是 b 的标准误差。 b 的标准误差通常可由统计软件包得到,但也可由下式计算:

$$s_b = \sqrt{\frac{\text{RSS}/(n-2)}{\sum (x_i - \bar{x})^2}}$$

假设检验

对于我们的脂肪含量和热量的例子,由回归系数 b 的值得到:

$$t = \frac{b}{s_b} = \frac{15.3}{1.85} = 8.24 \quad (10.8)$$

也可由相关系数 r 出发由下面的公式得到相同的 t 值:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.910}{\sqrt{\frac{1-0.829}{16-2}}} = 8.24 \quad (10.9)$$

习题

回顾(习题 10.1—1.31)

- 10.1 你怎样定义两变量间的相关系数?
- 10.2 回归这个词通常指向回移动。统计分析中回归分析指的是怎样的移动?
- 10.3 “简单”回归分析意味着什么?
- 10.4 a. Francis Galton 在他的研究父母和孩子的身高时发现了什么?
b. 这个研究表明了什么效应?
- 10.5 a. 举一个你可以用相关分析和回归分析的两个变量的例子。
b. 回归分析中首先计算的两个统计量是什么?
- 10.6 作一个散点图的目的是什么?
- 10.7 a. 在一个散点图上,哪个变量在 x 轴上,哪个在 y 轴上?
b. 你想作关于反对抽烟的商业广告的播放次数和看电视的高中生的戒烟率的数据的散点图。哪个变量应作为 x 轴?

- c. x 轴的变量叫做什么?
- d. 两个变量是正相关时, 散点图看上去是什么样子?
- 10.8 假定你作了一个关于城市街区文盲率与毒品有关的犯罪率的散点图。
 - a. 你预期的散点图看上去是什么样子?
 - b. 这些点看上去朝哪个方向散布, 沿 x 轴朝上还是朝下?
 - c. 对于统计上敏锐的观测者来说, 数据点朝这个方向意味着什么?
- 10.9 从一个散点图中看到在 100 个点中有 3 个点远离主要的点群。如果把这三个点移去, 这将会如何影响相关系数(如果有影响的话)?
- 10.10
 - a. 相关系数的一些其他名字是什么?
 - b. r 最大和最小的可能值是什么?
 - c. 相关系数 1.00 或 -1.00 哪个更强? 解释你的答案。
 - d. 相关系数 r 和 ϕ 或 V 有何区别?
- 10.11 在表 10.2 中, 对角线上的相关系数都等于 1.00, 为什么?
- 10.12
 - a. 如果相关系数 r 落在 0.75 和 1.00 之间, 两变量间的相关程度有多强?
 - b. 如果相关系数 r 落在 -0.70 和 -0.30 之间, 两变量间的相关程度有多强?
 - c. 如果相关系数 r 落在 0 和 0.25 之间, 两变量间的相关程度有多强?
- 10.13 如果你可以设计一个比现在更好的社会, 下面的变量间理想的相关系数将为多少? 用一个数假和一句话来描述它。
 - a. 收入水平和纳税水平。
 - b. 受教育的年限和文盲的人数。
 - c. 一个人的身高和收入的多少。
- 10.14
 - a. 给通过散点图点的中心的直线起一个名字。
 - b. 这条直线传递了两变量的什么信息?
 - c. 如果这条直线有一个正的斜率, 这条直线的方向如何?
 - d. 如果这条直线有一个负的斜率, 这条直线的方向如何?
 - e. 斜率为负的一条直线传递了两变量间相关性的什么信息?
 - f. 直线的倾斜度暗示了什么?
- 10.15 回归直线的 y 截距是 $x = 0$ 时直线和 y 轴的交点。
 - a. 脂肪/热量的例子中 y 截距是什么?
 - b. 解释这个数使得一个节食者能够明白。
- 10.16 由最小二乘得到的回归直线是距散点图中所有点距离最近的一条。术语最小二乘指的是什么?
- 10.17 回归直线和回归方程可以用来由自变量 x 的值预测因变量 y 的值。我们用什么符号来表示因变量的预测假?
- 10.18 在研究脂肪含量对热量的影响时, 哪些变量可以作为残差变量?
- 10.19 总平方和可以看作由两部分组成。
 - a. 这两部分分别是什么?
 - b. 这两部分如何计算?
- 10.20 写出脂肪/热量问题的回归方程。描述方程的每一个部分及怎样用方程理解脂肪含

量和热量的关系。

- 10.21 如果发现散点图中的所有点都在回归直线上,对于下面几个问题你能得出什么样的结论?
- 自变量和因变量的关系?
 - 残差变量对因变量的效应?
 - x 和 y 的相关系数?
- 10.22 举一个你想知道回归直线的截距是否等于 0 的两个变量的例子。
- 10.23
- 比较回归分析和相关分析的优点。
 - 在分析两个数值变量时,这两种方法各有什么特别的地方?
- 10.24 r^2 和因变量的比例之间的关系可归于自变量吗?
- 10.25 “多多益善”可以解释为一个宴会上人越多这个宴会就越快活。宴会的大小和快活程度是成正相关还是负相关?
- 10.26
- 在回归中希腊字母 β 代表什么?
 - 什么时候用它?
- 10.27 说出两个在回归分析中可以用来判断两变量间的关系在统计上是否显著的方法。
- 10.28 哪两个理论统计量对于计算用以判断一个样本中的关系是否显著的 p -值很有用?
- 10.29
- 在一个实验中扩展 x 变量的值会对相关系数有何影响?
 - 这对相关系数的 p -值有影响吗?
- 10.30
- 什么是虚拟变量?
 - 何时用虚拟变量?
 - 在研究打高尔夫球时有风天和无风天的得分的关系中,你怎样用一个虚拟变量?
- 10.31
- 我们何时用所谓的 logistic 回归?
 - 举一个用虚拟变量解决的问题使得这种方法对该问题很有帮助。

解释(习题 10.32—10.48)

- 10.32 假定你有了相邻的 48 个州的犯罪率的数据,你想知道不同类型的犯罪之间是否相关。如果将偷盗犯罪率作为因变量,抢劫犯罪率作为自变量进行回归分析,你会得到:

$$\text{偷盗} = 2682 + 1.49 \text{ 抢劫} \quad (t = 2.05, df = 46, p = 0.023)$$

来源: *Bureau of the Census, Statistical Abstracts of the United States; 1995, 115th ed., Washington, D. C., 1995.*

- 从这个分析中你能对偷盗率和抢劫率的关系作出什么结论?
 - 从这个分析你对这个关系不能得出什么结论?
- 10.33 考虑习题 9.29 的征兵数据的另一个方法是对 1 月份令自变量 $x = 1$, 对 2 月份令 $x = 2$, 直到 12 月份 $x = 12$ 。令因变量 y 是每个月征兵数的平均值。
- 如果你对 12 个月的数据点 (x, y) 作一个散点图, 并进行回归分析, 在抽签确实是随机的情况下, 你期望得到什么样的截距、 x 的回归系数和相关系数?
 - 如果分析给出如下结果, 你对抽签征兵可得出什么结论?

$$\text{平均征兵数} \approx 230 - 7.1 \text{ 月份} \quad (r \approx -0.87)$$

c. 相关系数等于 $t = -5.50$ 。这个值有可能完全出于偶然吗?

- 10.34** 有时似乎一些东西吃起来口味越好, 对我们就越有害。表 10.6 是关于不同类型的冻巧克力酸奶的数据。第一列数据表示酸奶的脂肪中的热量的百分比, 第二列数据表示由一些经过培训的品味师以 0 到 100 的记分对酸奶的评价。关于口味作为因变量, 脂肪中热量的百分比作为自变量的到了下面的结果:

$$\text{口味} \approx 37 + 1.6(\text{脂肪中热量的百分比})$$

$$(r = 0.74, t = 3.11, df = 8, p = 0.0073)$$

表 10.6 习题 10.34 的数据

品牌	来自脂肪的热量的百分比	口味记分
Breyer	24	83
Honey Hill Farms	33	85
Elan	21	80
Crowley Silver Premium	20	78
Edy's/Dreyer Inspirations	25	74
Häagen-Dazs	21	71
Kemps	20	65
Lucerne	23	63
Yoplait Soft	20	61
Albertsons	12	51

来源: "Low-fat frozen desserts: Better for you than ice cream?" *Consumer Reports*, vol. 57, no. 8 (August 1992), pp. 483-487.

- 作这些数据的散点图。
- 作出回归直线。
- 关于两变量间的关系, 你能得出什么结论?
- 我们为什么更喜欢吃位于回归直线上方的甜食而不是下方的?

- 10.35** 巧克力和香草味冻点心的价钱之间有什么差异吗? 在“消费者报告”杂志(*Consumer Reports*)搜集到的数据中, 得到的巧克力点心平均花费为 29.4 分, 而香草味点心的平均花费为 30.4 分。(来源: "Low-fat frozen desserts: Better for you than icecream?" *Consumer Reports*, vol. 57, no. 8 (August 1992), pp. 483-487.) 要知道在两个均值之间在统计上是否有显著的差异, 你可以通过对甜食的类型引进一个虚拟变量进行回归分析。虚拟变量的所有的巧克点心赋值为 1, 香草味冻点心赋值为 0。把花费作为因变量, 甜食类型作为自变量的回归分析得到下面的结果:

$$\text{花费} = 29.4 + 1.0 \text{ 类型}$$

- 你怎样从两个均值直接得到截距和斜率?
- 斜率 $b = 1.0$, 由这个值可以变换成 $t = 0.18$ ($df = 42$)。甜食的类型对于花费的影响在统计上是显著的吗?

10.36 许多因素都会影响贫穷,教育也许是其中的一个。研究这两个变量的关系的一个方法是搜集被调查者的受教育的年限和他们的收入。在这个习题中,数据并非对于每个人的,而是对每个州。从普查局我们知道了每个州的成年人受过9年或更少的教育的百分比和收入低于官方的贫穷线的人的百分比。

对这50个州和哥伦比亚地区,作低于贫困线的人的百分比对受9年或更少教育的人的百分比的回归可以得到:

$$\text{低于贫困线的人的百分点} = 4.6 + 0.8 \text{ 低教育的百分点}$$

$$(r = 0.70, t = 6.72, df = 49, p < 0.0001)$$

- 回归系数等于0.8告诉了我们什么?
- 如果51个观测的样本是来自于一个相关系数等于0的总体,得到样本相关系数大于或等于0.70的概率有多大?
- 这些数据的散点图如图10.13所示。为什么有些点在回归直线上方而有些点在回归直线下方?
- 负的残差最大的几个州是新泽西、康涅狄克、夏威夷、罗得岛和弗吉尼亚。为什么这些州会落在回归直线的下方?
- 正的残差最大的几个州是密西西比、路易斯安那、蒙大拿和犹他。为什么这些州会落在回归直线的上方?

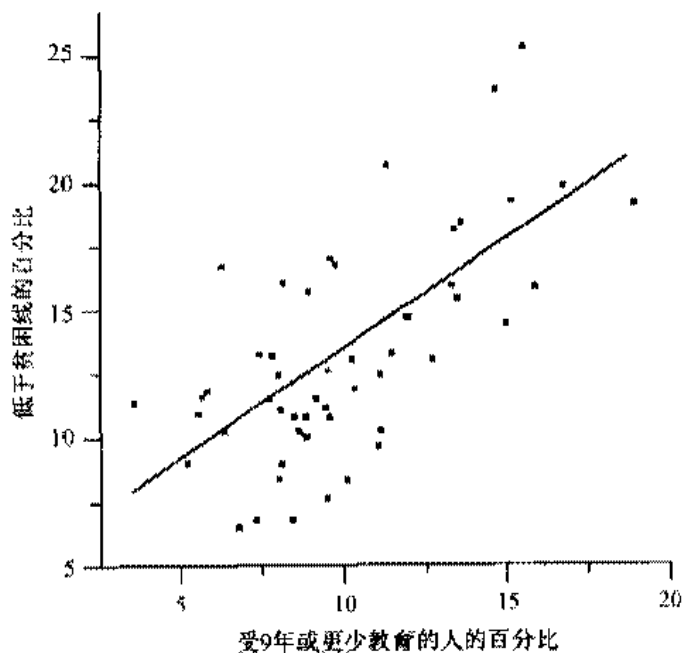


图 10.13 低于贫困线的人和受9年或更少教育的人的散点图(习题10.36)。(来源:1990 U. S. Census data reported in *The Chronicle of Higher Education*, vol. 34, no. 1 (August 26, 1992), p. 4)

10.37 如果以州为单位,我们把低于贫困线的人口的百分比作为因变量,把至少有大学文凭

的人的百分比作为自变量进行回归分析,得到:

贫穷的百分点 = $26.9 - 0.7$ 受过大学或更高教育的人的百分点

($r = -0.62$, $t = -5.54$, $df = 48$, $p < 0.0001$)

- 为什么受教育程度变量的系数为负的并不令人奇怪?
- 这个关系能仅归于偶然吗?
- 两个变量间的关系是因果关系吗?
- 为什么分析这量变量时,去掉华盛顿特区是有意义的?

- 10.38** 许多人高中毕业后就不再读大学了。在这里我们得到了从1980年到1990年18岁到24岁的年轻人为在校大学生的百分比。(来源: *Data from annual Census Bureau surveys of 60000 householders as reported in The Chronicle of Higher Education*, vol. XXXIX, no. 1 (August 26, 1992), p. 12) 我们把黑人、西班牙裔和白人分开来看。为了简化分析,我们把1980年编号为0,1981为1,直到1990编为10。如果我们将每年考入大学的百分比作为因变量,把年份从0到10作为自变量进行回归分析,可得到下面的三条回归直线:

黑人: 百分比 $\approx 26.5 + 0.42$ 年份 $r = 0.71$

西班牙裔: 百分比 $\approx 29.9 - 0.08$ 年份 $r = -0.22$

白人: 百分比 $\approx 31.1 + 0.75$ 年份 $r = 0.97$

- 在一个图上作出这三条回归直线。
 - 对黑人来说,系数等于0.42是什么意思?
 - 在这一阶段,这三个组哪一个的大学入学率年增长最快?
 - 描述这三条直线;它们彼此有何不同?
 - 为什么你认为西班牙裔的回归直线的斜率是负的,从而在80年代这10年间18岁到24岁的西裔的大学入学百分比呈下降趋势?
- 10.39** 在一个回归问题中, Sam 发现自变量 X 变化一个单位, 因变量 Y_1 平均变化 10.2。Anne 发现自变量变化一个单位, 另一个因变量 Y_2 平均变化 4.2。
- 在一个散点图上画出这些数据, 哪一条回归直线更陡?
 - 自变量变化一个单位, 所有的因变量都增加 10.2 吗?
 - 你认为因变量 Y_1 和 Y_2 有一定的关系吗?
- 10.40** 一个著名的统计学家搜集和分析了父亲和他们的孩子的身高(单位:英寸)的关系。(来源: *Class data, introductory statistics course, Swarthmore College, 1992.*) 她发现了下面的结果:

孩子身高 = $1.52 + 0.75$ 父亲身高

($r = 0.59$, $r^2 = 0.34$, $t = 2.90$, $p = 0.0051$, $n = 18$)

- 这个统计分析的结果告诉你两变量间的什么关系?
 - 总体的回归系数的置信区间是从 0.20 到 1.30。怎样解释这个区间?
- 10.41** 在研究化学药品对动物的作用时,通常通过逐渐加大对各组动物的药物的剂量,并观察每一组中有多少个动物发生反应。表 10.7 中的数据说明在研究狄氏剂(一种白色晶体状的杀虫剂 $C_{12}H_8Cl_6O$)的效果时,每组中有多少个动物对于不同的剂量发生了反应。
- 剂量的增加看上去是怎样影响对狄氏剂起反应的老鼠的比例?

b. 在发生反应的比例对剂量率的回归分析中(没有考虑每组中动物的数目), 得到如下结果:

$$\text{反应比例} = 0.08 + 0.13 \text{ 剂量率} \quad (t = 9.24, p = 0.006)$$

这些数据告诉你两变量间关系如何?

c. 关于两变量间的关系, 这些数据没有告诉你什么?

表 10.7 习题 10.41 的数据

剂量率(ppm)	反应的比例	动物的数量
0.00	0.11	156
1.25	0.18	60
2.50	0.43	58
5.00	0.73	60

来源: A. I. T. Walker, E. Thorpe, and D. E. Stevenson, "The toxicology of dieldrin (HEOD): I. Long-term rat toxicity studies in mice," Food and Cosmetics Toxicology, vol. 11 (1972), pp. 415-432.

10.42 图 10.14 是从 1936 年到 1972 年间康涅狄克州每 10000 个居民中患恶性黑色素瘤的数目的散点图。黑色素瘤是一种皮肤肿瘤, 它含有一种黑色素并可能致癌。关于这

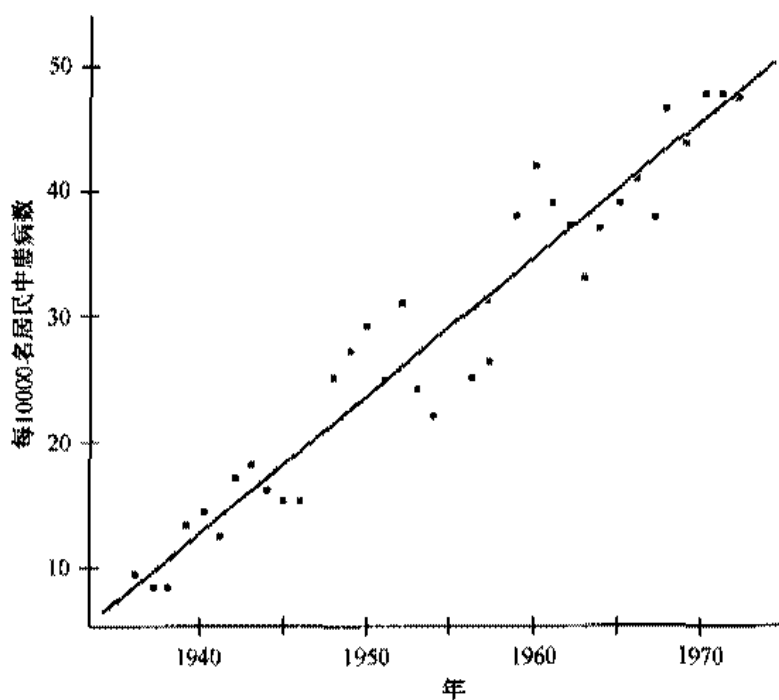


图 10.14 1936—1972 年间康涅狄克州患恶性黑色素瘤的人数(习题 10.42)。来源: A. Houghton, E. W. Meunster, and M. V. Viola, "Increased incidence of malignant melanoma after peaks of sunspot activity," The Lancet, April 8, 1978, pp. 759-760, as reported in D. F. Andrews and A. M. Herzberg, Data: A collection of Problems from Many Fields for the Student and Research Worker, New York: Springer-Verlag, 1985, p. 201.

些数据的回归分析得到的回归直线如图所示。这条直线的方程是：

$$\text{发生率} = -2,127 + 1.1 \text{ 年} \quad (r = 0.963)$$

(截距是一个很大的负数是因为用年份作为自变量取值很大,它从1936变到1972,所以这条直线必须向左延伸很长才能得到实际的 y 的截距。)回归分析结果如表 10.8 所示。

- 描述散点图中的规律性,这对进一步分析这些数据有何建议?
- 回归直线的方程告诉你两变量间有什么关系?
- 表中的数据告诉你两变量间有什么关系?

表 10.8 习题 10.42 的数据

来源	平方和	自由度	均方	F-比	p-值
年	5131	1	5131.0	453	0.0000
残差	396	35	11.3		
总计	5527	36			

来源: A. Houghton, E. W. Meunster, and M. V. Viola, "Increased incidence of malignant melanoma after peaks of sunspot activity," *The Lancet*, April 8, 1978, pp. 759-760, as reported in D. F. Andrews and A. M. Herzberg, *Data: A collection of Problems from Many Fields for the Student and Research Worker*, New York: Springer-Verlag, 1985, p. 201

- 10.43** 英国运动员 Roger Bannister 在 1954 年创造了一项纪录,他在一次田径运动会上首次在 4 分中内跑完了 1 英里的路程。从那以后,一直到 1993 年,男子 1 英里的世界纪录已被刷新了 12 次;那时,阿尔及利亚的 Noureddine Morceli 以 3:44.39 跑完了 1 英里,他比 Bannister 快了 15 秒。把纪录时间作为因变量,创造纪录的年份作为自变量的散点图表明有一个显著的直线关系。两变量的相关系数为 $r = -0.968$ 。为了使回归分析的结果更容易处理,我们首先把所有的纪录的时间都减去 3 分钟,把年份都减去 1900。Morceli 的时间和年份就变为了 44.39 和 93。经过这些变换以后,回归直线的方程是:

$$\text{纪录} = 70.07 - 0.3468 \text{ 年份}$$

- 基于这个回归分析,你预期的纪录在 10 年内会有何变化?
 - 在 39 年时间内,与实际改变了 15 秒相比,你预期的世界纪录的变化如何?
 - 与你预测的相比, Morceli 跑得是快还是慢?
 - 到 2000 年,你预期的 1 英里的世界纪录是多少?
- 10.44** 来自美国 *News and World Reports* 的大学年鉴有关于学院和大学的大量数据。排名的意义是引人争论的,但这些学校确实想排名高一些。看一下表 10.9 中前 12 名的大学(不包括加州理工学院)及他们每年花在每个学生上的钱。这些花费数字是用年财政预算除以学生总数得来的,这比学费和其他费用要高,因为学校还用其他收入资助学生。

在这些数据的散点图上,有些学校落在回归直线上方,有些在下方。关于花费的排名的回归分析可以得到回归直线的方程为:

$$\text{排名} = 17 - 0.34 \text{ 花费} \quad r = -0.60$$

相关系数为负的是因为排名最高的 Harvard 大学其数值秩最小,所以落在散点图的底部,排名为 12 的西北大学在图的上部。

表 10.9 习题 10.44 的数据

大学	平方和	每年在每个学生上的花费(千美元)
哈佛(Harvard)	1	36
普林斯顿(Princeton)	2	2
耶鲁(Yale)	3	39
麻省理工学院(MIT)	4	33
斯坦福(Stanford)	5	36
杜克(Duke)	6	26
达特茅斯(Dartmouth)	7	30
芝加哥(Chicago)	8	37
康奈尔(Cornell)	9	21
哥伦比亚(Columbia)	10	31
布朗(Brown)	11	20
西北(Northwest)	12	25

来源: *America's Best Colleges 1994 College Guide*, U. S. News and World Report, pp. 20-21.

- 一个学校落在回归直线的下方意味着什么?
- 落在回归直线下方的所有学校有什么共同点吗?
- 如果两个学校每年在每个学生上的花费相差 \$3000, 其排名的平均差异是多少?

10.45 在 1907 年一项关于 16 艘轮船的研究中, 船的吨位区间从 192 吨到 3246 吨, 船员的人数从 5 人到 32 人。船员人数关于船的吨位的回归分析得到下面的结果:

$$\text{船员人数} = 9.5 + 0.00062 \text{ 吨位} \quad (r = 0.87, t = 6.79, df = 14, p < 0.0001)$$

(来源: R. Floud, *An Introduction to Quantitative Methods for Historians*, London: Methuen, 1973, Table 4.1)

- 这些数据告诉你两变量间的什么关系?
- 假定两艘轮船吨位相差 1000 吨, 船员人数平均相差多少?
- 对于最小的船估计的船员数是多少, 对于最大的船估计的船员数是多少?

10.46 *The New York Times* 曾经委托一个实验室分析纽约市不同的商店里 12 片比萨饼的热量和脂肪含量。这些饼重量从 5.25 盎司到 10.5 盎司, 热量从 366 卡到 613 卡, 脂肪从 11 g 到 15 g。(来源: *The New York Times*, september 14, 1995; p. C1.) 热量关于脂肪(克)的回归分析结果是:

$$\text{热量} = 280 + 13.4 \text{ 脂肪(克)} \quad (r = 0.78, t = 4.01, df = 10, p = 0.0012)$$

- 这些数据告诉你两变量间的什么关系?
- 假定你已用热量和脂肪除以每片馅饼的重量。你认为这对分析有什么影响?

10.47 不同学科间的平均学术工资不同, 从农业的 \$36900 到图书馆学的 \$23600 (1984 年关于 24 个领域的研究数据)。(来源: M. Bellas, and B. F. Reskin, "On comparable worth,"

Academe, vol. 80 (1989), no. 5, pp. 83-85.)。不同工作领域的妇女比例也不平衡;女护士占 94%,而女工程师只占 5%。在平均工资作为因变量,妇女的百分比作为自变量的回归分析中,

$$\begin{aligned}\text{平均工资} &\approx 34300 - 120 \text{ 妇女百分点} \\ (r &= -0.829, t = -7.10, df = 2, p < 0.0001)\end{aligned}$$

- a. 作出回归直线的图。
- b. 这个分析的结果告诉你两变量间的什么关系?
- c. 是否有其他变量也许会解释这个关系?
- d. 数据的散点图表明数学、社会和人类学、音乐、新闻业、英语、外国语和戏剧明显地落在回归直线下方。社会工作、护士、生命科学、农业和工程则明显的落在回归直线的上方。如何解释这些模式?

- 10.48** 在第四章末是一些关于某些国家由于患肝硬化病的死亡率数据。对于同样这几个国家,我们还有每年消费的纯酒精(单位:夸特/人)数。因为通常认为过度饮酒会对肝脏有害。在这些国家中,卢森堡年耗酒量高达 13.3 夸特/人,而以色列则很低,为 1.0 夸特/人。美国则居中为 9.2 夸特/人。关于这些数据的回归分析得到的回归直线的方程为:

$$\text{肝硬化死亡} = 2.1 + 2.09 \text{ 夸特酒} \quad (r = 0.45, t = 2.67, df = 28, p = 0.006)$$

这些回归分析的结果告诉你消费的酒精和肝硬化死亡有何关系?

分析(习题 10.49—10.74)

- 10.49** 在 Springer 的网点(<http://www.springer-ny.com/supplements/ivrsen/>)可以发现关于本书的数据文件。打开叫做 Baseball Team Scores 的数据文件。第一列显示的是在整个赛季每个队获胜的比赛的次数,第三列是每次比赛的平均得分。
- a. 两变量间的关系强度如何?
 - b. 这个统计关系是显著的还是出于偶然?
 - c. 作回归分析并看看得分(自变量)是如何影响获胜次数(因变量)的?
 - d. 如果一个队每次比赛可以多得一分,这个队将多获胜几场比赛?
- 10.50** 一个城市到另一个城市的道路距离通常比它们之间的直线距离要长。在一个英国城市的样本中,对于这些距离的回归方程是:

$$\text{道路距离} = 7 + 1.17 \text{ 直线距离}$$

(来源: Neville Hunt, "A tale of six cities," Teaching Statistics, vol. 16, (1994) no. 1, pp. 5-8.)关于英国城市间的直线距离和道路距离,这个方程告诉你了什么?

- 10.51** 对美国城市的一个随机样本,道路距离/直线距离数据如表 10.10 所示。

表 10.10 习题 10.51 的数据

	Cheyenne	Fargo	Los Angeles	Oklahoma City	St. Louis
Atlanta	1482, 1235	1394, 1105	2121, 1940	833, 765	565, 476
Cheyenne	—	780, 553	1124, 894	694, 560	942, 790
Fargo		—	1808, 1430	870, 782	850, 647
Los Angeles			—	1339, 1205	1836, 1590
Oklahoma City				—	500, 465

来源: Road Atlas, Boston: Rand McNally, 1991.

- 对于这些数据, 计算回归方程。
 - 你也许会预料到回归直线的截距等于 0, 这是为什么?
 - 你能拒绝认为总体的截距等于 0 的零假设吗?
 - 你能解释为什么在美国的道路距离和直线距离的趋势使得某些对城市间两个距离几乎相等, 而对另外一些城市对之间的距离却相差很大?
- 10.52 在美国我们可以听到许多关于离婚数目不断上升的报导。分析这种现象的一种方法是比较离婚的人数和结婚的人数, 因为人们必须先结婚然后才能离婚。下面的数据是从 1890 年每隔 5 年直到 1980 年的结婚和离婚的人的数据。年份变量已重新取值, 1 代表 1890, 2 代表 1895, 直到 19 代表 1980, 这样往计算机中输入数据时更容易些。

年份	1	2	3	4	5	6	7	8	9	10
结婚	570	620	709	842	948	1008	1274	1188	1127	1327
离婚	33	40	56	68	83	104	170	175	196	218

年份	11	12	13	14	15	16	17	18	19
结婚	1596	1613	1667	1531	1523	1800	2159	2153	2413
离婚	264	485	385	377	393	479	708	1036	1182

来源: National Center for Health Statistics, Public Health Service, in The World Almanac 1986, p. 779.

- 在一个散点图上画出这些结婚和离婚的数据。
 - 评价散点图的形状。对这些数据作回归分析和相关分析有意义吗?
出于数学上的兴趣, 对每一个观测取对数。在计算机上的文件中再加入一列结婚数据的对数和一列离婚数据的对数。
 - 在一个散点图上描述这两个对数变量。
 - 评价散点图的形状。
 - 对每一年用离婚数除以结婚数, 作这个比例变量关于时间变量的图。
 - 这个散点图告诉了我们什么?
- 10.53 在 Springer 的网(<http://www.springer-ny.com/supplements/ivrsen/>)上, Baseball Team Scores 这个数据文件包含 28 个棒球队 1996 年赛季的数据。数据中的列包含下列变量:
- 这个队获胜的比赛次数。
 - 队平均得分(对投手能力的一个度量)。
 - 每次比赛的平均得分。

4. 偷垒的总次数。
5. 本垒打的次数。
6. 每队击球得分的平均数。

最终关心的是每个队在整个赛季获胜的场次,所以第一列的变量是因变量。

- a. 用变量 1 和其他每个变量进行相关分析,并找出其他变量对决定获胜次数的重要性。
- b. 以变量 1 为因变量和其他每一个变量为自变量作回归分析,并找出自变量增加 1 个单位对获胜的场次有何影响。

- 10.54 下面的数据是关于 20 个国家的受教育的人的百分比和人均收入的样本。这些国家包括:阿富汗,玻利维亚,柬埔寨,智利,古巴,厄瓜多尔,加纳,圭亚那,象牙海岸,北朝鲜,马里,马拉维,尼泊尔,巴基斯坦,菲律宾,塞内加尔,南非,坦桑尼亚,乌干达和也门。

国家	1	2	3	4	5	6	7	8	9	10
受教育的百分比	6	43	50	87	80	71	30	77	9	77
人均收入	61	165	125	645	398	208	289	311	246	86

国家	11	12	13	14	15	16	17	18	19	20
受教育的百分比	10	6	6	22	80	6	46	11	30	6
人均收入	46	72	73	107	246	158	600	174	92	66

来源: Arthur S. Banks, Cross - Polity Time Series Data, Cambridge, MA: MIT Press, 1971, pp. 237 - 255, 269 - 282.

- a. 两变量如何相关?
- b. 如果两个国家的人均收入(自变量)相差 100,其受教育的百分比(因变量)有多大差异?
- c. 如果两个国家的受教育的百分比(自变量)相差 10%,则其人均收入(因变量)相差多少?

- 10.55 在 Springer 的网点 (<http://www.springer-ny.com/suplements/ivrsen/>) 上, Baseball Individual Scores 这个数据文件包含 1996 年赛季两个俱乐部 480 个棒球手的数据。数据中的列包含下列变量:

1. 击球的次数。
2. 得分数。
3. 击中的次数。
4. 本垒打的次数。
5. 击球得分数。
6. 平均击球得分数。
- a. 作为中间趋势的度量,哪些变量使用平均值,哪些用中位数?
- b. 研究这些变量中某些变量间的关系。

- 10.56 在对人口增长和谷物的产量增加的关系的讨论中,1858 年的一本教科书有下面的一

段话:

最近的一段研究工作中的一段话说明了许多不同的食物的产量的变化,它还表明食物的增长是人口增长的两倍。所以,Malthusian(马尔萨斯)理论对于法国的情况证据不足……对于谷物来说,我们的农业统计给出了……的数字[表 10.11]。(来源:H. C. Carey, *Principles of Social Science*, vol. 2, Philadelphia: Lippincott, 1858, p. 54.)

表 10.11 习题 10.56 的数据

年份	人口(百万)	产量(百万升)
1760	21	94.5
1784	24	115.8
1813	30	132.4
1840	34	182.5

来源:H. C. Carey, *Principles of Social Science*, vol. 2, Philadelphia: Lippincott, 1858, p. 54

- 作产量关于人口的回归。
 - 用分析的结果分析引用的这段话。
 - 作人口关于年份的回归及产量关于年份的回归。
 - 构造一个新的变量: 比例 = 产量/人口, 并作比例关于年份的回归。
 - c 和 d 部分的分析告诉你什么?
- 10.57 从表 10.12 中关于 Calabrian 黑手党的数据,可以研究黑手党(coscas)组织选择其领袖时,是否使得其领袖的年龄与组织成员的平均年龄相关。分析两变量间的关系。

表 10.12 习题 10.57 的数据

Cosca 名字	成员平均年龄	首领的年龄
Cataldi-Marafioti	37	42
Nirta-Romeo	40	67
Ursino-Jerino	34	53
Ruga	31	29
D'Agostino	39	54
Mazzaferro	32	38
Aquino-Seali	33	36
Cordi	35	29
Maeri	39	43

来源:Pino Arlacchi, *Mafia Business*, London: Verso, 1986, p. 132. Brought to our attention by Matthew Werner

- 10.58 考虑量近你花费在 10 件礼物上的钱,这些礼物你送给那些曾送给你礼物的那些人,用这些数据作一个散点图。模仿下面 Fred 这样的例子,他收到了一份 \$5 的礼物,送了一份 \$10 的礼物。

名字	给他人的礼物的花费	他人给我礼物的花费
Fred	\$10	\$5

你可以用任何你喜欢的(礼品)交换,不管是实际的还是虚构的。分析两变量间的关系。

- 10.59 从一个回归分析得出,在一个州立大学预期的由平均积分点(GPA)度量的成功可由下式得出:

$$\text{大学 GPA} = 0.6 + 0.74(\text{中学 GPA})$$

- 用你自己的中学的 GPA,如果你上了大学计算你预期的 GPA 是多少。
 - 就你现在的情况估计你的 GPA。改变这个回归方程使得它更适合你的个人情况。
 - 新的方程看上去像什么?
 - 这个回归方程适合于每一个人吗?为什么不是?
 - 一个招生办公室的官员正在看 Elmer Ebert 的中学成绩,他的中学 GPA 是 1.9。在大学 GPA 低于 2.0 的就不可能毕业。如果 Elmer 被录取了,你预测他会毕业吗?
 - 如果 Elmer 及其他 GPA 比较低的都不被大学录取,回归方程可能有何变化?
- 10.60 下面是关于参加篮球训练营对于得分百分比的影响的数据(表 10.13)所作一个平方和的表。训练营是一个虚拟变量,不参加为 0,参加为 1。

表 10.13 习题 10.60 的数据

来源	???? 的和	比例
训练营	90999	????
残差变量	???????	0.21
总计	115189	????

- 表中一些项是由一个粗心的喝咖啡的人不小心给弄模糊了。为你这个邋邋的朋友修复这个表。
 - 计算训练营变量和篮球手的表现变量的相关系数 r 。
 - 训练营看上去对表现有影响吗?
- 10.61 给四个雄鼠和四个雌鼠的肝脏注入油酸。表 10.14 是注入和吸收到体内的数量。
- 用注入量作为自变量,吸收量作为因变量作这些数据的散点图。用不同的符号代表雌鼠和雄鼠。

表 10.14 习题 10.61 的数据

注入	吸收	性别
29.3	1.82	雌
25.5	0.84	雌
26.3	1.09	雌
31.0	1.45	雌
20.6	1.56	雄
17.9	0.93	雄
23.6	1.54	雄
25.4	1.76	雄

来源: C. Soler-Agilaga and M. Heimberg, "Comparison of Metabolism of free fatty acid by isolated perfused livers from male and female rats," *Journal of Lipid Research*, vol. 17 (1976), pp. 605-615.

- 描述对每个性别的小鼠注入和吸收的关系。
- 对每个性别计算回归直线。

d. 两组观测数据中的关系有何差别?

e. 假定对一个雌鼠和一个雄鼠的注入值均为 25, 这两个鼠的因变量的预测值有何差异?

f. 雌鼠和雄鼠的吸收量之间看上去有何差异吗?

10.62 如果你从婴儿起在每个生日都量了你的身高, 并作这些数据的散点图, 这些点将不会落在一条直线上。但是短期内的增长数据有时可以用线性回归来分析。下表是 Count de Montebillard 的儿子从 1762 年到 1789 年间其年龄和身高的数据。

年龄(年)	3	4	5	6	7	8	9
身高(cm)	98.8	105.2	111.7	117.8	124.3	130.8	137.0

来源: R. E. Scammon, "The first serial study of human growth," *American Journal of Physical Anthropology*, vol. 10 (1927), pp. 329 - 336, as reported in R. L. Sandland and C. A. McGulchrist, "Stochastic growth curve analysis," *Biometrics*, vol. 35 (1979), pp. 255 - 271.

a. 作这些数据的散点图。

b. 这些数据有线性趋势吗?

c. 计算这些回归直线的方程。

d. 对这个例子怎样解释回归系数的值?

e. 用下一年的身高减去当年的身高计算 Count 的儿子每年增长多少, 并计算每年平均增长多少。

f. 解释回归系数和每年平均增长的身高间的联系。

10.63 在一个 10 个州的样本中, 每个州中关于接受医疗补助人口的百分比和医院中每 100000 人的床位数的数据如表 10.15 所示。

表 10.15 习题 10.63 的数据

州	获医疗资助的百分比	每十万人床位数
阿肯色 (Arkansas)	11.3	430
佛罗里达 (Florida)	8.0	392
印地安那 (Indiana)	6.3	382
缅因 (Maine)	10.8	335
密西西比 (Mississippi)	16.8	457
新罕布什尔 (New Hampshire)	4.0	290
北达科它 (North Dakota)	7.7	507
罗得岛 (Rhode Island)	11.7	319
犹他 (Utah)	6.3	255
威斯康辛 (Wisconsin)	8.0	342

来源: Medicaid data: U. S. Department of Commerce, Bureau of the Census and the Health Care Financing Administration, Form - 2082. Hospital bed data: American Association of Retired Persons, Reforming the Health Care System; State Profiles 1990, Washington D. C.: AARP, 1991. These data are reprinted in the report Medicaid Hospital Payment Congressional Report, The Prospective Payment Assessment Commission, C - 91 - 02, October 1, 1991, pp. 27 and 39.

a. 为什么不同的州之间每 100000 人床位数彼此不同, 在北达科它州高达 507 个, 而在犹他只有 255 个?

- b. 以接受医疗补助的百分比作为自变量, 以每 100000 人中床位数作为因变量作这些数据的散点图。以州的名字来标记这些点。
- c. 评价你在数据中看到的模式。
- d. 分析两变量间的关系。

10.64 死刑的威慑作用是一个广受争论的问题。表 10.16 是自 1950 年起的 10 年间, 这个国家由于杀人而被执行死刑的人数和杀人率的散据。这些数据对于死刑的威慑作用的评价有何补充的地方?

表 10.16 习题 10.64 的数据

年份	执行死刑的次数	杀人率
1950	68	5.3
1951	87	4.9
1952	71	5.2
1953	51	4.8
1954	71	4.8
1955	65	4.5
1956	52	4.6
1957	54	4.5
1958	41	4.5
1959	41	4.6

来源: W. C. Bailey and R. D. Peterson, "Murder and capital punishment: A monthly timeseries analysis of execution publicity," *American Sociological Review*, vol. 54 (1989), p. 740.

10.65 在一项全国毒物计划(National Toxicology Program)的研究中, 大约 100 个雌鼠被喂了乙二醇, 然后观察它们的幼仔。在这个研究中用了四种不同的剂量, 表 10.17 是接受了特定剂量的每一组中幼仔的平均数目及幼仔中畸型的百分比和胎儿的平均重量。

表 10.17 习题 10.65 的数据

剂量(g/kg)	幼仔平均大小	畸型的百分比	胎儿的平均重量(g)
0.00	11.90	0.3	0.972
0.75	11.50	9.3	0.877
1.50	10.40	39.0	0.764
3.00	9.83	57.0	0.704

来源: C. J. Price, C. A. Kimmel, R. W. Tyl, and M. C. Marr, "The developmental toxicity of ethylene glycol in rats and mice," *Toxicological Applications in Pharmacology*, vol. 81 (1985), pp. 113 - 127, in P. J. Catalano and L. M. Ryan, "Bivariate latent variable models for clustered discrete and continuous outcomes," *Journal of the American Statistical Association*, vol. 87 (1992), pp. 651 - 668.

- a. 用剂量作为自变量, 另外三个变量分别作为因变量作三个散点图。
- b. 这些散点图说明了什么?
- c. 利用回归分析和相关分析来分析剂量分别和其他三个变量的关系。
- d. 比较三个相关系数, 你能看出剂量对三个因变量哪个更重要吗?
- e. 幼仔的大小及重量是大约 25 个观测的平均值。对每一个剂量如果用每个个体的初始值而不是用平均值, 这对分析有什么影响?

10.66 一些州对其住院病人的医疗资助费用的支付,可以用两种方法。一种是基于花费的支付方案,另一种是预期支付系统。表 10.18 是这些州在四个不同时间用预期支付方法的百分比。分析两变量间的关系。

表 10.18 习题 10.66 的数据

年份	百分比
1977	14
1981	32
1985	84
1991	92

来源:Medicaid Hospital Payment Congressional Report, *The Prospective Payment Assessment Commission*, C-91-02, October 1, 1991, p. 44.

10.67 在一次对身体中的脂肪的百分比和年龄的研究中,得到了下面的数据。分析这些数据。

年龄	23	23	27	27	39	41	45	49	50
脂肪的百分比	9.5	27.9	7.8	17.8	31.4	25.9	27.4	25.2	31.1

年龄	53	53	54	56	57	58	58	60	61
脂肪的百分比	34.7	42.0	29.1	32.5	30.3	33.0	33.8	41.1	34.5

来源:R. R. Mazeness, W. W. Peppler, and M. Gibbons, "Total body composition by dualphoton (^{153}Gd) absorptiometry," *American Journal of Clinical Nutrition*, vol. 40 (1984), pp. 834-839.

10.68 表 10.19 是在一些安第斯北部南美洲热带地区的荒山隔离地带鸟的种类和这些地区的海拔高度(千英尺)的数据。这里的数据只包括海拔低于 5000 英尺的地区。分析这些数据并看一看鸟的种类数是否和海拔高度有关。

表 10.19 习题 10.68 的数据

地区	种类数	海拔(千英尺)
Chiles	36	4.1
Las Papas - Cocunuc	30	3.8
Sumapaz	37	3.5
Parmillo	11	1.5
Pamplona	11	2.3
Cachira	13	2.4
Tama	17	2.0
Batallon	13	2.2
Merida	29	4.9
Perja	4	2.5
Cende	15	1.8

来源:F. Vuilleumier, "Insular biogeography in continental regions: I. The northern Andes of South America," *American Naturalist*, vol. 104 (1970), pp. 373-388.

10.69 下表展示了关于大不列颠、挪威和瑞典某些地区的年平均温度和乳腺癌死亡率的数据。

地区	1	2	3	4	5	6	7	8
温度	51.3	49.9	50.0	49.2	48.5	47.8	47.3	45.1
死亡率	102.5	104.5	100.4	95.9	87.0	95.0	88.6	89.2

地区	9	10	11	12	13	14	16	16
温度	46.3	42.1	44.2	43.5	42.3	40.2	31.8	34.0
死亡率	78.9	84.6	81.7	72.2	65.1	68.1	67.3	52.5

来源: A. J. Lea, "New observations on distribution of neoplasms of female breast in certain European countries," British Medical Journal, vol. 1 (1965), pp. 448-490.

a. 分析这些数据。

b. 这是个偶然的联系或者你是否可以认为有其他变量可以解释这两个变量为什么相关?

10.70 表 10.20 是 10 年间加州每 100000 人中的犯罪率和这个州的年人口数。

表 10.20 习题 10.70 的数据

年份	杀人	强奸	抢劫	殴打	总计	总人口(百万)
1983	10.5	48.2	342.3	374.6	775.6	25.1
1984	10.5	45.7	328.3	379.9	764.4	25.6
1985	10.7	43.0	331.1	388.2	773.0	26.1
1986	11.4	45.3	346.0	526.1	928.7	26.7
1987	10.7	44.2	304.4	568.5	927.8	27.4
1988	10.5	41.9	307.2	574.0	933.6	28.1
1989	11.0	41.6	335.1	599.5	987.2	28.8
1990	12.1	43.0	380.5	619.8	1055.4	29.6
1991	12.6	42.2	408.2	616.7	1079.7	30.6
1992	12.5	40.7	418.1	632.5	1103.8	31.3

来源: California Department of Justice, as reported in The Economist, March 19, 1994, p. 31.

a. 为什么分别以年份列或以人口列作为自变量没有多大差别?

b. 作殴打犯罪率和年份的散点图并描述你看到的模式。

c. 你怎样解释散点图中的模式?

d. 作强奸犯罪率关于年份的回归分析, 并报告你的结果。

e. 作杀人犯罪率关于年份的回归分析, 并报告你的结果。

f. 作抢劫犯罪率关于年份的回归分析, 并报告你的结果。

10.71 收集两个数量变量的数据, 分析这些数据并报告你的结果。

10.72 下表是华盛顿州 Garrison 海湾关于 16 个小颈蛤两次测量的数据(单位: 毫米)。

蛤	1	2	3	4	5	6	7	8
长度	530	517	505	512	487	481	485	479
宽度	494	477	471	413	407	427	408	430

蛤	9	10	11	12	13	14	15	16
长度	452	468	459	449	472	471	455	394
宽度	395	417	394	397	402	401	385	338

来源: *D. F. Andrews and A. M. Herzberg, Data: A Collection of Problems from Many Fields for the Student and Research Worker, New York: Springer - Verlag, 1985, p. 336.*

- 在一次关于这两个变量的研究中,应选哪个变量作自变量,哪个作因变量?
- 作这些数据的散点图。
- 两变量间关系的强度如何?
- 散点图中有一点与其他点相距甚远。除去这点后再计算关系强度。
- 左下角的点关系强度有大的影响吗?

10.73 地方报纸每周都公布那些在地方法院申请结婚证的人的名字和年龄。这里是一周内关于新郎和新娘的年龄,记作(新郎年龄,新娘年龄):

(37, 30) (30, 27) (65, 56) (45, 40) (32, 30) (28, 26) (45, 31) (29, 24)
 (26, 23) (28, 25) (42, 29) (36, 33) (32, 29) (24, 22) (32, 33) (21, 29)
 (37, 46) (28, 25) (33, 34) (17, 19) (21, 23) (24, 23) (49, 44) (28, 29)
 (30, 30) (24, 25) (22, 23) (68, 60) (25, 25) (32, 27) (42, 37) (24, 24)
 (24, 22) (28, 27) (36, 31) (23, 24) (30, 26)

来源: *The Philadelphia Inquirer, September 10, 1995, p. MD 12 - d.*

- 如果每个新郎和新娘都同岁,穿过这些点的回归直线的斜率和截距等于什么?
- 如果每个新郎都比他的新娘大 5 岁,穿过这些点的回归直线的斜率和截距等于什么?
- 如果每个新郎都比他的新娘大 10%,穿过这些点的回归直线的斜率和截距等于什么?
- 对于这些实际年龄作出回归直线。
- 从这条回归直线,你对新郎和新娘的年龄模式可得出什么结论?
- 对于这些数据,这个散点图比习题 3.20 多告诉你什么?

10.74 这个习题研究自变量 x 与因变量 y 的部分变异有关这个思想。假定你有如下的数据:

$x:$ 1 2 3 4

$y:$ 3 2 5 6

- 说明 y 的总变异为 10.0。
- 回归直线的方程为 $y = 1.0 + 1.2x$ 。计算预测值 y , 并说明这些预测值的变异为 7.2。
- 计算 4 个残差并说明这些残差的变异为 2.8。
- 在 y 的总变异中与 x 有关的比例是多少,与残差变量有关的比例是多少?
- 在 x 和 y 之间的相关性有多强?

C H A P T E R 11



- 11.1 方差分析:对比事物的平均值
- 11.2 问题 1. 犯罪率和地区之间的关系
- 11.3 问题 2. 关系有多强?
- 11.4 问题 3. 这个关系是纯属偶然的吗?
- 11.5 问题 4. 是因果关系吗?
- 11.6 方差分析:鸟瞰回顾
- 11.7 配对分析:每个单元两个观测
- 11.8 小 结

ANOVA:一个分类变量

和一个数量变量

的方差分析



许多调查表明,犯罪是当今人们关注的主要问题之一。因此在政治辩论中、新闻报导里以及邻居朋友之间的争论中都少不了要谈论到这个问题。但是关于这个问题的重要性,人们的看法往往不一致。我们不禁要问:是犯罪率高得太危险了,还是我们在杞人忧天?

关于犯罪的问题有许多。其中一个问题是犯罪率是否与地区有关,即住在一个地方是否比住在另一个地方安全?另一个问题是犯罪数量是否在上升。

在本章中我们主要讨论一类特殊的犯罪——暴力犯罪,它包括:谋杀、强奸、抢劫和团体暴力罪等。我们要讨论的问题是:在各个地区遭到暴力侵犯的几率是否一样大。如果我们发现各地的暴力犯罪数确有不同,那么什么地方高,什么地方低?这些问题不只是理论问题。人身伤害随时可能会发生。

要回答这个问题,我们需要知道在某个时间段内各地的暴力犯罪数。但是要想得到去年宾夕法尼亚州有多少暴力犯罪发生是不容易的。虽然经常计算了已经记录在案的暴力犯罪,但是我们有理由相信有大量的暴力犯罪因为这样或那样的原因而没有备案。另一个获得数据的方法是抽取一个样本,调查其中的人是否当过暴力犯罪的牺牲者。但是这种调查有着抽样

调查普遍存在的误差(你可能记得在第二章关于收集数据中我们已经对其中的几个问题做过讨论)。尽管如此我们还是能从适当的统计样本中得到比警察局所能提供的数据更精确的数据。(例如性暴力犯罪案件一般倾向于不报告警察局。)

本章的数据是从 FBI 报告提供的 1986 年到 1992 年有关 48 个大陆州的综合犯罪报告中获得的。因为各州的人口数目和犯罪数目相差甚远,所以我们用暴力犯罪率(每 100000 个人中发生暴力犯罪的次数)来作为衡量标准。例如:1986 年宾夕法尼亚的暴力犯罪率是 359 次每 100000 人。我们把所有的州划分为 7 个地区——新英格兰、中大西洋、中西部、南方、西南方、落基山地区和太平洋岸——并对比它们在犯罪率上的不同。表 11.1 给出了它们各自的犯罪率。

表 11.1 1986 年 48 个州的暴力犯罪率

州名	犯罪率 (每十万人)	地区	州名	犯罪率 (每十万人)	地区
缅因	147	新英格兰	西弗吉尼亚	164	南方
新罕布什尔	140		北卡罗来纳	476	
佛蒙特	149		南卡罗来纳	675	
马萨诸塞	557		佐治亚	588	
罗德岛	336		佛罗里达	1036	
康涅狄格	426		肯塔基	334	
纽约	986	中大西洋	田纳西	540	西南方
新泽西	572		阿拉巴马	558	
宾夕法尼亚	359		密西西比	274	
俄亥俄	423	中西部	阿肯色	395	
印地安纳	308		路易斯安纳	758	
伊利诺依	800		俄克拉何马	436	
密歇根	804		德克萨斯	659	
威斯康辛	258		亚利桑纳	658	
明尼苏达	285		新墨西哥	726	
依阿华	235		怀俄明	293	落基山区
内布拉斯加	263		克罗拉多	524	
密苏里	578		蒙太拿	157	
北达科它	51		爱达荷	222	
南达科它	125		犹他	267	
堪萨斯	369		内华达	719	
特拉华	427	南方	华盛顿	437	太平洋岸
马里兰	833		俄勒冈	550	
弗吉尼亚	306		加利福尼亚	920	

我们现在的任务是用各地区所包含的州的数据来研究这 7 个地区的犯罪率是否不同。注意,虽然我们已经确定要利用各州的数据来研究地区的差异,我们也可以选择其它的地理区域作为分析的单位。例如可以利用 3000 个左右的县的暴力犯罪率。我们并不清楚,用州的数据是否比用县的更好;用州数据的一个理由是观测数不那么大。

在进行分析时,因变量是暴力犯罪率,它是一个数量型变量(它的值属于一个区间);而地区是自变量,它是一个分类变量。每个地区可以用一个数来代替,但要注意:如果南方用 6 来代替,即使我们是南方人,这也不意味着它是标以 3 的新英格兰的 2 倍。因为我们要研究的是

地区对暴力犯罪率的影响,也就是研究分类型自变量对数量型因变量的影响。

自变量是只有两个值的分类型变量的特殊情况已经在第 7 章中研究过了;它出现在用 t -检验来研究两个均值差异的问题中。在第 10 章中也研究过自变量是虚拟变量的情况。而在这里自变量是有 7 个值的分类型变量,所以适用于两个值的 t -检验就不能胜任这项工作了。

停下来想一想 11.1

我们想知道失业率是否因为地区的不同而不同。这个问题与我们讨论的暴力犯罪率问题如何类似?请给出这个问题中的变量和它们的类型(分类型,数量型)。

你能构造出其它类似的问题,但利用不同的变量吗?

11.1 方差分析:对比事物的平均值

方差分析是用来对比因变量在不同组中的平均值的统计方法。

研究分类型自变量对数量型因变量的影响时,我们用的统计方法称为方差分析(analysis of variance),简称为 anova。尽管表面上看起来差别很大,方差分析与第 10 章中的回归分析其实是密切相关的。它们都可以认为是一个更一般的统计模型的特殊情况。

方差分析最初是在 20 年代发展起来的,是一种在许多领域普遍应用的统计工具。它特别经常地用于分析心理学、生物学、工程和医药的实验数据。在实验中我们常常把自变量叫做处理变量,把因变量叫做响应变量。在农业实验中,处理变量可能是分别用在一块玉米地里的不同类型的肥料,而响应变量则是由不同肥料所得的产量。当然,就像下面将要看到的那样,方差分析也可以用于处理观测数据。

11.2 问题 1. 犯罪率和地区之间的关系

散点图

我们要问的第一个问题就是数据中的两个变量之间是否存在关系;即两个地区的犯罪率是否真的有差别。我们用与相关分析和回归分析中相同的方法来回答这第一个问题。为了显示暴力犯罪是否因地区的不同而不同,我们先把数据展示在一个散点图中,其中地区是自变量被标在水平轴上,暴力犯罪率是因变量被标在竖直轴上。图 11.1 就是这个数据的散点图。图中的每个点代表 48 个州之一;这些州又被分在了几个地区中。

方差分析与回归分析的主要区别在于:方差分析中的沿水平轴的自变量是分类变量,而回归分析中沿水平轴的自变量是数量变量。如果自变量是分类变量,那么我们可以随便把它的各个类(值)以任何顺序放在水平轴上我们想放的地方。在这个图中,自变量的值(地区)只是被简单地按照字母顺序排在了水平轴上。如果我们用其它的方式来排列这些地区,比如从东

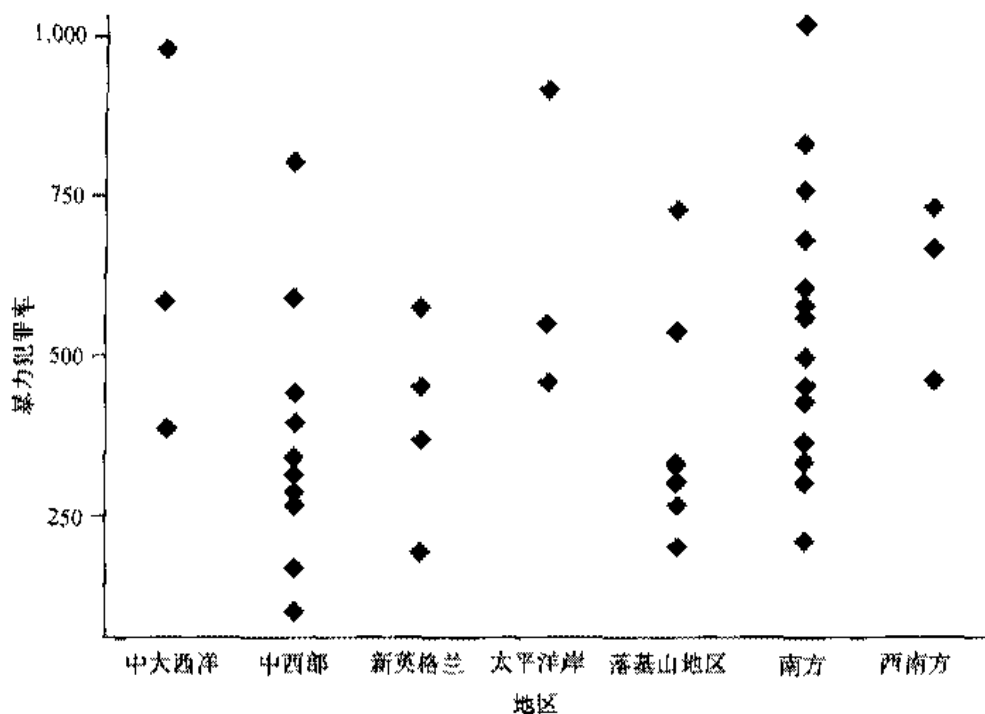


图 11.1 暴力犯罪率的散点图。

北到西南,这个点图的模式就会有所不同。既然这个点图的模式是任意的,像在回归分析中所做的那样画一条穿过点的线就毫无意义。对于数量型变量来说,水平轴上值的位置是由从低到高的数值决定的,因此只有一种方式来放置这些点并且画一条穿过这些点的直线。在散点图的水平轴任意放置分类变量值和固定放置数量变量的值之间的区别是方差分析与回归分析之间的本质区别。

略微看一下散点图你就会发现各个地区的暴力犯罪率的确是有明显差别的。而且即使在同一地区,各州也明显不同。从图中可以看出新英格兰暴力犯罪率的水平普遍低于其它各地区,而中大西洋和太平洋岸的暴力犯罪率似乎总的比较高。这些区别至少说明地区与犯罪率之间是有关系的。如果所有这七个地区的暴力犯罪率大小相似,则可以认为地区与犯罪率之间没有关系。

盒子图:更简单地了解数据

为了更容易找出各地区之间犯罪率的不同我们需要一种比散点图更简单的图。散点图中的 48 个点提供了数据的全部的信息,但是这些点太多了以至很难看清两个变量之间的关系。简化方式有许多种,其中一种就是对每个地区做一个盒子图。在盒子图中,把在一个区域的各州数据用 5 个数(暴力犯罪率的中位数,第 25 和第 75 分位点,最大值和最小值)代替。盒子图如图 11.2 所示。

盒子图把每个地区的数据减少到了 5 个,因此 7 个地区的 48 个原始观测就被缩减成了 35 个数。这个减少并不多,但至少每个地区的数据都被用同样的方式表示出来了。如果每一群中的观测比这个数据提供的要多,那盒子图的简化作用就更明显了。盒子图使地区之间的可

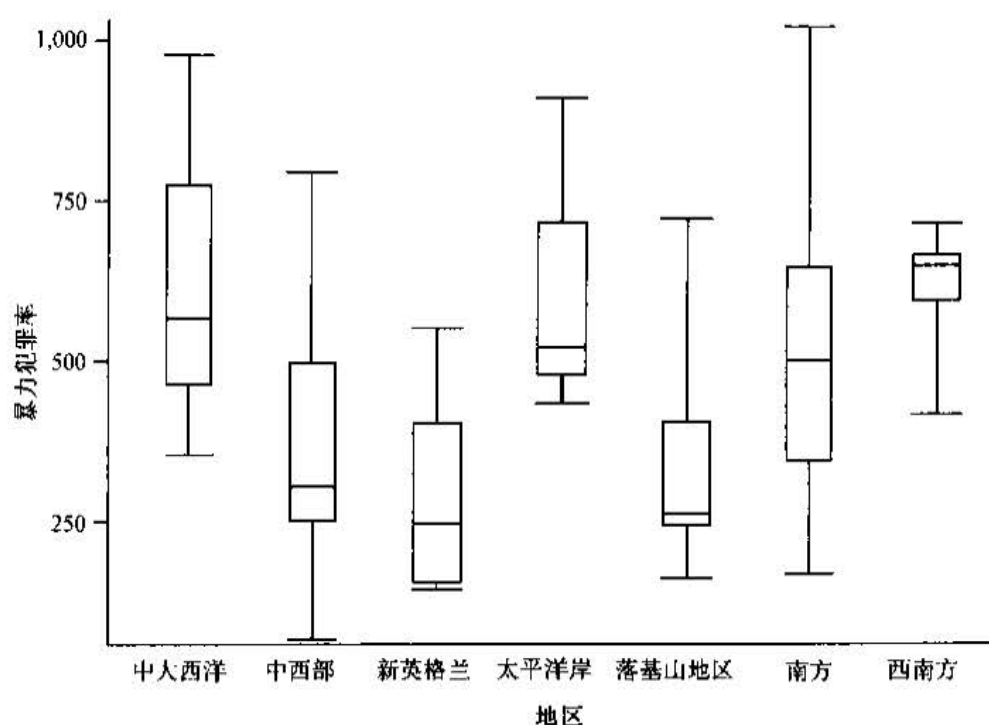


图 11.2 图 11.1 中暴力犯罪率数据的盒子图。

比性增强了。因此各个地区的盒子图被放在一起就使我们更容易看出地区之间的不同。

比较基于数据的盒子图揭示了哪些关于暴力犯罪的信息? 首先应该来对比不同地区的中位数,因为它们代表中心值。当我们细看盒子中间代表中位数的横线时,就会注意到西南部、中大西洋和太平洋岸三个地区的中位数最高。因此可以断定这三个地方的平均暴力犯罪率最高。同样可以认为新英格兰的平均暴力犯罪率最低,紧接着是落基山地区。

盒子图的另一个特征是盒子的高度不同。例如西南部和落基山地区的盒子比其它地区要短。这就意味着该地区所含各州之间的暴力犯罪率比其它地区要更相似些。

11.3 问题 2. 关系有多强?

盒子图比散点图更能显示各地区之间暴力犯罪率的不同和两个变量间存在关系。但我们还想知道这两个变量之间关系的强度,以及这个关系是否可能出于偶然。要回答这些问题我们还需要做进一步的工作,即利用方差分析。

主要是在研究各组之间的差异时用均值进行数学运算更方便,所以在正规的方差分析中用的不是中位数,而是每一组观测的均值。

方差分析的名字在某种程度上会使人产生误解。一个更适合的名字恐怕应该是均值分析:先根据自变量(地区)分组,再求出每一组的因变量的平均值,我们关心的是因变量(暴力犯罪)的均值在由自变量(区域)所定义的组之间是否不同。于是,虽然我们的兴趣在均值上,但

在判断均值之间是否有差异时要借助于方差。

要回答第二个问题,首先就要计算各地区的平均暴力犯罪率,和所有州的暴力犯罪率的总平均。图 11.3 给出了这些均值。图 11.3 是和图 11.1 相同的散点图,唯一不同的就是加了一条用以显示所有州的暴力犯罪的总平均值($\bar{y} = 460$)的水平线和一条联接 7 个地区暴力犯罪率均值的折线。(既然暴力犯罪率是因变量或 y -变量,我们就用字母 y 来作为均值的记号)注意均值的范围是从最低的新英格兰(292)到最高的中大西洋(639)。

地区变量

方差分析的前提是某个州的暴力犯罪率完全是由这个州所在的地区和一些对各州都有影响的因素决定的。这两个因素完全决定了一个州的暴力犯罪率(逻辑上不会是别的。)

一种理解其含义的方法是思考一个州如何得到其暴力犯罪率的。首先,想像所有的州有相同的暴力犯罪率。要估计这个共同的犯罪率最好的办法就是将 48 个州的观测值加在一起求平均,这个值在图 11.3 中被记为 \bar{y} 。数值上 \bar{y} 值等于 460。这样如果没有其它变量影响的话所有州的暴力犯罪率都应该是 460。

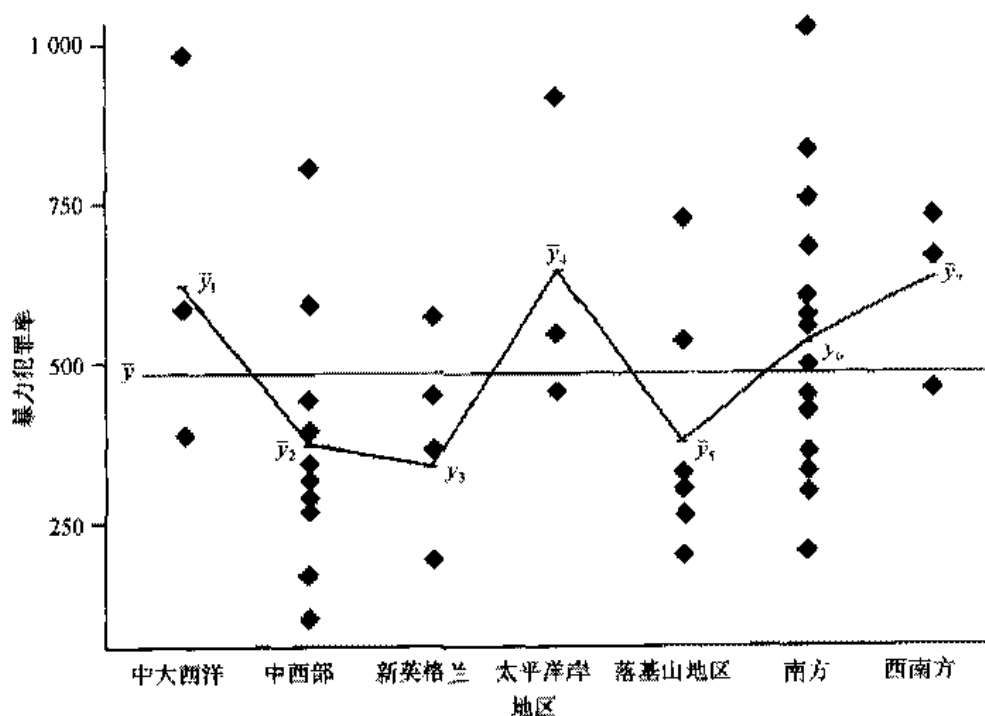


图 11.3 暴力犯罪率的散点图及总均值和地区均值。

下面再想象各州的暴力犯罪率被各自所在的地区的效应所影响。这个影响使每一个地区的暴力犯罪率从总均值变化到地区的一个共同值。因为属于同一个地区的各州受到的影响相同,所以这些州的暴力犯罪率应该相同。该共同值的合理估计是该地区的观测平均。例如中大西洋的三个州纽约、新泽西和宾夕法尼亚将有同样的暴力犯罪率 639,它是三个州观测值

986,572 和 359 的平均。其它地区值的求法也类似,它们在图 11.3 中是用一条折线联接在一起的。

这样地区变量对各州的影响等于地区均值与总均值之差。例如,地区变量的影响使宾西法尼亚的暴力犯罪率从总均值 460 变到了 639,中间相差 $639 - 460 = 179$ 。地区变量对中大西洋的其它两个州同样也有影响。把每个州的这个差值求平方然后再相加得到值 662241。这个值可以用来度量地区变量的效应,我们叫它地区(自变量)平方和。

求自变量(这里是地区)平方和的公式为:

$$(\text{组均值} - \text{总均值})^2 \text{ 之和}$$

这里是对所有观测求和。

残差变量

但是同一个地区每一个州的暴力犯罪率并不相同,也就是一定有其它变量影响了州的暴力犯罪率使它偏离地区均值。其它变量的总影响称为残差变量。例如,宾夕法尼亚的暴力犯罪率是 359,而地区均值是 639。是残差变量使这个比率从 639(在没有残差变量的影响下)变成了 359(这个州的观测值)。残差变量的效应可以用观测值与地区均值之差来度量。

下面我们计算所有这些差值的总和有多大。因为这些差值有正有负,所以它们和的均值是 0,这对我们没有任何帮助。但是如果我们先对某一地区的差异求平方然后再相加,这个值可以用来度量残差变量在这一地区的效应。把所有地区的这个值加在一起就得到了残差变量对所有州的效应,这个和叫做残差平方和,在本例中它的值是 2145613。因此在这个数据中残差对暴力犯罪率影响的量——除地区变量之外的所有变量的联合效应——是 2145613。

残差变量有时称为误差变量(error variable)。在这里,误差可不意味着变量中有错误。方差分析中许多原始的工作是处理多次测量同一个量时所得到的数据。它先假设有一个真值存在,而一个观测值不等于真值的部分是由测量误差引起的。这个误差项现在称为残差变量。

残差变量,有时称为误差变量或“所有其它”变量,它是为除自变量之外所有能够对因变量产生影响的变量所起的名字。

残差平方和定义为,对所有观测的

$$(\text{观测} - \text{组的均值})^2 \text{ 之和}$$

地区变量和残差变量的总效应:总平方和

各州的暴力犯罪率不尽相同,因为它们同时受地区变量和残差变量的影响。这样一个州观测到的暴力犯罪率的值与总均值 460 的差异是地区变量和残差变量共同影响的结果。例如,在宾夕法尼亚观测值是 359,它与总均值的差是 $359 - 460 = -101$ 。这样,地区变量和残差变量对宾夕法尼亚暴力犯罪率的联合效应等于 -101。用同样的过程可以求出这 48 州各自的差异值。

为了把观测与总均值之间的差异值归并成一个量,我们先求出它们各自的平方,然后再把它们相加。这个差异的平方和称为总平方和。在这里暴力犯罪率的总平方和等于 2804254,亦即所有影响暴力犯罪率的变量的联合效应是 2804254。

自变量和残差变量的联合效应是总平方和,公式为对所有观测的:

$$(\text{观测} - \text{总均值})^2 \text{ 之和}$$

这两个变量的联合效应引人注目地等于我们前面计算的两个变量各自的效应之和。地区变量的效应是 662641,残差变量的效应是 2145613; $662641 + 2145613 = 2804254$ 。

测量关系的强度

表 11.2 给出了用平方和来度量的变量的效应。像这么大的数很难对比,所以表的第三列给出了两个效应各自在总效应中所占的比例。地区变量的效应所占的比例是 $662641/2804254 = 0.24$,或占 24% 的总效应。这个比例 0.24 称为 R^2 。(这个数可以与回归分析中相关系数的平方直接进行比较。)

表 11.2 地区变量和残差变量的效应

来源	平方和	比例
地区	662641	0.24
残差	2141613	0.76
总数	2804254	1.00
$R = 0.49$		

在对犯罪率的影响中地区变量的效应只占总效应的四分之一,剩下的是残差变量的效应。回顾散点图 11.1 和盒子图 11.2,不难看出残差变量的影响的确很大。在这 7 个地区中观测值的变化范围都相当大,而且州与州之间差异也很大。差异主要是来自残差变量。这意味着要想全面地了解犯罪率,我们不得不考虑导致这么大残差的其它变量,并把它们作为附加的自变量进行分析。我们在 13 章中将会简单地介绍这种分析。至于 R^2 的计算则为本章末的公式 11.2。

R^2 的平方根当然就是 R 。既然已知 R^2 是 0.24,取平方根后, R 就应该是 0.49 了。这个数可以用来测量自变量和因变量之间关系的强度。(这个数可以直接与回归分析中的相关系数 r 进行比较;它们之间主要的区别是 R 没有负值。) R 的变化范围是从 0 到 1。因为 R 的值是 0.49,地区变量和暴力犯罪率之间的关系强度是中等。

对变化量的解释程度

我们常说自变量解释了(产生了,引起了或是造成了)因变量变化的百分之多少。例如,暴力犯罪率变化的 24% 可以用地区变量来解释。那么这些话是什么意思?

Pythagoras 三角^①

让我们用稍微数学化的眼光来重新回顾一下表 11.2 中的数据。我们如何在一个图中表示出这三个平方和? 我们可以用分成两块的面饼图来表示, 其中用较大的一块代表残差变量, 用较小的一块代表地区变量。我们当然也可以用不同种类的条形图来表示。

这里我们要介绍另一种图示方法。这三个数都是平方和, 但我们只把它们看成普通的平方。然后我们跟从聪明的希腊数学家的想法(即勾股定理): 当两个数的平方和等于第三个数的平方时, 这三个数就能构成一个直角三角形, 它的三个边的长度就是这三个数。

两个数的平方和等于第三个数的平方的公式如下:

$$662641 + 2141613 = 2804254$$

它可以被改写成:

$$814^2 + 1463^2 = 1674^2$$

我们于是可以画一个斜边为 1674, 两个直角边分别为 814 和 1463 的直角三角形。

这个三角形有一些很好的性质。首先, 它修改了我们所做的平方的效应。因为利用平方就强调了差异; 一平方就使值变得非常大。我们或许不应给远离均值的观测加这么大的权。平方根改变了对效应的比例。现在看来地区变量的重要性(814)差不多是残差变量重要性(1463)的一半。而如果用平方和来测量, 地区变量只能解释犯罪率变化的四分之一。

其次, 由地区自变量和斜边构成的左下角(θ)的度数是这个直角三角形的另一个特点。用与这个角相邻的直角边的长度除以斜边的长度所得的值就是这个角的余弦值, 用公式表示就是 $\cos(\theta) = 814/1674 = 0.49$ 。从数学用表中可以查到余弦值等于 0.49 的角的度数是 61 度。

由斜边和地区自变量的边构成的角的余弦值引人注目地等于自变量和因变量之间的相关系数 R 。这个角的度数越小, 它的余弦值或相关系数 R 就越大。

当角的度数是 0 时就没有残差的影响, 相关系数是 1。同样当它是 90 度时, 相关系数是 0。很清楚, 这个余弦值能够解释自变量的效应、残差变量的效应以及它们之间的相关性。



表示平方和的三角形

如果地区变量和残差变量对犯罪率没有影响, 则所有的州将会有相同的暴力犯罪率亦即

^① 即中国的勾股定理, 勾股定理比 Pythagoras 定理要早许多年。

州与州之间的暴力犯罪率没有变化。此时这个共同的数最好的估计就是总平均值 $\bar{y} = 460$ ——表 11.3 中的第三列。表的最后一行说当值全都相等时,它们是没有变化的。

表 11.3 暴力犯罪率在不同假设之下的变化情况

州	地区	没有变量影响 时的犯罪率 (总均值)	只有地区变量的 影响时的犯罪率 (地区均值)	地区和残差都有 影响时的犯罪率 (观测值)
缅因	新英格兰	460	292	147
新罕布什尔		460	292	140
佛蒙特		460	292	140
马萨诸塞		460	292	557
罗德岛		460	292	336
康涅狄格		460	292	426
纽约	中大西洋	460	639	986
新泽西		460	639	572
宾夕法尼亚		460	639	359
俄亥俄	中西部	460	385	423
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
加利福尼亚	太平洋岸	460	636	920
比例的变化情况	0	662641	2804254	
$R^2 = 662641/2804254 = 0.24$		$R = 0.49$		

当地区变量影响这个比率时,这些比率的值就不全都等于总均值 460,而是各个州的值等于它所处地区的均值。这样,如果地区是唯一对暴力犯罪率有影响的变量,观测数据就是表 11.3 中第四列的样子;在那里每个地区中各州的值相等;比率值的变化来自地区变量。像以前一样,我们在一列中度量这些数的变化量;这一列中的每一项减去这列的总平均值,把所得的值平方,然后相加,这就得到了由地区变量引起的 48 个州的值的变化情况,它等于 662641。

最后一列是我们实际观测到的、受到残差变量影响的数据。这些比率与只受地区变量影响时相比有了更多的变化。

要发现残差变量到底能引起多少的附加的变化,我们用每个州的暴力犯罪率减去总均值后求平方,再计算这些平方的和,得到总变化是 2804254。而地区变量的总效应是 662641。从表最下面一行来看,地区变量引起的变化只占 48 个州总变化的 24%,我们把这个量称为 R^2 。因此残差变量解释了总变化中剩下的 76%。但是我们并不知道究竟是什么原因(可能是气候、贫困、社会因素等)引起了地区与地区之间暴力犯罪率的不同,因此用解释这个词似乎不很合适。我们可以说地区变量与暴力犯罪率变化的 24% 有关联(associated),而残差变量与剩下的 76% 有关联。

11.4 问题 3. 这个关系是纯属偶然的吗?

在统计分析中要问的第三个问题是:两个变量之间的关系是否不仅存在于样本中,而且也存在于在总体中。但是这个问题只有当数据是总体的一个样本时才能得到回答。这个例子给出的是所有 48 个州的总体数据而非样本数据,所以在进行统计推断时提问的方式稍有不同:地区与暴力犯罪率之间的关系是偶然发生的还是原本就存在的。

所发生的事件产生了观察到的犯罪率。它们是否仅仅是偶然机会事件?如果这些犯罪是由偶然因素引起的,则各个地区之间的暴力犯罪率的变化就只是随机的。

停下来想一想 11.2

在研究工作中常常要区分各种作为自变量的分类变量,比如教育水平、性别、年龄、种族、住处、收入、自然背景、地区等等。回答为什么地区差异会造成暴力犯罪率不同这个问题已经是很困难的,为什么我们在基于诸如性别的人口学变量来解释差异时应该很小心?为什么在表述一个简单而直接的自变量时会有困难?为什么人们还是要这么做?

零假设

在研究两个变量有无关系时的一个典型零假设问:是否两个变量在总体中没有关系。在方差分析中,零假设通常是用术语描述在按照自变量的值分成的类中因变量均值的情况。例如:在暴力犯罪率的问题中零假设描述的是这 7 个地区的犯罪率相等,用公式表达就是:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$$

在逻辑上,7 个均值相等的备选是它们之中至少有一些不相等。这样如果我们要拒绝零假设,我们就要证明至少有两个均值不相等,而不必证明所有的均值都不同。

零假设的内容是因变量与自变量之间没有关系,即我们所发现的关系纯属偶然因素造成。要拒绝零假设(或不拒绝),我们必须从数据出发来对比我们发现的效应的值。如果对比中发现自变量(地区)的效应与残差变量的效应相比较,我们就拒绝零假设。如果对比中发现自变量(地区)的效应与残差变量的效应相比较小,我们就不能拒绝零假设。用另一种说法来表达就是:当 R^2 较大时拒绝零假设,如果 R^2 小就不能拒绝零假设。

在这个例子中我们总共处理 7 组数据以及 48 个观测。按照统计推理的规则(和统计表),在这么多组和这么多观测的情况下 R^2 必须大于 0.28 时才能拒绝零假设。

F 变量的 p -值

为了判断零假设的正确与否,我们必须找到观测数据的 p -值。为此我们就要把 R^2 变换成四个标准统计变量之一。我们已经在前面的章节中提到了这四个变量中的三个,它们分别是 z -得分, t -得分和 χ^2 -得分。在方差分析中要用到的是第四个—— F -变量(见公式 11.5)。在

把 R^2 的值转换成相应的 F -值之后,可以通过统计软件或者查 F -分布表(附录中的统计表 5)来找到 p -值。如果我们得到的 F -值大于表中给出的临界值就拒绝零假设。我们当然也可以得到观测到的 F 的 p -值即观测到的 R^2 的 p -值。如果 p -值较小,比方小于 0.05,我们就拒绝零假设。无论哪一种方法,样本均值之间的差异很不可能仅仅用偶然机会来解释。

对于暴力犯罪的数据,计算机输出的分析结果如表 11.4。这类表被称为方差分析表(它与我们在回归分析中得到的表属于同一类;参见表 10.5)。认真注意该表的结构,我们看到如何从这个大量数字的矩阵中计算 F -变量的值。

方差分析表提供了包括用来计算 R 和 F -值的数据。用自由度对自变量和残差变量的效应先做一个调整然后再进行对比就可以得到 F 的值了。通常方差分析表还包括 F 的 p -值。

表 11.4 中的一些数来自表 11.2,而另一些则是新的。其中第四列包含了地区变量和残差变量的自由度。在这个例子中共有 7 个地区,所以地区变量的自由度是 6;自变量的自由度总是比它的类型个数少一个。总的自由度也要比观测的个数少一个,因此 48 个观测的自由度是 47(见公式 11.3)。残差变量自由度是用观测个数减去自变量的类型数而得到;本例中是 $48 - 7 = 41$ 。

表 11.4 方差分析表

来源	平方和	比例	自由度	均方	F -比	p -值
地区	662641	0.24	6	110440	2.11	0.072
残差	2141613	0.76	41	52234		
总计	2804254	1.00	47			

停下来想一想 11.3

有一群老人分别住在 5 个疗养院中,每个疗养院中有 20 人。给他们服用抗抑郁剂。用方差分析来判断老人服药的剂量与所住的病房是否有关。作为自变量的病房及残差变量的自由度分别是多少?

F 的值就是为了比较自变量和残差变量的效应。在计算中我们对比的不是平方和而是平方和除以它们各自的自由度以得到每个自由度上的平方和(见公式 11.4)。这个量称为均方(mean square),之所以这样是因为它代表的是效应在自由度上的平均。表中地区变量的均方为 $662641/6 = 110440$ 。类似地,残差变量的均方为 $2141613/41 = 52234$ 。这两个均方的值仍然很大而且看上去似乎没有什么意义,但它们是简化过程中必要的一步。

如果数据只受到偶然因素的影响而地区变量并没有影响到它们的值,则正式的理论表明这两个均方应大概相等。因此我们就想对比均方看看它们是否相似。当然我们可以用两者相减看差值是否接近于 0;也可以用一个除另一个看比值是否接近于 1。

我们最后选定用自变量的均方除以残差变量的均方。在这个例子中地区变量的均方大约是残差变量均方的两倍。更精确地说,就是这个比值为 $110440/52234 = 2.11$ 。该比值就是自

由度为 6 和 41 的 F 变量的观测值。

F 变量得到大于等于 2.11 这个值的概率是 0.072。也就是说,在两个变量之间没有关系的前提下随机抽样 1000 次,其中有大约 72 次的 F 值因为偶然因素而大于等于 2.11。因此在 7 组 48 个观测的情况下得到 R^2 大于等于 0.24 的概率是边缘统计显著的。这个值虽然没有达到魔术般的显著水平 0.05,但也差不多了。

现在我们来看一看为什么这种分析被称做方差分析。效应是用平方和来计算的,计算方差的分子也是平方和。在第 4 章中为了寻找方差,我们曾用平方和除以 $n - 1$ 这个合适的自由度。类似地,当我们用平方和除以它们各自的自由度得到均方时,我们得到方差。而当我们实际上是用对比方差的办法来对比均值时,我们得到 F 值。

停下来想一想 11.4

在停下来想一想 11.3 的例子中的 F -比值是 3.50,其 p -值是 0.01,是统计显著的。关于各病房中使用抗抑郁剂的情况你能得到什么结论? 关于这一统计结果我们不能说什么? 这个分析中忽略了什么我们应该感兴趣的东西?

表 11.5 暴力犯罪率的均值和七个地区所含的州数

地区	暴力犯罪率的均值	州数
中大西洋	639	3
中西部	375	12
新英格兰	292	6
太平洋岸	636	3
落基山地区	364	6
南方	526	14
西南方	620	4
总计	460	48

超出 F 检验:比较均值

到目前为止,这 7 个地区暴力犯罪率的均值差异的 p -值是 0.072,这并不是一个很有说服力的结果。零假设说均值相等。备选假设则是均值不全相等,它的意思可以是两个均值不等、多个均值互不相等或所有均值都不相等。方差分析中 F 的值可以说明在实际上均值不全都是相等的。但是哪些地区的均值不相等? 在太平洋岸比在中大西洋区更安全吗? 是否除了大湖区之外其它 6 个地方的暴力犯罪率都相同? 如果我们想知道哪些均值互相不同,在方差分析中我们总要问这些问题。

如果分析中只有两个均值,而且有显著的 F 值(或是从关于两个均值差别的 t -检验中得到的 t 值),我们就可以下结论说这两个均值是不同的。我们可以类似地比较所有可能的一对均值来分辨出到底那些不同。对于这里的 7 个均值有种可能的 21 对去比较。如果我们做了 21 次独立的、不同的统计检验,那么即使所有均值都相等,统计理论会说:以 0.05 的显著水

平,我们会有 5% 的时候犯错误,并且发现 5% 的对即使实际上没有区别也是统计显著的。因为 21 次的 5% 大约是 1 次,也就是说,我们期望能遇到 1 次纯粹由于偶然而导致的均值差异过大。在此虽然不是所有的检验都独立,但是仍然会出现我们进行多次统计检验时发生的一些问题。

为了找到是哪一个均值与众不同,我们在表 11.5 中列出了每个地区的均值和包含的州数。(在表 11.3 中也提供了均值。)但是仍然很难说哪一个均值在统计意义上有不同,哪一个没有不同。图 11.4 使地区均值的差异看起来更直观。从图中可以看出中大西洋、太平洋岸和西南方这三个地区均值最高,而中西部和落基山地区的均值居中,新英格兰均值最低。

把其中几个地区的均值进行对照,我们注意到中大西洋和新英格兰这两个地区的均值差异达到 $639 - 292 = 347$ 。这个不同转换成 t -值是 2.14,自由度是 41,它在 $p = 0.02$ 的水平上是统计显著的。同样,如果我们继续用 t -检验去继续两个均值的对比,我们会发现太平洋岸和新英格兰、西南方和新英格兰、南方和新英格兰之间均值的差异都是统计显著的。这些不同都是总体的 F 值至少是边缘显著的反映,因为总体的 p -值是 0.072,它很接近通常的临界值 0.05。

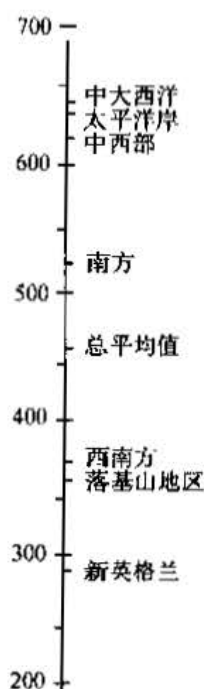


图 11.4 7 个地区的平均暴力犯罪率和 48 个州的总平均值。

11.5 问题 4. 是因果关系吗?

这里有一些统计方法可以帮助我们回答这个问题。这些方法一般是引入其它的变量,并注意这些变量的引入是否可以解释现存的关系。但是目前用这种方法回答第四个问题也只是猜测而已。确实很难想象是地区本身决定了暴力犯罪率。地区之间暴力犯罪率的不同可能被一些其它变量如生活在都市区域人口比例、贫困水平和人口密度等来解释。尽管无法列举出犯罪率不同的原因是什么,我们仍然可以用这个分析结果来预测哪个州的犯罪率高,哪个州的犯罪率低。

11.6 方差分析:鸟瞰回顾

因为方差分析的步骤很多,所以在介绍一个新的方差分析方法之前我们有必要在此对它做一个简单的回顾。在研究分类型自变量和数量型因变量之间关联的过程中的一部分是方差分析。在这里,我们在此研究的是地区和暴力犯罪率两个变量。其它还有诸如宗教信仰与收入的关系、不同教育方法与学生的学习水平的关系等例子。

方差分析是基于计算因变量在按照自变量的各类的均值之间的差异程度和每一类中观测值的差异程度。我们所得到的方差分析的结果是基于各种平方和的大小。表 11.4 是一个典型

的计算机输出的结果(当然,对不同的计算程序,表的外观势必也会发生变化)。

我们可以从方差分析表中得到几个结论。只要分类型自变量的平方和不是 0,数据中涉及的两个变量之间就有关系。这个平方和比另一个平方和大得越多,就表明两个变量之间的关系越强。这个情况也可以从用总平方和除分类变量的平方和的结果中反映出来。这个比值告诉我们因变量的变化在多大程度上能由分类变量所解释,它的值可以小到 0,也可以大到 1。

F -比检验及其 p -值告诉我们因变量在各类中的均值是否有显著差异。如果 F 值大而因此 p -值小,我们就拒绝声称无区别的零假设,并认为在实际中两个变量之间是有关系的。通常当 p -值小于 0.05 时就可以拒绝零假设了。一些新的统计软件和某些便携计算器能够给出精确的 p -值。

有时我们会看到 F^* 和 F^{**} ;在脚注中会解释一个星号表示它的 p 值小于 0.05,而两个星号则表示 p -值小于 0.01。我们可以用统计表找到近似的 p -值。统计表的缺点是它无法提供精确的 p -值;它一般只能给出 p 是小于某些值的。精确的 p -值能够提供更多的信息,因为我们能知道它究竟比 0.05 或比 0.01 小多少,也可以知道在拒绝零假设时的把握有多大。

11.7 配对分析:每个单元两个观测

在本章我们提出的另一个问题是:犯罪率会随时间变化吗?这类例子的数据通常都是起始-结束数据。这个名字暗示它有一个起始测量。例如:先用某种标准测量一个对象的态度,然后让他接受某种刺激比如电影,最后再用同样的标准测量他的态度,这样就得到了一对起始-结束的数据。因为这两次观测来自同一个对象,所以它们又称为配对数据。

t -检验

如果在同一个实验单元上进行重复测量,就会得到在每个单元中有两个观测值的数据。而处理这类数据时所用的分析方法就叫做配对分析。

表 11.6 提供了 48 个大陆州暴力犯罪率的配对数据。它是由已经在表 11.1 中给出的 1986 年的数据和六年后即 1992 年的数据组成的。为了了解从 1986 到 1992 这六年间暴力犯罪率是否有了变化,一个自然的想法是对照这两个时间点的平均得分。如果我们只是简单地对比两个时间点的总体均值:用 1992 年的均值 579 减 1986 年的均值 460,得到的结果是二者相差 119,转换成 t 值为 2.13,自由度是 94,相应的 p -值是 0.04。这对于显著水平是 5% 的双边检验不能算显著,因为它要求的 p -值应该小于 0.025。在这个例子,自变量是时间,残差变量是其它变量的总效应。

其它变量中有一个是州变量。例如,纽约是一个具有两个最高值 986 和 1130 的州;而密西西比则是一个拥有两个最低值 274 和 418 的州。虽然它们的犯罪次数在每 100000 居民中都同样增加了 114 次犯罪。但是二者的值却不相同,一个高一个低,这个不同中包括了残差变量的影响。因为在计算 t 值时,残差变量的效应出现在分母上,所以两个州之间的差异就使 t 值变小了。

为了绕开这个问题,让我们看一看在两个时间点上暴力犯罪率的差别,毕竟,我们的目的

是研究变化。纽约和密西西比都增加了 114, 因此它们的增加值相同。下表中的第 4 列给出了两个时间点上暴力犯罪率的差值。

表 11.6 1986 年和 1992 年 48 个州的暴力犯罪率(次/100000 人)

州名	1986 犯罪率	1992 犯罪率	差值	地区	州名	1986 犯罪率	1992 犯罪率	差值	地区
缅因	147	132	-15	新英格兰	西弗吉尼亚	164	214	50	南方
新罕布什尔	140	126	-14		北卡罗来纳	476	703	227	
佛蒙特	149	111	-38		南卡罗来纳	675	976	301	
马萨诸塞	557	777	220		佐治亚	588	764	176	
罗得岛	336	395	59		佛罗里达	1036	1258	222	
康涅狄格	426	494	68		肯塔基	334	546	212	
纽约	986	1130	144	中大西洋	田纳西	540	769	229	西南方
新泽西	572	630	58		阿拉巴马	558	892	334	
宾夕法尼亚	359	432	75		密西西比	274	418	144	
俄亥俄	423	534	111		阿肯色	395	588	193	
印地安纳	308	519	211		路易斯安纳	758	1000	242	
伊利诺值	800	519	194		俄克拉何马	436	636	200	落基山
密歇根	804	825	21	中西部	德克萨斯	659	838	179	
威斯康辛	258	282	24		亚利桑纳	658	701	43	
明尼苏达	285	346	61		新墨西哥	726	976	241	
依阿华	235	281	46		怀俄明	293	329	36	
内布拉斯加	263	355	92		克罗拉多	524	610	86	
密苏里	578	757	179	南方	蒙大拿	157	175	18	太平洋岸
北达科它	51	89	227		爱达荷	222	298	76	
南达科它	125	199	74		犹他	267	306	39	
堪萨斯	369	520	151		内华达	719	770	51	
特拉华	427	643	216		华盛顿	437	564	127	
马里兰	833	816	-17		俄勒冈	550	534	262	
弗吉尼亚	306	386	80		加利福尼亚	920	1161	241	

来源 F. B. I. Uniform Crime Reports for the United States

为了找出两个时间点上的数据是否有区别,我们求平均。差异的均值等于 119;这说明暴力犯罪率还是有一些变化的。为判断这一变化是否只是由偶然因素引起的,我们建立零假设:差异的总均值是 0。我们可以用第 7 章中提到的检验单个均值的 t -检验方法来检验这个零假设。对于该数据, t 值是 8.74 自由度为 47,与之相应的 p -值小于 0.0001,现在我们有拒绝零假设的一个有利证据。 t 值的计算公式见公式 11.6。

符号检验:只回答是或否

对于暴力犯罪率是否随时间变化的问题,另一种比较简单的解决方法是符号检验。与 t 检验不同,符号检验主要考虑的不是两次观测时数值改变了多少而是数值变化的趋势(增加或减少)。它的逻辑是简单的:如果实际上两次观测中数值没什么变化,则它们的差值就只是在随机变化,没有规律。如果是这样,在我们的例子中差值是正或是负的可能性应该一样,即在

这 48 个差值中差不多有 24 个正数, 24 个负数。可实际上, 我们的差值中只有 4 个负数, 而正数却有 44 个之多。

现在我们回忆第 4 章中提到的二项分布来研究差值的符号, 并找出正负差值的数目。这个问题类似于扔 48 次硬币, 得到 4 次正面和 44 次反面。如果从 1986 年到 1992 年暴力犯罪率只是在随机地发生着变化, 那么差值是正的概率应该是 0.5, 同样是负的概率也是 0.5。这样我们可以用二项分布来检验“差为正的的概率是 0.5”的零假设。因此要先求出正差值大于等于 44 个的概率。

二项分布表中不包括有 48 个观测这么多的情况, 但是我们可以通过计算 z -得分来得到 p -值。这里 $z = 6.21$, 得到 p -值小于 0.0001。因此可以很有把握地拒绝犯罪率没有变化的零假设。

对配对数据做 t -检验得到的 p -值比用二项分布和正态近似做的符号检验的 p -值要小。这没有什么奇怪的。因为 t -检验用的是观测到的实际差值; 它含有的信息量比差值的符号所能提供的要多。此外, t -检验有个附加前提, 这就是原始得分要服从正态分布。数据所提供的可用信息越多, 结果就越显著。

符号检验的一个优点是用起来方便。它要求的计算量比 t -检验要少得多。而且对于小样本有现成的统计表可以直接查二项分布表。如果我们不能确定数据是否服从正态分布或是否应该用配对的 t -检验时, 用符号检验是比较安全的。

11.8 小 结

11.1 方差分析: 对比事物的平均值

当我们研究一个(多个)分类型自变量对一个数量型因变量的影响时, 我们可以利用称为方差分析的统计方法(简记为 anova)。这个统计方法比较因变量在自变量的每一个值上的均值。

11.2 问题 1. 暴力犯罪率与地区的关系

我们先做了一个散点图, 其中用自变量做横轴, 因变量做纵轴。如果把观测按自变量的取值分成几个组, 我们还可以用盒子图来进一步对比各组中因变量的不同。

11.3 问题 2. 关系有多强?

下面, 我们计算因变量在全部数据上的总均值和因变量对自变量每个值(组)内的均值。像在回归分析中一样, 我们假设观测值和总均值之间的差异是由自变量和残差变量共同造成的。

为了找到自变量的总效应, 我们先用每一组的均值减去总均值, 然后我们用这些差的平方分别乘以各自组中的观测个数, 再对所有乘积求和。找残差变量的总效应时, 先用观测减去所在的组的均值, 然后再对这些差值的平方求和。这个和称为残差平方和。为了用一个数来表示观测值与总均值之间的区别, 我们求区域变量的均值和残差变量的均值之差的平方, 然后求

和。这个平方了的差的和称为总平方和。

为了更好地理解这些平方和的意义,我们分别计算这两种效应(自变量的效应和残差变量的效应)在总效应中占的比例。用自变量的平方和除以总平方和就是自变量所占的比例了,这个比例称为 R^2 ,它可以用来测量因变量的变化中有多少是由自变量引起的。它可以直接与回归分析中相关系数的平方进行比较。

R^2 的平方根是 R ,它是用来衡量自变量与因变量之间关系强度的量。 R 的取值范围是从 0 到 1,它的作用等同于在回归分析和相关分析中用于衡量两个数量型变量之间关系强度的相关系数 r 。

方差分析能告诉我们因变量的变化中有百分之几与自变量有关,又有百分之几与残差变量有关。有时我们用解释,产生,造成,引起等词来代替“有关”。

11.4 问题 3. 这个关系是纯属偶然的吗?

我们要检验零假设:在自变量的各组中的因变量的总体均值相等。如果自变量的效应比残差变量的效应显著,我们就拒绝零假设。反之则不拒绝零假设。

在检验中,为了寻找 p -值,要把 R^2 转换成 F -变量的值。自变量的自由度是它的值的个数减 1,残差变量的自由度是观测个数减去自变量的值的个数。自变量的平方和与残差变量的残差平方和分别被它们各自的自由度除,就得到所谓的均方。而 F -比的值正是用自变量的均方除以残差变量的均方得到的。一旦我们得知各组中因变量的均值不相等时,一个自然的问题就是哪些均值互不相同。

11.5 问题 4. 是因果关系吗?

尽管我们无法查明到底是什么因素引起了各地区之间暴力犯罪率的不同,我们还是可以预测哪些州的犯罪率高,哪些州的犯罪率低。

11.6 方差分析:鸟瞰回顾

本节主要是回顾方差分析的全部过程。

11.7 配对分析:每个单元两个观测

对于配对数据,我们用其中一个观测减去与之相配的另一个观测来观察数据是否有变化。为了看统计的显著性,我们利用对一个均值的 t -检验来检验零假设:总体中变化的均值等于零。为了快速检验这个变化,我们能计算正的和负的差有多少,并应用二项分布来看差值中出现正号的概率是否等于出现负号的概率。如果我们的数据不符合正态分布,用符号检验是比較好的。

补充读物

Iversen, Gudmund R., and Helmut Norpoth. *Analysis of Variance*, 2nd ed. (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-001). Newbury Park, CA: Sage, 1987. 简明方差分析入门。

Toothaker, Larry E. *Multiple Comparison Procedures* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-089). Newbury Park, CA: Sage, 1993. 在检验了总的显著性之后, 如何比较某些类的均值。

公 式

方差分析

在计算方差分析时利用手算和台式计算器时, 数据是按表 11.7 的形式排列的——每一组(类)测值占一列, 每一个观测都有两个下标, 一个标出这个观测所在的行, 另一个标出它所在的列。方差分析的计算公式就是针对这样的数据排列形式给出的。另一方面, 当数据以表 11.1 的计算机文件的形式给出, 即每一列是一个变量, 每一行是来自同一个单元不同变量的观测值, 其中一列是因变量, 另一列是用于鉴别该观测属于哪一组的自变量。下面的公式是针对第一种情况给出的, 它的结果形式由表 11.8 给出。这些公式可直接用来处理比较小的数据集; 当数据集比较大时最好用计算机来处理。

表 11.7 方差分析数据的排列形式

	1	2	组	k
	y_{11}	y_{21}	...	y_{k1}
	y_{12}	y_{22}	...	y_{k2}
	y_{13}	y_{23}	...	y_{k3}
	\vdots	\vdots	...	\vdots
	y_{1n_1}	y_{2n_2}	...	y_{kn_k}
均值	\bar{y}_1	\bar{y}_2	...	\bar{y}_k
	总均值 \bar{y}			

方差分析中要用到的平方和的计算公式如下:

$$\begin{aligned} \text{分类型自变量的平方和} &= \sum n_i (\bar{y}_i - \bar{y})^2 \\ \text{残差平方和} &= \sum \sum (\bar{y}_{ij} - \bar{y}_i)^2 \end{aligned} \quad (11.1)$$

$$\begin{aligned} \text{总平方和} &= \sum \sum (\bar{y}_{ij} - \bar{y})^2 \\ R^2 &= \frac{\text{分类型自变量的平方和}}{\text{总平方和}} \end{aligned} \quad (11.2)$$

表 11.8 一个分类变量的方差分析表

来 源	平方和	自由度	均方	F -比	p -值
分类变量	CSS	$k - 1$	CMS	F	p
残差变量	RSS	$n - k$	RMS		
总 计	TSS	$n - 1$			

公式中的 k 是组数(自变量的值的个数), n 是数据中的总观测数。平方和的自由度由下面公式给出:

$$\begin{aligned}\text{分类变量的自由度} &= k - 1 \\ \text{残差的自由度} &= n - k \\ \text{总自由度} &= n - 1\end{aligned}\quad (11.3)$$

用平方和除以它们各自的自由度就得到了均方:

$$\begin{aligned}\text{分类变量的均方} &= \frac{\text{分类变量的平方和}}{k - 1} \\ \text{残差的均方} &= \frac{\text{残差平方和}}{n - k}\end{aligned}\quad (11.4)$$

最后, F -比的计算公式如下:

$$F = \frac{\text{分类变量的均方}}{\text{残差的均方}} \quad (11.5)$$

其中 F -比的自由度是 $k - 1$ 和 $n - k$ 。

这些计算结果常被总结成方差分析表。(如表 11.8; 其中 C 代表分类型自变量, R 代表残差变量。)

配对数据

先从第 i 个观测单元中计算两次观测的差

$$d_i = y_{B_i} - y_{A_i}$$

其中 y_{B_i} 是第一次的观测值(起始), y_{A_i} 是第二次的观测值(结束)。然后求 d_i 的均值 \bar{d} 以及标准差 s , 最后把它们按下面公式转换成 t -变量的值:

$$t = \frac{\bar{d}}{s/\sqrt{n}} \quad n - 1 \text{ d.f.} \quad (11.6)$$

在前面的例子中上式为:

$$t = \frac{118.47}{93.93/\sqrt{48}} = 8.74 \quad 47 \text{ d.f.}$$

习题

回顾(习题 11.1—11.17)

- 11.1 考虑课文中暴力犯罪率的那个主要例子:
- 我们如何计算某一个州的暴力犯罪率?
 - 为什么我们在分析过程中宁愿用暴力犯罪率而不是暴力犯罪数?
 - 为什么在研究暴力犯罪时不直接研究各州本身而是要把这些州归到几个地区来研究?
- 11.2
- 从表 11.1 中找出暴力犯罪率最高的 5 个州。
 - 哪 5 个州的暴力犯罪率最低?
 - 从表 11.1 和课文中其它的图表中能否看出哪些地区暴力犯罪和地区相比较重或较轻?
- 11.3 假设我们已经知道了各个州女性的暴力犯罪率和男性的暴力犯罪率。现在我们想研究暴力犯罪率在各州的差别。
- 什么是自变量,什么是因变量?
 - 我们要比较什么类型的变量(分类变量、顺序变量还是数量变量)?
- 11.4 方差分析,简称 anova,是用于研究哪种数据的统计方法?
- 11.5 请举出一个你感兴趣的用方差分析研究两个变量关系的例子。
- 11.6 在暴力犯罪率问题的方差分析中,哪两个因素完全决定各个州的暴力犯罪率?
- 11.7 在研究暴力犯罪率时,如果地区变量和残差变量都对该犯罪率无影响,那么各个州的暴力犯罪率将会出现什么情况?
- 11.8 假设我们先用各个州的暴力犯罪率减去暴力犯罪率的总均值,再对所得差值的平方求和。
- 这个和的名字是什么?
 - 我们如何计算某个州的残差变量的值?
 - 我们把所有残差项的平方的和叫什么?
- 11.9 “从样本中得到的分类型变量和数量型变量之间的关系在产生样本的总体中是否依然成立”是我们问的第几个问题?
- 11.10
- 在方差分析中应该是在 R^2 较大时拒绝零假设呢?还是在它较小时拒绝零假设?
 - 除了 R^2 的实际值之外,还有什么能决定是否应该拒绝零假设?
 - 在方差分析中,为了寻找数据的 p -值需要把 R^2 转换成哪种统计变量?
- 11.11
- 如果自变量有 6 个值,这就意味着我们应该把数据分成 6 组。此时自变量的自由度是几?
 - 如果样本中共有分成 6 组的 50 个观测,那么残差变量的自由度是多少?
- 11.12
- 什么叫一个变量的均方?
 - 如何计算均方?
 - 我们如何将自变量的均方与残差变量的均方进行比较?
 - 该均方比较的结果称做什么?
- 11.13
- 请给出一个配对数据的例子。

- b. 为什么研究者对这类数据感兴趣?
- 11.14 符号检验是干什么用的?
- 11.15 a. 为什么残差变量被戏称为“不感兴趣的变量”?
- b. 为什么把残差变量叫做误差变量使人感到起错了名³⁸?
- 11.16 a. 和符号检验相比,为什么人们用 t -检验更多?
- b. 符号检验的优点是什么?
- 11.17 如何把配对数据的分析也看作一个回归分析,而在该回归分析中我们研究是否回归直线是一条截距为 0、斜率为 1 的 45 度线?

理解(习题 11.18—11.28)

- 11.18 对 1994 年 48 个大陆州的人均收入的数据进行方差分析;把这 48 个州分成 8 个地区;结果如表 11.9:

表 11.9 习题 11.18 的数据

来源	平方和	自由度	均方	F -比	p -值
地区	195.0	7	27.86	5.33	0.0002
残差	209.1	40	5.23		
总计	404.1	47			

来源: *Bureau of Census, Statistical Abstract of the United States: 1995 (115th edition), Washington, DC: U. S. Government Printing Office, 1995*

- a. 在分析中哪一个是自变量, 哪一个是因变量?
- b. 从这个方差分析表中你能得出什么结论?
- c. 关于地区与收入有哪些你感兴趣的问题是方差分析无法回答的?
- 11.19 在奥克兰成长研究会中, 研究人员用一种衡量好体格的变量来评估中学女生。在一个含有 35 个中产阶级家庭女生的样本中该变量的均值是 56.6, 方差是 13.5。在另一个包含 43 位工人家庭女生的样本中这两个值分别是 48.6 和 14.2。(来源: *G. H. Elder, Jr., "Appearance and education in marriage mobility," American Sociological Review, vol. 34 (1969), p. 524.*)
- a. 这两组女孩体格的差异有多大?
- b. 这个差异转换成 F 值是 6.40, 自由度是 1 和 76, 相应的 p -值为 0.013。这两个均值之间是否有显著的不同?
- 11.20 巧克力和香草冷冻小吃的价格有区别吗? “消费者报告”(Consumer Reports)收集的数据是平均每份巧克力小吃的价格是 29.4 美分, 而香草的是 30.4 美分。为了了解这两个均值的差异是否显著, 对它们的差异做 t 检验得到 $t = 0.18$, 自由度是 42。(来源: *"Lowfat frozen desserts: Better for you than ice cream?" Consumer Reports, vol. 57, no. 8 (August 1992), pp. 483-487.*)
- a. 请用数学方法证明在分类型变量只有两个值(巧克力和香草)的情况下 $t^2 = F$ 。
- b. 这两种小吃价格的差异显著吗?
- 11.21 在表 11.10 的基础上回答下列有关地区变量和残差变量对暴力犯罪率影响的问题。

表 11.10 习题 11.21 的数据

变量	效应	比例
自变量(地区)	自变量平方和	R^2
残差变量(其它)	残差平方和	$1.00 - R^2$
总计	总平方和	1.00

- 什么是自变量?
- 残差变量都由什么组成?
- 为什么因变量中由残差变量引起的那一部分变化在总变化中占的比例等于 1.00 减去 R^2 ?
- 如果地区变量没有影响,那么残差变量的平方和等于几?
- 如果残差变量没有影响, R^2 的值应该等于几?
- 如果地区变量和残差变量都对暴力犯罪率没有影响,那么暴力犯罪率的值应该是多少?

11.22 如果 F -变量的值超过了统计表中给定的临界值 F 或 F 的 p -值小于 0.05 ,你认为分类型自变量和数量型因变量之间的关系是什么样的?

11.23 从 Minneapolis 的 10 所中学随机抽取高年级学生,分别计算他们的 GPA 得分。这个数据的方差分析给出的结果列在表 11.11 中。在不做进一步计算的情况下,根据表中提供的数据你认为各中学之间的学习情况有差别吗?

表 11.11 习题 11.23 的数据

来源	平方和	自由度	均方	F -比	p -值
中学	1450	9			
残差变量	9000	91			
总计	10450	100			

11.24 一胎小猪中公猪和母猪的个数是喂养人所关心的。表 11.12 分别给出了 6 胎小猪中公猪和母猪的数目。平均一胎中公、母数目差异的 t 值为 -0.29 ,自由度是 10 , $p = 0.39$ 。

表 11.12 习题 11.24 的数据

第几胎	母猪数	公猪数
1	4	5
2	6	4
3	5	5
4	3	6
5	4	4
6	6	5

来源: S. M. Free, Jr., "Response: The consultant's forum," *Biometrics*, vol. 33(1977), no. 3, p. 561.

- 这个研究中所做的零假设是什么?
- 你能就此下结论说在产生此样本的总体中公猪的平均数和母猪的平均数有差异吗?
- 这个分析中没有包括哪一方面的数据?

11.25 "美国新闻和世界报道"杂志(U.S. News and World Report)要对各高校 12 个领域培养研究生的系做年度评审。他们选取了一些专家分别为这些系用 5 分制进行评分。每个系的得分是这些专家所打的分的平均。例如:按照这个程序斯坦福(Stanford)大学

生物系的得分是 4.9。有 6 所大学出现在所有被评价的 12 个系的名单上。6 所大学对其所有 12 个系评分的均值如下: 哥伦比亚 (Columbia) 4.13, 哈佛 (Harvard) 4.55, 普林斯顿 (Princeton) 4.44, 斯坦福 (Stanford) 4.78, 加州大学伯克莱分校 (University of California at Berkeley) 4.79 以及威斯康辛大学麦迪逊分校 (University of Wisconsin at Madison) 4.24。为了检测各高校之间的得分是否有显著不同, 我们做了一个单因子方差分析, 由表 11.13 给出了这一结果。

表 11.13 习题 11.23 的数据

来源	自由度	平方和	比例	均方	F-比	p-值
大学	5	4.40	0.40	0.88	8.85	0.000002
残差变量	66	6.57	0.60	0.10		
总计	71	10.97	1.00			

a. 从表中看来高校之间的得分有差异吗?

b. 要估计各高校之间得分的差异到底有多大, 你还需要做什么别的分析?

- 11.26 公路损失数据研究所 (The Highway Loss Data Institute) 收集了各种品牌汽车的单位保险赔偿金额。这里的数据是 1991—1993 年的, 它们是在把所有品牌投保金额的平均数定在 100 的基础上计算出来的得分。用这种方法得到的数的值从 Chevrolet Suburban 牌的 44 一直到 Hyundai Elantra 牌的 201。Suburban 牌汽车的得分比所有品牌的平均得分要少一半还多, 而 Elantra 牌的汽车的得分比总平均多了一倍。小型、中型、大型三种汽车的保险赔偿金额是不是确有不同? 在随机抽样中, 5 辆小型汽车的平均赔偿金是 155, 12 辆中型汽车的这个平均值为 95, 而 5 辆大型汽车的平均值则是 60。用汽车型号作为自变量, 赔偿金额作为因变量的方差分析结果见表 11.14。

表 11.14 习题 11.26 的数据

来源	自由度	平方和	比例	均方	F-比	p-值
型号	2	23110	0.75	11555	28.20	< 0.0001
残差变量	19	7786	0.25	410		
总计	21	30897	1.00			

来源: The Highway Loss Data Institute, as reported in Motor Trend, vol 47, no 1 (January 1995), p. 77.

a. 以型号为横轴, 赔偿金额为纵轴做图, 并在图上标出这三个均值。

b. 方差分析的结果使你如何看待汽车型号和赔偿金额之间的关系?

c. 什么是这些结果的可测解释?

- 11.27 在一项涉及范围广泛的问题的对比调查中, 样本来自几个欧洲国家。对样本中的人所提的问题之一是“在 1985 年一年中你共有几次持续时间在 4 天或以上的假日旅行?” (来源: Jacques - Rene Rabier, Helen Riffault, and Ronald Inglehart, Euro - barometer 25: Holiday Travel and Environmental Problems, April 1986. ICPSR ed. An Arbor, MI.: Inter - University Consortium for Political and Social Research, 1988. Codebook p. 20.) 其中丹麦人平均有 1.06 次, 法国人均有 1.11 次, 爱尔兰人均是 0.81 次, 而葡萄牙人则平均有 0.41 次。用方差分析来检验这些均值之间的差异是否显著, 得到 $F = 85.77$, 自由度是 3 和 4019, $p < 0.0001$ 。

a. 从这个结果看来, 各个国家人的假日旅行次数相同吗?

- b. 为了更好地理解各国人旅行次数为什么不同,你还需要知道什么其它信息?
- 11.28 在习题 11.27 的关于欧洲城市的研究中,被访者被提问及是否他们对所过生活的满意程度。记分标准为:非常满意是 1,相当满意是 2,不太满意是 3,很不满意是 4。结果发现丹麦人的平均满意程度是 1.41,法国人的是 2.16,荷兰人是 1.65,西德人是 1.88,而总均值是 1.77。对这个数据进行方差分析得到 $R^2 = 0.16$, $F = 250$ 并具有自由度 3 和 3995。通过这个结果你认为这 4 个国家的人对生活现况的满意程度有区别吗?

分析(习题 11.29—11.46)

- 11.29 史前的金豺(golden jackal 或 *Canis aureus*)雌性与雄性的下鄂一样长? 下面是从大英博物馆得到的这种豺的数据(单位:毫米)。

雌性	110	111	107	106	110	105	107	106	111	111
雄性	120	107	110	116	114	111	113	117	114	112

来源: C. F. Higham, A. Kyngum, and B. F. J. Manly, "An analysis of prehistoric canid remains from Thailand," *Journal of Archaeological Science*, vol. 7(1980), pp. 149 - 165.

- a. 求出这两组的平均长度。
- b. 对两个均值的差异做一个 t 检验或方差分析,并判断均值之间的差异是否统计显著。
- 11.30 请到 Springer 的网站(<http://www.springer-ny.com/supplement/iversen>)查找与本书相关的文件。其中有一个数据文件“Singers”包括了纽约合唱团中的女高音、女低音、男高音和男低音歌唱家身高的数据。(来源: J. M. Chamber et al., *Graphical Methods for Data Analysis* Boston: Duxbury, 1983 p. 350)
- a. 这四种歌唱家的身高是不是确有不同?
- b. 这其中的不同是否只是一个偶然?
- c. 是否嗓音类型影响身高? 这其中有没有别的变量的影响在内?
- 11.31 表 11.15 列出了受过训练的品尝家用 0 到 100 的评分对几种香草冷冻小吃的口味评比结果。

表 11.15 习题 11.31 的数据

冷冻酸奶	冰奶	冷冻小吃
87	83	33
74	76	31
70	76	31
68	70	31
68	58	27
67	52	10
64	50	
64	47	
63		
57		
54		
50		
48		

来源: "Low-fat frozen desserts: Better for you than ice cream?" *Consumer Reports*, vol. 57, no. 8 (August 1992), pp. 483 - 487. F

- a. 用此数据做一个散点图,你能从这个图上看出来什么?
- b. 计算每一类点心的平均得分,看一看这三种点心的口味有无区别?
- c. 用方差分析来找出小吃类型和口味之间的关系强度,并判断三种小吃得分的均值是否有显著差异?

11.32 冷战期间,有关的国家把大量的资金投入国防建设当中。不同的国家分配资金的方式也不同;每一年的分配情况是一些复杂的变量的反映。我们在此只想利用配对数据方法对比冷战期间几个国家和地区在教育 and 国防上的资金分配情况(见表 11.16)。

表 11.16 习题 11.32 的数据(1969 年的数据)

国家和地区	教育经费占国民收入的百分比	国防经费占 GNP 的百分比
澳大利亚	4.0	3.6
台湾	3.8	8.8
匈牙利	4.4	3.5
韩国	3.8	4.0
挪威	6.3	2.9
苏丹	4.9	6.0
美国	6.3	7.8
南斯拉夫	5.1	5.4

来源: Bureau of the Census, Statistical Abstract of the United States; 1972, 93rd ed. Washington, DC: U. S. Government Printing Office, 1972, pp. 809, 831

- a. 这两种经费所占百分比的平均值的差异是多大?
- b. 把这个差异的平均转换成 t 值。(对于这个数据,无论是否用配对分析的方法结果都差不多。)
- c. 如果这些国家和地区只是从众多的国家和地区中抽出的随机样本,你能从这个 t 值上得出什么结论?

11.33 北方城市的种族歧视严重呢,还是南方城市的种族歧视严重?为了回答这个问题,选取了 10 个北方城市。而且对每一个北方城市选取一个在种族结构、平均收入和工业的数量和种类上与之差不多的南方城市进行配对。这样在每一对城市中这些变量可能产生的影响就可以被抵消了。然后对每个城市的种族歧视状况用百分制进行评分。种族歧视是基于一些指数,比如:自由居住立法数,种族混合学校的数目,不同种族收入的差别,不同种族失业情况的差别等等。得分越高种族歧视就越严重。表 11.17 给出了种族得分。(对于这个数据,用正确的配对分析方法和用不正确的对比南北方均值的方法所得到的结果差异很大。)

表 11.17 习题 11.33 的数据

	配对的城市									
	1	2	3	4	5	6	7	8	9	10
南方	72	52	59	36	67	25	80	41	62	55
北方	79	68	45	45	59	38	75	56	72	60

- a. 这些配对的城市在这个变量上有什么差别吗?(一个进行对比的方法是对这个数据做散点图,用南方/北方做自变量,城市的得分做因变量。)
- b. 关于这个国家的两部分城市的差别,差异的均值告诉你什么?

c. 这个差的平均与 0 相比是否显著的不同?

- 11.34 关于 60 年代美国人口从东、北方向西、南方迁移的情况已经不是一个新话题了。表 11.18 给出了那个时期从具有 20 万或以上人口的标准都市统计区域(SMSA)中通过随机样本估计出来的人口变化所占的百分比。这些 SMSA 是按照人口普查区域划分的。这组数据的总平方和是 12738, 四组观察值中的残差平方和为 8822。

表 11.18 习题 11.34 的数据

地区			
西部	中北部	东北部	南方
20.4	9.2	6.1	36.7
32.1	21.4	5.2	24.0
89.0	13.6	10.8	19.4
41.2	22.4	-0.2	85.7
39.0	15.2	7.6	40.0
3.3	12.8	12.6	16.1
32.4		4.7	8.8
			18.9
			14.4
			30.1

来源: *Bureau of the Census, Statistical Abstract of the United States; 1972, 93rd ed. Washington, DC: U. S. Government Printing Office, 1972, pp. 838 - 878*

- a. 这些数据做一个散点图, 其中要包括四组的均值。
- b. 从图中看来, 这四个组所含的城市的生长情况相同吗?
- c. 请给出该数据的 F -值。
- d. 基于 F 的值来判断各地区的城市增长百分比的均值之间的差异是否显著?
- 11.35 研究者们报道了一个含有 22 名体操运动员和 21 名游泳运动员的样本的研究。基于几个观测变量来计算每个运动员成年身高预测。结果预测体操运动员的平均高度是 5.48cm/年, 游泳运动员的平均高度是 8.00cm/年。二者的标准差分别为 0.32cm/年和 0.50cm/年。分析表明, 体操运动员的平均值要显著低于 ($p < 0.05$) 游泳运动员的平均值。(来源: *G. E. Theintz, H. Howard, U. Weiss, and P. C. Szononko, "Evidence for a reduction of growth potential in adolescent female gymnasts," The Journal of Pediatrics, vol. 122 (1993), no. 2, pp. 306 - 313*)
- a. 在该研究中什么是自变量?
- b. 在该研究中什么是因变量?
- c. 请给出这两个均值差异的精确 p -值。
- d. 为什么这个精确的 p -值所能提供的信息比作者所用的 $p < 0.05$ 要多?
- 11.36 下面的数据是在几个欧洲国家中, 合同中规定的带工资的每年休假天数。澳大利亚 25, 比利时 25, 丹麦 25, 西班牙 32, 芬兰 30, 法国 30, 大不列颠 20, 冰岛 24, 爱尔兰 18, 意大利 25, 挪威 21, 荷兰 25, 葡萄牙 30, 瑞典 40, 瑞士 23。(来源: *Juliet B. Scor, The Overworked American: The Unexpected Decline of Leisure, New York: Basic Books, 1991, p. 82.*)
- a. 对这个数据做一个盒子图。
- b. 请猜测、估计或查找以下北美国家雇员的带工资休假天数: 美国、加拿大、墨西哥、

古巴、海地。类似于欧洲,用这些数据做一个盒子图。

c. 根据两组国家的盒子图你可以从哪些方面进行对比?

11.37 表 11.19 是一组气候带与杀人率数据的方差分析表。

表 11.19 习题 11.37 的数据

来源	自由度	效应	比例	均方	F-比	p-值
气候带	7	88866				0.002
残差变量	40					
总计		218031				

a. 请把这个表的空白处填满。

b. 气候带对因变量有影响吗? 解释你的回答。

c. 要找出到底是哪些气候带之间有区别,哪些气候带之间无区别,你还需要做些什么?

11.38 表 11.20 是分别用标准方法和新方法治疗头痛时的止疼时间(单位: min)。我们希望知道这两种治疗方法的效果是否不同? 如果对这两种方法的均值做普通的 t 检验, 结果是不显著的。($t = 1.15$, 自由度 = 18, $p = 0.13$) 这个 t 值是用两组标准差的平均值计算的。一些病人的极端反映严重影响了这个标准差, 尤其是第 3 号病人和第 10 号病人, 他们的得分差别太大了。解决这个问题一个办法是——利用数据是配对的这一特征, 计算两种方法止疼时间上的差值。

表 11.20 习题 11.38 的数据

病人	标准治疗方法	新方法
1	8.4	6.9
2	7.7	6.8
3	10.1	10.3
4	9.6	9.4
5	9.3	8.0
6	9.1	8.8
7	9.0	6.1
8	7.7	7.4
9	8.1	8.0
10	5.3	5.1
均值	8.43	7.86

来源: A. J. Gross and V. A. Clark, Survival Distributions: Reliability Application in the Biomedical Sciences, New York: John Wiley & Sons, 1975, p. 232

a. 求出每一个人用两种方法治疗后止疼时间的差别, 再对这 10 个差别求均值。

b. 差别的均值与两个均值的差别如何进行比较?

c. 在检验总体平均差异等于零的零假设时 t 的值是多少?

d. 为什么对于这组数据, t 的自由度只有 9?

e. 这个新的 t 值的 p -值是多少? 这两种治疗方法的效果相同吗?

11.39 a. 请用符号检验的方法分析习题 11.38 的数据。

b. t -检验与符号检验有何区别?

- 11.40 a. 请把习题 11.23 的表填满。
b. 现在你认为两个变量之间的关系如何?
- 11.41 表 11.21 汇总了儿童看护花费的数据。请用方差分析的方法来判断不同看护方式之间的收费是否也不相同。

表 11.21 习题 11.41 的数据

婴儿	看护	每小时花费(美金)
1	亲属	4.90
2	保姆	7.00
3	亲属	5.00
4	日托	6.60
5	私人家庭	5.35
6	保姆	7.50
7	私人家庭	5.50
8	日托	6.75
9	亲属	5.25
10	私人家庭	5.15
11	保姆	7.55
12	日托	6.67
13	亲属	5.10
14	私人家庭	5.35
15	保姆	7.40
16	日托	6.75

来源: Sandra L. Hofferth, *Urban Institute*.

- 11.42 下面是一个由 Charles Darwin(查尔斯·达尔文)收集的著名的数据。它是 15 对植物高度的数据(单位:英尺)。每一对是由一个异花受粉的植株和一个自花受粉的植株组成。Darwin 想知道这两组植物的高度是否有所不同?

数据对	1	2	3	4	5	6	7	8
异花受粉	23.5	12.0	21.0	22.0	19.1	21.5	22.1	20.4
自花受粉	17.4	20.4	20.0	20.0	18.4	18.6	18.6	15.3

数据对	9	10	11	12	13	14	15
异花受粉	18.3	21.6	23.3	21.0	22.1	23.0	12.0
自花受粉	16.5	18.0	16.3	18.0	12.8	15.5	18.0

来源: Charles Darwin, *The Effect of Cross- and Self-fertilization in the Vegetable Kingdom*, 2d ed., London: John Murray, 1876, p. 451

- a. 求出每一对数据中植物高度的差, 并检验零假设: 差异均值等于零。
b. 把这些数据看成两组独立的观测, 并用普通的 t 检验来判断这两组之间是否确有差异。
c. 比较这两种数据分析方法。

11.43 用两个心理变量来对几个社会打分：“口头社交焦虑”的程度和是否存在“口头解释疾病”。结果不存在口头解释的社会有如下焦虑程度得分：

6 7 7 7 7 8 8 9 10 10 10 10 12 12 13

$(\bar{y} = 8.9, s = 2.14)$

存在口头解释的社会有如下焦虑程度得分：

6 8 8 10 10 10 11 11 12 12 12 12 13 13 13 14 14 14 15 15 15 16 17

$(\bar{y} = 12.2, s = 2.73)$

(来源: J. W. M. Whiting and I. L. Child, Child Training and Personality, New Haven: Yale University Press, 1953, p. 156)

把观测当做数量型数据来分析数据。

11.44 软木塞是用软木树的树皮制成的。软木在树的南边和北边的沉积量是否相同？下面的数据是 28 棵软木树的软木沉积量(单位:克)。在树的两边软木的沉积量有差别吗？

树	1	2	3	4	5	6	7	8	9	10	11	12	13	14
南边	72	60	56	41	32	30	39	42	37	33	32	63	54	47
北边	76	66	64	36	35	34	31	31	31	27	34	74	60	52

树	15	16	17	18	19	20	21	22	23	24	25	26	27	28
南边	91	56	79	81	78	46	39	32	60	35	39	50	43	48
北边	99	47	70	68	67	37	34	30	67	48	39	37	39	57

来源: C. R. Rao, "Tests of significance in multivariate analysis," Biometrika, vol. 35(1948), pp. 58 - 79.

11.45 初中和教高中的老师花在教室的时间相同吗？表 11.22 给出了各国的一个样本中教师每年授课的时数。

表 11.22 习题 11.45 的数据

国家	初中	高中
德国	761	673
爱尔兰	792	792
意大利	612	612
挪威	666	627
西班牙	900	630
瑞典	576	528
美国	1042	1019
均值	764	697

来源: OECD, from The New York Times, May 28, 1995, p. E7

- a. 分析这些数据。
- b. 通过把这些考虑为配对数据, 我们是否受益？

11.46 下面的数据是习题 10.73 中提到的新郎和新娘申请结婚证时的年龄,它的形式是(新郎的年龄,新娘的年龄):

(37,30) (30,27) (65,56) (45,40) (32,30) (28,26) (45,31) (29,24) (26,23) (28,25)
(42,29) (36,33) (32,29) (24,22) (32,33) (21,29) (37,46) (28,25) (33,34) (17,19)
(21,23) (24,23) (49,44) (28,29) (30,30) (24,25) (22,23) (68,60) (25,25) (32,27)
(42,37) (24,24) (24,22) (28,27) (36,31) (23,24) (30,26)

来源: The Philadelphia Inquirer, September 10, 1995, p. MD-12d

- a. 把这些数据看成配对数据,并用适当的 t 检验来判断新郎新娘的年龄是否有显著的不同。
- b. 再用符号检验的方法来判断是否年龄有显著的不同。

C H A P T E R 12



12.1 用词作为值的两个顺序变量

12.2 把数目的排序作为值:Phillies 表现如何?

12.3 小 结

两个顺序变量

的秩方法

12

什么使得某些人对政治选举更感兴趣呢？也许与他们感觉与某一个政党的关系有多密切有关。

棒球队的名次怎样随时间变化的呢？好的队伍是否每个赛季都很好呢？类似于这样的问题及其他许多问题可以用顺序变量来回答。

至此提到的许多变量都是分类或数量变量，而你也已习惯于区分它们。你是否曾经想过如何度量这样的变量，比如班级的排名，赛跑中的运动员，或者对某件事的态度？这些变量既不是分类变量，也不是数量变量，而是顺序变量(rank variable)。

正如它的名字所暗示的，顺序变量是以数量(多还是少)术语来比较个体的某些特征的变量。例如，选民可能对选举的结果非常感兴趣、有点兴趣或者不怎么感兴趣。社会地位可以作为一个顺序变量，人们可以被分为上层阶级、中产阶级或下层阶级。研究人们态度的社会心理学家利用顺序变量来评价人们的观点是否强烈。态度变量可以以一个取值的刻度进行排序，包括强烈反对、反对、中立、同意、强烈赞成。不仅人们的态度可以被认为是不同，而且根据顺序变量的值还可以认为某些个体或群组比其他的个体或群组的態度更强烈。对于这些顺序变量，我们用词(words)来表示变量的值。

在整个赛季，我们在体育版而上寻找我们喜爱的棒球队的排名。这个队可能排第一、第二或者比较靠后。在 Kentucky Derby 一年一度的赛马会上，根据赛马穿过终点线时的先后顺序排名。我们知道哪匹马获胜，哪匹第

二、第三,直到最后一匹。类似地,在选拔运动员时,职业运动队对运动员进行排名,首选最好的运动员。另外一个关于顺序变量的比较特殊的例子是中国 18 世纪的地方官员对每年的收成从 1 到 10 进行排序。10 是最好的。对于这些顺序变量,我们用数(number)来代表变量的值。

有时,我们特意地把数量数据转化为顺序变量数据。比如两个数量变量初始的散点图表明一个非线性的趋势,从而用一条直线来拟合这些数据是没有意义的。但是如果我们把这两个变量的数据转化为顺序变量,也许就会出现线性趋势。然后,我们就可以用本章所述的秩方法来分析直线数据了。

关于一个数量变量和一个顺序变量的数据也可以用这种方法来分析。如果我们想要比较 Kentucky Derby 赛马的最后排名和它们的经济价值,我们可以通过把经济价值转化为顺序变量,然后看看这两个变量有多强的相关性。

对于顺序变量的两个观测,我们可以判断它们相同还是不同。另外,我们还可以判断一个观测是否比另一个观测多(或少)。

顺序变量要比分类变量复杂。对于顺序变量,我们对变量的值进行排序。对于顺序变量的观测,不仅仅是一个观测和另一个观测不同,而且一个观测比另一个观测是多还是少。但是注意,对顺序变量我们并不知道一个观测比另一个观测多多少或者少多少。在一次赛马中,我们并不知道第二匹马是紧跟在第一匹马后边还是相差几个马长。对于态度变量,我们不知道持同意观点的人和持中立态度的人的差异是否和持中立态度的人与反对态度的人的差异相同。

顺序变量有时称为次序变量(ordinal variable),因为要对值排次序。顺序变量并不如分类变量和数量变量那么常用。

停下来想一想 12.1

举一个顺序变量的例子并列这个变量的取值。你举的变量为什么是顺序变量?

12.1 用词作为值的两个顺序变量

为什么一些人对政治选举感兴趣,而另一些人则不感兴趣?在一份关于美国政治的著名报告中,作者想知道这种兴趣是否与人们与某一个主要的政党的密切程度有关。表 12.1 是 1956 年总统选举时美国人民在这两个变量上的分布。这次选举是民主党的 Adlai Stevenson (史蒂文森)与在任的共和党领袖 Dwight Eisenhower (艾森豪威尔)竞选。Stevenson 曾经是伊利诺依州的州长,并在 4 年前选举中失败。Eisenhower 已做了 4 年的总统,他是二战时著名的军事将领。朝鲜战争后,国家处于和平时期,披头士(Beatles)也即将问世。人们对即将进行的选举有什么看法呢?

表 12.1 政党身份和对 1956 年总统选举的兴趣

		政党身份			总计
		中立	弱	强	
兴 趣	非常感兴趣	104	150	262	516
	有点感兴趣	178	273	237	688
	不太感兴趣	133	228	125	486
总 计		415	651	624	1690

来源: Angus Campbell, Philip E. Converse, Warren E. Miller, and Donald E. Stokes, *The American Voter: An Abridgement*, New York: John Wiley & Sons, 1964, p. 84.

在抽样调查中,人们被问到他们是强烈支持民主党或者共和党,还是比较弱地支持民主党或者共和党,还是中立的,以及他们对 1956 年的总统选举感兴趣的程度(非常感兴趣、有点感兴趣或不太感兴趣)。在这个表中,政党身份是自变量(x -变量),置于横轴,而兴趣变量是因变量(y -变量),置于纵轴。把强烈支持民主党或者共和党的人均归于具有强烈的政党倾向,而比较弱地支持民主党或者共和党的人归于具有弱的政党倾向。

问题 1. 身份和兴趣间的关系?

从右向左粗略的看一下这张表,我们发现从强的政党倾向到弱的政党倾向,他们对选举感兴趣的程度也在下降。这表明对于这些数据,政党倾向和对选举的兴趣这两个变量间有一定的关系。

这些数据当然可以用不同的图来描述。图 12.1 是这些数据的四种不同的表示方法。图 12.1a 是一个有相同宽度和不同高度的条形图,它说明中立的人们的数目比支持某一政党的人们的数目少。它还说明较弱的政党倾向的人中对选举不太感兴趣的人比另外两组的多。图 12.1b 是一个有不同宽度和相同高度的条形图,它说明中立的人比另两类的人少,而且强政党倾向的人在不太感兴趣的人中的百分比比较小。图 12.1c 是一系列圆,每个圆的面积相应于表中相同格子中的观测数。在这个图中,很难看出总数,但图的大小很清楚地说明了在哪些组中人比较多,哪些组中人较少。最后,图 12.1d 是 9 个棱柱表示的频率,棱柱的不同高度相应于观测的个数。在这个图上很难看出总数,但要比较行或列之中的频率却很简单。

表 12.1 也可以作一下修改,变成数据的百分比如表 12.2 所示。表 12.2 中三列百分比和图 12.1 都说明这些数据中两个变量间存在一定的关系。

表 12.2 不同政党身份的人的兴趣的百分比

		政党身份		
		中立	弱	强
兴 趣	非常感兴趣	25%	23%	42%
	有点感兴趣	43	42	38
	不怎么感兴趣	32	35	20
总 计		100%	100%	100%

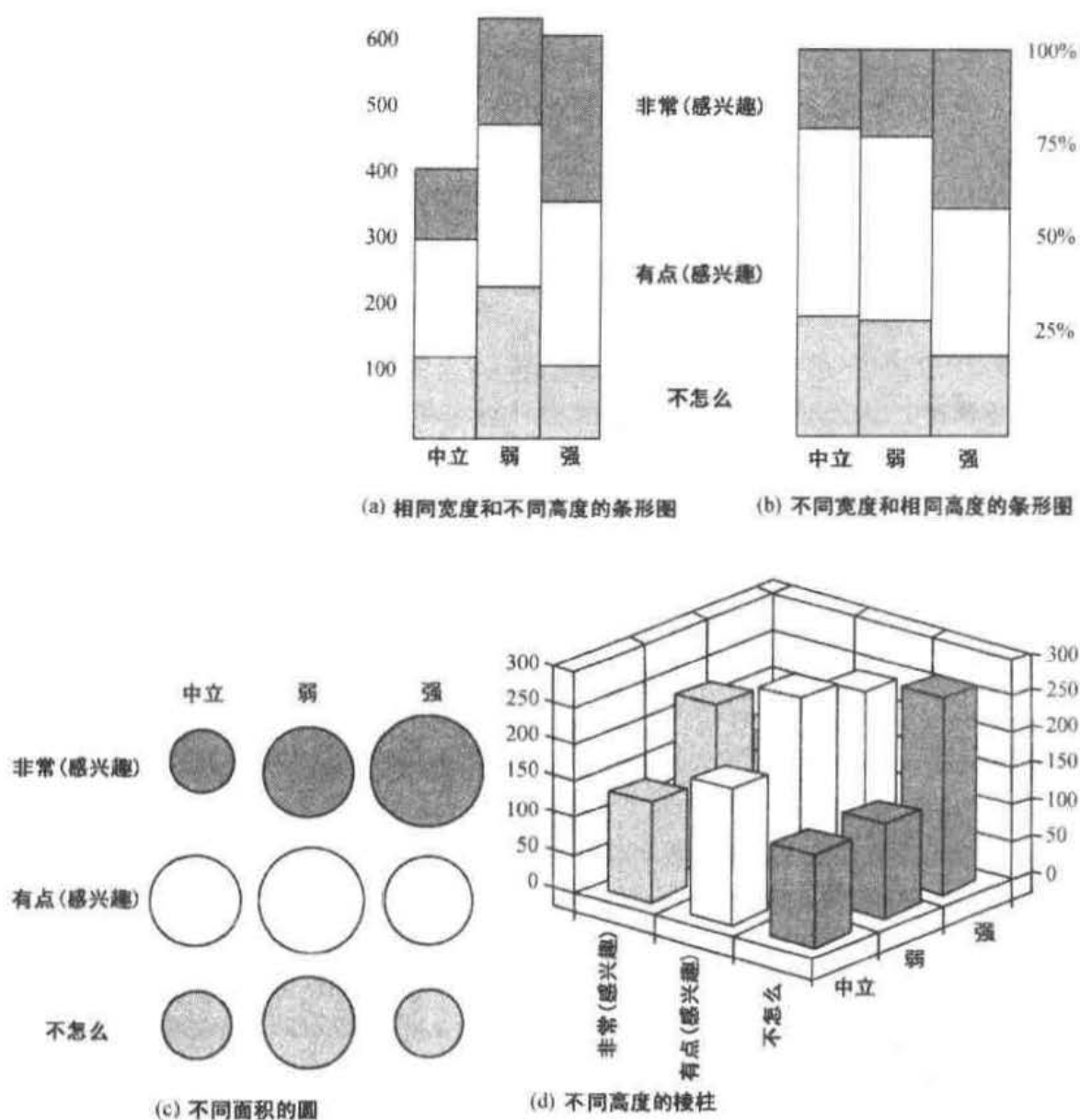


图 12.1 表 12.1 中数据的四种图。

停下来想一想 12.2

在一项关于性别角色(sex-role identity)和职业成就的关系的研究中,对 161 个被雇佣的亚裔美国妇女进行了一项心理性别角色测验。男性化得分(高、中、低)与职业成就(高、低)进行比较。在男性化得分比较高的 45 个妇女中,有 81% 的职业成就得分比较高,19% 的得分比较低。在男性化得分居中的 45 个妇女中相应的百分比为 71 和 29。在男性化得分比较低的妇女中,相应的百分比为 51 和 49。(来源: Esther Ngan-Ling Chow, "The influence of sex-role identity and occupational attainment on the psychological well-being of Asian American women," *Psychology of Women Quarterly*, vol. 11 (1987), pp. 69-81.)

你怎样说明这两个顺序变量间的相关关系? 你怎样用比较通俗的话来总结两个变量间的关系?

问题 2. 相关的程度?

系数 γ 度量了两个取值为词的顺序变量的相关程度。

常用来度量两个顺序变量的相关程度的是一个称作 γ 的系数。对于选举兴趣的数据, $\gamma = 0.21$ 。像其他表示相关程度的系数一样, γ 的值落在 -1 和 1 之间。 $\gamma = 0.21$ 表明对于这些数据, 两变量间有一个比较弱的正的相关性。

和其他表示相关程度的系数一样, γ 的定义与预测有关。对于顺序变量的两个个体, 我们试图基于它们在自变量中的排序来预测它们在因变量中的排序。

假定 Julia 是一个政党倾向很强的人, 并对选举很感兴趣, Paul 是中立的并对选举有点感兴趣。如果身份变量是从左向右水平排列, 而兴趣变量是从下向上垂直排列, 因为 Julia 落在 Paul 的右边, 所以她在身份变量上排名比 Paul 高(图 12.2)。同样的道理, 她在兴趣变量上的排名也比 Paul 高。

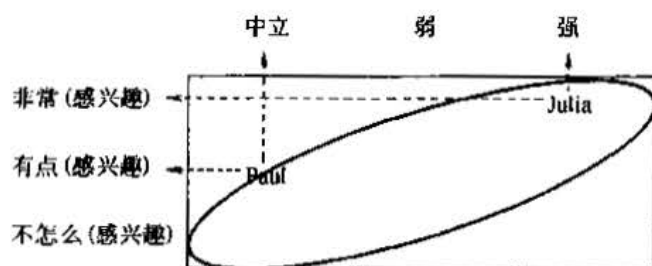


图 12.2 Julia 在两个变量上都比 Paul 排名高。

假定我们仅知道 Julia 和 Paul 在身份变量上的排序。我们能预测他们在兴趣变量上的排序吗? 假定所有的人都落在图 12.2 中从左下角到右上角的椭圆内, 那么除了那些和 Julia 在一个或两个变量上取值相同的以外, Julia 将比其他所有的人在两个变量上的排名都高。所以, 如果我们知道 Julia 在自变量上排名较高, 那么我们就可以准确地预测她在因变量上的排名也较高。在这种准确预测的情况, $\gamma = 1.00$ 。

当预测都落在从左上角到右下角的椭圆内时, 如图 12.3 所示, 就会发生相反的极端情况。在这里, Julia 是一个政党倾向很强的人, 且对选举她不怎么感兴趣。Paul 是一个中立分子, 并对选举有点感兴趣。现在由于 Julia 落在 Paul 的右边, 所以在身份变量上她比 Paul 排名高, 但由于 Paul 在 Julia 上方, 所以他比 Julia 在兴趣变量上排名高。对于任意一对落在这个从左上角到右下角的椭圆内的观测都会发生相同的情况, 除非他们在一个或两个变量上有相同的值。如果我们知道了自变量的排名, 那么我们就可以准确地预测因变量的排名。对这种情形, $\gamma = -1.00$ 。

数一数图 12.2 和图 12.3 中这两类观测每类有多少对, 就可以计算 γ 的值。 γ 告诉我们如果我们知道了两个观测在自变量上的排名比不知道这一点, 能够多大程度上更好地预测因变量的排名。对于 $\gamma = 0.21$, 知道他们在自变量上的排名比不知道这点对预测因变量提高了 21% 的准确性。 γ 的更加详细的解释和计算方法见本章末的公式 12.1。

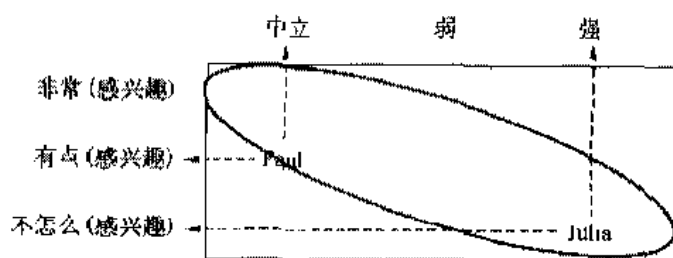
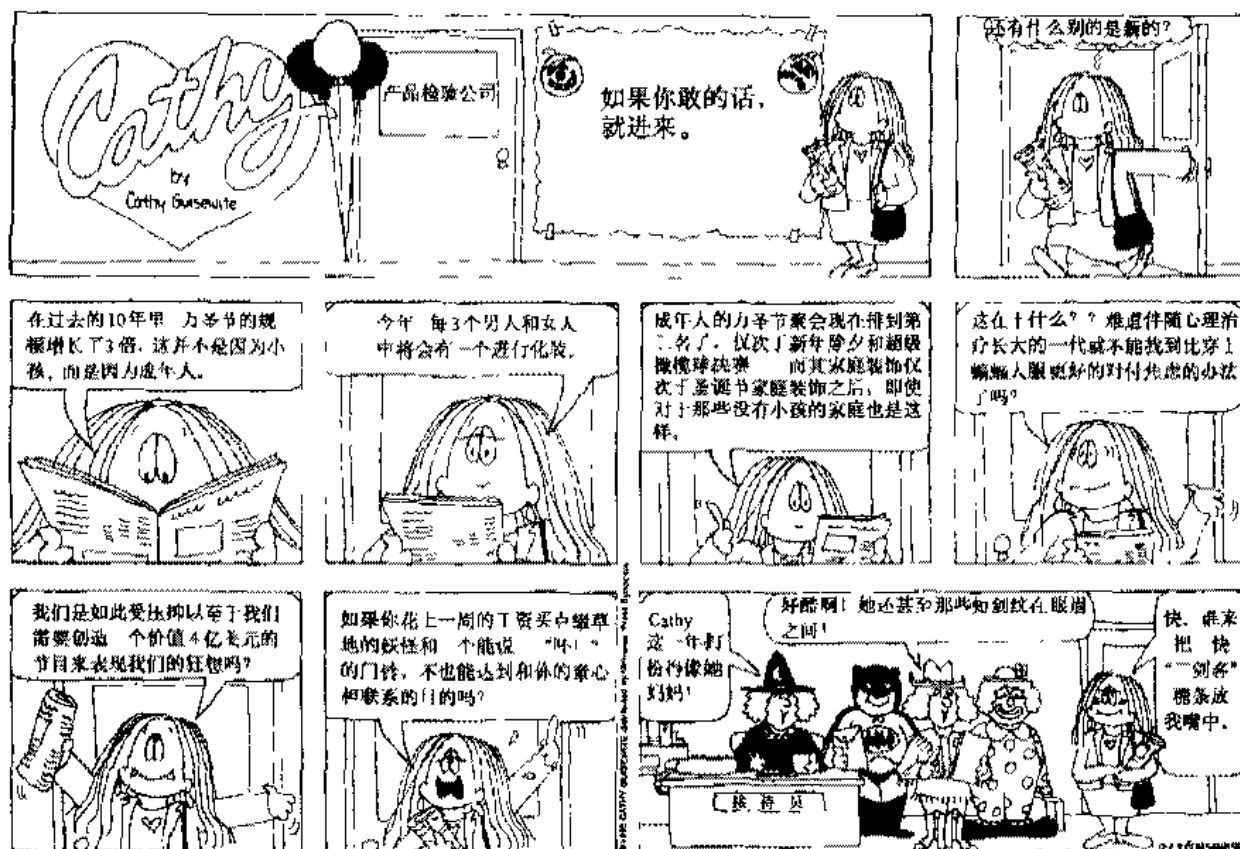


图 12.3 Julia 在政党倾向上排名高,而 Paul 在兴趣上排名高。



来源: "Cathy" 1995 Cathy Guisewite. Reprinted with permission of Universal Press Syndicate All right reserved

问题 3. 总体的关系?

和通常一样,零假设假定在总体中两变量间没有关系,而且如果 p -值很小就拒绝它。 p -值是从一个两变量没有关系的总体中得到一个 $\gamma \geq 0.21$ 的样本的概率。

要计算 p -值首先把观测到的 γ 值转化为一个标准正态的 z -变量的值。对于这个例子, $z = 6.47$ 。利用统计软件或统计表,我们发现对于这样大小的 z , p -值小于 0.0001。这个概率很小,所以把得到 $\gamma \geq 0.21$ 的数据归于偶然性是不可信的。从两变量间没有关系的总体中得到 $\gamma \geq 0.21$ 的样本的可能性小于 1 比 10000。 p -值这样小,所以我们有足够的证据拒绝零假设,并认为对所有成年人的总体这两个变量是相关的。公式 12.2 和 12.3 说明怎样由 γ 值得到

z -变量的值。

问题 4. 是因果关系吗?

仅从数据我们不能判断这个关系是否是因果关系。在这里,甚至先有哪个变量都是有疑问的。人们是因为是狂热的某一政党的支持者才变得对选举感兴趣了呢,还是由于对选举感兴趣才变成某一政党的狂热支持者的呢?

停下来想一想 12.3

若两个顺序变量的相关系数 $r = 0.47$, 相应的 p -值是 0.15。用预测和统计上显著的术语,你怎样说明变量间的关系?

12.2 把数目的排序作为值: Phillies 表现如何?

随着时间的推移,如何比较棒球队间的名次? 好的队是否一直是很好,或者随时间有什么变化? 在这里,我们研究 1987 赛季和 1992 赛季之后国家俱乐部东部联盟中 6 支棒球队的名声变化;其中包括我们家乡的球队。表 12.3 说明 5 年期间开始和结束时各个队的排名。在这段时间内,这两列排名有很大变化。为了更好地了解这些数据,我们对它们进行某些统计分析。

表 12.3 1987 和 1992 赛季后国家东部联盟球队的排名

队	排名	
	1987	1992
Chicago Cubs	6	4
Montreal Expos	3	2
New York Mets	2	5
Philadelphia Phillies	4	6
Pittsburgh Pirates	5	1
St. Louis Cardinals	1	3

问题 1. 数据中的关系?

尽管这些变量仅仅是顺序变量,但因为我们有了这些变量的值,所以我们仍可以在散点图中描述这些数据。用 1987 年的数据作为一个轴,1992 年的数据作为另一个轴。在图 12.4 中,每个队在散点图中是一个点。对于观测比较少的情形,标记一下这些点是很有帮助的。在这里,我们用每个队的名字来标记。如果一个队在两年的排名都比较高,这个队将出现在图的左下角,但并没有队落在那里。如果一个队两年排名都比较低,这个队将落在图的右上角,我们的 Phillies 和 Cubs 队接近于这个角。

如果每个队的排名都没有变化,那么散点图看上去将是什么样子呢? 每个队在两个时间

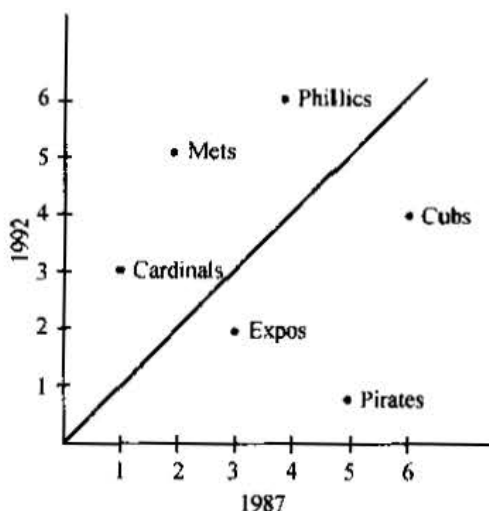


图 12.4 1987 和 1992 赛季后球队排名的散点图。

的排名将一样,所以 6 个点将落在穿过图的一条 45 度的直线上;如果真是这样,图 12.4 会显示表示数据的很明显的规律。但是我们的观测点在图中是到处散布的,几乎没有什么规律。这声明在这两年之间不存在什么关系。

问题 2. 关系强度?

为了纪念它的首创者,秩顺序相关系数通常记作 r_s ,它是用来度量取值为数的两个有数量值的顺序变量的相关程度的。

取数为其值的顺序变量的相关程度由一个称作秩顺序相关系数(rank order correlation coefficient)的系数来度量,通常记作 r_s 。下标“s”是为了纪念英国心理和统计学家 Charles Spearman;他早在 1900 年代就在这个领域作出了开创性的工作。由于这个原因,这个系数有时也称作 Spearman 秩相关系数。秩相关系数 r_s 的计算方法和计算数量变量的相关系数 r 的方法相同。我们把数值次序当成数量变量,所以 r_s 是 r 的一种特殊情况。为了利用统计软件计算 r_s ,我们简单地选择两个顺序变量,并计算其普通的相关系数 r 。这两个系数间有一点微小的差异。从本章后面的公式中你可以看到,用纸和笔计算 r_s 要比计算 r 简单得多。这是因为两个变量的取值包括了从 1 到我们观测的个数的所有整数,从而简化了 r 的计算公式而得到 r_s 的计算公式。

对于棒球队的数据, $r_s = -0.09$,这在 -1 到 1 之间的刻度上是一个很弱的关系。如果每个队的排名没有变化,数据点将落在 45 度的直线上,这时 r_s 将等于 1.00。如果六个队的排名完全颠倒了过来,数据点将落在从左上角到右下角的 45 度的直线上。对这种情形, r_s 将等于 -1.00。公式 12.4 说明怎样计算 Spearman 秩相关系数。

因为观测到的系数与 1 相差较大,所以我们认为排名有所变化。又由于这个系数接近于 0,所以我们可以得出结论认为这些排名的变化几乎没有什么趋势。如果说有点趋势的话,由于系数是负的,1987 年成绩比较好的队 5 年后稍有一点变差的趋势,而 1987 年比较差的队 5

年后有所好转。

停下来想一想 12.4

怎样利用 Spearman 相关系数对一个贪婪的赛马赌徒提供一些帮助?

问题 3. 相关性是由于偶然吗?

要知道排名的变化是否出于偶然,我们提出一个零假设,然后计算数据的 p -值。在这种情形,零假设认为两个变量的关系是仅出于偶然。

要计算 p -值,首先要把 r_s 的值变为相应的一个标准统计量的值。在这个例子中我们把它变为一个自由度为 $n-2=6-2=4$ 的 t -变量。我们得到 $t = -0.17$, 仅仅靠偶然得到这样大小的 t -值的概率或得到 $r \leq -0.09$ 的概率等于 0.44; 也就是说,在 100 个来自两个变量间不相关的总体的样本中 44 个样本将得到 $r_s \leq -0.09$ 。这个 p -值是不显著的。很明显,1987 年哪个队表现好或坏与 5 年后哪个队好还是坏没有什么关系。

问题 4. 是因果关系吗?

和通常一样,我们不能认为统计关系就是因果关系,但是在这里看来低相关性是出于偶然,所以因果关系就更无从谈起。

12.3 小结

当一组元素根据某种比较,比如大小、质量、年龄、或者速度进行排序时,就会生成顺序变量。顺序变量可以取值为词或数值。取值为词的顺序变量在一个词语程度的刻度上进行排序,比如非常、适中、稍微、一点也不。取值为数值的顺序变量在一个数值刻度上进行排序,例如 1、2 等直到所有观测的个数。顺序变量含有的信息量比分类变量多而比数量变量少,所以不能用分析这两种变量的方法来分析顺序变量。

12.1 用词作为值的两个顺序变量

取值为词的两个顺序变量的相关程度的一个常用的度量叫做 γ 。 γ 根据成对观测在两个变量上的相对次序,度量了这对观测的相似性。认为总体中两个变量没有关系的零假设可以通过把 γ 变为一个 z -变量的值然后计算 p -值来检验。通过计算两个顺序变量的相关系数,我们可以根据一个变量的次序来预测另一个顺序变量的次序。秩相关并不意味着两变量间有因果关系。

12.2 把数目的排序作为值: Phillies 表现如何?

如果观测是以数字排序的,每个数代表一个观测(除非观测中有结存在),变量间的相关程

度用 Spearman 秩相关系数度量, 通常记作 r_s 。其值为 0 表示由两个独立变量生成, 取值范围

补充读物

公 式

$$\begin{aligned}
 \text{不同的次序} &= \sum \text{所有的频率与右下方的频率的乘积} \\
 &= 104(273 + 237 + 228 + 125) + 150(237 + 125) \\
 &\quad + 178(228 + 125) + 273(125) = 258268
 \end{aligned}$$

从而 γ 的定义为:

$$\gamma = \frac{\text{相同的次序} - \text{不同的次序}}{\text{相同的次序} + \text{不同的次序}}$$

对于这个例子,

$$G = \frac{394256 - 258268}{394256 + 258268} = \frac{135988}{652524} = 0.21$$

我们可以这样来理解 γ 。总共有 652524 对可能的配对,对每一对,我们来预测哪个在因变量上排名较高。在没有任何知识的情况下,这就像掷硬币,我们正确或错误的可能性各占一半。652524 的一半等于 326262,所以我们预期将会有 326262 次预测是错误的。现在假定我们知道了在一对人中哪一个在自变量上的次序较高,我们预测这个人在因变量上的次序也较高。但是我们知道有 258268 个不同的排序,所以对这 258268 对人这样的预测是错误的,也就是说,我们将有 258268 次预测是错误的。所以知道了自变量的排序后,我们可以多预测对 $326262 - 258268 = 67994$ 对。我们改进的比例是 $\frac{67994}{326262} = 0.21$,这和计算 γ 的公式是一样的。

要检验零假设,即对总体的 $\gamma = 0$,我们计算检验的统计量:

$$z = \frac{G}{\sqrt{\frac{4(\text{行数} + 1)(\text{列数} + 1)}{9n(\text{行数} - 1)(\text{列数} - 1)}}} \quad (12.2)$$

这里 n 是表中观测的总数。对于这个例子:

$$z = \frac{0.21}{\sqrt{\frac{4(3 + 1)(3 + 1)}{9(1690)(3 - 1)(3 - 1)}}} = \frac{0.21}{0.0324} = 6.47$$

这个 z 的式子仅是一个近似,对观测很多的比较大的表,这个逼近是很好的。 z 的更精确的值可由下式计算:

$$z = \frac{G}{\sqrt{1 - G^2}} \sqrt{\frac{\text{相同的次序} + \text{不同的次序}}{n}} = \frac{0.21}{\sqrt{1 - 0.21^2}} \sqrt{\frac{394256 + 258268}{1690}} = 4.22 \quad (12.3)$$

第一个表达式的优点在于如果我们知道了 γ 、观测数和表的大小就可以计算 z 。而要由第二个公式计算 z ,我们既需要知道 γ ,也需要知道相同和不同的排序数目。在这个例子中,尽管这样得到的两个 z 值不同,但两个 z 得到的 p -值却都小于 0.0001,所以对于这两个不同的 z -值可以得出相同的结论。

(你也许想知道如果我们把表 12.1 作为分类变量的列联表来处理并用 χ^2 分析来检验显著性的结果如何。我们将会得到 $\chi^2 = 71.69$, $df = 4$, 由这个 χ^2 值得到的 p -值不会像由公式 12.2 和 12.3 得到的 z 的 p -值那样小。出现这种情况的原因是 χ^2 对初始表的行和列的顺序不敏感, 所以它没有充分利用数据中所有可利用的信息。但是即使这个 χ^2 值也是高度显著的, 所以如果只是检验显著性, 我们也可以用这个 χ^2 进行检验。)

Spearman 的 r_s

可以用两列秩代替普通相关系数 r 的公式(公式 10.1)中的 x 和 y 来计算秩顺序相关系数。因为两个变量的两列值是从 1 到 n 的整数, 计算 r 的公式可简化为公式 12.4。

我们两列顺序(即秩), 现在我们计算每一个观测的两个秩的差异, 然后把每个差异平方。这些计算如表 12.4 所示。从这些平方的和我们用下面公式得到 r_s :

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (12.4)$$

在这里, 分子中的 6 是总在此公式中的常数, n 是秩的对数。

表 12.4 r_s 的计算

x 中的秩	y 中的秩	差异	差异的平方
x_1	y_1	$d_1 = x_1 - y_1$	d_1^2
x_2	y_2	$d_2 = x_2 - y_2$	d_2^2
\vdots	\vdots	\vdots	\vdots
x_n	y_n	$d_n = x_n - y_n$	d_n^2
		和	$\sum d_i^2$

表 12.5 计算 r_s 的例子

队	1987	1992	差异	差异的平方
Chicago	6	4	2	4
Montreal	3	2	1	1
New York	2	5	-3	9
Philadelphia	4	6	-2	4
Pittsburgh	5	1	4	16
St. Louis	1	3	-2	4
			和	38

$$r_s = 1 - \frac{6 \times 38}{6(6^2 - 1)} = 1 - 1.09 = -0.09$$

对于棒球队的数据, 基于表 12.5 中的平方和我们可得到下面的结果。要知道我们是否可以拒绝认为总体的秩顺序相关系数等于 0 的零假设, 我们计算下面的 t -检验的统计量:

$$t = \frac{r_s}{\sqrt{1 - r_s^2}} \sqrt{n - 2} = \frac{-0.09}{\sqrt{1 - (-0.09)^2}} \sqrt{6 - 2} = -0.17 \quad df = 4 \quad (12.5)$$

然后我们就可以用这个 t -值来计算 p -值了。在这里, $p = 0.44$, 显然, 对于这样小的 t -值我们

不能拒绝零假设。

习题

回顾(习题 12.1—12.10)

- 12.1 a. 什么是顺序变量?
b. 从你自己的生活经历中举两个顺序变量的例子。
- 12.2 顺序变量可以取值为词或数值。
a. 举一个取值为词的顺序变量的例子和一个取值为数值的顺序变量的例子。
b. 这两类顺序变量通常的主要区别是什么?
c. 你所举的例子出现这种差异了吗?
- 12.3 a. 什么是 γ ?
b. γ 的最大和最小的可能值是多少?
c. 举出另一个具有相同的取值范围的统计量。
d. 看一下像表 12.1 和 12.2 那样的表, 你怎样判断 γ 的秩值是正的还是负的?
- 12.4 a. 举一个你可以用 γ 用来度量变量间的相关关系的两个变量的例子。
b. 一般怎样计算 γ ?
c. 如果 $\gamma = 0.75$, 你由自变量的信息预测因变量的能力有多大?
d. 如果对于这个 γ , p -值很小, 产生这个样本的总体中两变量间的关系如何?
e. 对于两变量间没有关系的零假设你能得出什么结论?
- 12.5 一般地需要哪些步骤来计算 γ 的 p -值?
- 12.6 a. 什么是秩顺序相关系数?
b. 为什么秩顺序相关系数记作 r_s ?
c. 什么时候用 r_s ?
d. r_s 和哪一个统计量最相似?
- 12.7 举一个可以用 r_s 度量两个变量关系强度的例子。
- 12.8 a. 顺序变量和分类变量的主要差异是什么?
b. 为什么说顺序变量比分类变量含有更多的信息?
c. 怎样区别顺序变量和数量变量?
d. 为什么说顺序变量比数量变量含有信息少?
- 12.9 a. 班级的排名是什么类型的变量?
b. 对一组学生来说, 为什么一群学生对班级的学业表现排名不如对 GPA 敏感?
- 12.10 关于两个数量变量的数据, 例如全国前 100 个公司的年销售量和利润, 我们为什么有时要把这些数据变为顺序变量的数据, 然后用顺序变量而不是初始变量来研究变量间的关系?

解释(习题 12.11—12.20)

- 12.11 简明受伤尺度(Abbreviated Injury Scale)是试图用于度量摩托车事故的严重性。变量

的值是:(1)轻微,(2)中等,(3)严重但不会危及生命,(4)严重,危及生命,有可能存活,(5)危急,不知道会不会存活,(6)致命,目前无法治疗。(来源:Andrew A. Weiss, "The effects of helmet use on the severity of head injuries in motorcycle accidents," *Journal of the American Statistical Association*, vol. 417(1992), p. 496)解释是否这个受伤尺度是一个名义上的、次序的还是数量变量?

- 12.12 1984年夏季奥运会是不寻常的,因为几个东欧国家都抵制参加。在这个问题中,我们研究参加了那年冬季和夏季奥运会的国家获奖牌的情况,看一看在夏季和冬季奥运会上获得的奖牌数是否有什么关系。因为大的国家获得的奖牌一般也较多,我们按国家获得的奖牌数进行排名。共有12个国家既在夏季奥运会又在冬季奥运会上获得了奖牌,所以数据包含两列1到12这12个数。对于这些数据, $r_s = 0.78$ ($t = 3.81$, $df = 10$, $p = 0.002$)。(来源:The World Almanac and Book of Facts, 1988, pp. 834, 837.)

a. 这个正的秩相关系数意味着什么?

b. 如果两变量间的相关系数为负的,你会感到奇怪吗?

c. 这些数据告诉我们那年冬季和夏季奥运会上获得的奖牌数有什么关系?

- 12.13 在1950年 Detroit(底特律)进行的一项研究中,研究者报导了官僚的(与企业家相对)的39岁以下、家庭收入在\$6000以上的太太的情况,他们发现了下面的社会阶层和孩子的数目的分布情况,如表12.6所示。对于这个表, $\gamma = 0.24$ ($z = 1.34$, $p = 0.09$)。对于这一类妇女,你从分析中可得出社会阶层和这些阶层母亲的孩子个数间有什么关系?

表12.6 习题12.13的数据

		社会阶层				总计
		下下	下上	中下	中上	
孩 子 数 目	3+	4	7	12	9	32
	2	2	13	5	5	25
	1	10	9	7	6	32
	0	4	6	7	4	21
总计		20	35	31	24	110

来源:Daniel R. Miller and Guy E. Swanson, *The Changing American Parent: A Study in the Detroit Area*, New York: John Wiley & Sons, 1958, p. 205

- 12.14 在习题10.34中我们研究了调味品和对于不同的冷冻巧克力酸奶中来自脂肪的热量百分比的关系。在这两个变量的散点图上,Albertsons 酸奶的值比其他酸奶的值小许多。这个数据点可能会对分析产生比较大的影响;减小这样点的影响的一个方法是把两个变量的值变为秩(即顺序)。在表12.7中是两个变量的秩的值。(有结存在的地方用秩的平均值,所以秩不是从1到10连续的。)对于这两个变量, $r_s = 0.68$ ($t = 2.62$, $df = 10$, $p = 0.015$)。

表 12.7 习题 12.14 的数据

酸奶的品牌	排名	
	脂肪的热量的百分点	口味
Breyers	9.5	8
Honey Hill Farms	9.5	10
Elan	8	5.5
Crowley Silver Premium	7	3
Edy's/Dreyer Inspiration	6	9
H'agen - Dazs	5	5.5
Kemps	4	3
Lucerne	3	7
Yoplait Soft	2	3
Albertsons	1	1

- a. 对于这两个顺序变量的关系你能得出什么结论?
 - b. 这个结论与你在习题 10.34 中分析相同的数据得出的结论有何不同?
 - c. 用顺序变量时, 极端值数据 Albertsons 酸奶有什么变化?
- 12.15 研究人员发现和煦日子的长短与葡萄的甜度的相关系数是 $r = 0.81$ 。来自欧洲的一些葡萄的样本表明阳光越充足, 葡萄就越甜。
- a. 如果研究人员想说明他们的这个发现可以用于抽取这个样本的所有葡萄的总体, 需要作什么工作?
 - b. 你认为发现的结果将会如何?
 - c. 这里要考虑的零假设是什么?
- 12.16 一个特定的 r_s 或更极端的值偶然发生的概率为 0.021。你怎样对你的一个不懂统计的朋友解释这句话的意思?
- 12.17 如果我们对全国 25 个最好的大学篮球队在某一个赛季获胜的次数进行排名, 并对这几个队下一个赛季也进行排名, 得到 $r_s = 0.79$ 。
- a. 你怎样对你的一个不懂统计的朋友解释这个结果? 是否一个赛季排名靠前的球队下个赛季排名仍然较靠前?
 - b. 如果对两倍多的 (即 50 个) 队在两个赛季进行排名, 得到 r_s 仍等于 0.79, 进行假设检验的 p -值将增大、不变还是减小?
- 12.18 新闻杂志 The Economist (1993 年 6 月 12 日, 62 页) 对英国的 12 个地区在许多变量上进行了排名。其中一个变量是死亡率。East Anglia 死亡率最低, 所以排第一, Scotland 死亡率最高所以排第 12。另一个变量是住房花费。Northern Ireland 花费最低, 所以排第一, Great London 毫不令人惊奇地花费最高, 所以排第 12。对于死亡率和家庭花费这两个顺序变量, $r_s = -0.76$ 。
- a. 相关系数为负的告诉你两变量间关系如何?
 - b. 你为什么认为两变量间关系有这么强?
 - c. 要看一看两列秩的趋势是否是偶然出现的, 我们可由相关系数得到 t -变量的值为 -3.70 , $p = 0.0021$ 。零假设是什么? 对于零假设你能得出什么结论?
 - d. 两变量间的关系是因果关系吗? 或者你能认为有其他变量会解释这个关系吗?

- 12.19 如果你对全国 10 个最大的城市在 1980 年和 1985 年进行排名,对这两组秩会得到 $r_s \approx 0.95$ 。
- 如果每个城市在一段时间内人口增长同样的数字,这个增长会改变城市的排名吗?
 - 如果每个城市在一段时间内人口增长百分比相同,这个增长会改变城市的排名吗?
 - 如果对于这两组秩你知道 $r_s \approx 0.95$,对于不同城市的增长趋势如何评论?
 - 要知道这两组排名是否是偶然出现的,你可以由 r_s 得到一个 t -变量的值,得到 $t = 8.75, df = 8, p = 0.00001$ 。关于这些排名你能得出什么结论?
- 12.20 联合国发展计划在巴基斯坦经济学家 Mahboub ul Haq 的指导下发展了人类发展指数 (HDI)。这个指数超越了过去 GNP 的范畴。它提供了包括购买力、期望寿命和受教育的程度的一个度量方法。在人口超过 1 百万的 130 个国家中, Niger 的 HDI 最低,排第一,日本的 HDI 最高排第 130。图 12.5 是随机选取的 13 个国家在两个变量 GNP 和 HDI 上排名的散点图。
- 描述你在散点图中看到的趋势。
 - 两组秩的相关系数为 $0.89, p = 0.00002$ 。你能认为两组秩的观测相关系数是出于偶然吗?
 - 如果每个国家在两个变量上的排名都相同,这些点将落在一条 45 度的直线上。美国落在这条直线的下方而法国落在直线的上方的可能的原因是什么?

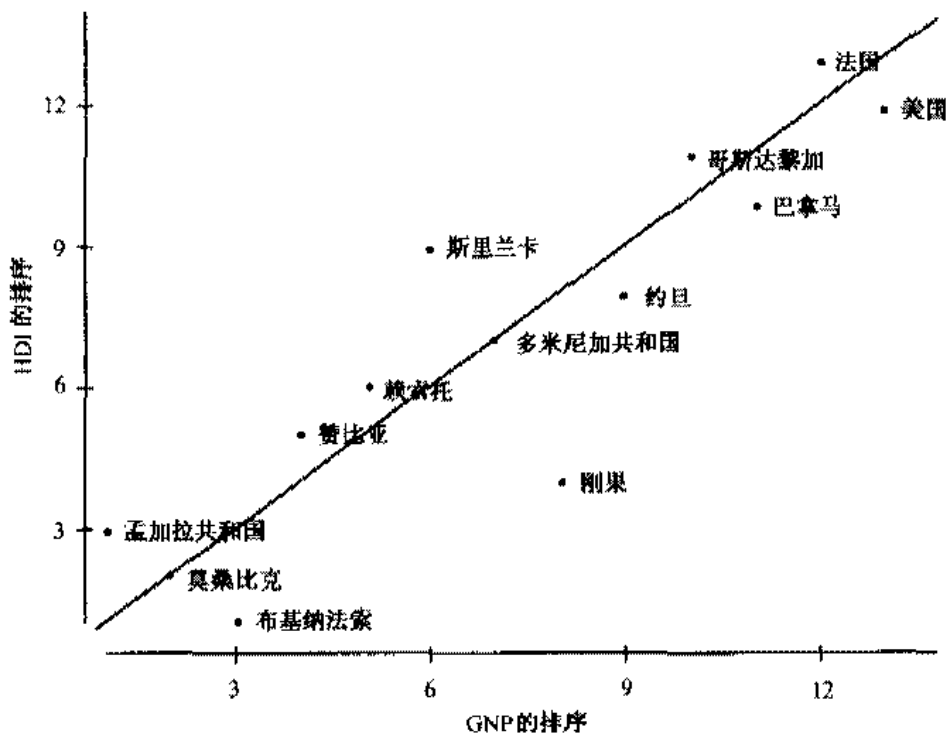


图 12.5 在国民生产总值 (GNP) 和人类发展指数 (HDI) 上 13 个国家的随机样本。

分析(习题 12.21 - 12.34)

12.21 在一次全国性的调查中,其中一个问题是询问被调查者关于流产的态度以及他们认为流产问题的严重性(表 12.8)。

表 12.8 习题 12.21 的数据

		对流产的态度			总计
		反对	中立	同意	
重 要 性	最重要的之一	85	167	89	341
	重要	80	638	357	1075
	不太/一点也不重要	56	583	366	1005
	总计	221	1388	812	2421

来源: John Scott and Howard Schuman, "Attitude strength and social action in the abortion dispute," American Sociological Review, vol. 53(1988), p. 788.

- 对于这些数据,态度和严重性间有关系吗?
- 关系的强度如何?
- 在所有的美国成年人的总体中这两个变量间是否存在一定的关系?
- 这个关系是因果关系吗?

12.22 表 12.9 中的数据表明 1950 和 1970 年不同类型的产品在一般杂志和全国性农业杂志上的广告花费的排名。1950 年正经历二战后向和平年代的调整期,到 1970 年调整期结束。排名数字越小,该产品花在广告上的钱就越多。

表 12.9 习题 12.22 的数据

产品种类	在 1950 年的排名	在 1970 年的排名
衣服、鞋袜等	3	6
汽车	2	2
酒精类饮料	5	1
建筑材料	8	13
消费者服务	11	4
食物和食物类产品	1	3
家庭用器	4	11
家具	6	9.5
工业材料	7	8
保险	13	12
收音机、电视等	12	9.5
烟	9.5	5
旅游	9.5	7

来源: U. S. Bureau of the Census, Statistical Abstract of the United States: 1972, 93rd edition, Washington, DC: U. S. Government Office, 1972, p. 759.

- 仅看这些数据,产品的促销活动有什么变化吗?
- 如果从 1950 年到 1970 年两列排名没有变化,这两组排序的关系强度如何?
- 作这些数据的散点图,并在图上标记出这些产品。

d. 从图中你能看到什么规律? 这个规律与你对 1950 年到 1970 年时期所了解的有什么关系?

e. 两列排名间的关系强度如何?

f. 这个关系的出现能是出于纯偶然的吗?

12.23 搜集最近的冬季和夏季奥运会的数据。

a. 把在两次奥运会上都获奖牌的国家按它们所获奖牌的总数进行排名。

b. 计算两列排名的秩顺序相关系数。

c. r_s 的值统计显著地不等于 0 吗?

d. 这次分析的结果与习题 12.12 中夏季奥运会被抵制的那次分析的结果有何不同?

12.24 对于习题 10.57 中 Calabrian 黑社会的数据, 可以研究黑社会组织在选择他们的领袖时是否可能使领袖的年龄和组织成员的平均年龄有关。这些数据可能不是线性相关的, 所以我们将原始年龄变为顺序变量。

a. 把两个年龄变量变为顺序变量。

b. 分析两个顺序变量的关系。

c. 把这些顺序(秩)的分析结果和对习题 10.57 中关于原始数据分析的结果进行比较。造成这两个结果差异的可能的主要原因是什么?

12.25 在习题 10.52 中我们研究了 100 年间离婚数目作为因变量与结婚的数目作为自变量的关系。初始数据的散点图显示了一个很强的非线性趋势。所以我们通过取每个观测的对数把数据化为顺序变量数据。数据表明从 1890 年每隔 5 年到 1980 年结婚和离婚的数目(单位:千人)。对年份变量以 1 代表 1890, 2 代表 1895, 直到 19 代表 1980, 这样输入数据时容易些。

另一个方法是把所有的观测按照每个变量进行排序, 然后研究这两列秩的相关关系。

年份	1	2	3	4	5	6	7
结婚	570	620	709	842	948	1008	1274
离婚	33	40	56	68	83	104	170

年份	8	9	10	11	12	13	14
结婚	1188	1127	1327	1596	1613	1667	1531
离婚	175	196	218	264	485	385	377

年份	15	16	17	18	19
结婚	1523	1800	2159	2153	2413
离婚	393	479	708	1036	1182

来源: National Center for Health Statistics, Public Health Service, in The World Almanac 1986, p. 779.

a. 把这两个变量的观测值从 1 到 19 排序。

b. 作这两列秩(顺序)的散点图。

c. 评价散点图的形状。

d. 分析这两列秩的关系。

- 12.26 在习题 10.54 中我们研究了人均收入在 \$ 2000 以下的国家中, 20 个国家的人们受教育程度和人均收入的数据。原始数据的散点图并没有表明两变量间有明显的线性关系。所以我们按照每个变量进行排序, 然后研究两个顺序变量的关系。这些国家是: 孟加拉国、博茨瓦纳、柬埔寨、智利、古巴、埃及、加纳、圭亚那、象牙海岸、北朝鲜、马达加斯加、毛里塔尼亚、莫桑比克、巴基斯坦、菲律宾、圣多美、南非、坦桑尼亚、乌干达和扎伊尔。

国家	1	2	3	4	5	6	7
受教育的百分点	25	30	48	90	96	44	30
人均收入	119	544	100	1950	840	686	420

国家	8	9	10	11	12	13	14
受教育的百分点	86	24	99	53	17	14	24
人均收入	457	1100	570	279	466	220	280

国家	15	16	17	18	19	20
受教育的百分点	88	50	98	66	25	40
人均收入	772	300	1296	240	240	127

- 把这两个变量的观测值从 1 到 20 排序。
 - 作这两列秩的散点图。
 - 评价散点图的形状。关于两列顺序变量的相关性作相关分析有意义吗?
 - 计算两列顺序变量的相关系数。
 - 这个秩相关系数显著的不为 0 吗?
- 12.27 表 12.10 是关于妇女的年龄和患乳腺癌的概率的数据。

表 12.10 习题 12.27 的数据

年龄	患乳腺癌的概率
年轻(39 岁以下)	0.0005
中青年(40 - 49)	0.015
中老年(50 - 59)	0.024
老年(60 - 69)	0.036
最老的(70 - 80)	0.042

来源: *National Cancer Institute, American Cancer Society, as reported in The Philadelphia Inquirer, January 18, 1993, p. D1.*

- 如果年轻排为 1, 最老的排为 5, 这些概率也从 1 到 5 排序。度量这些变量的关系强度的相关系数等于多少? 用一、两句话描述计算的过程。
 - 这个表对于媒体所描述的在美国妇女间的“流行乳腺癌”有什么启示?
 - 为什么在这个国家过去 40 年中, 患乳腺癌的人数可能上升了, 然而对于一固定的年龄段的患乳腺癌的概率却没有变化?
- 12.28 一项关于 8 年级学生成绩的全国性调查表明, 社会经济地位不同的孩子成绩也不同。

这些变量是社会经济地位(高和低)和能力(高、中/混合、低)。各类的百分比数据如表 12.11 所示。

表 12.11 习题 12.28 的数据

		社会经济地位(SES)	
		低(%)	高(%)
能 力	高	13	39
	中/混和	50	47
	低	37	14
	总计	100	100

来源: U. S. Department of Education, National Center for Educational Statistics, National Education Longitudinal Study of 1988

a. 在一个图中用圆(像图 12.1c 那样)或条形图(像图 12.1a 或 b 那样)来表示这些数据。

b. 对两变量间的关系,从这些数据你能得出什么结论?

12.29 多少世纪以来,人们参加生产劳动的时间有多长(不幸的是,不包括家务劳动)?

富于创造性的历史学家和人口统计学家已搜集了每年工作的小时的估计数。大部分数据是关于英国农民和制造业的工人的数据。其中部分数据如表 12.12 所示。

表 12.12 习题 12.29 的数据

年份	每个工人每年工作的时间
1200	1620
1300	1440
1400	2300
1500	3200
1600	1980
1700	—
1800	3300
1900	1900

来源: Juliet B. Schor, The Overworked American: The Unexpected Decline of Leisure, New York: Basic Books, 1991, p. 45

a. 在一个散点图上描述这些数据。

b. 这个散点图告诉你什么?

c. 把这些数据变为两个顺序变量的数据。

d. 如果你想知道在不同世纪,工作时间是否有变化,零假设是什么?你怎样利用顺序数据检验零假设。

e. 你能拒绝零假设吗?

f. 你能用其他什么统计方法来研究这些数据?

12.30 搜集你所选的两个顺序变量的数据。

a. 两变量间有一定的关系吗?

b. 关系强度如何?

c. 产生样本的总体中两变量间是否存在关系?

d. 这个关系是因果关系吗?

- 12.31 评估教育成就国际联合会(The International Association for the Evaluation of Education Achievement)在 1991 年公布了一项关于不同国家 12 和 13 年级的学生在科学方面的表现。在生物和化学方面这些国家的排名情况如表 12.13 所示。

表 12.13 习题 12.31 的数据

国家	排名	
	生物	化学
新加坡	1	3
英国	2	2
匈牙利	3	5
波兰	4	7
香港	5	1
挪威	6	8
芬兰	7	13
瑞典	8	9
奥地利	9	6
日本	10	4
加拿大	11	12
意大利	12	10
美国	13	11

来源: *International Association for the Evaluation of Educational Achievement*.

- 用生物和化学作为两个变量作这些数据的散点图并用国家的名字来标记这些点。
- 描述你在散点图中看到的趋势,例如包括在化学方面比在生物方面排名高的国家等。
- 两列顺序变量的关系强度如何?
- 出现这个关系是纯属偶然的吗?
- 这个关系是因果关系吗?

- 12.32 在对政府统计组织,国际机构及其他组织中工作的利用国际统计数字的统计学家进行的一项调查中,经济学家杂志对 10 个国家的统计部门进行了排名。在同一篇文章中,这份杂志还列出了每 10000 个人中统计学家的人数及政府在按每个统计学家的人均预算(美元)的数据。这些数据如表 12.14 所示。

表 12.14 习题 12.32 的数据

国家	统计部门的排名	每 10000 人中统计学家人数	政府统计预算(美元/人)
加拿大	1	1.6	8.20
澳大利亚	2	2.0	9.00
荷兰	3	2.0	7.60
法国	4	1.7	6.00
英国	5	0.9	4.20
德国	6	1.9	8.00
美国	7	0.6	8.80
意大利	8	1.4	5.00
西班牙	9	1.2	4.20
比利时	10	1.3	3.60

来源: *The Economist*, September 11, 1993, p. 65

- a. 把最后两列数变为顺序变量的值, 值最大的排为 1, 次大的排为 2, 等等。
- b. 统计部门的排名与统计学家的人数有关吗?
- c. 统计部门的排名与花费有关吗?

12.33 秋季的每个星期, 体育专栏作家们通过美联社 (Associate Press)、教练们通过今日美国报 (USA Today) 和有线电视网 (CNN) 对大学的橄榄球队进行排名。这些排名, 每周都有所变化, 表 12.15 表示在 1994 年 11 月末这些排名的变化。这些是在 Texas A&M 和 Auburn 被美联社从排名上删除后的前 12 个队排名; 这两支队由于被全国大学生体育协会 (NCAA) 惩罚, 所以没有对其进行排名。

表 12.15 习题 12.33 的数据

大学	排名	
	11 月 28 日	11 月 21 日
Nebraska	1	1
Penn State	2	2
Alabama	3	3
Miami	4	4
Colorado	5	5
Florida	6	6
Florida State	7	7
Colorado State	8	10
Kansas State	9	8
Oregon	10	9
Ohio State	11	11
Utah	12	12

来源: The New York Times, November 28, 1994, p. C2

- a. 作这些排名的散点图。
- b. 评价散点图中的规律。
- c. 计算秩相关系数, 看看这两列秩在多大程度上相吻合。
- d. 这样的排名是否出于偶然?

12.34 许多政治上的紧张状态是由于国家间争论是否加入欧洲共同市场引起的。在 1960 年代, 挪威考虑要加入, 表 12.16 是 1965 年间 286 个挪威人在两次不同时间表达的观点的变化。

表 12.16 习题 12.34 的数据

		1965			总计
		正式成员	松散联系	不参加	
1969	正式成员	100	45	15	160
	松散联系	24	35	23	82
	不参加	4	10	30	44
	总计	128	90	68	286

来源: Henry Valen and Willy Martinussen, Velgere og politiske frontlinjer (Voters and Political Front Lines), Oslo: Gyldendal Norsk Forlag, 1972, p. 214

- a. 分析两次不同时间代表的两个变量的关系。
- b. 比较表右上角的 $45 + 15 + 23 = 83$ (人)与表左下角的 $24 + 10 + 4 = 38$ (人), 你能了解到什么?

CHAPTER 13



13.1 偏 ϕ : 三个分类型变量

13.2 数值型变量的多元回归

13.3 用一个哑元作多元回归

13.4 双因子方差分析

13.5 建立因果关系

13.6 小 结

多元分析



社会科学家常常寻找人们的性别与他们如何投票之间的某种关系；不过很可能实际上并不是性别影响了投票。也许还有其它的一些因素在起作用。在本章中我们特别地把收入作为另一个影响投票的分类变量来研究。

一份小吃中的所含热量数量是由很多因素决定的。在第十章中我们考察了脂肪含量对热量的影响，现在我们想考察是否胆固醇和钠也对热量有影响。多元回归分析就能帮助我们解决这类问题。

一个人开车去上班，从两条不同的路线驾驶会花去不同的时间，而且驾驶时间还与是否处与交通的高峰时段有关。那么路线的选择和时段的选择是怎样影响驾驶时间的呢？双因子方差分析能帮助我们解决这类问题。

对于大多数的问题而言，问题的结果或者因变量是由多个自变量来影响决定的。因此，在因变量的统计分析中，我们经常地使用多于一个自变量，对多个自变量的相关作用进行分析就是多元统计分析（multivariable statistical analysis）。

多元统计分析
考察两个或多个自变量对一个因变量的相关的影响。

通过对多个自变量的分析，我们总可以逐个分析其中每个自变量与因变量之间的关系。在关于投票倾向的研究中，我们可以先考察年龄因素，然后是性别因素，再然后是种族因素等等。但是，同时研究所有自变量对因变量的效应会更为有效和有益。那样我们就可以明白某一个变量在其它变量同时存在于分析之中时的效应。

在多元分析中，残差变量对因变量的影响减少了。这是因为我们将所有自变量的作用同时从残差变量中分离出来，而非逐一分离。

在多元统计分析中,我们想要回答这四个有关统计关系的问题:问题 1,是否某一个变量在数据中有效用?问题 2,因变量与所有自变量之间的关系有多强?每一个自变量对因变量的效用有多大?从而,我们可以知道哪一个自变量更重要,哪一个自变量不那么重要。我们经常也考虑这样的问题 3,每一个自变量与因变量之间的关系在统计意义上是否显著?

问题 4,两个变量之间是不是有因果关系?到此为止,我们仅研究过仅有一个自变量和一个因变量的情形,对于问题 4 我们还不能做什么工作。然而,通过多元分析有时有可能断定两个变量之间的关系是否为因果关系。有时候当我们把其它变量考虑进来做多元分析后,我们会发现原来似乎与因变量有关的变量实际上不产生效用。如果多元分析的结果是两个变量之间的关系消失了,那么原来那种关系就不是一种因果关系。

统计方法的选择总是决定于所考虑的变量的性质。分类型变量需要一种分析方法而数值型需要另一种分析方法。在本章中,我们考虑统计学家们发明的三种多元分析的方法。

因变量	自变量	方法
分类型	分类型	偏 ϕ 法
数值型	数值型	多元回归
数值型	虚拟型	多元回归
数值型	两个分类型	双因子方差分析

为了说明什么是偏 ϕ 法(partial ϕ coefficient),首先让我们考虑所有变量都是分类型变量的情形。大多数多元方法都是为数值型变量设计的,我们也将在此讨论这些变量。自变量可以是分类型也可以是数值型。对于数值型自变量,我们采用多元回归分析(multiple regression analysis);它是第 10 章介绍的简单回归方法的延伸。这里我们只考虑一例有三个自变量的事件,然而一个分析中往往会有多于三个的自变量。多元回归分析对用来表示分类型变量的所谓虚拟变量同样适用;我们还可以对数值型和哑元的组合做多元回归分析。最后,当所有自变量都是分类型时,我们常常用方差分析来替代多元回归分析。

13.1 偏 ϕ : 三个分类型变量

表 13.1 显示了一则有关两个分类型变量性别与投票之间关系的例子。现在我们来回答问题 2,两个变量之间的关系的强度用 $\phi = 0.21$ 来衡量。因为 ϕ 不为 0,我们就知道在这些数据中这两个变量是相关的。

这是否意味着性别对于投票产生了一种因果效应,还是有其它变量作用于此,从而我们所见的关系并非因果关系呢?例如,收入是不是一种决定男女不同投票的潜在因素呢?表格左上单元中 205 名妇女参与投票支持民主党可能是因为她们的性别是女性并且是民主党人,但也可能因为她们有相近的收入。表格的其它三个单元也有类似情形。或许事实是他们分别属于四个不同收入阶层;或许是我们这儿的 205 名低收入群众都是民主党妇女;或许那 167 名位于收入第二阶层的人都是共和党妇女,其它两个群体也有类似情况。

两个变量之间的关系消失了,我们得出结论:他们之间原有的关系是虚假的而非因果的。

停下来想一想 13.1

你能否想出这样一个有关两个变量的例子,当我们控制第三个变量时,一个变量对另一个变量的效应会消失?

偏 ϕ

在表 13.2 中我们计算出每个收入群体的 ϕ 值为 0.00。如果是取值多于两个的控制变量(例如取值为贫穷、中等、中上等、富裕的变量),那么我们画四个子表,并分别计算它们的 ϕ 值。

因为对大量的 ϕ 值难以解释,我们通过算平均值的方法来概括 ϕ 值。这种由几个部分计算的平均系数称为偏系数。在我们的例子中,由于每个偏 ϕ 为 0.00,这两个 ϕ 的平均值显然是 0.00。这样当我们控制了收入时,性别与投票之间的偏 ϕ 就是 0.00。

偏 ϕ 系数表明了当一个或多个其它变量被控制时,两个变量之间关系的强度。偏 ϕ 可以认为是每个部分内两个变量之间 ϕ 的平均值,而这些部分是根据控制变量来划分的。

当我们控制收入时,偏 $\phi = 0.00$ 也可被称为性别与投票的各组中的平均 ϕ (average within-group ϕ)。之所以这样称谓是因为我们把被研究对象按收入分成不同部分,对每个部分算出 ϕ 值,然后对这些 ϕ 值取平均。控制第三个变量时两个变量之间关系的偏 ϕ 可以按照公式 13.1 计算。

作为以上结果的总结:

$$\phi(\text{性别与投票之间}) = 0.21$$

$$\phi(\text{控制收入时的性别与投票之间}) = 0.00$$

因为偏 $\phi = 0.00$,所以我们认为当考虑收入因素时,性别和投票之间的关系消失了。偏 $\phi = 0.00$ 就表明了性别和投票之间没有因果关系。而当我们控制其它变量时,因果关系并没有消失。综上所述,两个变量之间原有的关系是虚假的。

我们还可以将收入作为自变量来研究收入和投票之间的关系。先个别研究,然后控制性别因素来研究。考察收入和投票之间的关系时,我们先将数据罗列在一个表示这两个变量频率的表中(表 13.3),在这里收入和投票之间的 ϕ 值为 0.45。为了控制性别因素,我们分别为女性和男性画两个分表(表 13.4)。我们先对每个分表计算出 ϕ 值,然后把这两者做平均。当控制性别因素时,收入和投票的偏 ϕ (或组内 ϕ) 为 0.40。这个均值是对女性 $\phi = 0.48$ 和男性 $\phi = 0.29$ 的加权平均(这个均值更接近于女性 ϕ 的主要原因是例子中女性的人数比男性多)。

表 13.3 收入与投票之间的关系

投票		收入		合计
		贫穷	富裕	
	民主党	177	146	323
	共和党	57	346	397
	合计	228	492	720

$$\phi = 0.45$$

表 13.4 控制性别因素时收入与投票之间的关系

投票		性别 = 女性			投票		性别 = 男性		
		贫穷	富裕	总计			贫穷	富裕	总计
	民主党	153	52	205		民主党	24	94	118
	共和党	44	123	167		共和党	7	223	230
	总计	197	175	372		总计	31	317	348

$$\phi = 0.48$$

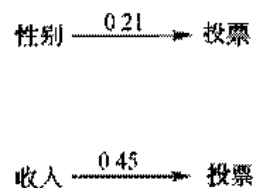
$$\phi = 0.29$$

控制性别因素时性别与收入的平均 $\phi = 0.40$ 。

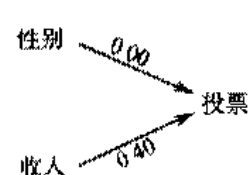
原来收入与投票之间的 ϕ 为 0.45, 当我们控制性别后, 收入与投票的偏 ϕ 大体保持着原有水平(0.40)。从而收入因素并未消去, 它仍然是一个重要的因素, 所以在收入与投票之间可能存在着某种因果关系。我们之所以用可能二字是因为也许还有其它的控制因素使得这种关系消失, 那将意味着这种关系根本不是因果关系。

对于性别与投票和收入与投票之间关系强弱的分析分别总结在图 13.1 和图 13.1a 中。而图 13.1b 则体现了我们做多元分析时发生的情形。就多元分析而言每个箭头表示了分析中其它变量在场时相应变量的 ϕ 。这样当性别是唯一的自变量时, 它与投票之间的 ϕ 值为 0.21。但是, 当收入因素出现时, 性别与投票之间的偏 ϕ 为 0.00, 这样两个变量之间的关系消失了。相似地, 当收入是唯一自变量时, 它与投票之间关系的强度为 0.45, 而当性别因素出现时, 这种强度稍稍减少到 0.40。

(a) 两个二元变量分析



(b) 一个多变量分析

图 13.1 两个二元变量分析的 ϕ 和多变量分析的偏 ϕ 。

当控制收入时, 性别与投票的偏 ϕ 跟原来的 ϕ 是不同的, 原因是性别与收入和收入与投票之间有关系。在表 13.2 的两个列联表中, 我们发现在第一个表中贫穷的人倾向于支持民主党而富裕的人倾向于支持共和党。在第二个表中我们发现女性一般比较贫穷而男性一般比较

富裕。此处起作用的因果机制是性别影响收入而收入影响投票。我们所观察到的性别与投票之间的关系是性别对收入的效应和收入对投票的效应的附属品。这三个变量之间的关系表示在图 13.2 中。

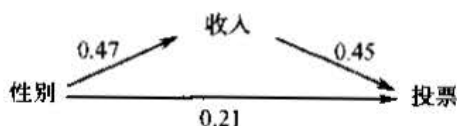


图 13.2 性别、收入和投票之间关系的强度(ϕ)。

13.2 数值型变量的多元回归

多元回归是一种用于研究多个数值型自变量与一个数值型因变量之间关系的统计方法。

对于一个数值型因变量与多个数值型自变量,我们可以用一种名为多元回归分析(multiple regression analysis)的统计方法来同时分析因变量与所有自变量之间的关系。

多元回归分析是目前多元统计中最常用的方法。由于它需要大量的计算,所以在计算机统计软件程序广泛普及之前,多元回归分析用得不是很多。今天,统计软件使得迅速地做多元回归分析成为可能,从而我们现在可以研究以前很难考察的很多复杂关系。但是,多元回归分析容易被许多没有足够统计知识的人误用。在多元回归分析中很多东西会发生错误,尤其是当分析没有被正确操作时,所得的结果可能会严重误导。

多元回归对市场学、广告学、公众关系及其它许多应用研究领域变得越来越重要。例如,广告人试图使客户产品的特殊人口统计特征与促销活动媒体的对象特征相匹配。如果 Jaguar 汽车的拥有者是 50 多岁的富人,并且大多数是白人男性,那么它就不适宜在 Playboy、Ebony 和 Vague 这些杂志做广告。多元回归分析有助于分析年龄、收入和种族在购买 Jaguar 这件事上所扮演的角色,从而影响到广告策略的选择。

另外一个关于健康职业(the health Professions)研究的例子也从多元分析受益匪浅。医学研究者用多元分析的方法去识别关于预测患病率、治疗和康复率时的特征。例如,最近的研究表明对于某类乳腺癌,如果其它变量保持不变,仅切除癌体(lumpectomies)和全乳房切除(mastectomies)从长期康复率的预测来看是一样成功的。

问题 1. 数据中的关系是什么?

在第 10 章中,我们研究了几种小吃中热量与脂肪含量的关系。在同样的食品中我们还会发现胆固醇的含量和钠的含量,所以在这里我们把脂肪、胆固醇和钠作为自变量而把热量作为因变量做多元回归分析。通过这三个变量,我们可以更好地理解热量是怎样被决定的。即使我们只有为数甚少的观测值,这个例子仍然能说明多元回归中所出现的情形。

首先让我们检查一下这些数据是否适合于用多元回归,以及数据中是否有某种关系。每个自变量应该与因变量有关系但自变量之间应该相互没有关系。做这件事的一个好办法是看看每对变量的散点图。脂肪与热量有较强的正相关,而胆固醇与热量关系较弱。第三个变量钠与热量有正相关,但又比脂肪的正相关要弱。所有这三种关系都适合线性关系,故我们可以做回归分析。但脂肪跟热量和钠都有关系,这可能产生一些困难(由于它们自己之间有关系,称其为独立的自变量就可能会有误)。

问题 2a. 这种关系的形式是什么? 偏回归系数

当我们通过做多元回归分析考察脂肪、胆固醇和钠是怎样共同决定热量时,我们发现这样的方程:

$$\text{热量} = 21.3 + 12.6 \text{ 脂肪} - 0.11 \text{ 胆固醇} + 0.18 \text{ 钠}$$

这个方程告诉我们当另两个变量被考虑进来时,这三个变量各自是怎样跟因变量发生联系的。第一项 21.3 是截距,它表明我们估计一种不含脂肪、胆固醇和钠的食品会有 21.3 卡热量。然而这种假想的食品其它值都取 0,位于数据范围之外,所以从这种角度说 21.3 是毫无意义的数。但是,在代入三个自变量的实际值时,我们需要 21.3 这个值以获得正确的热量值。

方程中剩下的三个数是偏回归系数。它们显示了当两个变量在分析中出现并被控制时另外一个变量所起效应的信息。脂肪的偏回归系数 12.6 告诉我们:当我们控制胆固醇和钠时,两份在脂肪含量上相差 1 克的食物相差 12.6 热量。换一种说法是两份具有相同胆固醇含量和钠含量而脂肪含量相差 1 克的食物相差 12.6 热量。我们通过选取在胆固醇和钠含量上具有相同值的食物来保持这两个变量为常数,然后我们考察当这些食物在脂肪含量上相差 1 克时会发生什么情况。因为系数 12.6 是正的,我们不难知道脂肪含量较高的食物热量值也会较高。

为了使我们确信两份食物将相差 12.6 热量,有一个办法就是计算出数来。设两份食物都含有 100 毫克的胆固醇和 300 毫克的钠,但是,一份含有 11 克脂肪而另一份含有 10 克脂肪。对这两份食物所含热量的估计是:

$$\text{食物 1 的热量} = 21.3 + 12.6(11) - 0.11(100) + 0.18(300) = 202.9$$

$$\text{食物 2 的热量} = 21.3 + 12.6(10) - 0.11(100) + 0.18(300) = 190.3$$

$$\text{差} = 12.6$$

预计第一份食物的热量值为 202.9,第二份食物为 190.3。这两份食物在脂肪含量上相差的 1 克转变为热量值相差 12.6 卡。对于胆固醇和钠含量的其它取值会得到同样结果。

下面是对偏回归系数 12.6 的另一种理解。为了控制胆固醇和钠这两个因素,我们把数据分成小组使得每组中的观测值对胆固醇和钠均取相同值。然后我们在每组中画出热量对于脂肪的散点图,并做回归分析。这样我们将在每组中分别得到一个回归系数,对所有系数取平均,则得到均值 12.6。

相似地,偏回归系数 -0.11 表明了控制脂肪和钠含量时胆固醇的效应。两份具有相同脂肪和钠含量而胆固醇含量相差 1 毫克的食品将平均相差 0.11 卡。这个系数是负的,所以胆固醇含量越多的食品热量值越少。最后,偏回归系数 0.18 表明了控制脂肪和胆固醇的含量时钠的效应。两份具有相同脂肪和胆固醇含量而钠含量相差 1 毫克的食品,钠含量较多的那一种

会平均多出 0.18 热量。

偏回归系数是当我们控制其它所有自变量并使它们保持常值时,某一个变量的系数。它是当我们按分析中所有其它自变量的值分组后各个组内回归系数的平均值。

比较回归系数 三个偏回归系数 12.6, -0.11 和 0.18 的大小是不同的,但我们不可能通过比较它们而得出任何有关某个变量比其它变量更重要的结论。这些系数有不同的单位,比较这三个系数就像比较苹果和橘子一样。尽管每单位脂肪有 12.6 卡热量,每单位胆固醇有 -0.11 卡热量,每单位钠有 0.18 卡热量,因为这些单位的不同,它们大小之间的差别就无从得知了。

改变回归系数 图 13.3 比较了我们对所有三个变量做多元回归分析与做单一回归分析时每个变量的回归系数是怎样改变的。脂肪的系数减小了一些,胆固醇的系数几乎消去,钠的系数也减小了。这表明由于脂肪和钠的出现,胆固醇几乎与热量无关。而且图中体现了上述分析中对变化比例的解释(在多重相关系数的部分我们将讨论这一点)。

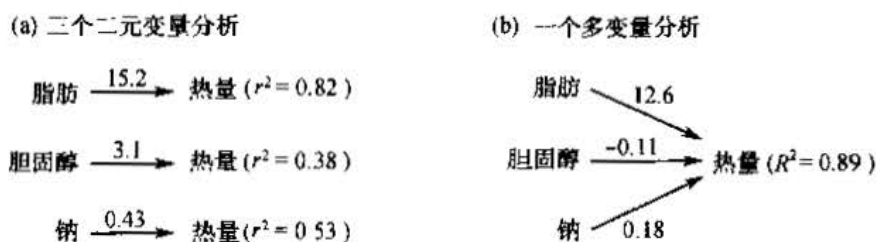


图 13.3 三个二元变量分析的回归系数和一个多元分析的偏回归系数。

上述系数改变的原因是脂肪、胆固醇和钠三个自变量之间互相有关系。我们在图 13.3 关于自变量的分枝图中可以看到这种相互关系。在计算这些关系的强度时,我们发现这样三个相关系数:

$$\begin{aligned}\text{脂肪与胆固醇的 } r &= 0.69 \\ \text{脂肪与钠的 } r &= 0.59 \\ \text{胆固醇与钠的 } r &= 0.41\end{aligned}$$

共线性存在于两个或多个相互关联的自变量之间。

每当我们引入一个与分析中变量有相互关系的变量时,回归系数就发生了改变。没有哪一个变量会保持唯一的系数值;系数会随着我们所用的其它变量而改变。这种现象称为自变量之间的**共线性**(collinearity)。我们试图避免共线性,但这常常是不可能的。为了研究因变量,我们常常可以选择自变量的值。在这种实验情形下,避免共线性还有一点希望。当我们能决定时,我们去选择那些使自变量之间没有相互关系的值。

问题 2b. 这些关系的强度有多大? 偏相关系数

就像一个偏回归系数表明了有其它变量出现时某个特别变量的效应一样,一个偏相关系数(partial correlation coefficient)表明了当我们控制其它变量时两个变量之间关系的强度。我们可以控制一个或多个其它变量。控制其它变量时两个变量之间的偏相关系数按公式 13.1 计算。

对于小吃数据, $r = 0.91$ 用来表示热量与脂肪之间关系的强度。当我们控制胆固醇和钠时,我们发现热量与脂肪之间的偏相关系数为 0.82。相似地,热量与胆固醇之间的 $r = 0.62$,但当我们控制了其它两个变量时,偏相关系数降为 -0.05 。最后,热量与钠之间的 $r = 0.73$,但当我们控制了其它两个变量后,偏相关系数降为 0.58。当我们控制脂肪和钠时,热量与胆固醇之间的关系几乎消失。因此热量与胆固醇之间的关系被认为是一种虚假的关系。这个结果跟回归系数产生的结果是一致的。

问题 2c. 总体关系的强度有多大? 多重相关系数

对小吃数据做多元回归是否有意义? 脂肪、胆固醇和钠这三个变量是怎样决定热量的? 这些问题可以通过考察所谓多重相关系数 R 来回答。

多元回归分析赋予每个自变量一个偏回归系数。这些系数与变量可以共同组成一个公式,正如我们在前面问题 2a 中所见。对所有小吃:

$$\text{热量} = 21.3 + 12.6 \text{ 脂肪} - 0.11 \text{ 胆固醇} + 0.18 \text{ 钠}$$

这个方程看起来很动人,但是对一份像家常面包圈这样实际日常生活中的小吃,它起多大的作用呢? 当我们把面包圈中脂肪、胆固醇和钠的数值代入方程并把每项加起来,结果是否正好是一份面包圈中所含热量的值呢? 对于一个面包圈,脂肪 = 8,胆固醇 = 25,钠 = 210。我们将这些数目代入回归方程,算出预测值:

$$21.3 + (12.6)(8) - (0.11)(25) + (0.18)(210) = 156.7 \text{ 热量}$$

但是,查表 10.1 中的热量变量,我们发现面包圈具有 164 热量而不是 156.7 热量,所以我们没能得到准确值。然而,我们已经很接近了。我们对数据文件中每种其它的食品可以从回归方程中通过相同的计算得出热量的预测值。卡路里的预测值越接近实际值说明我们的分析做得越好。

考察预测值与实际值接近程度的一种方法是对两组数据画如图 13.4 那样的散点图。对每种食品我们把实际的和预计的热量值画成一个点。如果预测值与实际观测值相等,那么相应的点会位于图中标明的 45 度线上。两个数值的差别越大,相应该点会离这条直线越远。

这个图表明除了几个离群点外大多数点聚集在 45 度线附近。一个离群点是图中右部表示苹果饼的点,它位于直线的下方并与直线之间有一段距离。这表明预计热量值比实际热量值低一些。即苹果饼具有比由脂肪、胆固醇和钠预测值更多的热量值。这个方程对于预计苹果饼所含热量不太适合说明我们需要一个或多个另外的自变量(也许“罪魁祸首”是糖或肉桂)。尽管如此,总的说来由回归方程得出的结果还是给人以很深的印象:大多数预测值跟观测值相当接近。

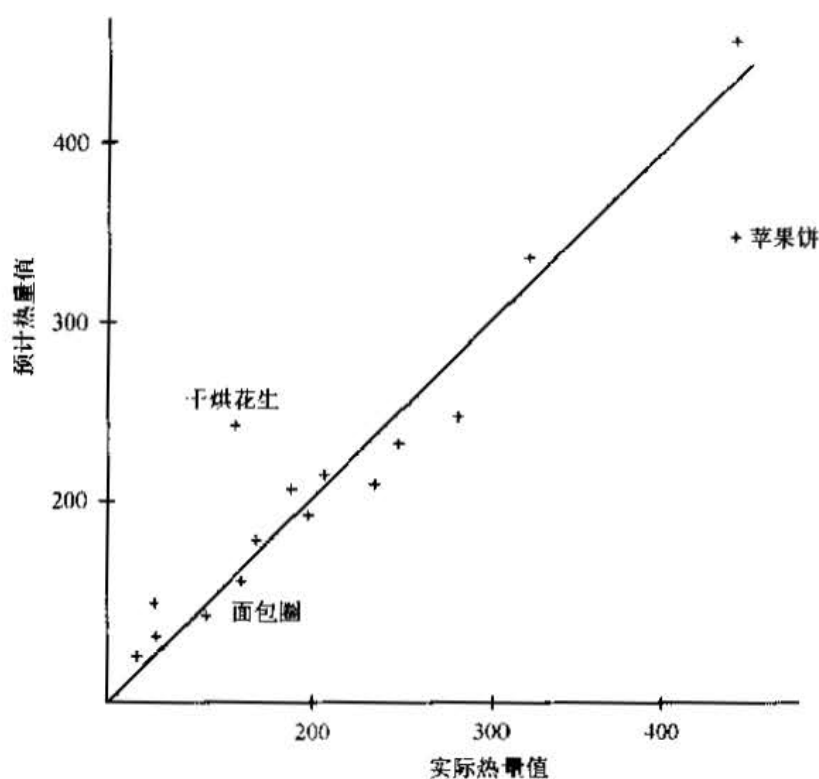


图 13.4 小吃中预计和实际热量值的散点图。

为了考察预测值和观测值吻合的程度,我们用第 10 章介绍的两个数值型变量之间的相关系数。回想一下,如果预测值与观测值相等,则这些点会落在 45 度线上,那么相关系数等于 1.00;而这些点越是分散,相关系数就越小。对于图 13.4 中的数据,相关系数等于一个相当高的值 0.94。这是因变量的观测值与预测值之间的相关系数。由于这些预测值是由三个自变量算出的,我们说已经发现了因变量的观测值与所有自变量(脂肪、胆固醇、钠)的联合效应之间的关系。这个相关系数是如此特别,我们给它一个新的名字和记号,称为**多重相关系数**(multiple correlation coefficient),用 R 表示。(回忆任何两个变量之间的普通相关系数用小写 r 表示。)

多重相关系数 R 度量因变量的观测值与由自变量由回归方程算得的预测值之间关系的强度。 R 的值域从 0 到 1。

多重相关系数的平方 R^2 等于 0.89,它意味着三个自变量共同解释了热量值差异的 89%,残差变量解释了剩下的 11%。另外一种说法是预计热量值的差异为观测热量值的差异的 89%。观测值之间的差异更大是因为它们中包括有残差变量的效应。

对于热量值差异的解释在图 13.3 中体现出来。这个图表明:单独地说差异的 82% 归因于脂肪,38% 归因于胆固醇,53% 归因于钠。这三个变量联合起来比三个变量中任何一个对差异的解释要多,但对预测的改进并不大。例如,单独地分析,差异的 82% 归因于脂肪,我们可

以这样设想,另外两个变量产生另外的 7% 的效应从而总计达到 89% 的效应。然而单独地分析胆固醇和钠的值均远大于 7%,而且三个变量单独分析的值 82%、38% 和 53% 的总和远远超过 100%。我们之所以不能这样把三个变量各自的百分值加起来去分析它们总的效应是因为三个变量之间是相关的,即它们是共线性的。

问题 3. 总体中的关系?

当我们只有一个观测数据样本并且想从中得出关于产生样本的总体的结论,我们就必须做某种统计推断。就如在先前的章节中所做一样,我们需要做假设检验或者构造置信区间。两者之中,假设检验是多元回归分析的一种更常用方法,我们对总的 R 和每个变量单独的回归系数都做这项工作。

R 的假设检验 零假设“总体中的多重相关系数 R 为 0”表示自变量合起来对因变量没有效应。这一点是难以想像的,我们对于变量的了解是如此的少以致于我们选择的所有自变量联合起来对于因变量毫无效应。这个没有效应的零假设常常没有什么意思,所以常常被拒绝。

在小吃的例子中,样本的 R 为 0.94。由假设检验,我们想问的是:一个如此大的或者更大的 R 值能否来自多重相关系数为 0 的总体的一个样本? 即在小吃的总体中脂肪、胆固醇和钠是否真的对热量没有效应? 我们通过 $R=0.94$ 的 p -值来回答这个问题。

我们需要用四个标准统计变量之一的 F 变量来计算 R 的 p -值。在分析中用三个自变量和 16 个观测值对 R 为 0.94 算出 $F=31.3$,自由度为 3 和 12 (F 及自由度的计算公式在本章末由公式 13.2 给出)。 F 为 31.3 或更大的概率非常小只有 0.000006;如果我们考察来自多重回归系数为 0 的 1000000 个不同的样本,单凭机会,只有 6 个样本的 R 为 0.94 或更大。这是不利于自变量没有效应的零假设的有力证据,故我们拒绝零假设。

每个变量的假设检验 这个 R 的值是高度显著的,但它并没有向我们展示是否所有三个自变量均具有统计上显著的效应或者是否只有它们中的一些或一个有效应。为了解决这个问题,我们对每个变量分别做假设检验。

在小吃的总体中脂肪含量是否与热量有关呢? 小吃中脂肪的偏回归系数为 12.6,总体数据中相应的系数可以为 0 吗? 把这个问题作为零假设,我们把 12.6 变为 t -变量的一个值。统计软件算出 $t=4.98$,自由度为 12, t 为 12.6 或更大值的概率为 0.00016。这样,在来自脂肪系数为 0 的总体的 100000 个样本中我们只能得到 16 个 t 为 12.6 或更大。所以,一个回归系数为 12.6 或更大的样本来自相应系数为 0 的总体的可能性很小,我们拒绝零假设并得出总体中脂肪含量与热量有关的结论。由于总体的回归系数不为 0,我们可以去估计这个系数,并得到偏总体回归系数的 95% 置信区间为 7.1 到 18.1。

类似地,胆固醇变量的偏回归系数为 -0.11。如果我们用处理脂肪变量相同的程序,可以算出胆固醇相关系数的 p -值为 0.44。这个 p -值如此大,故我们不拒绝没有效应的零假设,总体中相应的系数很可能为 0。我们得出热量与胆固醇可能无关的结论。最后,对钠变量,偏回归系数为 0.18 的 $t=2.46$ 。这个 t 的 p -值为 0.015,它已经足够小,从而使我们拒绝零假设并得出小吃中热量与钠变量有关的结论。

值得庆幸的是所有这些计算均可通过适当的统计软件在计算机上完成。我们在此不给出手工计算的公式,那简直太繁琐了!

我们在一个多元分析中常常去掉那些不显著的变量,从而使分析尽可能简化。除去胆固

醇变量,剩下两个变量脂肪和钠可以得到几乎相同的 R 值。所以把热量当作仅仅与脂肪和钠相关来研究并没有什么损失。

13.3 用一个哑元作多元回归

按照学院当局公布的数据,1994—1995 学年 Swarthmore 学院女性正教授的平均薪水为 \$71100 而男正教授的平均薪水为 \$76300。这就有了 \$5200 的差别,而使我们提出这样的问题:这个学院在付给教授薪水上是否存在性别歧视。当然上述两件事之间是有差别的。但是,在作出结论之前,我们在性别与薪水关系研究中先控制其它的变量。

小吃的例子当中只有数值型变量,这里我们希望像第 10 章一样把数值型变量与分类型变量结合起来。我们不是通过比较均值来研究性别与薪水之间的关系,而是用一个哑元来代替性别去研究数据。正如我们在第 10 章所做的一样我们对每名女性赋值 0,对每名男性赋值 1。数据的散点图看上去如图 13.5 一样。薪水标示在纵坐标上,单位为 \$1,000;性别标示在横坐标上,取 0 和 1 两个值。由于学院没有公布每个人的薪水数据,我们无从得知散点图的实际形状。

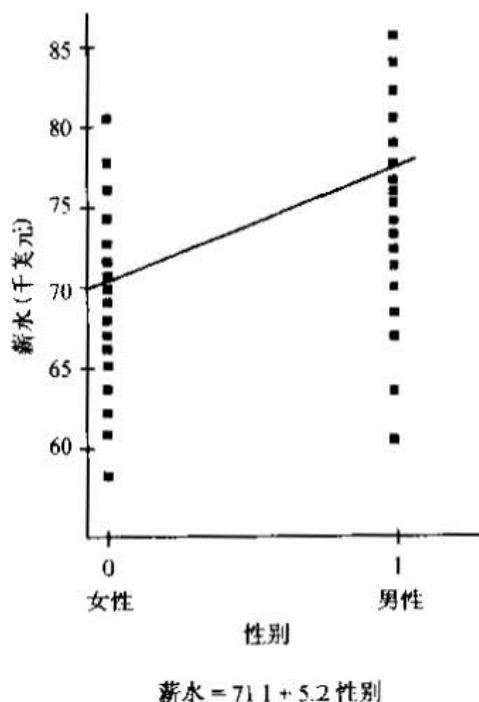


图 13.5 薪水与性别的散点图。

散点图把女性的薪水表为左边的点,男性的薪水表为右边的点。图中还画出了通过这些点的回归线。这条直线在 71.1 处与薪水的纵轴相交,即回归直线的截距为 71.1,它等于女性的平均薪水。这条直线的斜率是性别的回归系数。此处斜率为 5.2,表示女性平均薪水与男性的差别。当我们有两组点时,就像这里一样,回归直线总是是连接两组因变量均值的直线。

薪水上的差别表明在性别与薪水两个变量之间存在着某种关系。这是否意味着学院在付薪水中有歧视呢?在找到其它的可以被控制的有关变量以前我们无从得知。

统计学定理中没有什么可以指导我们去选择控制变量。这种选择基于我们对所研究事物的常识和手中的数据。本例中一个可能的控制变量就是年龄。也许男性的年龄普遍地大一些,工龄长一些,从而获得较高的薪水。这样,为了找出控制年龄时性别的效应,我们应当将薪水作为因变量而性别和年龄都作为自变量来做多元回归分析。当然,这个分析也能得到控制性别时年龄的效应。

假如我们拥有每位教授薪水、年龄和性别的数据,做多元回归分析将得出以下结果:

$$\text{薪水} = 40 + 0.0 \text{ 性别} + 0.5 \text{ 年龄}$$

这个结果中最让我们惊讶的是控制年龄时性别的偏回归系数为 0.00。在我们控制年龄之前,性别的回归系数是 5.2。由于现在这个系数为 0.00,性别对薪水的效应消失了。一个男性与一个和他同龄的女性之间没有薪水的差异。

为了理解对年龄的控制,我们可以把数据分成不同的年龄组(所有 40 岁的,所有 41 岁的,所有 42 岁的等等),然后对每组画出性别和薪水的散点图。图 13.6 显示了几个这样的散点图。由于在每个散点图中我们只考察了一个特别的年龄组,所以每个散点图中观测值的数目都比数据中的观测值少。所有这些人的年龄相同,故数据点在每个散点图中的分布和年龄没有关系。

对于每个散点图,回归直线分别穿过了女性与男性的薪水平均值。每组内的回归线几乎都是近似水平的,效率为 0.00,这意味着每组中性别与薪水之间没有关系。所有年龄组内的斜率之平均值为 0。这是控制年龄时性别的偏回归系数。由于斜率均值为 0,直线平均为水平的,故当我们控制年龄时男性和女性的薪水没有差异。有时候像我们对待偏 ϕ 一样把偏回归系数当作组内回归系数的平均值是很有帮助的。

幸运的是当我们想要计算偏回归系数时并没有必要把数据真的分成很多组。之所以这样说是因为如果观测值样本一开始时就很小,那么在每组内将没有足够多的数据可供分析。相反地,如果数据一开始时就很多,即使对较小的组也会太过繁琐难以处理。我们在此描述的步骤可以转化为计算偏回归系数的数学方程和公式。对于有很多变量的情形这些公式很繁琐,不过把它们编到统计软件包里用起来就容易了。

另外一种说明问题的方法是在年龄和薪水的散点图上用不同的记号分别表示男性和女性的数据点(图 13.7),然后在每个组我们用薪水对年龄做回归。为了做到这一点,我们必须简化数据以使所有的女性比所有的男性年轻。一种控制年龄的方法就是选择年龄相同的男性和女性。假设我们选择了总体的平均年龄。如果延长这两条回归直线,我们对所有男性和女性的预计薪水会是一样的。这是因为图中两条线的方程是一样的,具有相同的截距和斜率,唯一的差别是表示女性的直线位于表示男性的直线的左下方。

对这个例子中薪水差异的简单解释是男性年龄平均上比女性大从而比女性得到更多的薪水。但是,我们常常不能如此容易地找出合适的控制变量,或者对于究竟应该用哪个变量认识太晚。如果我们在一个抽样调查中收集数据时没有收集将来可能会用作控制变量的数据,以后再回头去找被访者收集额外数据是非常困难甚至是不可能的。因此,在为收集数据作准备的时候,事先考虑清楚以确定收集到所有在分析中打算用到的控制变量的数据,这一点是非常

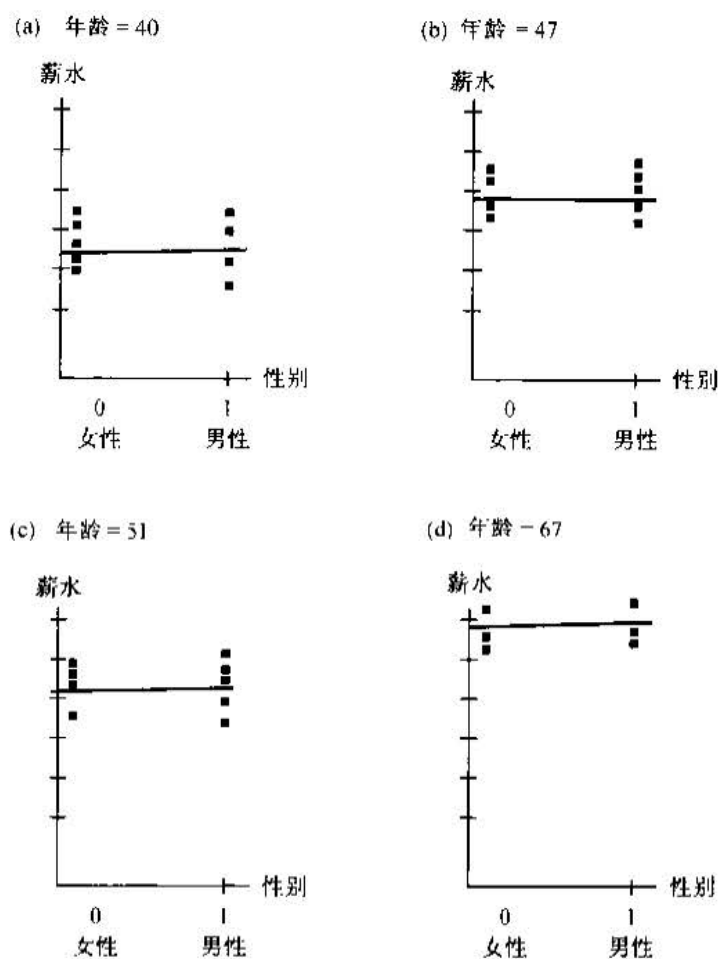


图 13.6 几个年龄组中薪水和性别的散点图。

重要的。

现在你可以发现哑元在多元回归分析中是非常有用的, 因为可以通过哑元把数值型和分类型变量同时当作自变量分析。此例中我们考察的分类型变量只有两个类别, 但是许多其它的分类型变量会有多个类别。例如, 宗教信仰就可能有以下四个类别: 天主教、犹太教、新教和其它。对于多个类别我们用多个哑元, 好的统计软件程序会自动构造哑元。

13.4 双因子方差分析

设想下面关于一位女士开车上班的例子。Sally Jones 女士可以选择从 Main Street 或是从 High Street 去上班, 她也可以选择交通高峰时段或是在其它时段去上班。她想知道的是: 她怎样可以最快地到达。作为一个统计模型, 这个问题的两个自变量是路线和一天中不同的时段, 而因变量是从家到工作地点驾驶时间的长度。是否某条路线比另一条更好? 是否某个时段比另一个更好? 或者是某条路线在某个时段更好, 而另一条路线在另一个时段更好? 研究

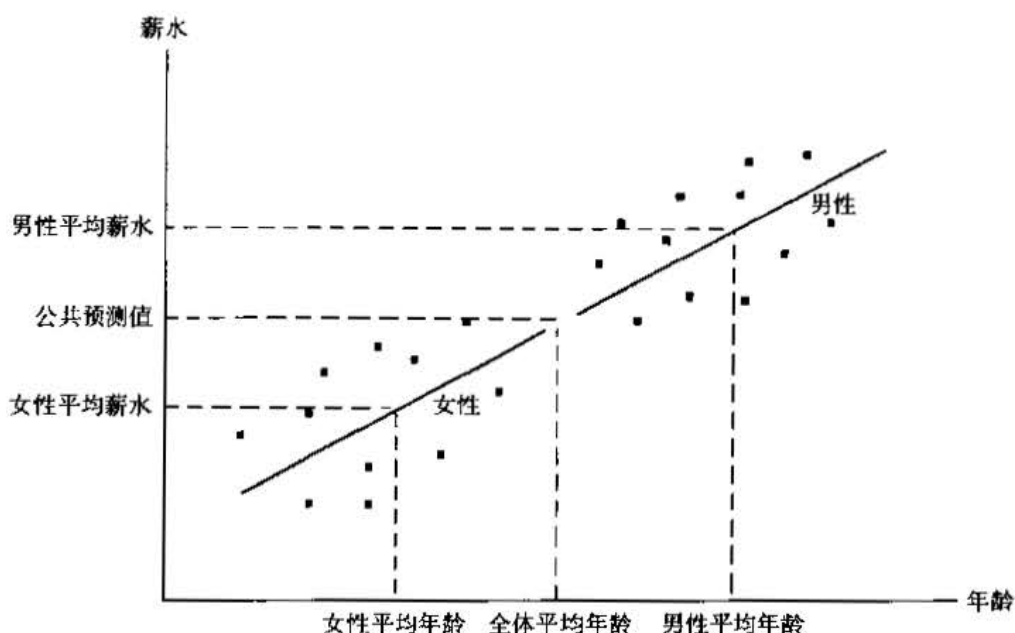


图 13.7 具有女性和男性回归直线的年龄与薪水散点图。

路线和时段对驾驶时间效应的一种方法是在不同的时段到每条路线上驾驶几天并测出每次所花时间。

双因子方差分析是有关两个分类型自变量对一个数值型因变量效应的分析。

路线和时段都是分类型变量,都有两个值。路线取值为 Main Street 和 High Street,时段取值为高峰或非高峰,驾驶时间是一个数值型变量,以分钟计算。当自变量都是分类型而因变量是数值型时,我们可以对分类型变量引入哑元做多元回归分析。然而,我们常常用双因子方差分析来代替它。在第 12 章中我们讨论了只有一个自变量的方差分析。这里的问题需要双因子方差分析是因为有两个自变量都影响因变量。当有三个分类型自变量时我们就用三因子方差分析,以此类推。

为了收集数据, Jones 女士在接下来的四周时间内随意选择不同的路线和时段驾驶。这个计划使她对每种可能的驾驶组合有 5 次不同的驾驶机会。驾驶时间的原始数据中有 20 个观测值(在变量的每种组合中有相同数目的观测值使双因子分析变得容易些,本书中我们不讨论观测值数目不同的情形)。

驾驶的平均时间显示在表 13.5 中。这个表显示驾驶时间的总体均值为 20 分钟。从 Main Street 驾驶的平均时间为 22 分钟,这个值比总平均时间多了 2 分钟;从 High Street 驾驶的平均时间为 18 分钟,这个值比总平均时间少了 2 分钟。相似地,在高峰时段驾驶的平均时间为 23 分钟,这个值比总平均时间多了 3 分钟;在非高峰时段驾驶的平均时间为 17 分钟,这个值比总平均时间少了 3 分钟。

表 13.5 不同路线和时段的平均驾驶时间

时 段	路线			
		Main Street	High Street	均值
	高峰	25 分钟	21 分钟	23 分钟
	非高峰	19 分钟	15 分钟	17 分钟
	均值	22 分钟	18 分钟	20 分钟

仅对于时段的单因子分析

研究时段与驾驶时间长度之间的关系就是研究一个分类型自变量与一个数值型因变量之间的关系。从第 11 章中我们得知,这需要单因子方差分析。表 13.6 显示了分析的结果。时段的平方和是 180.00,总平方和为 315.98,由此得出 $R^2 = 180.00/315.98 = 0.57$,则时段因素解释了驾驶时间差异的 57%,这个数的平方根 $R = 0.75$,故我们认为两个变量之间具有较强关系。

表 13.6 时段的单因子方差分析

来源	自由度	平方和	均方	F-比	p-值
时段	1	180.00	180.00	23.87	0.00012
残差	18	135.98	7.554		
总计	19	315.98			
$R^2 = 180/315.98$ 和 $R = 0.75$					

我们知道高峰时段驾驶多花 3 分钟时间而非高峰时段驾驶少花 3 分钟时间,但我们不知道这些差别在统计意义上是否显著,也许这些差别仅仅是随机的。 F -比等于 23.87,这个 F 的 p -值很小,为 0.00012。如果在驾驶时间上无差别的话,在 100000 次中只有 12 次的 F -值是这样大或者更大。这意味着如果没有差别的零假设为真,这些数据是不太可能的。因此我们拒绝零假设,则两个时段上驾驶时间的差别在统计意义上是显著的。

仅对于路线的单因子分析

为了研究路线与驾驶时间的关系,我们也做单因子方差分析。表 13.7 显示了分析的结果: $R^2 = 0.25$, $R = 0.50$ 。故路线因素解释了驾驶时间差异的 25%,两个变量之间具有中度关系。

表 13.7 路线的单因子方差分析

来源	自由度	平方和	均方	F-值	p-值
路线	1	80.00	80.00	6.10	0.024
残差	18	235.98	13.110		
总计	19	315.98			
$R^2 = 80.00/315.98 = 0.25$ 和 $R = 0.50$					

类似地,我们知道从 Main Street 驾驶多花 2 分钟时间而从 High Street 驾驶少花 2 分钟,但我们不知道这些差别在统计意义上是否显著,也许这些差别仅仅是随机的。 F -值等于 6.10,

这个 F 的 p -值很小,为 0.024。这不如时段因素那样显著,但是如果路线之间没有差别的零假设为真,这些数据也是不太可能的。因此我们也拒绝零假设,则两条路线驾驶时间的差别在统计意义上是显著的。

时段和路线的双因子分析

我们可以做一个驾驶时间的双因子分析来代替两个单因子分析,这就像用两个自变量的多元回归分析去代替对每个自变量分别做的两个一元回归一样。把时段和路线作为自变量的双因子分析减少了残差变量的效应。在将时段作为自变量的单因子分析中,路线被包括在残差变量中;相似地,在将路线作为自变量的单因子分析中,时段被包括在残差变量中。但在做双因子方差分析时这两个自变量同时从残差变量中分离出来。

我们可以用一种更有组织的方式列出驾驶时间的均值。

总平均:	驾驶时间的均值 = 20 分钟
行效应:	高峰时段的效应 = 23 分钟 - 20 分钟 = 3 分钟 非高峰时段的效应 = 17 分钟 - 20 分钟 = -3 分钟
列效应:	Main Street 的效应 = 22 分钟 - 20 分钟 = 2 分钟 High Street 的效应 = 18 分钟 - 20 分钟 = -2 分钟

这些差异是两个自变量的值对驾驶时间的效应。它们也称为这两个分类型变量的行效应和列效应。例如,高峰时段是表 13.5 的第一行,在高峰时段驾驶的效应是 3 分钟,在那时驾驶比总平均时间要多花 3 分钟。因为在高峰时段和非高峰时段的驾驶时间是不同的,则时段因素对数据产生了效应。相似地,因为从不同路线驾驶时间是不同的,路线因素也有效应。

如果知道总体均值、两个行效应和两个列效应,我们就可以算出表 13.5 中剩下四个单元的均值。计算过程显示在表 13.8 中。例如,高峰时段从 Main Street 驾驶要花 $20 + 3 + 2 = 25$ 分钟。这是因为驾驶时间的均值为 20 分钟,在高峰时段驾驶会多花 3 分钟,从 Main Street 驾驶要多花 2 分钟,总计是 25 分钟。剩下的三个单元也同样计算。

进一步,如果我们计算表中第一行两个单元的均值,Main Street 和 High Street 的 +2 与 -2 抵消了,得出高峰时段的平均时间为 $20 + 3 = 23$ 分钟。如果计算表中 Main Street 列两个单元的均值,我们发现 3 与 -3 抵消了,得出从 Main Street 驾驶的平均时间为 $20 + 2 = 22$ 分钟。对非高峰时段的行和 High Street 列可以采用同样的计算。

表 13.8 由总均值、行效应和列效应计算不同路线和不同时段平均驾驶时间

		路线		均值
		Main Street	High Street	
时 段	高峰	$20 + 3 + 2 = 25$ 分钟	$20 + 3 - 2 = 21$ 分钟	$20 + 3 = 23$ 分钟
	非高峰	$20 - 3 + 2 = 19$ 分钟	$20 - 3 - 2 = 15$ 分钟	$20 - 3 = 17$ 分钟
	均值	$20 + 2 = 22$ 分钟	$20 - 2 = 18$ 分钟	20 分钟

均值还可以用图表示出来。在图 13.8 中,两条街标在横轴上,四个单元的均值用小方块表示。为了在均值之间保持联系,两个高峰时段的点用一条线连接起来;两个非高峰时段的点

用另一条线连接起来。当然,也可以在横轴上标出两个时间段并用直线把两条街的均值连接起来。这样画图的好处是图中每个均值点直接与表 13.5 中均值的位置有关。

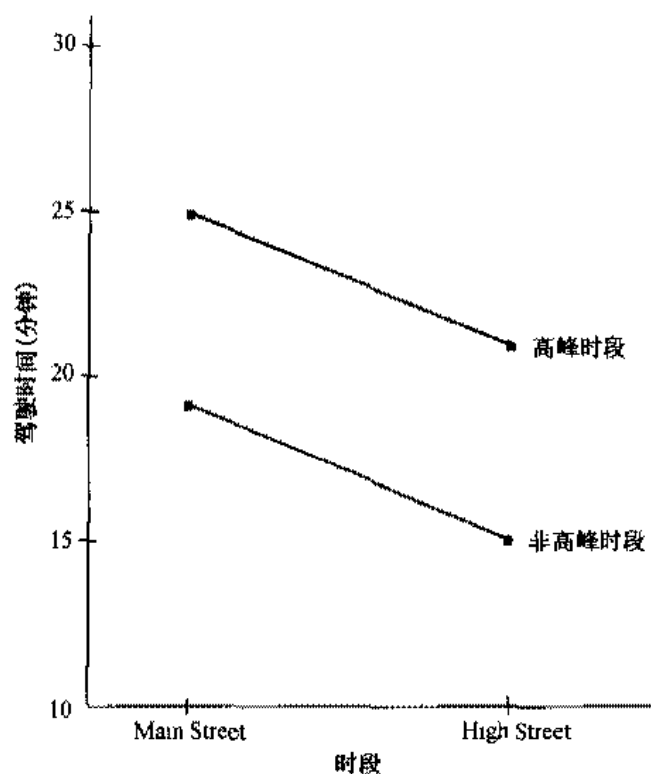


图 13.8 两条路线和两个时段的平均驾驶时间。

残差变量 不同的驾驶时间和均值如表 13.9 所示。我们早已知道在高峰时段从 Main Street 驾驶的平均时间为 25 分钟,但是从数据中我们发现每次驾驶并不是刚好 25 分钟。表 13.9 中每个单元内的观测值都不同的原因是除了路线和时段外还有其它变量影响驾驶时间。例如,某些天里 Jones 女士被阻塞在一辆学校巴士后而其它时候道路上汽车不多。

表 13.9 20 次不同驾驶的驾驶时间

		路线		均值
		Main Street	High Street	
时	高峰	26.0	18.8	23 min
		24.2	20.0	
		26.5	22.7	
		24.1	22.3	
		24.2	21.2	
		(均值 = 25)	(均值 = 21)	
段	非高峰	19.9	17.4	17 min
		16.7	17.3	
		20.7	12.7	
		20.5	15.6	
		17.2	12.0	
		(均值 = 19)	(均值 = 15)	
均值		22 min	18 min	20 min

正如我们知道的,所有其它变量的效应都在残差变量之中。在该例中,第一次驾驶花了

26 分钟,而不是均值的 25 分钟;因此残差变量加了一分钟到这次驾驶。残差变量在一次驾驶上的效应就是观察的时间和在该单元中的平均时间之差。残差变量在其它次高峰时段沿 Main Street 上的驾驶的效应为 -0.8, 1.5, -0.9 和 -0.8 分钟。类似地,我们能找到残差变量在 Sally Jones 的其它次驾驶上的效应。

问题 1. 数据中是否有任何关系? 我们已经从对每个自变量的分析中知道:Jones 女士是否在高峰或非高峰时段驾驶和从哪条路线驾驶是有差异的。因此我们可以进行下面问题的研究。

问题 2. 这些关系的强度有多大? 在表 13.6 中, $R^2 = 0.57$ 和 $R = 0.75$ 表示时段与驾驶时间的关系。类似地,从表 13.7, $R^2 = 0.25$ 和 $R = 0.50$ 表示路线与驾驶时间的关系。除开每一个自变量与因变量关系的强度,我们还可能得出自变量的联合效应与因变量之间关系的强度。这就是在回归分析中给予多重相关系数的。

我们已经得出时段变量的平方和(180.00)与路线变量的平方和(80.00)。这两个自变量的联合效应变成 $180.00 + 80.00 = 260.00$ 。这两个自变量解释了驾驶时间差异的 $260.00 / 313.98 = 82\%$,而残差变量解释了差异中余下的 18%。由于 $R^2 = 0.82$,所有自变量效应的多重相关系数 $R = 0.91$ 。于是,这两个变量合起来与驾驶时间有较强关系。

总计平方和与两个自变量的联合效应之差为 55.98,这是残差变量在双因子分析中的效应。如果我们计算了每个残差并对它们进行平方,这个平方和也将等于 55.98。双因子分析的残差平方和比任何单因子分析的残差平方和都小,这并不稀奇,因为残差变量现在不包括两个自变量中的任何一个。这些平方和列在表 13.10 中。

表 13.10 时段和路线的双因子方差分析

来源	自由度	平方和	均方	F-比	p-值
时段	1	180.00	180.00	54.66	0.000001
路线	1	80.00	80.00	24.29	0.00012
残差	17	55.98	3.293		
总计	19	315.98			

问题 3. 这些关系在统计意义上显著吗? 正如在多元回归分析中,我们可以检验是否两个自变量联合起来有显著效应,还可以对每个变量分别做检验。

对两个变量, $R^2 = 0.82$ 。由此得出 F-变量的值为 39.48,自由度为 2 和 17。这个 F 的 p-值等于 0.0000004;这即是说如果两个自变量和因变量之间没有关系,在 1000000 个 F-值中只有 4 个会这么大或更大。p-值还说明如果没有关系存在, R^2 几乎不可能大于等于 0.82。因此我们拒绝零假设。

对每个变量分别做的检验显示在表 13.10 中。跟以前的两个方差分析相比残差平方和少了一个自由度,但净效应仍然是这里的残差均方要比表 13.6 和 13.7 中的两个残差均方小。因此,双因子分析中 F-值比两个单因子分析中要大。大的 F-值产生了更小的 p-值。时段的 p-值从 0.00012 变为 0.000001 而路线的 p-值从 0.024 变为 0.00012。所有两个变量都与驾驶时间之间有统计意义上显著的关系,而双因子分析的 p-值比单因子分析中的更小。

考虑交互效应,再进行研究

接下来的一年 Sally Jones 女士重复了她的研究。新的平均驾驶时间列在表 13.11 中。在

高峰时段的驾驶时间仍然比总平均时间多 3 分钟而在非高峰时段仍然少 3 分钟。从 Main Street 驾驶仍然多花 2 分钟而从 High Street 驾驶仍然少花 2 分钟。行和列效应都没有改变。

表 13.11 一年后不同路线和时段的平均驾驶时间

		路线		
		Main Street	High Street	均值
时 段	高峰	26 分钟	20 分钟	23 分钟
	非高峰	18 分钟	16 分钟	17 分钟
	均值	22 分钟	18 分钟	20 分钟

但是,与以前的数据相比两条路线和两个时段里还是有一些新的情况,差异在四个单元的均值之间出现。在高峰时段从 Main Street 驾驶现在要花 26 分钟而不是一年以前的 25 分钟。相似地,在高峰时段从 High Street 驾驶和在非高峰时段从 Main Street 驾驶都比以前少花 1 分钟,而在非高峰时段从 High Street 驾驶多花 1 分钟。描绘四个单元里均值的图如图 13.9 所示。

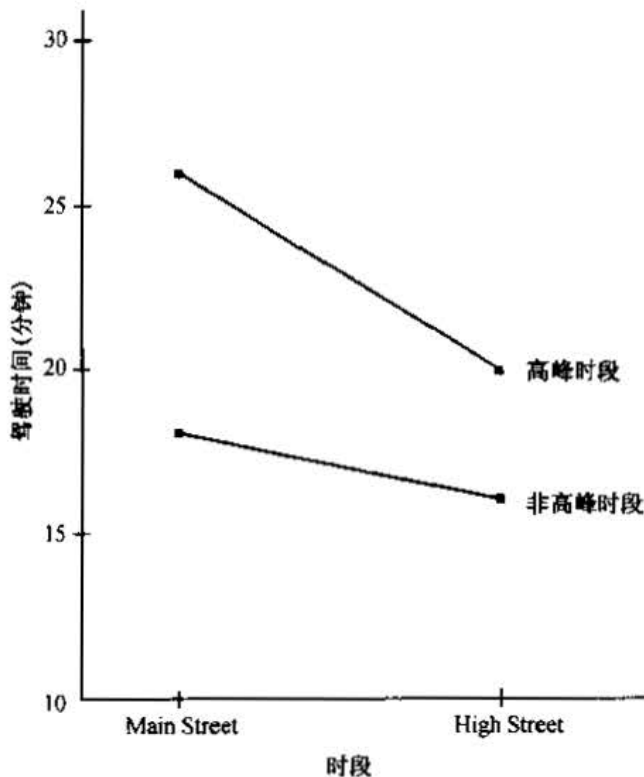


图 13.9 一年后两条路线和两个时段的平均驾驶时间。

交互效应发生时,两个变量的联合作用对因变量产生了它们各自效应之和以外的附加效应。

图 13.8 和图 13.9 之间最显著的差别是图 13.8 中表示两个时段的两条直线是平行的而在图 13.9 中它们是不平行的。图 13.8 中的两条平行线表明从 Main Street 和 High Street 驾驶时高峰时段与非高峰时段之间驾驶时间的差别都是 6 分钟,而图 13.9 中的两条直线表明从 Main Street 驾驶这两个时段之间相差 8 分钟,从 High Street 驾驶只相差 4 分钟。现在在高峰时段从 Main Street 驾驶的联合效应比这两个因素各自效应的和要大。这个路线和时段的组合给驾驶时间多加了 1 分钟,这种附加效应称为交互效应 (interaction effect)。这两个变量联合产

生了它们各自效应之外的附加效应作用在因变量。

在图 13.9 中有三个变量影响驾驶时间。像以前一样,它们中有路线变量和时段变量,但是现在还有路线/时段交互变量。交互变量的平方和等于 20.00,并且表 13.12 中列出了其它平方和。交互变量解释了 $20.00/335.98 = 6\%$ 的驾驶时间差异。这样交互变量与驾驶时间的关系不是很强, $R = 0.24$ 。公式 13.3 显示了怎样计算双因子方差分析需要的不同平方和。公式 13.4 显示了怎样得出与每个平方和的相应自由度。

表 13.12 时段、路线,以及交互的时段/路线联合作用的双因子方差分析

来源	自由度	平方和	均方	F-值	p-值
时段	1	180.00	180.00	54.66	0.000001
路线	1	80.00	80.00	24.29	0.00012
时段/路线的交互作用	1	20.00	20.00	5.72	0.03
残差	16	55.98	3.499		
总计	19	335.98			

新数据中三个变量是否都跟驾驶时间有统计意义上显著的关系呢?表 13.12 回答了这个问题。所有三个变量的 p -值都很小,这意味着它们都与驾驶时间有统计意义上显著的关系。交互变量刚好好在 5% 的水平上有较为显著的效应,而另外两个变量有非常显著的效应。

13.5 建立因果关系

因果关系组成的确定依靠于统计意义上显著的结果和研究者和其他参与者的共识。多数我们用统计方法研究的实际因变量受到几个其它变量的影响,而且必须用多元统计方法分析。多元方法的一个主要特点是使得我们有可能在其它变量在场时考察某个特别的自变量的效应。如果仅由一个自变量得出的关系会在其它变量出现时消失,则原来的关系就不会是因果关系。另一方面,如果我们控制其它变量时一种关系没有消失,我们仍然不能认为这种关系是因果的;这是因为我们永远不可能控制每一个可能的其它变量;我们没有其它所有可能变量的数据。这样,即使用多元统计方法,证明因果关系仍是一个难以捉摸的工作。

停下来想一想 13.2

耶鲁大学精神病学教授 Kyle Pruett 在一项研究中建议:如果在婴儿出世的头六个月里父亲参与照顾孩子,孩子们在四年级智力和运动神经的发展测试中会得到较好的分数。这项研究被写在这样一篇新闻报道中(Marc Schogol, "A father's hand," The Philadelphia Inquirer, May 31, 1995, p. H-1)。Pruett 教授研究了婴儿和他们的家庭(包括一项长达 10 年的跟踪调查)来测试儿童们的成绩。Pruett 指出,孩子们生命头六个月里父亲的照顾使他们在学习和体育方面有更好的表现。

基于多元回归分析、双因子方差分析和因果关系方面的知识你对这个结果做何评价?你能给出这项研究结果的另外一个解释吗?还有什么其它信息在评价这个结果上起重要作用?如果 Pruett 已经建议说父亲早期照顾与后来儿童发展的优势有关,这是否会平息你可能提出的主要异议?

13.6 小 结

多元分析是用来同时分析几个自变量对一个因变量的影响。作为分析的结果,自变量常常能按照它们对因变量的影响大小来排序。

在多元分析中我们常常可以决定两个变量之间的关系是否为因果的。如果两个变量之间的关系在多元分析后消失了,则我们假定它不是一种因果关系。

13.1 偏 ϕ : 三个分类型变量

两个分类型变量之间关系的强度是用 ϕ 来衡量的(见第九章)。对于在一个变量取常值的每个子部分中,仍有可能为一个分类型自变量与一个分类型因变量计算 ϕ 。每个 ϕ 描述了控制变量取特定值时一个自变量和这个因变量之间关系的强度。

控制一个变量意味着在其它自变量对因变量效应的研究中消去这个变量的影响。我们控制一个变量的效应就是使被控制变量保持常数。我们将控制变量的数据按其取值分组来保持其为常数。

由被控制变量各个小组中的 ϕ 计算出来的系数平均称为偏 ϕ 系数。偏 ϕ 系数表明了当第三个或更多变量的效应被控制时两个变量之间关系的总体强度。

13.2 数值型变量的多元回归

对数值型变量,多元分析就是计算偏回归系数和偏相关系数。这些系数都是相对于第 10 章中两个变量之间的回归系数和相关系数而言的。

两个或多个数值型自变量的偏回归系数是在除一个自变量外的其它自变量保持常数时计算出来的。偏回归系数可以组合成一个回归方程来估计所有自变量对因变量的联合效应。

偏系数会随着分析中考虑进不同的变量而改变。当自变量之间有着和自变量与因变量之间类似的相互关系时,偏系数的值会发生变化。自变量之间具有的相关性称为共线性。

多元回归分析中因变量的预测值与实际值之间关系的强度用多重相关系数 R 来衡量。 R^2 给出了由所有自变量共同解释的因变量差异的总和。

为了确定一个多元回归分析的结果是否为统计意义上显著的即是否样本的结果可以用到总体上去,我们常常不用置信水平而用假设检验的方法。常用的零假设是总体中的多重相关系数为 0。通过把样本的多重相关系数 R 转变为具有适当自由度的 F -变量值,我们可以用统计表格或软件得出观测数据的 p -值。 p -值表示从一个多重相关系数等于 0 的总体中得出观测值 R 或更大 R 值的样本的概率。

在多元回归分析中也可对每个自变量的回归系数分别做假设检验。

13.3 用一个哑元作多元回归

可以通过将分类型变量转化为哑元而把分类型自变量用于回归分析中。分类型变量常常被赋以 0 和 1 两个值。例如对于性别,男性可等于 0,女性等于 1。

13.4 双因子方差分析

为了研究两个分类型自变量对一个数值型因变量的效应,我们常常用所谓双因子方差分析的方法。由于双因子方差分析可以考虑两个分类型变量各自效应之外的交互效应,所以它要优于两个单因子方差分析。两个自变量对因变量的联合作用称为交互效应。通过同时考虑多个自变量和它们的交互效应,残差变量的效应减少了,从而更易于建立统计意义上的显著性。

13.5 建立因果关系

证明因果关系意味着检验所有可能对因变量产生效应的自变量。由于这是不可能的,所以宣称一种因果关系常常是一项基于对因变量的知识和易获得程度的推测性决策。

补充读物

Achen, Christopher. *Interpreting and Using Regression* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-029). Beverly Hills, CA: Sage, 1982. 多重回归的利用。

Asher, Herbert. *Causal Modeling*, 2nd ed. (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-003). Beverly Hills, CA: Sage, 1983. 利用回归考察可能的因果模型。

Berry, William D., and Stanley Feldman. *Multiple Regression in Practice* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-050). Newbury Park, CA: Sage, 1985. 多重回归的利用。

Bray, James H., and Scott E. Maxwell. *Multivariate Analysis of Variance* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-054). Beverly Hills, CA: Sage, 1985. 多元方差分析入门。

Fox, John. *Regression Diagnostics* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-079). Newbury Park, CA: Sage, 1991. 利用数据来看是否它们违反了利用回归分析所需之潜在假设。

Knoke, David, and Peter J. Burke. *Log - Linear Models* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-020). Beverly Hills, CA: Sage, 1980. 多元分类变量分析入门。

Wildt, Albert R, and Olli T. Ahtola. *Analysis of Covariance* (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-012). Beverly Hills, CA: Sage, 1978. 分类变量和数量变量二者作为自变量的回归分析。

公 式

偏 r (或 ϕ)
$$r_{12} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad \phi_{12} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

不同的平方和由下列方法算得：

$$\text{行变量平方和} = \sum n_{i.} (\bar{y}_{i.} - \bar{y})^2$$

$$\text{列变量平方和} = \sum n_{.j} (\bar{y}_{.j} - \bar{y})^2$$

$$\text{交互变量平方和} = \sum \sum n_{ij} (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y})^2 \quad (13.3)$$

$$\text{残差变量平方和} = \sum \sum \sum (y_{ijk} - \bar{y}_{ij})^2$$

$$\text{总计平方和} = \sum \sum \sum (y_{ijk} - \bar{y})^2$$

这些 n 表示不同行、列和单元中观测值的数目。这些公式能够直接地用于很小的数据集,否则,双因子方差分析最好在计算机上做。

对于 r 行、 c 列和每个单元内有 m 个观测值的情形由以下表达式算得不同的自由度：

$$\text{行变量的自由度} = r - 1$$

$$\text{列变量的自由度} = c - 1$$

$$\text{交互变量的自由度} = (r - 1)(c - 1) \quad (13.4)$$

$$\text{残差变量的自由度} = rc m - rc$$

$$\text{总计自由度} = rc m - 1$$

均方由平方和除以相应的自由度算得。 F -比率由每个均方除以残差变量均方算得。所有这些数都列在如表 13.12 的方差分析表中。

当在计算机上做分析时,数据列在如表 13.14 的计算机文件中。对于因变量 y 的每个值,我们将它填入相应的行列中。软件程序会自动考虑交互变量的构造。

表 13.14 双因子
方差分析数据在计算机
文件中的设置

y	行	列
y ₁₁₂	1	1
y ₁₁₁	1	1
y ₁₁₃	1	1
y ₁₁₄	1	1
y ₁₁₅	1	1
y ₁₂₁	1	2
y ₁₂₂	1	2
y ₁₂₃	1	2
y ₁₂₄	1	2
y ₁₂₅	1	2
y ₂₁₁	2	1
y ₂₁₂	2	1
y ₂₁₃	2	1
y ₂₁₄	2	1
y ₂₁₅	2	1
y ₂₂₁	2	2
y ₂₂₂	2	2
y ₂₂₃	2	2
y ₂₂₄	2	2
y ₂₂₅	2	2

习题

回顾(习题 13.1—13.17)

- 13.1 a. 多元统计分析与一元统计分析有何区别?
b. 为什么多元统计分析常常比多个一元统计分析有用?
- 13.2 如果两个变量之间的关系在多元分析后消失了,你能对这种关系做何假定?
- 13.3 在一个分析中控制一个变量意味着什么?
- 13.4 如果控制第三个变量后两个变量如性别与投票之间关系的 ϕ 由原来的 0.32 变成 0.00,你对这两个变量能得出什么结论?
- 13.5 a. 什么是偏 ϕ 系数?
b. 如果控制变量有四个值如:高、中、低、无,你怎样算得两个变量之间的偏 ϕ ?
c. 如果样本中“高”值大大多于其它值,这会影响偏 ϕ 吗?
- 13.6 什么时候你能用多元回归分析方法?

- 13.7 如果数据中的自变量相关,那么它们之间存在有_____。
- 13.8 如果散点图中通过两组数据点的回归线是水平的,你能得出有关一个组中的点相对于另一个组中的点对因变量效应的什么结论(例如,性别对视觉准确性的效应)?
- 13.9 a. 什么是哑元?
b. 从本章中找出或者自己举一个哑元的例子?
c. 构造一个哑元为什么有用?
d. 哑元的使用上有什么重要的限制?
- 13.10 a. 定义并描述一个相关系数 R 必须怎样?
b. R^2 告诉我们了什么?
c. $1 - R^2$ 告诉我们了什么?
- 13.11 为了确定由样本得出的一个结果是否可以用于总体,你必须做什么?
- 13.12 a. 你怎样向一位不懂统计学的朋友解释双因子方差分析?
b. 举一个用双因子方差分析方法的虚构例子。
c. 在哪一方面,一个双因子方差分析优于两个单因子方差分析。
- 13.13 双因子方差分析中定义的残差变量是什么?
- 13.14 举出一个要用多于双因子方差分析的问题——例如,一个有三个自变量的问题。
- 13.15 a. 双因子方差分析中交互作用意味着什么?
b. 举一个交互作用的例子。
- 13.16 a. 为什么近年来多元回归分析变得流行?
b. 为什么说多元回归分析可能会很“危险”?
- 13.17 即使关于几个重要自变量做了多元分析,为什么仍然如此难以断定是否某个自变量是一个具有因果关系的变量?

理解(习题 13.18—13.37)

- 13.18 当用 48 个大陆州的偷窃率对抢劫率做回归时,我们得出:

$$\text{偷窃} = 2682 + 1.49 \text{ 抢劫} \quad (t = 2.05, \text{ d.f.} = 46, p = 0.023)$$

为了研究这个关系是否为因果关系,我们想控制州人均收入。当用偷窃率对抢劫率和人均收入做回归时,我们得出:

$$\begin{aligned} \text{偷窃} = 3880 + 2.23 \text{ 抢劫} - 0.06 \text{ 收入} \\ (t = 2.75, \quad (t = -1.86, \\ p = 0.004) \quad \text{d.f.} = 45, \quad p = 0.035) \end{aligned}$$

(来源: *Bureau of the Census, Statistical Abstracts of the United States: 1995, 115th ed., Washington, D.C., 1995.*)

- a. 原则上当控制收入时,我们怎样得出偷窃的系数 2.23?
b. 第二项分析阐明了有关偷窃率与抢劫率之间关系的什么信息?
- 13.19 在习题 9.9 中,我们研究了众议院中两次投票之间的关系,并得出 $\phi = 0.62$ 。当控制第三个变量政党,并对同样的数据进行分析时,我们得出两次投票之间的偏 ϕ 为

0.26。

a. 控制政党因素时我们怎样分析?

b. 我们怎样算出偏 ϕ ?

c. 偏 ϕ 的值 0.26 告诉了我们有关两次投票之间原有关系的什么信息?

- 13.20 从康涅迪格州肿瘤登记处的用年龄调整过的黑瘤率表明从 1936 到 1972 年黑瘤病患者在增加,这些患病率似乎随着每年太阳黑子的相应数目而改变。黑瘤病按每 10000 人群中患者的数目来度量,其范围为 8 到 46 人,时间变量计 1936 为 1,1937 为 2,一直到 1972 为 37。太阳黑子的数值变化范围为 5 到 190。不同回归分析得出下列结果:

$$\text{黑瘤患病数} = 26.0 + 0.03 \text{ 太阳黑子}, \quad r = 0.13$$

$$\text{黑瘤患病数} = 7.1 + 1.10 \text{ 时间}, \quad r = 0.96$$

$$\text{黑瘤患病数} = 6.2 + 0.02 \text{ 太阳黑子} + 1.10 \text{ 时间}, \quad R = 0.97$$

(来源: A. Houghton, E. W. Munster, and M. V. Viola, "Increased incidence of malignant melanoma after peaks of sunspot activity," *The Lancet*, April 8, 1978, pp. 759 - 760, as reported in D. F. Andrews and A. M. Herzberg, *Data: A Collection of Problems from Many Fields for the Student and Research Worker*, New York: Springer-Verlag, 1985, p. 201.)

a. 描述黑瘤患病数与太阳黑子之间的关系。

b. 描述黑瘤患病数与时间之间的关系。

c. 描述黑瘤患病数与太阳黑子和时间之间的关系。

d. 对这些数据做多元分析的好处是什么?

- 13.21 国会成员是小部分可以决定自己薪水的人之一。然而投票支持薪水增长可能会被选民否决。有迹象表明距离下次大选越近,国会成员越不可能投票赞成薪水增长。在 1991 年的一次有关薪水增长的投票记录中,议员们被按照是否投票反对薪水增长,和是否在 1992 年的大选中他或他的政党被再次选出(不是所有议员同时被考虑再次选出)来分类。这些都是具有两个类别的分类型变量,所以我们可以算出代表它们之间关系的 ϕ 和偏 ϕ :

$$\phi(\text{再次选举,投票}) = 0.37$$

$$\phi(\text{再次选举,投票} \mid \text{控制政党时}) = 0.37$$

投票是指那些再次选举时倾向于投票反对薪水增长的议员。(来源: Roll call as reported in *The New York Times*, July 19, 1991, p. A13.) ϕ 的这两个值告诉你什么?

- 13.22 关于葡萄酒质量的研究有很长的历史并且主要依据主观判断。Princeton 经济学家 Orley Ashenfelter 教授,研究了作为葡萄酒质量衡量标准的拍卖价格与每年从 10 月到 3 月的冬季降水毫米数,从 4 月到 9 月的生长季节平均温度摄氏数,从 8 月到 9 月的收获季节降水量之间的关系。对 Bordeaux 葡萄酒的回归分析得出下列结果:

$$\text{质量} = -12.1 + 0.0012 \text{ 冬季降水} + 0.62 \text{ 温度} - 0.004 \text{ 收获季节降水}$$

(来源: Article in *The New York Times*, March 4, 1990, pp. A1, A22. This work was done for wines up through 1989, and according to the equation the Bordeaux wines of 1989 should be of an excellent quality. This wine was still too young to be judged at that time, and it is thought that one test of this analysis will be how well the wines

of 1989 actually turn out to be when they reach maturity)

- 这个等式就告诉你关于 Bordeaux 葡萄酒的质量与温度和降水变量之间关系的什么信息?
- 这个等式就有关这些变量之间的关系没有能够告诉你什么信息?

13.23 为了研究巧克力与香草小吃之间是否有口味等级的差异及冰奶、冻酸奶及冻甜点之间的差异,我们通过双因子方差分析得出表 13.15 的结果。你从这些结果中能得出什么有关甜点口味的结论?

表 13.15 习题 13.23 的数据

来源	平方和	自由度	均方	F-值	p-值
Q 种类	9248	2	4624	31.81	0.0000
巧克力/香草	14	1	14	0.10	0.75
交互作用	477	2	238	1.64	0.21
残差	5524	38	145		
总计	16613	43			

(来源:“Low-fat frozen desserts: Better for you than ice cream?” Consumer Reports, vol. 57, no. 8 (August 1992), pp. 483-487)

13.24 在习题 10.36 和 10.37 中用 50 个州的数据分别考察了低于贫穷水平人们的百分比与教育程度只有或低于九年级的百分比和教育程度为大学或更高的百分比之间的关系:

$$\text{贫穷百分比} = 4.6 + 0.8 \text{ 低教育程度百分比} \\ (r = 0.70, t = 6.72, \text{自由度为 } 49, p\text{-值} < 0.0001)$$

和

$$\text{贫穷百分比} = 26.9 - 0.7 \text{ 大学或更高教育程度百分比} \\ (r = -0.62, t = -5.54, \text{自由度为 } 48, p\text{-值} < 0.0001)$$

当在多元回归分析中对所有教育程度变量都分析了后,得出下面结果:

$$\text{贫穷百分比} = 14.6 + 0.6 \text{ 九年级百分比} - 0.4 \text{ 大学百分比} \\ (t = 4.41, \quad (t = -3.02, R = 0.75, \\ p = 0.0001) \quad p = 0.002)$$

(来源:1990 U.S. Census data Reported in The chronicle of Higher Education, vol. 34, no. 1 (August 26, 1992), p. 4)

- 为什么这次分析中的回归系数跟习题 10.36 和习题 10.37 中相应的系数不同?
- 从每个一元回归到这个多元回归,相关性上的改进是否足够大以至于值得我们做这个多元分析?(我们希望你从直觉上回答,而不是从统计上回答。)
- 还有什么其它变量你希望用于更好地理解是什么决定贫穷人们在不同州的百分比?

13.25 图 13.10 的散点图显示了来自四个年级的学生样本的自变量和因变量的数据。数据点上“fr”表示新生,“so”表示二年级,“jr”表示三年级,“sr”表示四年级。对于每个问题详细解释你的回答。

- 当你用因变量对自变量做回归分析时,回归系数将为正还是为负?

- b. 相应的相关系数将为正还是为负?
- c. 同一偏相关系数将是小、中还是大?
- d. 当年级受到控制,你研究自变量与因变量之间的关系时,自变量的偏回归系数是正还是负?
- e. 相应的偏相关系数为正还是为负?
- f. 同一相关系数是小、中还是大?

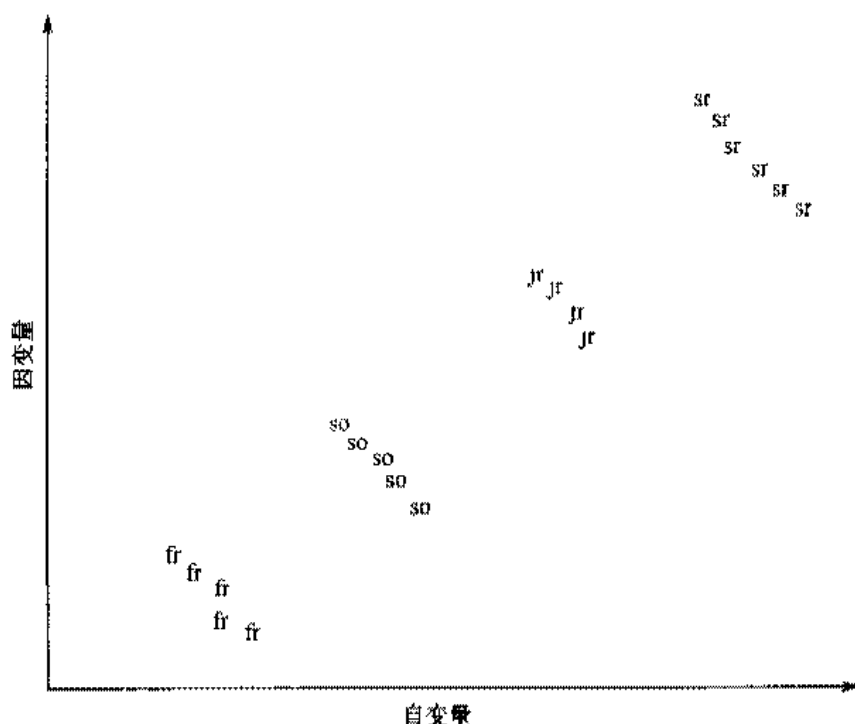


图 13.10 习题 13.25 的数据。

- 13.26 按照 *Bulletin of the American Association of University Professors* (vol. 79, no. 2 (March/April (1993), p. 71) 中的数据, 在 1991—1992 学年, Swarthmore 学院的 7 名女性正教授的平均薪水为 \$ 63700 而 59 名男性正教授的相同均值为 \$ 70900。如果我们把性别变为一个哑元, 用 0 表示女性, 1 表示男性, 则由收入对性别所作回归的回归线得出等式:

$$\text{收入} = 63.7 + 7.2 \text{ 性别}$$

从这个等式和这些均值看来, 似乎学院对女性和男性教授有不同的支薪标准。但是, 在接受这种对两个均值之间差异的解释之前, 我们需要控制其它可能相关的变量。假设我们在多元回归分析中控制教授的年龄, 又假设在收入对年龄和性别做分析时性别的系数为 0.0。

- a. 向你的一位聪明但又不懂统计的朋友详细阐释控制第三个变量意味着什么和控制年龄时我们怎样得出性别的偏回归系数为 0.0。
- b. 由性别的回归系数从一元回归分析中的 7.2 到多元回归分析的 0.0 的变化, 我们可

以知道什么?

- 13.27 在田纳西州的 Nashville,我们发现星期几与孩子们事故次数有某种统计关系。
- 举出一个变量,如果用作控制变量,则上述关系会消除,并解释原因。
 - 这个分析表明了关于星期几对因变量的因果效应?
- 13.28 看一看这个有关课本成本与课本页数和印刷数目之间关系的回归分析,并回答下列问题。

$$\text{课本成本} = 30 + 0.05 \text{ 总页数} - 0.0001 \text{ 印刷数目}$$

- 30 表示什么?
 - 0.05 和 0.0001 这两个数叫什么?
 - 阐述 0.05 和 0.0001 这两个数通过总页数和印刷数目告诉了我们有关课本成本的什么信息?
 - 如果用课本成本对页数做一元回归分析,你会得出与多元回归分析相同的系数 0.05 吗?
- 13.29 为什么说试图比较分析中偏回归系数的大小就像比较苹果和橘子一样?(你可以用每加仑汽油汽车可行驶的里程数的例子来解释你的答案。)
- 13.30 某大学五个分校区的全职教员指控管理处存在偏见,因为他们的薪水平均上比本部的全职教员少 \$10000。
- 作为管理委员会的统计顾问,你希望研究哪三个变量以决定是否地区性是解释薪水差异的唯一原因?
 - 你对这项研究的结果有何预感?你是否认为还有其它变量解释这个差异?
 - 你认为如果做两个变量而非多个变量的分析会有什么差别?
 - 你推荐哪种分析?为什么?
- 13.31 当希望研究两个变量之间,例如习题 13.30 中的地区与薪水之间的统计关系的性质时,我们怎样决定哪些变量需要受到控制?
- 13.32 对某个 R 为 0.084,你得到一个 F 为 22.20,自由度为 4 和 15。得出这样大或更大的 F 的概率非常小($p < 0.0001$)。
- 你应该拒绝没有效应的零假设吗?
 - 这个结果对于产生样本数据的总体有何暗示?
- 13.33 如果一个关于三个自变量(例如,脂肪、蛋白质和钠)的多元回归分析有一个显著地不等于 0 的总体 R 系数,这是否意味着每一个单个变量也有显著效应?解释你的回答。
- 13.34 对于一个双因子方差分析,一个单元观测值与均值之间的差异如下:

$$\begin{array}{ccccccc} -1.5 & 2.0 & -0.5 & 0.0 & -2.0 & 1.0 & 1.0 \end{array}$$

为什么不是单元中的所有观测值等于单元的均值以使得这些差异等于 0?

- 13.35 在六种欧洲轿车的一次比较中,一份汽车杂志请了几个人从不同方面做评价,例如引擎、变速箱、安静度和座椅等等。所有评价都基于一个 0 到 10 的尺度,并且算出得分的均值。例如,Alfa Romeo 164L 的引擎的平均得分为 9.4。每种轿车都有 21 个平均得分,它们表示在图 13.11 里。这些轿车之间有什么不同?是否某种轿车比其它的

好？是否某种轿车比其它的差？

- 描述你在图中所见的模型。
- Mercedes 具有最高总均值 8.32, VW 具有最低总均值 8.05, 方差分析给出的结果列在表 13.16 中。你能得出有关这些轿车之间差异的什么结论？
- 这些数据像是配对数据一样因为每种轿车在每项性能指标上有一个观测值而这些性能指标又不是基于同一个尺度计分。例如, 车体结构平均分为 8.67 而引擎的平均分为 7.71。这些差异包括在表 13.16 中残差的自由度及平方和中, 应该除去它们的影响。因为每项性能指标有六个观测值而不是两个, 你不能用对于配对数据的方式来考察这些差异。但是, 你可以对表 13.17 中的结果做没有交互作用的双因子方差分析。为什么根据性能指标之间的差异从残差变量中除去变差似乎对评价轿车没有任何效应？

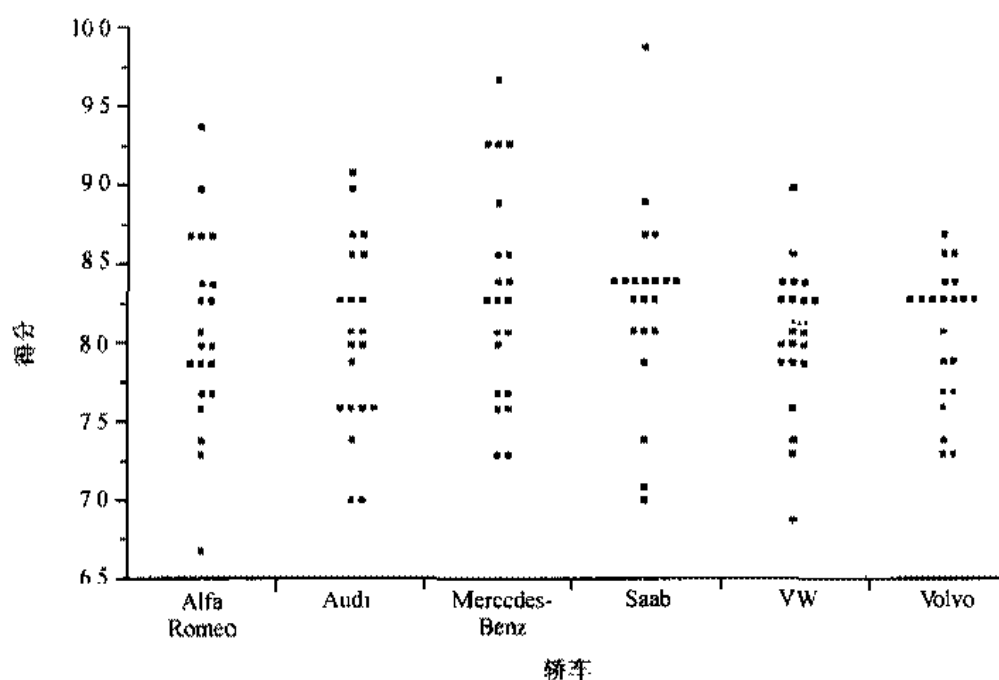


图 13.11 习题 13.35 的数据。(来源: "European influence," *Road and Track*, August 1991, pp. 64-84.)

表 13.16 习题 13.35b 的数据

来源	自由度	平方和	比例	均方	F-值	p-值
轿车	5	140	0.03	28.05	0.85	0.52
残差	120	3977	0.97	33.14		
总计	125	4117	1.00			

表 13.17 习题 13.35e 的数据

来源	自由度	平方和	比例	均方	F-值	p-值
轿车	5	140	0.03	28.05	0.86	0.51
性能指标	20	712	0.17	35.60	1.09	
残差	100	3265	0.80	32.65		
总计	125	4117	1.00			

- 13.36 在一项用政府数据对某 41 个城市的样本的分析中,把二氧化硫作为因变量,年平均气温和以千人计的人口规模作为自变量,我们得出:

$$\text{二氧化硫} = 91.7 - 1.3 \text{ 气温} + 0.02 \text{ 人口规模} \quad R = 0.64$$

对于气温的系数 -1.3 , $t = -3.23$ (自由度为 38), $p = 0.0013$ 。对于人口规模的系数 0.02 , $t = 3.74$ (自由度为 38), $p = 0.0003$ 。这些数告诉了你什么?

- 13.37 分析来自制造商有关 1996 型号汽车某样本的数据时,你会发现当你用城市内每加仑汽油所能走的英里数(mpg)对轿车重量(以千磅计)和马力分别及联合作回归分析时,得出:

$$\text{城市 mpg} = 40.5 - 6.27 \text{ 重量} \quad (r^2 = 0.77)$$

$$\text{城市 mpg} = 29.2 - 0.05 \text{ 马力} \quad (r^2 = 0.56)$$

$$\text{城市 mpg} = 40.4 - 6.23 \text{ 重量} - 0.0004 \text{ 马力} \quad (R^2 = 0.77)$$

这些结果告诉了你有关城市 mpg 与轿车的重量和马力之间关系的什么信息?

分析(习题 13.38—13.47)

- 13.38 在习题 9.42 中,我们得出在第一校区学校运动员的性别与大学毕业人数之间关系的 $\phi = 0.10$ 。在同一项研究中,种族与毕业之间关系的 $\phi = 0.22$,种族与性别之间关系的 $\phi = 0.14$ 。

- 算出控制种族时性别与毕业之间关系的偏 ϕ 。
- 算出控制性别时种族与毕业之间关系的偏 ϕ 。
- 这些 ϕ 和偏 ϕ 告诉你有关三个变量之间关系的什么信息?

- 13.39 习题 13.23 的结果没有给出关于不同小吃口味的全部信息。分析中的一个增补措施是考察每组味道的平均得分(表 13.18)。

- 把表中六个单元的均值表示在与图 13.10 相似的图中。
- 这个图比我们从前习题 13.23 的表 13.15 中所知道的增加了什么?

表 13.18 习题 13.39 的数据

种类		香精	
		巧克力	香草
	冻酸奶	71.3	54.1
	冰牛奶	55.2	64.0
	冻甜点	31.0	25.5
	总体	60.4	55.5
			总体
			67.3
			61.1
			27.3
			57.4

来源: "Low-fat frozen desserts: Better for you than ice cream?" Consumer Reports, vol. 57, no. 8 (August 1992), pp. 483-487.

- 13.40 表 13.19 中的数据矩阵表示了关于一个数值型因变量 Y 、一个数值型自变量 X 和一个替代具有两个类别的分类型变量的哑元自变量 D 的数据。

表 13.19 习题 13.40 的数据

Y	X	D
1	1	0
1	2	0
1	3	0
3	5	1
3	6	1
3	7	1

- 画一个 Y 对 X 的散点图。(Y 对 X 的回归分析得出一个直线方程 $Y = 0.3 + 0.4X$)
 - 控制变量 D 研究 Y 与 X 之间的关系意味着什么?
 - 对于这些数据当你用 Y 对 X 和 D 做了一个多元回归分析而分析结果的回归方程为 $Y = 1.0 + b_1x + 2.0D$ 时, 算出回归系数 b_1 的值。
 - 控制 D 相对于不控制 D , X 的回归系数的值告诉了 Y 与 X 之间的什么关系?
- 13.41 在一项薪水对四个自变量的多元回归分析中:
- $$\text{薪水} = 30000 + 2500 \text{ 大学教育年限} + 400 \text{ 工龄} \\ - 5000 \text{ 钟点工 / 固定工资}(1,0) + 1500 \text{ 男性 / 女性}(1,0)$$
- 对于两个具有 10 年工龄的领固定工资的 0 男性, 一个有 3 年大学学历, 而另一个有 4 年学历, 你对于他们薪水差异的预计为多少?
 - 如果这两人为女性, 你对薪水差异的预计又为多少?
 - 哪一项看来对薪水更为重要, 大学学历还是工龄?
- 13.42 男性和女性评分者被邀请以 1(最好)到 5(最差)的五个等级评价学术论文。三分之一的论文署名 John T. McKay, 另外三分之一署名为 Joan T. McKay, 剩下的三分之一署名为 J. T. McKay。署名为 John 的论文得到男性评分者 1.9 的平均分, 女性评分者 2.3 的平均分; 署名为 Joan 的论文得到的都是 3.0 的平均分; 署名为 J. T. 的论文得到男性评分者 2.7 的平均分, 女性评分者 2.6 的平均分。(来源: Quoted in L. Billard, "A different path into print," *Academe: Bulletin of the American Association of University Professors*, vol. 79, no. 3 (May/June 1993), pp. 28-29, from M. A. Paludi and W. D. Bauer, "Goldberg revisited: What's in an author's name," *Sex Roles*, 9(1983), 287-390.)
- 为什么你应该对这个问题用双因子方差分析?
 - 将均值像表 13.5 那样列出。
 - 均值是否为评分者的性别与作者性别之间的交互效应提供了某些证据?
 - 在女性和男性评分者给出的平均分中是否有明显差异?
 - 在署名为 John、Joan 和 J. T. 论文的平均得分之间是否有明显差异?
- (在这个习题里, 学术论文中没有足够信息来检验统计显著性。)
- 13.43 在习题 10.61 中, 我们对雌性和雄性老鼠分别做了油酸注入量和吸收量的分析。现在我们对数据做多元分析。4 只雌老鼠和雄老鼠的肝脏被注入油酸。表 13.20 显示了注入量和体内的吸收量。
- 把性别变量变成一个哑元, 0 表示雌性老鼠, 1 表示雄性老鼠。
 - 把注入量和性别作为两个自变量对吸收量作多元分析。
 - 注入量的系数与习题 10.61 中两个分别的系数相比有什么不同?

- d. 用 0 代替性别, 求出雌性老鼠的回归直线; 用 1 代替性别, 求出雄性老鼠的回归直线。
- e. 这两条直线与习题 10.61 中的两条回归直线相比有什么不同?
- f. d 中两条直线的垂直距离有多大?
- g. 为什么两条直线之间的距离可以解释为我们控制注入量时性别的效应?

表 13.20 习题 13.43 的数据

注入量	吸收量	性别
29.3	1.82	雌性
25.5	0.84	雌性
26.3	1.09	雌性
31.0	1.45	雌性
20.6	1.56	雄性
17.9	0.93	雄性
23.6	1.54	雄性
25.4	1.76	雄性

- 13.44 如果你对于蛋白质来源(牛肉或谷类食物)和蛋白质含量(低或高)对老鼠体重的增长效应感兴趣, 请分析表 13.21 中显示的四组中每组十只老鼠体重增长的数据。

表 13.21 习题 13.44 的数据

		蛋白质来源	
		牛肉	谷类食物
蛋白质	低	90, 76, 90, 64, 86, 51, 72, 90, 95, 78	107, 95, 97, 80, 98, 74, 74, 67, 89, 58
	高	73, 102, 118, 104, 81, 107, 100, 87, 117, 111	98, 74, 56, 111, 95, 88, 82, 77, 86, 92

来源: "George W. Snedecor and William G. Cochran, Statistical Methods, 6th edition, Ames: Iowa University Press, 1967, p. 347.

- 13.45 在一项统计研究中, 法学教授 David Baldus 和统计学家 George Woodworth 在 Georgia 州的法庭上分析了来自 2475 个案例的数据。他们考虑的变量中有受害者和被告的种族及是否判决死刑(表 13.22)。
- a. 画一张体现被告的种族与被害者种族之间关系的表格。
- b. 被告的种族与被害者种族之间关系有多强?
- c. 控制是否判决死刑后, 考察被告的种族与被害者的种族之间关系的强度。
- d. b 和 c 部分的答案联合起来告诉你什么?
- e. 哪种其它有关这些数据的分析可能会令人感兴趣?

表 13.22 习题 13.45 的数据

(a) 黑人被告					(b) 白人被告				
死刑	是 否 合计	被害者种族			死刑	是 否 合计	被害者种族		
		黑人	白人	合计			黑人	白人	合计
		18	50	68			2	58	60
		1420	178	1598			62	687	749
	合计	1438	228	1666		合计	64	745	809

来源: *Chance*, vol. 1, no. 1, p. 7, 1988.

13.46 习题 11.48 中我们比较了教师在初中和高中的教学小时数。现在我们也包括了小学的数据(表 13.23), 用双因子方差分析方法(没有交互作用)分析这些数据。

表 13.23 习题 13.46 的数据

国家	小学	初中	高中	均值
德国	790	761	673	741
爱尔兰	951	792	792	845
意大利	748	612	612	657
挪威	749	666	627	681
西班牙	900	900	630	810
瑞典	624	576	528	576
美国	1093	1042	1019	1051
均值	836	764	697	766

来源: *OECD*, from *The New York Times* May 28, 1995, p. E7.

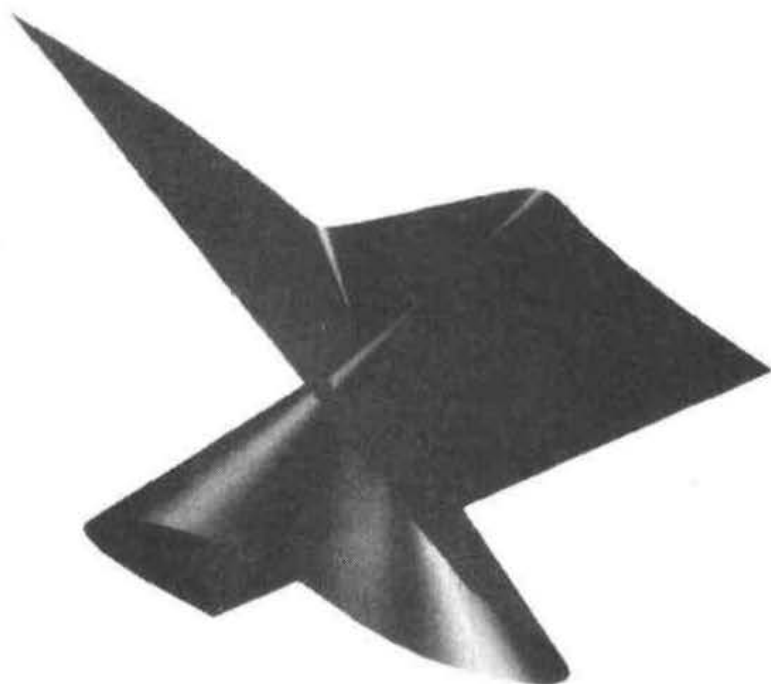
13.47 表 13.24 节显示了来自表 3.5 的中国食品数据。分析中国食品数据并跟 13.2 中小吃的多元分析结果进行比较。

表 13.24 习题 13.47 的数据

菜名(数量:杯)	热量(卡)	脂肪(克)	脂肪中的热量百分比	钠(毫克)
蛋卷(1 卷)	190	11	52	463
木须肉(4)	1228	64	47	2593
宫爆鸡丁(5)	1620	76	42	2608
糖醋里脊(4)	1613	71	39	818
西兰花炒牛肉(4)	1175	46	35	3146
曹将军鸡(5)	1597	59	33	3146
鲜橘牛肉(4)	1766	66	33	3135
酸辣汤(1)	112	4	32	1088
烙饼(5)	1059	36	31	3460
家常炒饭(4)	1484	50	30	2682
鸡丝炒面(5)	1005	32	28	2446
湖南豆腐(4)	907	28	27	2316
蒜绒虾(3)	945	27	25	2951
清炒小菜(4)	746	19	22	2153
川味虾(4)	927	19	18	2457

来源: *Center for Science in the Public Interest*, as given in *The Philadelphia Inquirer*, September 2, 1993, p. F7.

C H A P T E R 14



14.1 通向统计精妙的基石

14.2 小心地处理数据

14.3 数据和统计方法

14.4 怎么会出错

14.5 统计和专制

14.6 在高潮时结束

当我们在一个陌生地域的树林中砍出树上标明小路的记号时,我们经常是如此专心地和周围环绕的荆棘纠缠,而失去了我们(可能)迷失于其中的、较大森林中的小路。不过,现在我们战胜了迷途,来到一块空地上。我们可以基于已有的基础去期待未来的美好前景。

在本章中我们有两个主要目的。首先,让我们回顾一下所走的历程,你可以看看在通向统计文化的道路上你已经走出了多远。本书的每一章在技巧和知识理解上都是下一章的基石,你不能像阅读诗书那样将统计书倒过来读。在 14.1 节中,我们复习你已经掌握的统计知识。这些试金石标志着在统计意义上你从幼稚到成熟的飞跃。不像许多入门课本的作者那样抱有宏伟幻想,我们希望你已经获得了对统计学家所做工作的鉴赏能力,但我们并不期望你成为该领域的一个并非货真价实的专家。本章第二部分的设计是为了提醒你开发研究中潜在的问题以及公众对统计的使用和误用。

牙齿神话^① 和臼齿的价格(Tooth fairies and the price of a molar)

大多数统计信息都是零碎的、不可靠的并随便让记者及评论家用来发表令人印象深刻的断言。在 Family Circle 杂志报道的一份调查发现:牙齿神话为一颗牙齿平均付出 \$1 到 \$2,但是有时为每颗门牙和双尖牙付 \$20。这项来自全美儿童牙科学会(the American Academy of Pediatric Dentistry)的调查发现最为慷慨的保险是在 Houston(休斯顿),至少有一个小孩在那里因为一颗牙而得到了 \$50。(来源: The Philadelphia Inquirer, May 24, 1995, Family section, p. 1.)

统计知识能帮助焦虑的父母为他们牙齿松动的孩子作出正确的选择吗?仅仅从上面一点的信息我们很难得出什么结论。牙齿神话的样本好不好?哪种类型的度量(均值、中位数、众数、随便猜测)得出了平均 \$1 到 \$2 的结果?数据是否应该按牙齿类型分类?从总体来说, Houston 与其它城市的牙齿神话相比相差多远?特别地,这个 \$50 的牙齿神话又和别的相差多远?

对于每一种统计情形,我们都需要评价我们拼凑的图景对现实反映的精美程度。为了得出结论,我们必须知道统计方法是怎样被应用和被误用的。如果我们是知识丰富的统计消费者,我们就足以去理解和评价统计在多方面的应用及由统计研究得出的各种结论。

14.1 通向统计精妙的基石

统计学家所做的许多工作都是关注一个变量是否影响另一个变量。我们把这种关注概括为以下四个问题:

- 问题 1. 在数据中,变量之间是否有关系?
- 问题 2. 变量之间的关系有多强?
- 问题 3. 总体中是否有关系?
- 问题 4. 观测到的关系是一种因果关系吗?

纵观本书,我们在不同情形下得出了回答这些问题的技巧。为了奠定回答以上这些问题的基础,本书的前半部分列出了统计“数字游戏”的许多重要方面。本书后半部分以不同的方式将这些概念组织起来用以阐明几种重要的数据分析方法。

在第一章中,统计被定义为在面对随机性的情况下寻找规律。接下来的章节描述了统计工作的三个部分:数据收集、数据分析和由数据做推断。这一章中还介绍了变量的重要概念和它在某些元素集合上的取值。最后讨论了几种应用统计学的人。

第二章的重点是数据的收集。介绍了抽样误差的概念,并强调了得到一个好样本的关键所在。通过使你对坏血病的了解比你希望的更多,你了解了观测研究与实验研究之间的区别。我们还描述了数据矩阵和数据文件。

以可视的形式罗列数据是第三章的主题。在整本书中,我们都建议分析前将数据可视化。我们还介绍了 Tufte 对优秀图形的要求,即越简单越好。

在第四章中通过讨论对集中趋势的衡量(均值、中位数、众数)和对差异的衡量(主要是标

^① 西方人哄小孩的故事,讲如果把掉的牙齿放在枕头下,小精灵就会拿走牙齿而放一个钱币, fairy 在英文中有童话和仙女(精灵)二意;这里是指牙医保险——译者注。

准差和方差)我们介绍了数据的分析。对于标准得分和均值的标准误差的熟悉为各种更为复杂的数据分析铺平道路。在数据分析中丢失信息与追求简洁之间的矛盾显然是我们必须面对的。

第五章着重概率论。介绍了作为重要标准的具有钟型曲线的正态分布;同时也介绍了它的在每个标准差段落内曲线下方面积所占比例的独特性质。这些和其它的类似曲线奠定了评价所收集的统计量的显著性的基础。还简略提到了书中将要用到的四个主要理论统计量: z , t , χ^2 和 F 。 p -值意味着什么和人们怎样基于事件发生的概率来对数据做出决策这样的问题为假设检验提供了舞台。

第六章中,为了做结论,我们辨明了样本统计量与总体参数之间的差别和从样本统计量中估计参数的方法。我们还讨论了参数的点估计和区间估计方法。置信区间的概念提供了评价参数估计好坏程度的途径。

第七章详细讨论了怎样用假设检验的方法从样本数据得出有关总体参数的结论。讨论的主题包括:零假设检验后面的推理;决定是否拒绝零假设时可能出现的错误类型;怎样算出和使用 p -值;怎样算出相应的自由度。这些技巧可用于有关 t -检验和 z -得分的问题。该章还比较了假设检验与构造置信区间。在假设检验中我们问参数是否可能等于某一特定假;在构造置信区间时,我们通过获得我们希望包含参数真值的一个区间来估计参数的实际值。

第八章强调了怎样着手回答有关统计关系的四个重要问题。对于问题1,我们考察样本数据的模式。如果发现某种关系,则我们提出问题2。为了回答问题2,我们计算变量之间关系的强度。为了回答问题3,我们建立一个变量之间没有关系的零假设并检验这个假设看看是否拒绝它。关于因果关系的问题4常常是难以回答的。两个变量之间即使没有因果联系也可能存在某种关系(甚至是很强的)。并且,即使两个变量之间不是因果相关的,如果知道一个变量的值,我们仍有可能去估计另一个变量的值。关系的强度即使只告诉了一点或完全没有告诉我们因果关系,但它反映了可以在多大程度上用一个变量去估计另一个。

第九章讨论了用列联表及 χ^2 分析来研究分类型变量。不同的例子阐释了回答四个问题的不同方法。考察现实世界里变量之间是否有关系需要用到假设检验。为了算出某个样本的 p -值,我们把 ϕ 或者 V 系数(用于评估变量之间关系强度的一个统计量)转化为 χ^2 变量的值。为了算出与某个 χ^2 相应的 p -值,我们需要知道相应列联表的自由度。

第10章探讨数值型变量的相关和回归分析。散点图表明变量之间是正相关还是负相关;相关系数则是衡量关系的强度。相关系数的值域为 -1 到 $+1$ 。相关系数告诉了我们估计的好坏,但是不能用于评价两变量之间的因果关系。回归分析包括了在散点图上画一条经过数据点中心的回归线。这条线的斜率告诉我们在多大程度上一个变量随着另一个变量而改变。由这条直线的斜率和截距生成的回归方程可以用于从某个自变量的值去估计因变量的值。通过为分类型变量构造一个哑元(例如0,1)回归分析可以用于研究一个分类型与一个数值型变量之间的关系。

作为一种研究某个分类型自变量对数值型因变量效应的方法,在第11章我们介绍了方差分析(anova)。如果自变量对于因变量的效应相对于残差效应来说很大的话,我们就可以拒绝没有关系的零假设。我们可以用自由度算出 F -变量的值,从而得出 p -值。一旦发现了某种统计意义上显著的关系,如果又有多于两个的因变量,则我们检查哪些因变量的均值显著的不相同。简单的符号检验对于研究配对数据中出现的差异非常有用。

第12章着重讨论分析顺序变量的特殊方法。我们介绍了被称为秩标识的 γ 统计量。它用于衡量两个顺序变量之间关系的强度。像其它分析一样,顺序变量之间的关系也用没有关系的零假设来进行检验。 p -值通过将 γ 转变为 z 并用正态分布表得出。当变量具有数值秩时,变量之间关系的强度用 Spearman 秩相关系数衡量。这个系数跟数值型变量的 Pearson 相关系数相似。通过将 Spearman r_s 转变为 t -得分并算出 p -值,我们可以评价它在统计意义上的显著性。(同样,相关并不意味着因果关系。)

第13章简略地介绍了多元分析。它用于同时分析几个自变量对某个因变量的效应。自变量常常可以按照它们对因变量的作用来排序。在多元分析中,如果保持第三个变量为常数时两个变量之间的关系消失了,则我们假定原有关系不是因果关系。按照被控制变量定出的各个子部分的系数平均值称为偏系数。偏系数显示了当第三个或更多的变量被分离出去时,两个变量之间总的关系。通过将分析中每个变量的偏回归系数与相应的变量相结合我们可以得出相应的回归方程。如果恰当地选择自变量,多元回归对因变量真实值的估计是非常有力的。

统计的创新用途:音乐和神秘

许多不同职业的人们用统计款待他们的观众,用它来传递信息。当代作曲家 John Cage 把电脑产生的随机旋律写入了他的作品中。作家 Michael Crichton 通过应用统计增强了那些乐于观看像 *The Andromeda Strain*、*The Terminal Man*、*Jurassic Park* 和 *Coma* 之类恐怖电影的影迷的乐趣。

第13章中我们还回顾了双因子方差分析。它同时考虑两个分类型自变量对某个数值型因变量的独立和交互效应。当和多元回归分析一起时,双因子方差分析提高了单因子的准确度,因为它同时比较多个变量,从而减少了残差变量的效应。

确定因果关系的讨论贯穿于最后六章中。有关因果关系最安全的声明是否定的:证明某种关系不是或者可能不是因果的要比证明它为因果的容易得多。任何有关因果变量的假定在面对一个新出现的控制变量的挑战时总是脆弱的。

14.2 小心地处理数据

对于统计方法的熟悉可以帮助我们评价和理解统计分析的结果。做一个懂统计的人还可以帮助我们知道什么时候该怀疑一项统计论断。当读到本月的失业率为 6.7%,女性的收入比男性少 \$7000 或者人马座的人更会娱乐时,我们在接受这些数字作为事实上应该小心。我们已经有足够多的知识认识到许多的限制、忽略和错误能堆积在获得最终统计声明的道路上,也认识到统计研究的结果更经常地不等于只有在完美的统计世界中才会得到的精确的真值。

我们可以从形式上刻画为什么样本统计量的观测值不等于总体参数的真值:

$$\text{观测值} = \text{真值} + \text{非统计错误} + \text{随机性}$$

方程式左边的观测值可能是百分比、两个均值之间的差异或任何由数据算得的值。方程式右边是控制观测结果的三项。真值是我们统计梦想中的参数,是一个不受随机和错误影响的想象得分。非统计错误和随机性是两个非常不同的因子。由于统计学原来是作为一门主要处理公式和方程的数学科学,非统计错误传统上被列在统计学家的研究范围之外。我们常常假定这些是研究某一特定领域方法问题的人们所关心的事情。例如,心理学家常常比统计学家更加关心被实验者的种族、性别或年龄对其反应产生的影响。今天,对数据收集的考虑不再是统计学家研究范围之外的事了。在现今的数字探索中,当参与公式和方程的实际应用时,统计学家不能回避某些“非统计”问题。统计学家再也不能将自己的头埋在沙漠中,而让别人来对付方法领域中这些棘手的“仙人掌”。

另一方面,随机性是可以预料的对真值了解的障碍;由于其无法避免,它是统计学家可以忍受的。随机性是统计世界的一部分,它拥有其它错误所不能得到的某种程度的数学声誉。正如你已认识到的,统计学家们对于不可避免的随机性已建立起精细的防范和小心。

14.3 数据和统计方法

到目前为止我们已经非常清楚任何统计分析的结果都是基于:(1)数据的收集和(2)统计方法的运用(图 14.1)。这可能是本书最为重要的信息。

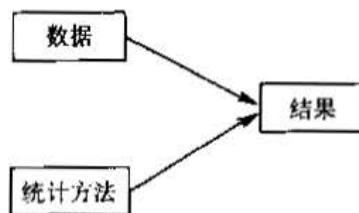


图 14.1 影响统计研究结果的因素。

假设我们看到报纸头条说女性收入比男性少 \$ 7000。这是一项统计分析结果。结果从何而来? \$ 7000 的差异并不是生活中的实物,不能仅仅因为我们看到它印在纸上就认为它是对真实世界的描述。这个结果是以收集到的特殊数据和用于分析数据的特殊方法为基础的。对于其它数据和其它方法可能会出现不同的结果。

为了理解一项统计结果,我们首先需要知道数据是怎样收集的。数据是否来自所有成人中的一个随机样本,或者来自某年国内税收局(Internal Revenue Service)归档的所有返回税?每种分别的收集模式都将影响到结果。除知道数据从何而来以外,我们还需要知道数据中的每个个体是用什么衡量的。收入数字是由工作所挣收入组成还是包括债券和股票的利息和分红?有些富豪并非挣得他们的收入,他们是否应被排除在分析之外?

当我们想要知道谁是研究对象和收入数字是怎样被决定时,我们需要知道数据是怎样被分析的,从而结果究竟意味着什么。是否每个男性比女性多 \$ 7000 呢?这是不可能的,因为我们知道无论是男性之间还是女性之间,收入的差异都是很大的。也许 \$ 7000 的差异是男性

和女性之间平均收入的差异。如果是那样的话,这个平均是指的均值、中位数或者其它形式的平均呢?无论用哪种平均,如果换成另外一种形式的平均,研究者都会获得另一个差异值。男性与女性之间收入均值的差异与相应中位数的差异是不同的。

我们还需要知道差异是否为统计意义上显著的。\$ 7000 的数量听起来似乎很大且意味深长,但是在知道一种或另一种统计推断的检验结果之前,我们不能真正地断定它究竟有多重要。也许它只是由随机性引起的偶然差异。

有了以上的背景我们将图 14.1 中的方框“增肥”成图 14.2 的样子。这个图是我们检验所遇到的任何统计结果的一个参照标准。在接受一个统计结果之前,我们可以借助这个图来构造关于它的问题。这个图鼓励我们怀疑数据的来源及变量是如何被度量的。它同时让我们去怀疑所用的统计方法。从数据中得出了什么样的统计量,是否控制了其它变量,用不同的统计检验方法结果是否还是统计意义上显著的等等?

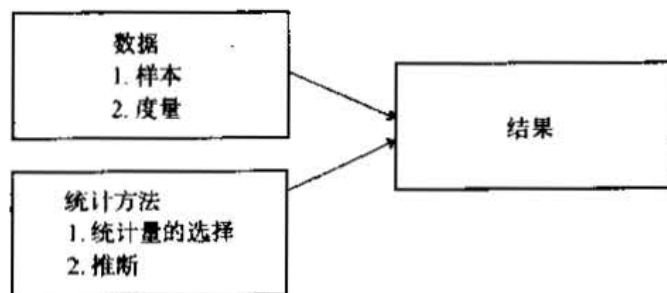


图 14.2 影响统计研究结果的因素细节。

这个图还可以用来检查研究中将会导致有问题的结果。由于统计结果依赖于数据和方法,如果数据或方法或两者同时有错误时,结果也会错。如果在数据收集或分析方法中有严重错误,则结果就不能接受。

14.4 怎么会出错

统计学是一个可能出现很多误用的领域。从问题原始思想的系统陈述到最后结果的输出,每一个步骤都可能出现错误。大多数统计的错误和误用都不是故意的,但是用不恰当的方法可能会产生对结果的有意歪曲。

数据收集中的危险

数据的收集是一个有两个步骤的过程。第一步为选择用于度量待考察变量的元素。这意味着来自一个更大总体的样本元素或整个总体。这个工作包括确定调查中被提问的对象或实验中要用到的元素。显然,这些元素并不一定是人。我们可以研究任何个体,例如:动物、土地、灯泡、棒球比赛或国家。

第二步是实际地收集数据。有时我们收集一个特定群体的数据,并希望我们的结果只用于这个群体。具有群体中每个个体的数据就消除了研究中的选择过程和选择所面临的各种问

题。例如,如果全美橄榄球联盟(the National Football League)的老板们想要知道每个球员的薪水,他们可以将某一时期的所有球员和他们的薪水列一张表。在这里,从样本到总体的推断是不恰当的,数据本身就是总体。老板们掌握的不是统计量而是参数了。

如果我们希望把结果应用于所收集的数据元素之外的元素,情形将会改变。当我们想要通过置信区间和(或)假设检验把结果推广到大的总体上时,我们必须采用正确的统计程序选择元素。对于实验,这意味着将对象随机地分配到处理和控制小组;对于调查,这意味着随机地选取样本。

即使对于恰当的随机样本,结果也只能应用于这个样本或通过统计推断把它推广到产生该样本的总体。这个事实对研究增加了重要的限制。假设一个药业公司将样品只给那些身为公司股东的医生,然后调查这些医生的一个样本,看有多少人不用其竞争对手的产品而喜欢用它的产品。这个公司绝不能在不透露她所选样本的特殊总体的情况下,宣称“调查中76%的医生喜欢我们的产品”。这个结果不能用于所有医生的总体,报告中也不能否定这一点。

我们相信研究者都懂得随机性的要求,但是达到随机性的要求则是另外一码事。在许多研究中,随机性条件几乎从未满足过。例如,大多数心理学研究是依靠大学生来获得他们的研究结果。(批评家们建议现代心理学应该称为大学二年级学生的心理学。)大学生难道是成人总体异或年轻人总体的一个随机样本吗?不太像吧。然而,他们是方便的、有文化的、合作的和令人感兴趣的被调查者。从而,忽略由于利用他们而可能产生的偏差是非常具有诱惑力的。因为学者们要承担尽量快速有效地出版成果的压力,他们不愿意为一个真正意义上的实验对象随机样本付出那样大的代价。

心理学并不是唯一的一个在方便和合作基础上获取研究对象的领域。医学中的许多研究依赖于与研究设备有关联的一些医院的病人的可用性。健康的人、用非传统保健方式治疗的人或没有得到医学治疗的病人都不包括在研究计划之列。这样,一些重要的控制群体往往没有包括在研究设计中。因此,医生就不知道在没有医务治疗的情况下不同条件的人有多少能自然康复或没有药物干涉的情况下多少人能在不同条件下生存。我们就可以开始问这样的问题:医药干涉是阻碍还是延长了生命的满意度和长度?例如,对于前列腺癌,回答就是不予理睬常常比手术效果要好。于是,直到充分考虑了收集随机样本的研究计划产生之前,我们都不能回答以上的重要问题。

尽管对随机样本的渴求,我们笔者本身常常也将假设检验用于并非来自某一总体的随机样本的数据。在这些情况下,零假设表述为数据模式是完全偶然的,而不表述为某些参数具有特定值。当这样的零假设被拒绝时,结论为有某些非偶然因素起作用。即使样本不是那么完美,这仍然是很有帮助的结论。

调查研究的特殊问题

调查中也易于出现数据收集的严重问题。获得恰当的随机样本往往是非常困难而昂贵的。假设你现在要访问费城十几岁的年轻妈妈们。这听起来好像是一个精确定义的群体。但是,你怎样列出属于这个群体的所有年轻妈妈并将之作为总体来获得样本呢?即便总体知道了,获取样本也可能出问题。如果你下次观察你们当地商店的市场研究调查员,你会发现他们不可能让每个路过的人都参加调查。他们会避免接近那些烦躁的、邋遢的或不友善的购物者。统计学家常常可以通过矫正那些代表性不足或过分的群体来克服这个问题。最近,对于改进

全美人口普查的建议包括调整计数方式以弥补在某些区域对某些类型人群的不足估计。这些精细操作的具体细节不在本章的讨论范围之列,但是懂得对于有问题的样本可以通过对数据的细致处理来弥补总是令人安慰的。

尽管许多研究是基于并非随机收集的样本数据,研究者常常也并非一开始就有意要误导大众。只是在获得被访者随机样本的道路上有太多的障碍。电话调查是当今许多调查组织数据收集的重要来源,这必须与千万家庭目前还没有电话这一事实抗争。没有任何随机电话会找到那些家庭的人,而他们占总人口的6%。

调查的环境 数据收集本身也会有问题。例如,在访问的研究中,我们发现被访者反应的差异与访问者和被访者在性别、种族和个人爱好上的相似或者相异有关。我们知道谁提问和以什么样的口气提问对回答有影响。一个吸烟者如果知道调查人也吸烟,她会比面对一个不吸烟的提问者更容易承认她是多么爱吸烟,尤其是一个好斗的吸烟者。访问的地点同样很重要。被访者在舒服和私下的环境里比在商店里或电话中或大街上更乐意长久地交谈。谈话中上下文的影响也不能完全被忽略,必须作为数据收集过程的一个方面。

问题的内容是什么和在什么时候提出? 调查和实验的问题总是对被访者产生影响。问题的形式和排列影响结果。两个吸烟者,一个被提问“你认为你在多大程度上喜欢吸烟?”,另一个被提问“你认为你喜欢还是不喜欢吸烟?”。前者会比后者给出更为肯定的答案。

那些假定被访者对于主题有了解的调查表比来自缺少相应知识的样本的调查表会得到更多肯定的答复。但是收集的数据也可能会缺乏适用性。例如,在回答“你认为美国国会对最近联合国维和部队在中东的任务应做何反应?”时,被访者可能会假装有这方面知识。而如果他们被事先告知维和部队的任务或允许在说不知情的情况下保留面子的话,这种假装就不会发生。就问题的排列而言,有关宗教信仰的问题排在前面会比排在后面更多地对结果产生宗教影响。一项对大学生社会生活和宗教信仰的访问研究表明,当被访者先被确定宗教信仰时,他们对于后面的问题更加保守。(来源:W. W. Charters and T. M. Newcomb, “Some attitudinal effects of experimentally increased salience of a membership group,” in E. Maccoby, T. Newcomb, and E. Hartley (Eds.), *Readings in Social Psychology*, 3rd ed., New York: Holt, Rinehart & Winston, 1958) 当某个问题有意地把被访者引向某个方向,则再将它当作一个中立的问题来从统计上研究它的答案就是不正确的。(当然我们会问什么是一个真正的中立问题?)这些好像是更加心理学和社会学的而非统计学的事情,但是在试图理解数据的统计分析结果时它们是有关的。

什么是为分析选择的变量? 进行一项调查时,研究者必须决定提什么样的问题。这些决定很大程度上是基于研究的理论引导,它还和其它一些因子有关。例如,诸如研究的历史背景,以前提过的问题和其他研究者的工作,其它研究组织做了什么,适用于假设检验的设备,某个特定领域的传统等等。例如,产生出公司年度报告用的统计分析时,我们主要比较季度销售额、营业利润、市场份额和与其它同类公司相比。然而这些分析不包括新产品百分比、产品从研究到开发的速度和业务交易周期这样的变量。但是在当今,这些变量已经成为一个成功企业的生机所在,应该衡量并引入报告。像 Minnesota Mining and Manufacturing 这样的新兴公司已经将这些变量纳入股东报告,这是他们公司的部分使命和他们公司成功的重要评估标准。

结果是如何编码和储存的? 事情并不仅仅停留在提出问题。一旦某个问题被回答了,词句就以编码的方式进入了研究者设计的分类系统。研究者也许会将“我像旁边人一样爱抽烟”这样的反应评为1—7记分尺度中的4分,却并不知道被访者到底想表达怎样一种意思。

即使是被访者自己对她的吸烟兴趣评分为4分,研究者仍然不会真正清楚被访者对自己的评价及怎样把这个4分与另一个人的4分做比较。更为糟糕的是这种被编码的回答可能会被制作计算机数据文件的输入人员错误地登记下来。这种类型的错误如果发生10%将会导致怎样的后果呢?幸运的是调查研究者一般要对数据的输入做常规的重新检查以力争找出这种错误。

另外,人们给出的答案仅仅是他们的回答而已,并不代表他们的行为。有关人们用牙膏习惯的一项研究发现如果按照人们所说的来衡量,美国的牙膏销量应该比目前实际的销量大3倍。其中的潜在原因就是大多数研究场合下人们都有某种程度的夸张,而不管这种夸张意味着什么。



漫画来源:“PC and Pixel.” 1996, Washington Post Writers Group. Reprinted with permission.

分析方法的误用

在第三章中我们介绍了几种生成图表时可能会产生的问题。例如,我们看到在没有正确选择行和列或表中数据的小数位数太多时,这样的表格常常会误导我们。统计图常常会有太多的无用信息,即图表垃圾,掩盖了数据的主要信息。具有变动基线的条形图可能很难读懂,从而容易产生误导。那些形状奇怪的条形,例如人形或油桶形,高度可能合适但面积却不正确,事实上这也产生了误导。

我们必须用正确的数量来进行计算。回忆前面的知识,偏斜分布,例如收入分布,最好是以中位数表示。但是如果用均值就会是一个统计方法上的误用。然而均值却经常地被运用。这是因为在大众有关平均的观念中没有区别均值、中位数和众数的概念,而统计学家用均值却可以产生更多、更复杂的统计结果。

总之,统计分析的结果依赖于用于计算的数据(这点已经详细讨论过)和分析它们所用的统计方法。例如,在有关两个变量之间关系强度的回归分析中,我们用最小二乘法。如果用绝对值而不是平方就会得到不同的结果。因此,统计方法和数据一样对最后结论有贡献。

这里有关计算机程序的一个术语就是按步就班。由于计算机统计软件包简化数据带来的极大帮助,无论数据对或错,计算机都会分析它所得到的任何数据。在程序的自动运行和用户之间没有任何检查和平衡机制的干涉,也没有提出对数据可能误用的警告,所以算出来的结果可能跟数据中实际存在的信息毫无关系。在笔者身上就曾发生过一次这样的“灾难”。当笔者

的研究助手 Gergen 错误地将被调查者社会保险号码错误地当成头九个变量的信息输入到数据文件的前九列;后继的变量由于位置都错了而全部得到错误的值。虽然一位细心的读者费神发现标准差为 15 的大学年龄样本得出的平均年龄为 50 岁,但是,计算机处理数据时没有遇到任何麻烦。用计算机的行话说就是,输入错误的数字就输出错误的结果!

大多数标准的计算机统计软件包都是假定数据来自简单随机样本的前提下使用公式的。然而,大型的、全国性的数据研究常常是用更为复杂的抽样方法收集的,不应该用标准软件包分析。

统计推断的误用

如果把假设检验和置信区间都考虑在内,结论有时会是错误的。回忆以前的统计知识,当我们处理正确的零假设时,统计传统允许在 100 次中接受 5 次或更少次数的错误结论 ($p < 0.05$)。严格地说,这些错误并非统计的误用,但是认识到我们常常犯错是非常重要的。统计与其它数学之间的区别在于,在统计中我们期望有时要犯错误;统计方法本身使它可以表述为重复实验多次后,我们所遇错误的频率。不像其它科学,统计从来不打算使自身完美无缺,统计意味着你永远不需要确定无疑。

例如,0.05 的 p -值承认结论在 100 次中会错 5 次。在所有零假设为正确的 100 次假设分析中,如果得出其中 5 次结果是显著的(则拒绝零假设),那么这些结果是纯粹由偶然因素造成的。显然,我们对此还不能确定,因为我们不知道某一特定零假设是正确还是错误。我们不可能揭去笼罩在这个称为真理的隐蔽实体上的不确定性面纱。即使我们知道错误的频率,也不知道错误到来的时刻。

数字的错误解释

Mercedes-Benz 公司的广告说:过去 15 年以来被注册的他们公司的车中有 97% 仍然在使用中,并且这个数比任何在这段时间内销售的其它公司汽车的相应数要大。我们怎样来解释这个 97% 呢?

一个数可以像一句话那样诉说一个故事。对于任何故事,我们都是在我们能理解的意义下来解释这些数的。汽车制造商希望我们从这个百分比中听出什么故事呢? 他们希望我们有这样的想法:这是一个很高的百分比,大多数他们售出的汽车仍在驾驶当中,这反映了优良的质量和购买他们公司汽车带来的好处。最后,希望我们购买他们生产的汽车。广告没有把这些话都说出来,但是这的确是他们想要我们得出的结论。

从统计的观点看,事情不是这么简单。首先,公司怎样得到以上这种信息? 每个州都有机动车注册办公室,公司可以与 50 个州的办公室联系,了解每个州中 15 年来有多少他们生产的汽车被注册。由于公司知道同时期他们卖出汽车的数量,他们就可以算出仍然注册在案的汽车的百分比。因为这些记录是用计算机处理的,他们就假设这些记录像当初输入计算机时一样准确。然而,各州都有其独特的收集和检索程序。从每个州获取注册信息或把有效数据组合起来得出概括统计量并都不是那么简单的事情。

其次,15 年来销售的汽车是怎样被认为处于使用当中? 这些汽车被卖出的时候就存在着很大的差异。如果在头 10 年只卖出了很少的汽车而后 5 年卖出了大量的汽车,则大多数车仍

在路上奔驰并没有什么希奇的。假若卖出的汽车逐年增长,则许多车将相对比较新,大多数仍然能驾驶。如果每辆车一年平均行驶 12000 英里,并且预计可以行驶 100000 英里,则汽车的平均寿命为 8 年。

因此,不知道 15 年来的销售状况,我们不可能清楚地解释这则广告声明。然而,这样的声明不断出现:我们却只被告知故意将我们引向某结论的几个数。但是,当我们审慎地去想一想这些数据的质量及还会有什么其它可供选择的解释时,这些数据表示的信息就不再明朗了。

14.5 统计和专制

现在从怀疑主义转到一个可能更严重的社会关注的问题;统计知识还有调节日常生活的势力。强大而一致的统计收集系统的一个显著毛病就是人们很容易受到政府和私营企业利益的监视。在完整而精细的统计网络的帮助下专制成为现实。历史上,为了州利益,统计分析被社会精英们用于监视民众,尤其是用于收税和征兵。基督徒都知道,新约全书中耶稣的诞生是以一个统计故事开始的。Joseph 被邀请到 Bethlehem 参加在 David 屋里举行的人口统计调查。

传统上,在自由民主社会的公民不愿意让政府的中央集权弄清个人的状况。像美国人民自由联合会这样的组织对保护个人权利,反对强加于人做过很大贡献。近些年来,随着人们对持续监视的适应,这方面的敏感度越来越弱。例如人们已习惯不经允许就拍下顾客照片的银行取款机、保留私人谈话详细记录的电话公司、商店跟踪顾客的隐蔽摄像机及单向镜子、记录从幼儿园到研究生表现的学校档案和在百万观众面前揭开个人隐私和创伤的谈论节目主持人。

停下来 想一想 14.1

统计信息在个人生活中有很强应用。下面列举的在性行为、生育控制和堕胎方面的信息的形式会怎样影响你的想法呢?

在美国,大约 5 千 5 百万的女性年龄在 13 到 45 岁之间,其中 1 千 1 百万的人通过节制性欲来避孕,她们大多数是十几岁和年轻的成年人。Alan Guttmacher 研究所关于美国女性怀孕倾向的调查表明,每年在 4 千 4 百万性生活活跃的可生育女性中,只有少于 6 百万的女性想要怀孕。

剩下的 3 千 8 百万性生活活跃的育龄女性都不想怀孕。每年,她们中 90% 的人可以通过不同的生育控制措施在月经周期里成功避孕。总的看来,这还只是表面现象。每年美国可生育女性通过禁欲和避孕措施成功避免卵子受精 6 亿次。

不想要孩子的可生育女性中只有 8% 的人在她们避孕的家庭计划上失败。但是哪怕这样小的百分比也意味着每年有 4 百万的女性不期而孕。统计上说,全美的每一位女性都有可能曾经或将要不期而孕。(来源: *Biology and Gender Study Group*, "The Importance of Feminist Critique for Contemporary Biology," unpublished manuscript, 1993.)

然而统计本身是智者的忠实奴仆,当被重要组织的铁腕领袖使用时,它可能会变成压迫的工具。这中间包括了像警察、税收等政府机构、企业和保险公司。医院记录、能力测验及其它

个人技术、个性、特点、性格和兴趣等的清单都是可以以不同方式存储和使用的统计衡量形式。安全部门的识别代码,像指纹、视网膜形状和声音特征等,都是以统计的评价为基础,都可以用来存储和检索个人的行踪。某人所打长途电话的种类和频率会产生一个模式档案如果这个模式被违反了,一个基于统计数目的系统就会自动吊销他的信用卡。包括血液、头发、精子和皮肤分析的罪犯查验系统依赖于统计输入,并已成为永久可查记录的一部分。作为打击破坏活动的一种方式,最近政府建议对所有航线乘客做背景检查以便从这些数据的统计编辑中产生恐怖分子的模式档案。一旦统计的精细测量工作启动了,我们的私人生活就不可避免的要被介入和监督。

作为一个有素养的统计信息使用者,我们的任务是确定收集、存储和发布统计信息的界线。谁应该在怎样的环境下寻找什么样的结果?人们应该在什么时候以什么样的方式防范计算机时代统计对我们生活的管理和控制。

14.6 在高潮时结束

对统计信息被用于威胁个人自由而感到遗憾并不能改变什么。但我们公民对统计的功用和限制了解得尽可能多则是有帮助的。这样,我们理解到生活在一个看似随机和混乱的世界里,但在所有这些随机之中存在着规律,而统计方法就是用来发现这些规律。我们还知道应该用批判的眼光看待统计分析中数据的质量,任何统计分析的结果都受到数据质量和分析数据的特殊方法的影响。

我们欣赏统计上有根据的推断在导出结论上远远比个人经验、专题评论组、非正式调查或个人逻辑(包括依靠类比、常识、权威信息为基础的原理来做结论)等浮夸矫饰的形式更能理解、有逻辑的一致性、而绝对没有固执的教条。以上每一种浮夸矫饰的形式都在说理辩论的发展史上占有重要地位,但每一种都在许多方面有弱点,而这些弱点正是统计的方法和推断所不会有的。

统计方法能用来理解公众的意见并帮助确立国家应采取的公共政策。统计已经在发展我们所购买的许多商品和服务上扮演了极重要的角色。例如,新车型不像旧的那么容易坏是因为其改进的质量,这要感谢当今存在的复杂的统计质量控制体系。统计对于医学实践和对付疾病的药物的可得性上也有很大的影响。浏览一下各章后面的解释和分析习题,我们可以发现统计用于非常广阔的应用领域。简言之,任何有收集的经验数据的场合都需要统计方法。统计与社会的幸福是如此密不可分,我们不敢想象在一个没有统计的世界我们将如何运作。

我们留给你的是这样一个希望和信念:对于所面临的越来越多的统计信息,你现在可以作出一个更为完美的决策了。我们认为这是面向新世纪教育的重要组成部分。

补充读物

Crossen, Cynthia. *Tainted Truths: The Manipulation of Facts in America*. Simon & Schuster, 1994. 在解释数值结果方面是本好书。

Eberstadt, Nicholas. *The Tyranny of Numbers: Mismeasurement and Misrule*. Washington, DC: AEI Press, 1995. 政府统计的作用。

Hooke, Robert. *How to Tell the Liars from the Statisticians*. New York: Marcel Dekker, 1983. 一本关于统计误用的幽默书。

Jaffe, A. J., and Herbert F. Spierer. *Misused Statistics: Straight Talk for Twisted Numbers*. New York: Marcel Dekker, 1987. 统计探索可能出错的许多情况。

Paulos, John A. *A Mathematician Reads the Newspaper*. New York: Basic Books, 1995. 在解释报纸中的数字方面的一本幽默而又有教益的书。

习 题

- 14.1 在报纸、新闻杂志、刊物或书本上找出一个统计研究的例子,用本章的某些观点评价它。
- 14.2 数据收集集中存在的哪些主要问题即使经过细致负责的计划仍然不能很容易克服?
- 14.3 a. 统计学被描述为一个很好运作的州的基本因素。请举一个历史事例。
b. 统计学是否以某种方式造成社会的个人隐私水平的减少? 请举一个历史事例。
- 14.4 a. 为什么理解诸如报纸和杂志之类上的统计报告对人们如此重要?
b. 在你的日常生活中举一个由于错误理解(例如报纸杂志上的)统计结果而导致负面效应的例子。
- 14.5 描述一个你认为广告人或其它媒体制作者有意给出不可信统计结果的事例。
- 14.6 检查用来对当地民众显示全国、地区或当地的他们关心的某种显著趋势的统计报告和图表。尝试通过改变这些结果的某些特征来显著地改变数据对读者所表达的内涵。如果可能的话,以多种方式进行尝试。写一个报告表明针对某些结果的某些观点是这些公布的数据所特有的;在表现形式、统计方法,数据收集等方面的改变怎样影响这些结果。
- 14.7 制作一个包括某些对你自己是重要因素的研究计划。(如果可能的话,在你自己课程的限度中,实际地收集数据或以想象的对象为基础建立一个数据集。这项习题还会加强你从处理“真实”数据而获得的“动手触摸”的感觉。)对于这项研究之所以重要的原因、这项研究所作的假设及假设检验的方式给出一个基本原理。描述你怎样选择你的样本,开发你的研究工具,产生适宜的研究设置,并考虑围绕着你努力的道德和社会事宜。描述你怎样收集数据,怎样组织它,怎样用图表来表现它和怎样对数据进行统计分析。

朋友们,统计就是一切! (Stat's all, folks! ^①)

^① 这是借用美国动话片的结尾“That's all, folks!”——译者注。

统计术语

Alternative hypothesis, 备择假设: 参数除零假设中所标之外的其它可能值。

Analysis of variance, 方差分析: 用于分析一个或多个分类型自变量与一个数值型因变量之间关系的统计方法。

Analysis of variance table, 方差分析表: 显示平方和、自由度、均方、 F -比和 p -值的表格。

Average absolute deviation, 平均绝对偏差: 观测值与均值之间差异的绝对值的平均。

Bar graph, 条形图: 用条形来表示变量每个值的观测数目的图。

Binomial distribution, 二项分布: 用来表示在 n 次只有两个结果的实验中得到 x 次某种结果和 $n - x$ 次另一种结果的理论分布。

Boxplot, 盒形图: 表示观测的最大、最小值, 以第 75 和第 25 百分位点画一个盒子并过 50 百分位点有一条线的图。

Categorical variable 分类型变量: 具有不同值的变量。这些值之间不能排序, 而且我们不能说某个值比另一个值更怎么样。

Census, 人口普查: 从一个人口总体收集数据的过程。

Central value, 中心值: 用来代表某一变量所有观测值的一个值。

Chartjunk, 图中垃圾: 图中不携带信息的多余元素。

Chi-square distribution, χ^2 分布: 用来从样本数据作推断的理论分布。

Confidence interval, 置信区间: 有希望可能包含总体参数的区间。

Constant, 常数: 只取一个值的量, 经常是一个参数。

Contingency table, 列联表: 表示两个分类型变量频率的表。

Control group, 控制组: 实验中随机选取的元素子集, 这些元素并没有像实验组那样被操纵。

Control variable, 控制变量: 为了考察两个变量之间是否有因果关系而引入的另一个变量。

Correlation coefficient r , 相关系数 r : 两个数值型变量之间关系强度的度量。

Critical value, 临界值: 一个或多个为样本统计量事先确定的值; 如果观测的统计量比该值更为极端则拒绝零假设。

Cramer's V , Cramer V : 见 V 。

Data analysis, 数据分析: 通过图表和计算来简化数据。

Data file (data matrix), 数据文件(数据矩阵): 一个数据表格, 它的列包含了变量的观测值, 它的行包含了元素的观测值。

Data density, 数据密度: 一个图中每平方英寸包含的观测值数目。

Degrees of freedom, 自由度: 为得出所有观测值所需的最小观测值数目。

Dependent variable, 因变量: 受一个或多个自变量影响的变量。

Dummy variable, 哑元: 只取两个值的变量, 常常为 0 和 1; 它用于表示分类型变量。

Element, 元素: 我们衡量一个变量的单位。

Error of type I, 第一类错误: 拒绝一个为真的零假设带来的错误。

Error of type II, 第二类错误: 不拒绝一个伪的零假设带来的错误。

Estimation, 估计: 试图得出一个参数的值。

Expected frequency, 期望频率: 在列联表的每个单元格计算出来的使两个分类型变量之间没有关系的频率。

Experimental data, 实验数据: 当控制某些变量的值时我们收集的数据。

Experimental design, 实验设计: 研究怎样设计实验和分析实验数据的统计理论分支。

F-distribution F -分布: 用来从样本数据作推断的理论分布。

Fiftieth percentile 第 50 百分位点: 把数据分成相等两个组的变量的值; 一个组中所有观测值小于这个数而另一个组中所有观测值大于这个数。

Frequency distribution, 频率分布: 一些数对的集合, 其中第一个分量是变量的值而另一个分量是这个观测值的数目, 常常以直方图的形式来表示。

Gamma, γ : 当变量的某些数值上有多个观测值时用来衡量两个顺序变量之间的关联。

Graphical excellence, 图优性: 在最短的时间内和最小的空间里, 花最少的笔墨对看图者表达最多的概念的图的质量概念; 即在图形中简洁、准确和有效地表达复杂的思想。

Group sum of squares, 组平方和: 方差分析中衡量一个自变量效应的平方和。

Histogram, 直方图: 用一些长方形表示数值型变量的分布, 用面积表示数值频率的图。

Hypothesis testing, 假设检验: 试图发现某个参数是否为特定值。

Independent variable, 自变量: 猜想对因变量有影响的先发生的变量。

Inference, 推断: 从样本数据得出的对总体的总结。

Intercept, 截距: 当自变量为零时对因变量的估计值。

Interaction effect, 交互作用效应: 两个自变量超出它们各自效应之外的联合效应。

Interquartile range, 四分位点间距: 某个变量的第 75 分位点减去它的第 25 分位点。

Interval estimate, 区间估计: 见置信区间 (Confidence interval)。

Interval variable, 区间变量: 见数值型变量 (Metric variable)。

Lineplot, 线距: 把变量的值标在直线上, 并将每个观测值记为直线上的一个点。

Logistic regression, 逻辑回归: 把分类型变量作为因变量的回归分析。

Mean, 均值: 将所有观测值的和除以观测值个数而得的一个变量值。

Measure of association, 关联的测度: 衡量两个变量之间关系强度的数, 其值域为 -1 到 $+1$ 。

Median, 中位数: 这个变量值将所有观测值分成两半, 一半小于此数而另一半大于此数。

Metric variable, 数值型变量: 有一个测度单位的变量; 我们可以说出一个值比另一个值大多少或者小多少。

Mode, 众数: 出现次数最多的变量值。

Multiple correlation coefficient R , 多重相关系数 R : 衡量一个因变量的观测值与预测值之间关系强度的相关系数。

Multivariate analysis, 多元分析: 对于两个或多个自变量对一个因变量效应的研究。

Nominal variable, 标识变量: 见分类型变量 (Categorical variable)。

Nonresponse error, 未响应误差: 抽样调查时当不是所有样本中的被采访者对部分或全部调查

有回应时发生的误差。

Normal distribution, 正态分布: 统计理论中广泛应用的一种特别的单峰、对称的理论分布。

Null hypothesis, 零假设: 有关参数取值的表述。

Observational data, 观测数据: 从反映真实世界的观测中收集的数据。

Odds, 优势: 分子为某事件的失败次数而分母为事件的成功次数的一个数值比。

One-sided (one-tailed) test, 单边(单尾)检验: 当样本统计量在某一特定方向与总体参数有差别时拒绝零假设的检验。

p -value p -值: 观测到某个样本统计量或更为极端的样本统计量的概率。

Parameter, 参数: 描绘总体中一个或多个变量的常数, 如均值、方差和回归系数等, 常常用希腊字母表示。

Partial coefficient, 偏系数: 控制一个或多个变量时, 度量两个变量之间关系的系数。

Percentile, 分位点: 将观测值分成两组使得某一百分比的观测值小于这个数的变量值。

Φ , ϕ : 分别只有两个取值的两个分类型变量之间关系强度的度量。

Pie graph, 圆饼图: 根据每个值的观测数目用扇形大小表示的变量分布的圆形图。

Point estimate, 点估计: 对总体参数进行估计的单个数值。

Poisson distribution Poisson 分布: 表示稀有事件发生次数概率的理论分布。

Population, 总体: 被研究的所有元素的总和。

Predicted value, 预测值: 由自变量估计的因变量的值; 在回归分析中是回归线上的一点。

Probability, 概率: 很长的时期内事件发生次数的比例。

Random sample, 随机样本: 每个元素都以某一(有时相等的)已知概率被选入的样本。

Range, 极差: 最大与最小观测值之间的差。

Rank order correlation, 秩顺序相关系数: 用于衡量把秩(排序)作为取值的两个顺序变量之间关系强度的系数。

Rank variable, 顺序变量: 按顺序排列其值的变量; 但我们不能比较一个值比另一个大多少。

Ratio variable, 比例变量: 见数值型变量(Metric variable)。

Regression analysis, 回归分析: 分析数值型变量之间关系的统计方法。

Regression coefficient, 回归系数: 回归直线的斜率; 表示当自变量相差一个单位时因变量的两个值平均相差多少。

Regression line, 回归直线: 总结两个数值型变量之间关系的直线。

Regression sum of squares, 回归平方和: 衡量自变量效应的平方和。

Residual, 残差: 从某个观测点到相应预测点的竖直距离; 衡量除自变量之外其它所有变量对因变量的效应。

Residual mean square, 残差均方: 残差的方差。

Residual sum of squares, 残差平方和: 衡量残差变量效应的平方和。

Residual variable, 残差变量: 除自变量之外其它所有变量对因变量的联合效应; “所有其它的”变量。

Response error, 响应误差: 由诸如提问的形式、问题的位置之类因素对被访者产生影响而导致的响应上的误差。

Sample, 样本: 从总体中抽选出的因素集合。

Sampling error, 样本误差: 如果选取了许多样本, 其中 95% 的样本得出的结果与总体的真值之间的差异。

Scatterplot, 散点图: 用点表示两个数值型变量观测值的图, 每个点代表一个观测数对。

Sign test, 符号检验: 考察一个变量两次测量之间变化的检验。

Significance level, 显著水平: 事先决定的拒绝一个为真的零假设的概率。

Simple random sample, 简单随机样本: 总体中每个元素都有同等机会被选入其中的样本。

Spurious relationship, 伪关系: 观测到的两个变量之间不是因果相关的关系。

Standard deviation, 标准差: 观测值与均值之间的平均距离; 由方差的平方根得出, 用与变量相同的单位来衡量。

Standard error, 标准误差: 从许多不同样本算出的某个统计量的标准差。

Standard score, 标准得分: 用观测值减去均值再除以标准差而算出的得分。

Standard normal variable, 标准正态变量: 均值为 0, 标准差为 1 的具有正态分布的变量。

Statistic, 统计量: 从样本观测值算出的如均值、方差或相关系数等数。

Statistical significance, 统计意义上显著: 样本结果拒绝零假设的状况。

Statistics, 统计学: 收集、分析数据并从中得出结论的一系列概念、原则和方法; 在随机中寻找规律的研究。

Stemplot, 茎叶图: 将观测值的较大部分列在左边直线上, 而最小的整数部分列在直线右边的图。

Table, 表格: 表示一个或多个变量观测值的频率、百分比或概率的排列。

t -distribution t -分布: 与正态分布有关的 t -变量的单峰、对称的理论分布。

Total sum of squares, 总计平方和: 度量所有变量效应的平方和。

Two-way analysis of variance, 双因子方差分析: 有关一个数值型因变量与两个分类型自变量之间关系的研究。

Two-sided (two-tailed) test, 双边(双尾)检验: 当样本统计量非常小于或者大于零假设中的总体参数值时都拒绝零假设的检验。

Unbiased estimate, 无偏估计: 从许多样本得出的样本均值等于总体参数的估计。

V , V : 当一个或所有变量有三个或更多的值时, 衡量两个分类型变量之间关系强度的统计量。

Values, 赋值: 被赋以一个固定值的组。

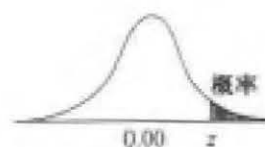
Variable, 变量: 可以取两个或多个值的特征, 品质或属性。

Variance, 方差: 观测值与均值之间变差平方的平均值, 用变量本身单位的平方来衡量; 标准差的平方。

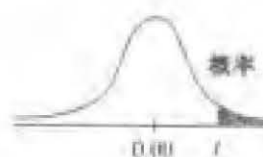
Variation, 变差: 一个数值型变量的一组观测值之间差异的总和。

统计表

统计表 1a 尾概率从 0.50 到 0.01 的 z 值



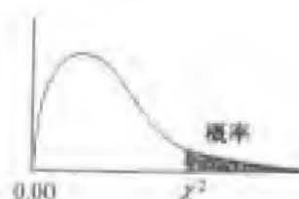
概率	0.50	0.40	0.25	0.15	0.10	0.05	0.025	0.010	0.005	0.001	0.0001
z	0.00	0.25	0.67	1.04	1.28	1.64	1.96	2.33	2.58	3.09	3.72

统计表 2 对于尾概率从 0.50 到 0.001 的 t 值

自由度	概率									
	0.50	0.40	0.25	0.15	0.10	0.05	0.025	0.010	0.005	0.001
1	0.00	0.32	1.00	1.96	3.08	6.31	12.71	31.82		
2	0.00	0.29	0.82	1.39	1.88	2.92	4.30	6.96	9.92	
3	0.00	0.28	0.76	1.25	1.64	2.35	3.18	4.54	5.84	10.21
4	0.00	0.27	0.74	1.19	1.53	2.13	2.78	3.75	4.60	7.17
5	0.00	0.27	0.73	1.16	1.48	2.02	2.57	3.36	4.03	5.89
6	0.00	0.26	0.72	1.13	1.44	1.94	2.45	3.14	3.71	5.21
7	0.00	0.26	0.71	1.12	1.41	1.89	2.36	3.00	3.50	4.78
8	0.00	0.26	0.71	1.11	1.40	1.86	2.31	2.90	3.36	4.50
9	0.00	0.26	0.70	1.10	1.38	1.83	2.26	2.82	3.25	4.30
10	0.00	0.26	0.70	1.09	1.37	1.81	2.23	2.76	3.17	4.14
11	0.00	0.26	0.70	1.09	1.36	1.80	2.20	2.72	3.11	4.02
12	0.00	0.26	0.70	1.08	1.36	1.78	2.18	2.68	3.05	3.93
13	0.00	0.26	0.69	1.08	1.35	1.77	2.16	2.65	3.01	3.85
14	0.00	0.26	0.69	1.08	1.35	1.76	2.14	2.62	2.98	3.79
15	0.00	0.26	0.69	1.07	1.34	1.75	2.13	2.60	2.95	3.73
16	0.00	0.26	0.69	1.07	1.34	1.75	2.12	2.58	2.92	3.69
17	0.00	0.26	0.69	1.07	1.33	1.74	2.11	2.57	2.90	3.65
18	0.00	0.26	0.69	1.07	1.33	1.73	2.10	2.55	2.88	3.61
19	0.00	0.26	0.69	1.07	1.33	1.73	2.09	2.54	2.86	3.58
20	0.00	0.26	0.69	1.06	1.33	1.72	2.09	2.53	2.85	3.55
21	0.00	0.26	0.69	1.06	1.32	1.72	2.08	2.52	2.83	3.53
22	0.00	0.26	0.69	1.06	1.32	1.72	2.07	2.51	2.82	3.50
23	0.00	0.26	0.69	1.06	1.32	1.71	2.07	2.50	2.81	3.48
24	0.00	0.26	0.68	1.06	1.32	1.71	2.06	2.49	2.80	3.47
25	0.00	0.26	0.68	1.06	1.32	1.71	2.06	2.49	2.79	3.45
26	0.00	0.26	0.68	1.06	1.31	1.71	2.06	2.48	2.78	3.43
27	0.00	0.26	0.68	1.06	1.31	1.70	2.05	2.47	2.77	3.42
28	0.00	0.26	0.68	1.06	1.31	1.70	2.05	2.47	2.76	3.41
29	0.00	0.26	0.68	1.06	1.31	1.70	2.05	2.46	2.76	3.40
30	0.00	0.26	0.68	1.05	1.31	1.70	2.04	2.46	2.75	3.39
32	0.00	0.26	0.68	1.05	1.30	1.69	2.04	2.45	2.74	3.37
34	0.00	0.26	0.68	1.05	1.30	1.69	2.03	2.44	2.73	3.35
36	0.00	0.26	0.68	1.05	1.30	1.69	2.03	2.43	2.72	3.33
38	0.00	0.26	0.68	1.05	1.30	1.69	2.02	2.43	2.71	3.32
40	0.00	0.26	0.68	1.05	1.30	1.68	2.02	2.42	2.70	3.31

统计表 2 对尾概率从 0.50 到 0.001 的 t 值(继续)

自由度	概率									
	0.50	0.40	0.25	0.15	0.10	0.05	0.025	0.010	0.005	0.001
42	0.00	0.25	0.68	1.05	1.30	1.68	2.02	2.42	2.70	3.30
44	0.00	0.25	0.68	1.05	1.30	1.68	2.02	2.41	2.69	3.29
46	0.00	0.25	0.68	1.05	1.30	1.68	2.02	2.41	2.69	3.28
48	0.00	0.25	0.68	1.05	1.30	1.68	2.01	2.41	2.68	3.27
50	0.00	0.25	0.68	1.05	1.30	1.68	2.01	2.40	2.68	3.26
55	0.00	0.25	0.68	1.05	1.30	1.67	2.00	2.40	2.67	3.25
60	0.00	0.25	0.68	1.05	1.30	1.67	2.00	2.39	2.66	3.23
65	0.00	0.25	0.68	1.04	1.29	1.67	2.00	2.39	2.65	3.22
70	0.00	0.25	0.68	1.04	1.29	1.67	1.99	2.38	2.65	3.21
75	0.00	0.25	0.68	1.04	1.29	1.67	1.99	2.38	2.64	3.20
80	0.00	0.25	0.68	1.04	1.29	1.66	1.99	2.37	2.64	3.20
90	0.00	0.25	0.68	1.04	1.29	1.66	1.99	2.37	2.63	3.18
100	0.00	0.25	0.68	1.04	1.29	1.66	1.98	2.36	2.63	3.17
200	0.00	0.25	0.68	1.04	1.29	1.65	1.97	2.35	2.60	3.13
500	0.00	0.25	0.67	1.04	1.28	1.65	1.96	2.33	2.59	3.11
无穷	0.00	0.25	0.67	1.04	1.28	1.64	1.96	2.33	2.58	3.09

统计表 3 对尾概率从 0.50 到 0.001 的 χ^2 值

自由度	概率									
	0.50	0.40	0.25	0.15	0.10	0.05	0.025	0.010	0.005	0.001
1	0.45	0.71	1.32	2.07	2.71	3.84	5.02	6.63	7.88	10.83
2	1.39	1.83	2.77	3.79	4.61	5.99	7.38	9.21	10.60	13.82
3	2.37	2.95	4.11	5.32	6.25	7.81	9.35	11.34	12.84	16.27
4	3.36	4.04	5.39	6.74	7.78	9.49	11.14	13.28	14.86	18.47
5	4.35	5.13	6.68	8.12	9.24	11.07	12.83	15.09	16.75	20.52
6	5.35	6.21	7.84	9.45	10.64	12.59	14.45	16.81	18.55	22.46
7	6.35	7.28	9.04	10.75	12.02	14.07	16.01	18.48	20.28	24.32
8	7.34	8.35	10.22	12.03	13.36	15.51	17.53	20.09	21.95	26.12
9	8.34	9.41	11.39	13.29	14.68	16.92	19.02	21.67	23.59	27.88
10	9.34	10.47	12.55	14.53	15.99	18.31	20.48	23.21	25.19	29.59
11	10.34	11.53	13.70	15.77	17.28	19.68	21.92	24.72	26.76	31.26
12	11.34	12.58	14.85	16.99	18.55	21.03	23.34	26.22	28.30	32.91
13	12.34	13.64	15.98	18.20	19.81	22.36	24.74	27.69	29.82	34.53
14	13.34	14.69	17.12	19.41	21.06	23.68	26.12	29.14	31.32	36.12
15	14.34	15.73	18.25	20.60	22.31	25.00	27.49	30.58	32.80	37.70
16	15.34	16.78	19.37	21.79	23.54	26.30	28.85	32.00	34.27	39.25
17	16.34	17.82	20.49	22.98	24.77	27.59	30.19	33.41	35.72	40.79
18	17.34	18.87	21.60	24.16	25.99	28.87	31.53	34.81	37.16	42.31
19	18.34	19.91	22.72	25.33	27.20	30.14	32.85	36.19	38.58	43.82
20	19.34	20.95	23.83	26.50	28.41	31.41	34.17	37.57	40.00	45.31
21	20.34	21.99	24.93	27.66	29.62	32.67	35.48	38.93	41.40	46.80
22	21.34	23.03	26.04	28.82	30.81	33.92	36.78	40.29	42.80	48.27
23	22.34	24.07	27.14	29.98	32.01	35.17	38.08	41.64	44.18	49.73
24	23.34	25.11	28.24	31.13	33.20	36.42	39.36	42.98	45.56	51.18
25	24.34	26.14	29.34	32.28	34.38	37.65	40.65	44.31	46.93	52.62
26	25.34	27.18	30.43	33.43	35.56	38.89	41.92	45.64	48.29	54.05
27	26.34	28.21	31.53	34.57	36.74	40.11	43.19	46.96	49.64	55.48
28	27.34	29.25	32.62	35.71	37.92	41.34	44.46	48.28	50.99	56.89
29	28.34	30.28	33.71	36.85	39.09	42.56	45.72	49.59	52.34	58.30
30	29.34	31.32	34.80	37.99	40.26	43.77	46.98	50.89	53.67	59.70
32	31.34	33.38	36.97	40.26	42.58	46.19	49.48	53.49	56.33	62.49
34	33.34	35.44	39.14	42.51	44.90	48.60	51.97	56.06	58.96	65.25
36	35.34	37.50	41.30	44.76	47.21	51.00	54.44	58.62	61.58	67.99
38	37.34	39.56	43.46	47.01	49.51	53.38	56.90	61.16	64.18	70.70
40	39.34	41.62	45.62	49.24	51.81	55.76	59.34	63.69	66.77	73.40

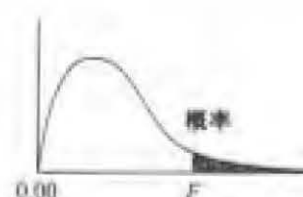
统计表 3 对尾概率从 0.50 到 0.001 的 χ^2 值(继续)

自由度	概率									
	0.50	0.40	0.25	0.15	0.10	0.05	0.025	0.010	0.005	0.001
42	41.34	43.68	47.77	51.47	54.09	58.12	61.78	66.21	69.34	76.08
44	43.34	45.73	49.91	53.70	56.37	60.48	64.20	68.71	71.89	78.75
46	45.34	47.79	52.06	55.92	58.64	62.83	66.62	71.20	74.44	81.40
48	47.34	49.84	54.20	58.14	60.91	65.17	69.02	73.68	76.97	84.04
50	49.33	51.89	56.33	60.35	63.17	67.50	71.42	76.15	79.49	86.66

对于自由度比表上所列的大的 χ^2 的值利用

$$\frac{\chi^2 - \sqrt{d.f.}}{\sqrt{2 d.f.}} = z$$

近似地服从正态分布的事实。

统计表 4 对尾概率从 0.10 到 0.001 的 F 值

		分子 d.f.								
分母 d.f.	概率	1	2	3	4	5	6	7	8	9
1	0.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	0.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
	0.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
	0.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
	0.001	405284	500000	540379	562500	576405	585937	592873	598144	602284
2	0.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	0.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
	0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	0.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	0.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39
3	0.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	0.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	0.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
	0.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
	0.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.65	129.86
4	0.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
	0.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
	0.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
	0.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
	0.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47
5	0.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
	0.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
	0.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
	0.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
	0.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24
6	0.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
	0.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
	0.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
	0.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
	0.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69
7	0.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
	0.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
	0.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82
	0.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
	0.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33

分子 d.f.

10	12	15	20	30	40	50	75	120	1000
60.19	60.71	61.22	61.74	62.26	62.53	62.69	62.90	63.06	63.30
241.88	243.91	245.95	248.01	250.10	251.14	251.77	252.62	253.25	254.19
968.63	976.71	984.87	993.10	1001.4	1005.6	1008.1	1011.5	1014.0	1017.8
6055.8	6106.3	6157.3	6208.7	6260.6	6286.8	6302.5	6323.6	6339.4	6362.7
605621	610668	615764	620908	626099	628712	630285	632390	633972	636301
9.39	9.41	9.42	9.44	9.46	9.47	9.47	9.48	9.48	9.49
19.40	19.41	19.43	19.45	19.46	19.47	19.48	19.48	19.49	19.49
39.40	39.41	39.43	39.45	39.46	39.47	39.48	39.48	39.49	39.50
99.40	99.42	99.43	99.45	99.47	99.47	99.48	99.49	99.49	99.50
999.40	999.42	999.43	999.45	999.47	999.47	999.48	999.49	999.49	999.50
5.23	5.22	5.20	5.18	5.17	5.16	5.15	5.15	5.14	5.13
8.79	8.74	8.70	8.66	8.62	8.59	8.58	8.56	8.55	8.53
14.42	14.34	14.25	14.17	14.08	14.04	14.01	13.97	13.95	13.91
27.23	27.05	26.87	26.69	26.50	26.41	26.35	26.28	26.22	26.14
129.25	128.32	127.37	126.42	125.45	124.96	124.66	124.27	123.97	123.53
3.92	3.90	3.87	3.84	3.82	3.80	3.80	3.78	3.78	3.76
5.96	5.91	5.86	5.80	5.75	5.72	5.70	5.68	5.66	5.63
8.84	8.75	8.66	8.56	8.46	8.41	8.38	8.34	8.31	8.26
14.55	14.37	14.20	14.02	13.84	13.75	13.69	13.61	13.56	13.47
48.05	47.41	46.76	46.10	45.43	45.09	44.88	44.61	44.40	44.09
3.30	3.27	3.24	3.21	3.17	3.16	3.15	3.13	3.12	3.11
4.74	4.68	4.62	4.56	4.50	4.46	4.44	4.42	4.40	4.37
6.62	6.52	6.43	6.33	6.23	6.18	6.14	6.10	6.07	6.02
10.05	9.89	9.72	9.55	9.38	9.29	9.24	9.17	9.11	9.03
26.92	26.42	25.91	25.39	24.87	24.60	24.44	24.22	24.06	23.82
2.94	2.90	2.87	2.84	2.80	2.78	2.77	2.75	2.74	2.72
4.06	4.00	3.94	3.87	3.81	3.77	3.75	3.73	3.70	3.67
5.46	5.37	5.27	5.17	5.07	5.01	4.98	4.94	4.90	4.86
7.87	7.72	7.56	7.40	7.23	7.14	7.09	7.02	6.97	6.89
18.41	17.99	17.56	17.12	16.67	16.44	16.31	16.12	15.98	15.77
2.70	2.67	2.63	2.59	2.56	2.54	2.52	2.51	2.49	2.47
3.64	3.57	3.51	3.44	3.38	3.34	3.32	3.29	3.27	3.23
4.76	4.67	4.57	4.47	4.36	4.31	4.28	4.23	4.20	4.15
6.62	6.47	6.31	6.16	5.99	5.91	5.86	5.79	5.74	5.66
14.08	13.71	13.32	12.93	12.53	12.33	12.20	12.04	11.91	11.72

统计表 4 对尾概率从 0.10 到 0.001 的 F 值(继续)

		分子 d.f.								
分母 d.f.	概率	1	2	3	4	5	6	7	8	9
8	0.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
	0.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
	0.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
	0.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	0.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77
9	0.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	0.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
	0.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
	0.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	0.001	22.86	16.39	13.90	12.56	11.71	11.18	10.70	10.37	10.11
10	0.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
	0.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
	0.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
	0.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
	0.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96
11	0.100	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27
	0.050	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
	0.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59
	0.010	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
	0.001	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12
12	0.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
	0.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
	0.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44
	0.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
	0.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48
13	0.100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16
	0.050	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
	0.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31
	0.010	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
	0.001	17.82	12.81	10.21	9.07	8.35	7.86	7.49	7.21	6.98
14	0.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
	0.050	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
	0.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21
	0.010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
	0.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58

分子 d.f.

10	12	15	20	30	40	50	75	120	1000
2.54	2.50	2.46	2.42	2.38	2.36	2.35	2.33	2.32	2.30
3.35	3.28	3.22	3.15	3.08	3.04	3.02	2.99	2.97	2.93
4.30	4.20	4.10	4.00	3.89	3.84	3.81	3.76	3.73	3.68
5.81	5.67	5.52	5.36	5.20	5.12	5.07	5.00	4.95	4.87
11.54	11.19	10.84	10.48	10.11	9.92	9.80	9.65	9.53	9.36
2.42	2.38	2.34	2.30	2.25	2.23	2.22	2.20	2.18	2.16
3.14	3.07	3.01	2.94	2.86	2.83	2.80	2.77	2.75	2.71
3.96	3.87	3.77	3.67	3.56	3.51	3.47	3.43	3.39	3.34
5.26	5.11	4.96	4.81	4.65	4.57	4.52	4.45	4.40	4.32
9.89	9.57	9.24	8.90	8.55	8.37	8.26	8.11	8.00	7.84
2.32	2.28	2.24	2.20	2.16	2.13	2.12	2.10	2.08	2.06
2.98	2.91	2.85	2.77	2.70	2.66	2.64	2.60	2.58	2.54
3.72	3.62	3.52	3.42	3.31	3.26	3.22	3.18	3.14	3.09
4.85	4.71	4.56	4.41	4.25	4.17	4.12	4.05	4.00	3.92
8.75	8.45	8.18	7.80	7.47	7.30	7.19	7.05	6.94	6.78
2.25	2.21	2.17	2.12	2.08	2.05	2.04	2.02	2.00	1.98
2.85	2.79	2.72	2.65	2.57	2.53	2.51	2.47	2.45	2.41
3.53	3.43	3.33	3.23	3.12	3.06	3.03	2.98	2.94	2.89
4.54	4.40	4.25	4.10	3.94	3.86	3.81	3.74	3.69	3.61
7.92	7.63	7.32	7.01	6.68	6.52	6.42	6.28	6.18	6.02
2.19	2.15	2.10	2.06	2.01	1.99	1.97	1.95	1.93	1.91
2.75	2.69	2.62	2.54	2.47	2.43	2.40	2.37	2.34	2.30
3.37	3.28	3.18	3.07	2.96	2.91	2.87	2.82	2.79	2.73
4.30	4.16	4.01	3.86	3.70	3.62	3.57	3.50	3.45	3.37
7.29	7.00	6.71	6.40	6.09	5.93	5.83	5.70	5.59	5.44
2.14	2.10	2.05	2.01	1.96	1.93	1.92	1.89	1.88	1.85
2.67	2.60	2.53	2.46	2.38	2.34	2.31	2.28	2.25	2.21
3.25	3.15	3.05	2.96	2.84	2.78	2.74	2.70	2.66	2.60
4.10	3.96	3.82	3.66	3.51	3.43	3.38	3.31	3.25	3.18
6.80	6.52	6.23	5.93	5.63	5.47	5.37	5.24	5.14	4.99
2.10	2.05	2.01	1.96	1.91	1.89	1.87	1.85	1.83	1.80
2.60	2.53	2.46	2.39	2.31	2.27	2.24	2.21	2.18	2.14
3.15	3.05	2.95	2.84	2.73	2.67	2.64	2.59	2.55	2.50
3.94	3.80	3.66	3.51	3.35	3.27	3.22	3.15	3.09	3.02
6.40	6.13	5.85	5.56	5.25	5.10	5.00	4.87	4.77	4.62

统计表 4 对尾概率从 0.10 到 0.001 的 F 值(继续)

		分子 d.f.								
分母 d.f.	概率	1	2	3	4	5	6	7	8	9
15	0.100	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09
	0.050	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.50
	0.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12
	0.010	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
	0.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26
16	0.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06
	0.050	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
	0.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05
	0.010	8.53	6.23	5.29	4.77	4.44	4.20	4.08	3.89	3.78
	0.001	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98
17	0.100	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03
	0.050	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
	0.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98
	0.010	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
	0.001	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75
18	0.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
	0.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
	0.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93
	0.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	0.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56
19	0.100	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
	0.050	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
	0.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88
	0.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
	0.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39
20	0.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
	0.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
	0.025	5.87	4.46	3.86	3.51	3.29	3.18	3.01	2.91	2.84
	0.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	0.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24
21	0.100	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
	0.050	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
	0.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80
	0.010	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
	0.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11

分子 d.f.

10	12	15	20	30	40	50	75	120	1000
2.06	2.02	1.97	1.92	1.87	1.85	1.83	1.80	1.79	1.76
2.54	2.48	2.40	2.33	2.25	2.20	2.18	2.14	2.11	2.07
3.06	2.96	2.86	2.76	2.64	2.59	2.55	2.50	2.46	2.40
3.80	3.67	3.52	3.37	3.21	3.13	3.08	3.01	2.96	2.88
6.08	5.81	5.54	5.25	4.95	4.80	4.70	4.57	4.47	4.33
2.03	1.99	1.94	1.89	1.84	1.81	1.79	1.77	1.75	1.72
2.49	2.42	2.35	2.28	2.19	2.15	2.12	2.09	2.06	2.02
2.99	2.89	2.79	2.68	2.57	2.51	2.47	2.42	2.38	2.32
3.69	3.55	3.41	3.26	3.10	3.02	2.97	2.90	2.84	2.76
5.81	5.55	5.27	4.99	4.70	4.54	4.45	4.32	4.23	4.08
2.00	1.96	1.91	1.86	1.81	1.78	1.76	1.74	1.72	1.69
2.45	2.88	2.31	2.23	2.15	2.10	2.08	2.04	2.01	1.97
2.92	2.82	2.72	2.62	2.50	2.44	2.41	2.35	2.32	2.26
3.59	3.46	3.31	3.16	3.00	2.92	2.87	2.80	2.75	2.66
5.58	5.32	5.05	4.78	4.48	4.33	4.24	4.11	4.02	3.87
1.98	1.93	1.89	1.84	1.78	1.75	1.74	1.71	1.69	1.66
2.41	2.34	2.27	2.19	2.11	2.06	2.04	2.00	1.97	1.92
2.87	2.77	2.67	2.56	2.44	2.38	2.35	2.30	2.26	2.20
3.51	3.37	3.23	3.08	2.92	2.84	2.78	2.71	2.66	2.58
5.39	5.13	4.87	4.59	4.30	4.15	4.06	3.93	3.84	3.69
1.96	1.91	1.86	1.81	1.76	1.73	1.71	1.69	1.67	1.64
2.38	2.31	2.23	2.16	2.07	2.03	2.00	1.96	1.93	1.88
2.82	2.72	2.62	2.51	2.39	2.33	2.30	2.24	2.20	2.14
3.43	3.30	3.15	3.00	2.84	2.76	2.71	2.64	2.58	2.50
5.22	4.97	4.70	4.43	4.14	3.99	3.90	3.78	3.68	3.53
1.94	1.89	1.84	1.79	1.74	1.71	1.69	1.66	1.64	1.61
2.35	2.28	2.20	2.12	2.04	1.99	1.97	1.93	1.90	1.85
2.77	2.68	2.57	2.46	2.35	2.29	2.25	2.20	2.16	1.09
3.37	3.23	3.09	2.94	2.78	2.69	2.64	2.57	2.52	2.43
5.08	4.82	4.56	4.29	4.00	3.86	3.77	3.64	3.54	3.40
1.92	1.87	1.83	1.78	1.72	1.69	1.67	1.64	1.62	1.59
2.32	2.25	2.18	2.10	2.01	1.96	1.94	1.90	1.87	1.82
2.73	2.64	2.53	2.42	2.31	2.25	2.21	2.16	2.11	2.05
3.31	3.17	3.03	2.88	2.72	2.64	2.58	2.51	2.46	2.37
4.95	4.70	4.44	4.17	3.88	3.74	3.64	3.52	3.42	3.28

统计表 4 对尾概率从 0.10 到 0.001 的 F 值(继续)

		分子 d.f.								
分母 d.f.	概率	1	2	3	4	5	6	7	8	9
22	0.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
	0.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
	0.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76
	0.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
	0.001	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99
23	0.100	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
	0.050	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
	0.025	5.75	4.38	3.75	3.41	3.18	3.02	2.90	2.81	2.73
	0.010	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
	0.001	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89
24	0.100	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
	0.050	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
	0.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70
	0.010	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
	0.001	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80
25	0.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
	0.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
	0.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68
	0.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
	0.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71
26	0.100	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
	0.050	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
	0.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65
	0.010	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
	0.001	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64
27	0.100	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87
	0.050	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
	0.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63
	0.010	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
	0.001	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57
28	0.100	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87
	0.050	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
	0.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61
	0.010	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
	0.001	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50

分子 d.f.

10	12	15	20	30	40	50	75	120	1000
1.90	1.86	1.81	1.76	1.70	1.67	1.65	1.63	1.60	1.57
2.30	2.23	2.15	2.07	1.98	1.94	1.91	1.87	1.84	1.79
2.70	2.60	2.50	2.39	2.27	2.21	2.17	2.12	2.08	2.01
3.26	3.12	2.98	2.83	2.67	2.58	2.53	2.46	2.40	2.32
4.83	4.58	4.33	4.06	3.78	3.63	3.54	3.41	3.32	3.17
1.89	1.84	1.80	1.74	1.69	1.66	1.64	1.61	1.59	1.55
2.27	2.20	2.18	2.05	1.96	1.91	1.88	1.84	1.81	1.76
2.67	2.57	2.47	2.86	2.24	2.18	2.14	2.08	2.04	1.98
3.21	3.07	2.93	2.78	2.62	2.54	2.48	2.41	2.35	2.27
4.73	4.48	4.23	3.96	3.68	3.53	3.44	3.32	3.22	3.08
1.88	1.83	1.78	1.73	1.67	1.64	1.62	1.59	1.57	1.54
2.25	2.18	2.11	2.03	1.94	1.89	1.86	1.82	1.79	1.74
2.64	2.54	2.44	2.33	2.21	2.15	2.11	2.05	2.01	1.94
3.17	3.03	2.89	2.74	2.58	2.49	2.44	2.37	2.31	2.22
4.64	4.39	4.14	3.87	3.59	3.45	3.36	3.23	3.14	2.99
1.87	1.82	1.77	1.72	1.66	1.63	1.61	1.58	1.56	1.52
2.24	2.16	2.09	2.01	1.92	1.87	1.84	1.80	1.77	1.72
2.61	2.51	2.41	2.30	2.18	2.12	2.08	2.02	1.98	1.91
3.13	2.99	2.85	2.70	2.54	2.45	2.40	2.33	2.27	2.18
4.56	4.31	4.06	3.79	3.52	3.37	3.28	3.15	3.06	2.91
1.86	1.81	1.76	1.71	1.65	1.61	1.59	1.57	1.54	1.51
2.22	2.15	2.07	1.99	1.90	1.85	1.82	1.78	1.75	1.70
2.59	2.49	2.39	2.28	2.16	2.09	2.05	2.00	1.95	1.89
3.09	2.96	2.81	2.66	2.50	2.42	2.36	2.29	2.23	2.14
4.48	4.24	3.99	3.72	3.44	3.30	3.21	3.08	2.99	2.84
1.85	1.80	1.75	1.70	1.64	1.60	1.58	1.55	1.53	1.50
2.20	2.13	2.06	1.97	1.88	1.84	1.81	1.76	1.73	1.68
2.57	2.47	2.36	2.25	2.13	2.07	2.03	1.97	1.93	1.86
3.06	2.93	2.78	2.63	2.47	2.38	2.33	2.26	2.20	2.11
4.41	4.17	3.92	3.66	3.38	3.23	3.14	3.02	2.92	2.78
1.84	1.79	1.74	1.69	1.63	1.59	1.57	1.54	1.52	1.48
2.19	2.12	2.04	1.96	1.87	1.82	1.79	1.75	1.71	1.66
2.55	2.45	2.34	2.23	2.11	2.05	2.01	1.95	1.91	1.84
3.03	2.90	2.75	2.60	2.44	2.35	2.30	2.28	2.17	2.08
4.35	4.11	3.86	3.60	3.32	3.18	3.09	2.96	2.86	2.72

统计表 4 对尾概率从 0.10 到 0.001 的 F 值(继续)

		分子 d.f.								
分母 d.f.	概率	1	2	3	4	5	6	7	8	9
29	0.100	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86
	0.050	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
	0.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59
	0.010	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
	0.001	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45
30	0.100	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
	0.050	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
	0.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57
	0.010	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
	0.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39
35	0.100	2.85	2.46	2.25	2.11	2.02	1.95	1.90	1.85	1.82
	0.050	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16
	0.025	5.48	4.11	3.52	3.18	2.96	2.80	2.68	2.58	2.50
	0.010	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96
	0.001	12.90	8.47	6.79	5.88	5.30	4.89	4.59	4.36	4.18
40	0.100	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79
	0.050	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
	0.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45
	0.010	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
	0.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02
50	0.100	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76
	0.050	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
	0.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38
	0.010	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78
	0.001	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82
60	0.100	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
	0.050	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
	0.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33
	0.010	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
	0.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69
80	0.100	2.77	2.37	2.15	2.02	1.92	1.85	1.79	1.75	1.71
	0.050	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00
	0.025	5.22	3.86	3.28	2.95	2.73	2.57	2.45	2.35	2.28
	0.010	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64
	0.001	11.67	7.54	5.97	5.12	4.58	4.20	3.92	3.70	3.53

分子 d.f.

10	12	15	20	30	40	50	75	120	1000
1.83	1.78	1.73	1.68	1.62	1.58	1.56	1.53	1.51	1.47
2.18	2.10	2.03	1.94	1.85	1.81	1.77	1.73	1.70	1.65
2.53	2.43	2.32	2.21	2.09	2.03	1.99	1.93	1.89	1.82
3.00	2.87	2.73	2.57	2.41	2.33	2.27	2.20	2.14	2.05
4.29	4.05	3.80	3.54	3.27	3.12	3.03	2.91	2.81	2.66
1.82	1.77	1.72	1.67	1.61	1.57	1.55	1.52	1.50	1.46
2.16	2.09	2.01	1.93	1.84	1.79	1.76	1.72	1.68	1.63
2.51	2.41	2.31	2.20	2.07	2.01	1.97	1.91	1.87	1.80
2.98	2.84	2.70	2.55	2.39	2.30	2.25	2.17	2.11	2.02
4.24	4.00	3.75	3.49	3.22	3.07	2.98	2.86	2.76	2.61
1.79	1.74	1.69	1.63	1.57	1.53	1.51	1.48	1.46	1.42
2.11	2.04	1.96	1.88	1.79	1.74	1.70	1.66	1.62	1.57
2.44	2.34	2.23	2.12	2.00	1.93	1.89	1.83	1.79	1.71
2.88	2.74	2.60	2.44	2.28	2.19	2.14	2.06	2.00	1.90
4.03	3.79	3.55	3.29	3.02	2.87	2.78	2.66	2.56	2.40
1.76	1.71	1.66	1.61	1.54	1.51	1.48	1.45	1.42	1.38
2.08	2.00	1.92	1.84	1.74	1.69	1.66	1.61	1.58	1.52
2.39	2.29	2.18	2.07	1.94	1.88	1.83	1.77	1.72	1.65
2.80	2.66	2.52	2.37	2.20	2.11	2.06	1.98	1.92	1.82
3.87	3.64	3.40	3.14	2.87	2.73	2.64	2.51	2.41	2.25
1.73	1.68	1.63	1.57	1.50	1.46	1.44	1.41	1.38	1.33
2.03	1.95	1.87	1.78	1.69	1.63	1.60	1.55	1.51	1.45
2.32	2.22	2.11	1.99	1.87	1.80	1.75	1.69	1.64	1.56
2.70	2.56	2.42	2.27	2.10	2.01	1.95	1.87	1.80	1.70
3.67	3.44	3.20	2.95	2.68	2.53	2.44	2.31	2.21	2.05
1.71	1.66	1.60	1.54	1.48	1.44	1.41	1.38	1.35	1.30
1.99	1.92	1.84	1.75	1.65	1.59	1.56	1.51	1.47	1.40
2.27	2.17	2.06	1.94	1.82	1.74	1.70	1.63	1.58	1.49
2.63	2.50	2.35	2.20	2.03	1.94	1.88	1.79	1.73	1.62
3.54	3.32	3.08	2.83	2.55	2.41	2.32	2.19	2.08	1.92
1.68	1.63	1.57	1.51	1.44	1.40	1.38	1.34	1.31	1.25
1.95	1.88	1.79	1.70	1.60	1.54	1.51	1.45	1.41	1.34
2.21	2.11	2.00	1.88	1.75	1.68	1.63	1.56	1.51	1.41
2.55	2.42	2.27	2.12	1.94	1.85	1.79	1.70	1.63	1.51
3.39	3.16	2.93	2.68	2.41	2.26	2.16	2.03	1.92	1.75

统计表 4 对尾概率从 0.10 到 0.001 的 F 值(继续)

分子 d.f.

分母 d.f.	概率	1	2	3	4	5	6	7	8	9
100	0.100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69
	0.050	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97
	0.025	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24
	0.010	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59
	0.001	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44
200	0.100	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66
	0.050	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93
	0.025	6.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18
	0.010	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50
	0.001	11.15	7.15	5.63	4.81	4.29	3.92	3.65	3.43	3.26
1000	0.100	2.71	2.81	2.09	1.95	1.85	1.78	1.72	1.68	1.64
	0.050	3.85	3.00	2.61	2.88	2.22	2.11	2.02	1.95	1.89
	0.025	5.04	8.70	3.13	2.80	2.58	2.42	2.30	2.20	2.13
	0.010	6.66	4.68	3.80	3.34	3.04	2.82	2.66	2.53	2.43
	0.001	10.89	6.96	5.46	4.65	4.14	3.78	3.51	3.80	3.13

分子 d.f.

10	12	15	20	30	40	50	75	120	1000
1.66	1.61	1.56	1.49	1.42	1.38	1.35	1.32	1.28	1.22
1.93	1.85	1.77	1.68	1.57	1.52	1.48	1.42	1.38	1.30
2.18	2.08	1.97	1.85	1.71	1.64	1.59	1.52	1.46	1.36
2.50	2.37	2.22	2.07	1.89	1.80	1.74	1.65	1.57	1.45
3.30	3.07	2.84	2.59	2.32	2.17	2.08	1.94	1.83	1.64
1.63	1.58	1.52	1.46	1.38	1.34	1.31	1.27	1.23	1.16
1.88	1.80	1.72	1.62	1.52	1.46	1.41	1.35	1.30	1.21
2.11	2.01	1.90	1.78	1.64	1.56	1.51	1.44	1.37	1.25
2.41	2.27	2.13	1.97	1.79	1.69	1.63	1.53	1.45	1.30
3.12	2.90	2.67	2.42	2.15	2.00	1.90	1.76	1.64	1.43
1.61	1.55	1.49	1.43	1.35	1.30	1.27	1.23	1.18	1.08
1.84	1.76	1.68	1.58	1.47	1.41	1.36	1.30	1.24	1.11
2.06	1.96	1.85	1.72	1.58	1.50	1.45	1.36	1.29	1.13
2.34	2.20	2.06	1.90	1.72	1.61	1.54	1.44	1.35	1.16
2.99	2.77	2.54	2.30	2.02	1.87	1.77	1.62	1.49	1.22

统计表 5 二项分布

对于不同的 π 值在 n 次试验中有 x 次成功的概率。对于大于 0.5 的 π , 查看对 $1 - \pi$ 的值及 $n - x$ 次失败的概率。



n	x	π										
		0.025	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
2	0	.9506	.9025	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500
	1	.0488	.0950	.1800	.2550	.3200	.3750	.4200	.4550	.4800	.4950	.5000
	2	.0006	.0025	.0100	.0225	.0400	.0625	.0900	.1225	.1600	.2025	.2500
3	0	.9269	.8574	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
	1	.0713	.1354	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750
	2	.0018	.0071	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750
	3		.0001	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250
4	0	.9037	.8145	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625
	1	.0927	.1715	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500
	2	.0036	.0135	.0486	.0975	.1536	.2109	.2646	.3105	.3456	.3675	.3750
	3	.0001	.0005	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500
	4			.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625
5	0	.8811	.7738	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0313
	1	.1130	.2036	.3280	.3915	.4096	.3955	.3602	.3124	.2592	.2059	.1563
	2	.0058	.0214	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125
	3	.0001	.0011	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125
	4			.0004	.0022	.0064	.0146	.0284	.0488	.0768	.1128	.1562
	5				.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0312
6	0	.8591	.7351	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156
	1	.1322	.2321	.3543	.3993	.3932	.3560	.3025	.2437	.1866	.1359	.0938
	2	.0085	.0305	.0984	.1762	.2458	.2966	.3241	.3280	.3110	.2780	.2344
	3	.0003	.0021	.0146	.0415	.0819	.1318	.1852	.2355	.2765	.3032	.3125
	4		.0001	.0012	.0055	.0154	.0330	.0595	.0951	.1382	.1861	.2344
	5			.0001	.0004	.0015	.0044	.0102	.0205	.0369	.0609	.0938
	6					.0001	.0002	.0007	.0018	.0041	.0083	.0156
7	0	.8376	.6983	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152	.0078
	1	.1503	.2573	.3720	.3960	.3670	.3115	.2471	.1848	.1206	.0872	.0547
	2	.0116	.0406	.1240	.2097	.2753	.3115	.3177	.2985	.2613	.2140	.1641
	3	.0005	.0036	.0230	.0617	.1147	.1730	.2269	.2679	.2903	.2918	.2734
	4		.0002	.0026	.0109	.0287	.0577	.0972	.1442	.1935	.2388	.2734
	5			.0002	.0012	.0043	.0115	.0250	.0466	.0774	.1172	.1641
	6				.0001	.0004	.0013	.0036	.0084	.0172	.0320	.0547
	7						.0001	.0002	.0006	.0016	.0037	.0078

奇数号练习题答案

1 统计学：随机性和规律性

- 1.1 早期的统计活动主要集中于州政府收集其公民的人口数据。因此,统计学被认为是“州政府的事情”。
- 1.3 a. 随机性发生在下一个观测值不能被准确预测之时。
b. 由许多观测值得出的模式。
c. 统计学通过考察差异变动从而获得规律性。
- 1.5 概率表示当某项研究重复时不同结果发生的频率。如果仅仅考虑随机性,一项观测的差异发生的概率也许很小。
- 1.7 a. 可以取不同值的某项特征。
b. 经验变量具有观测数据,而理论变量服从特定分布。
- 1.9 1月,2月,...;Florida,Texas,...;100 mph,101 mph,...;\$ 10000000,\$ 11000000,...;0,1,2,...。
- 1.11 观测多次时常数的值仍然不变。
- 1.13 人口普查局(The Census Bureau)和劳动统计局(Bureau of Labor Statistics)。
- 1.15 由于风暴的随机性,我们从来不知道是否某一特定的财产会被破坏。然而,对于某种强度的风暴,经济上的损失是大致相同的,这体现了其中的规律性。
- 1.17 a. 计算机使追踪大量统计结果成为可能。
b. 更加容易使观测者进行比较。
c. 抽样调查使得衡量音乐、电影、政客等的受欢迎程度成为可能。相反地,抽样调查对未来所发生的一切会有影响。
- 1.19 学生的课题。d. 莎士比亚可能用了较长的句子。
- 1.21 a. 长期以来婴儿死亡率已经下降了。白人和非白人的婴儿死亡率不同。
b. 与婴儿死亡率相同。
c. 因为数值上较小,所以是母亲死亡率。
d. 一定在记录到表格中之前报告某次死亡。时间越早,报告的死亡数越少。
e. 较老的数据较为不准确。出生数据可能更不准确,因为不是所有的婴儿出生都报告了。
- 1.23 学生的课题。

2 数据的收集

- 2.1 学生总体是否应包括大学二年级和四年级的学生呢？是否应包括半读的学生呢？是否应包括所有女学生和男学生呢？
- 2.3 没有接受实验处理和以反应变量来衡量的元素。
- 2.5 它可以被控制，并且大多数时候我们可以得出它有多大。
- 2.7 a. 给每个学生一个数作为号码，然后用随机数表来决定所需学生的号码。从所有学生的列表中找到一个随机的开始，然后每隔 k 个选一位学生。
b. 总体列表可能不完全，可能有拒绝回答的人和错误的回答。
c. 有偏结果。
d. 花时间和精力去寻求并获得被选者的合作，并以最佳的方式提问。
- 2.9 否。抽样误差不可能避免。它指的是如果研究重复多次，结果中出现差异的多少。
- 2.11 a. 我们由全部总体数据得出的数值。
b. 应用到全部总体中。
- 2.13 学生的课题。
- 2.15 a. 他们并没有说他们赞成这个决策。他们或者对这个决策没有意见或者反对它或者不知道这个决策。
b. 它是一个合适的随机样本吗？多少人回答了这些问题？这些问题是如何组织的？调查表中这些问题是如何排列的？
c. 20 次重复实验的 19 次中，样本的百分比位于总体真值的 2 个百分点之内。如果这个样本是 19 次中的一个，则总体真值百分比位于 54% 和 58% 之间。
- 2.17 这个反应可能是由刺激或其它因素产生的。
- 2.19 a. 不可能。
b. 它可能诱导被访者以一种并非他们自己感受的方式来回答问题。
- 2.21 显示研究中所收集数据的表格。
- 2.23 对于数据收集中存在的诸多问题，你最好考虑坏的或错误的数据。
- 2.25 a. 否。我们不能从这些数据中总结。在所有参与这项抽样调查的人中，Philadelphia/Trenton 的参与者的百分比少于全国其它考察地区。
b. 这个样本不是随机样本。它是自己选择的。
- 2.27 a. 这个抽样调查的结果只能用于郡中一部分地区的消防队员，不能用来推广到全郡。
b. 零。
- 2.29 其它服务可能有较高的百分比。这个百分比是在什么基础上计算的，它可用于哪种类型的人？
- 2.31 不经过当事人同意就在他身上做实验总是会导致麻烦。
- 2.33 a. 不是随机样本。
b. 寻找自愿者会使样本不随机。
c. 否。这些女性很多不是自愿者。
d. 男性可能会乐意接受。

e. 缺乏统计知识。

2.35 a. 实际的数字。

b. 相同。

2.37 哪段时期的薪水？电影的质量可能不符合 3 分标准。关于电影的质量存在两个问题。跟那一部电影相比较？可能没有看过 *Some Like It Hot*。重复这样的问题：我们最喜欢的是什么。不明确的轿车问题：中型轿车相对于大家庭货车还是其它类型？哪一年的车子？没有说出评分尺度趋向哪个方向。

2.39 学生的课题。

2.41 a. 表 A.1。

b. 期望寿命增加了。整个时期里，女性比男性期望寿命长。

表 A.1 练习 2.41a 的性别和寿命表

年	性别	期望寿命
1789	m	34.5
1789	f	36.5
1850	m	38.3
1850	f	40.5
1890	m	42.5
1890	f	46.0
1910	m	54.0
1910	f	56.6
1930	m	59.3
1930	f	62.6

3 数据的描述：图和表

3.1 获得简洁性相对于信息的损失。

3.3 a. 用刻度表示变量并用 x 或 $*$ 等符号标示观测值的直线。

b. 表示哪儿有许多观测值的聚集，哪儿只有少数几个观测值。

c. 对很多观测值都没有好处。

3.5 a. A. 地理区域的名称。B. 频率数字。C. 地理区域。D. 大约 17000。

b. 如条形的面积。

c. 不用高度必须用水兵的面积来表示频率。否则从这个图上难以看出每个地区有多少水兵。

d. 大多数部队在欧洲和东亚及太平洋地区；只有几支在西半球、非洲和南亚。

e. 地区的名字很混乱，应该再详细一些。

3.7 通过采用适当的纵向尺度，即使有许多的观测值，条形也不会很大。

3.9 单峰。

3.11 直方图的一半跟另一半不对称。

- 3 13 a 从本世纪初到世纪末,人口上升了而价格指数下降了。
b 它显示了一个世纪以来两个变量的表现。
c 我们应该为两个变量画一个散点图。
- 3 15 跟一个圈一样。小数可以去掉。行列的排列应使越大的数越往表格的左上角靠。需要比较的数应该按列排。
- 3 17 一个好的图可以传递来自制图人有关数据的信息。
- 3 19 a 没有丢失信息。原始观测值可以重新获得。
b 当样本很大时。
c 没有很多的大得分,也没有很多的成功。
d 其它变量的信息,如年龄和性别。
- 3 21 学生的课题。
- 3 23 a 中位数显示区域的差异。盒子的长度显示区域之间的比例的变差。
b 是。这个图清晰而准确地表现了数据。
- 3 25 a 当尺度从 0 英寸开始时,图 3.8 对于数据更为正确。在图 3.7 中更容易读出每年的确切英寸数。
b 该图以何种方式和数据更接近? 读该图是怎样地容易?
- 3 27 枝叶图不能用于分类型变量,因为它们没有数量观测值。
- 3 29 枝叶图对小样本合适,盒子图对大样本合适。盒子图显示 5 个可以在数据组之间比较的特别的数。而枝叶图只给出了数据的中心和分散程度的大概印象。枝叶图显示了原始数据,而从盒子图上不能重新获得原始数据。
- 3 31 a “最好”可能是指味道而与脂肪没有关系。
b 食品可以按照任何其它变量排序。
- 3 33 a 星期六最好,星期天和星期三最差。
b 他们画了垂直尺度,而且不包括原点。
c 这些天看起来将更为相似。
d 由于对尺度所知甚少,我们很难知道。
e 当回过头来看时,事情看起来并不那么糟糕。
- 3 35 a 图 A.1。
b 对于在职男性和女性,个人和工作的时间占去了他们大部分时间,还有一些时间则留给闲暇和家务劳动。管家把大量时间花在个人事物上,家务劳动和闲暇则是接下来的两项。
c 个人时间在三个组中大约相同。在职女性工作时间不像男性那样长,她们在家务劳动上花去更多的时间。
- 3 37 a 图 A.2。
b 单峰且几乎对称。
c 可能还有其它有关什么时候生孩子的考虑。
d 每个年龄组中母亲的总数目。
- 3 39 a 小,中和大的条形图将显示每一种尺寸有多少。
b 这 10 个尺寸是怎样与小,中和大三个值相关的?

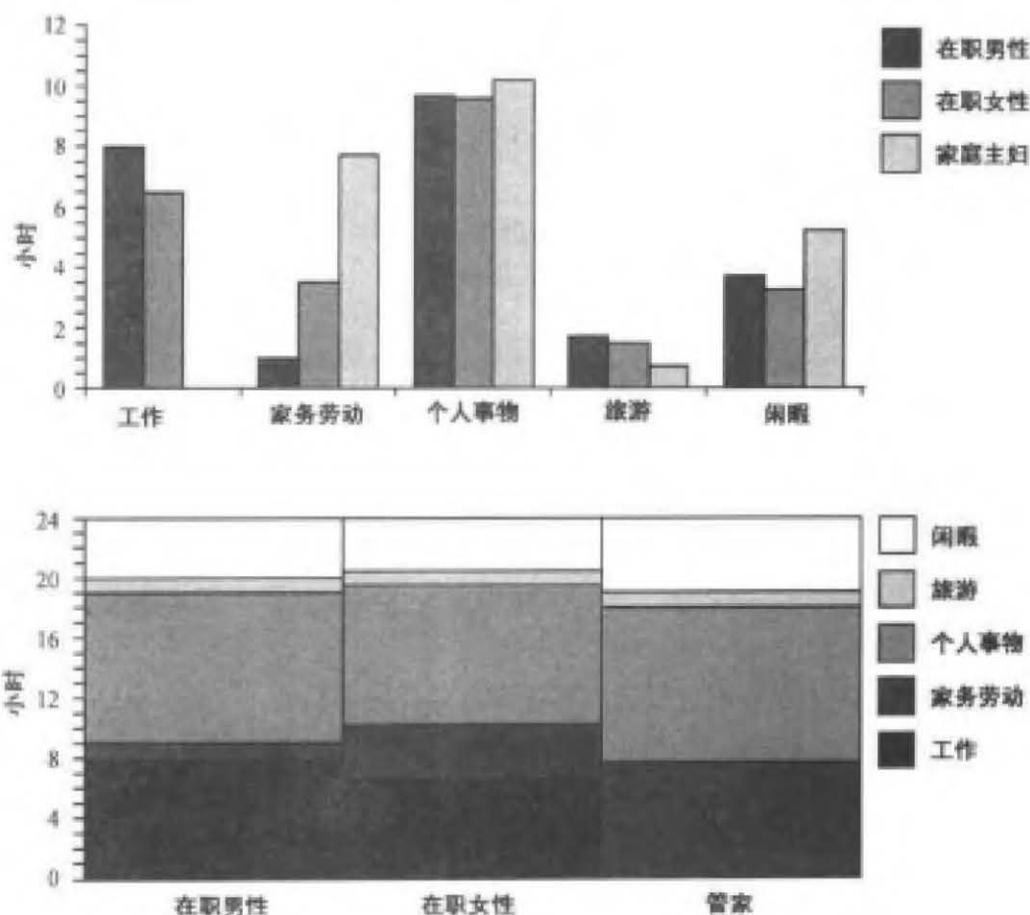


图 A.1 练习 3.35a 的图。

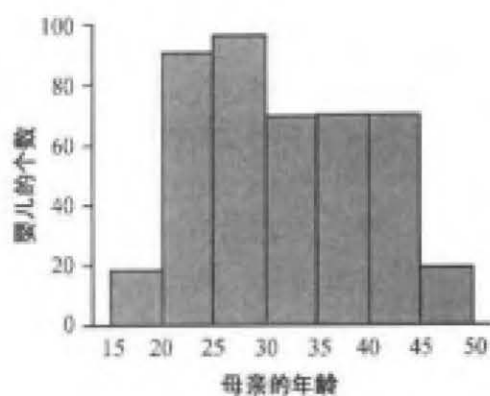


图 A.2 练习 3.37a 的直方图。

3.41 学生的课题。

3.43 学生的课题。

3.45 a. 图 A.3

b. 偏斜的形状表示大多数观测值在 18 到 25 岁之间。

c 那些大年龄学生的数目是不可思议的。

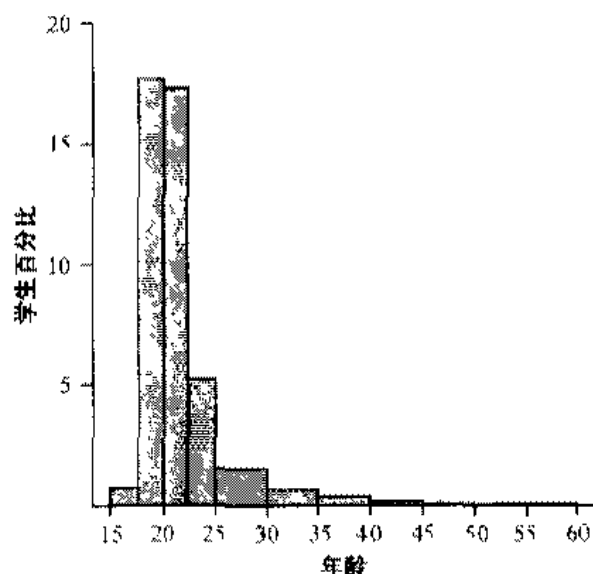


图 A.3 练习 3.45a 的直方图。

3.47 a 很偏斜,小的观测值大大多于大的观测值。

b 图 A.4。

c 盒子图简单易懂,但是含有较少的信息。盒子图显示第 25、50 和 75 百分位点及最小和最大观测值。

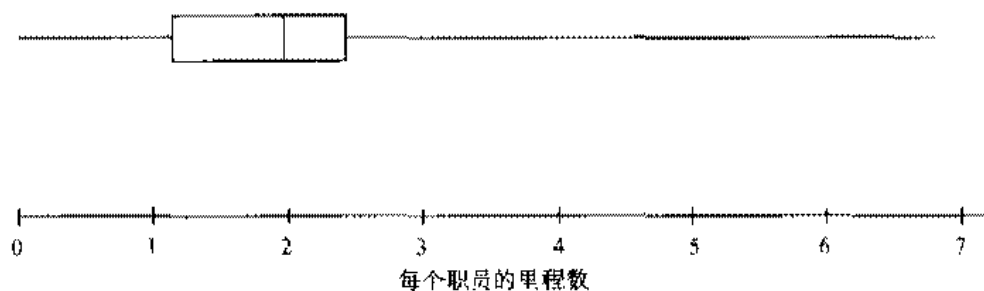


图 A.4 练习 3.47b 的盒子图。

3.49 a. 图 A.5

b 黑人和白人受害者的数目不知道。

3.51 学生的课题。

3.53 表 A.2。把表 3.6 转置后再来比较起因。国家按死亡率而不是字母顺序排队。为了便于比较没有小数。

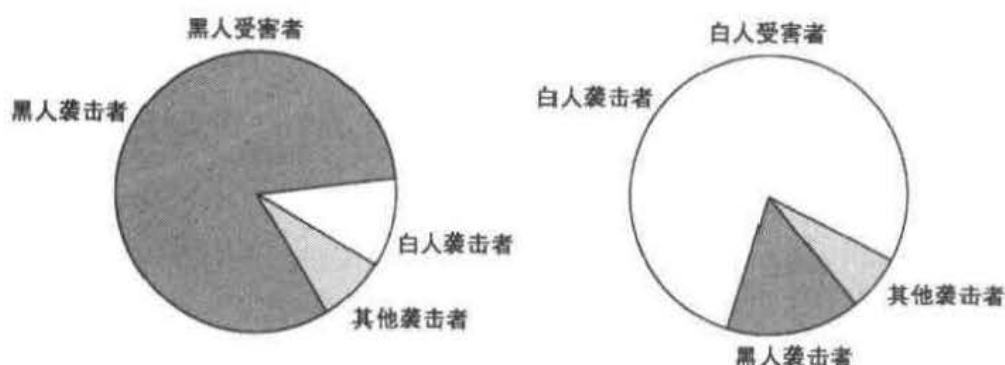


图 A.5 练习 3.49a 的圆饼图。

表 A.2 练习 3.53 表 3.6 的新表

事件类型	法国	奥地利	美国	挪威	意大利	荷兰
交通	24	35	23	17	23	18
自然	31	30	16	25	19	18
他杀	1	2	10	1	1	1
其它	22	9	11	5	4	4
总计	78	75	61	48	47	40

4 数据的描述:计算汇总统计量

- 34.1 只失去少量信息并获得简洁性。
- 4.3 众数是最经常出现的变量值。中位数是把观测值分成两组并使得一半的观测值小于这个数而另一半的观测值大于这个数的变量值。均值是数据直方图的重心。
- 4.5 某班学生的高度,一个众数主要由女性决定而另一个由男性决定。
- 4.7 a. 大多数收入的分布都是倾斜的,故中位数是此类分布更好的概括。
b. 如果那是大多数人的收入。
- 4.9 学生的课题。
- 4.11 a. 除最大和最小之外的其它观测值。
b. 最大和最小观测值。
- 4.13 s 。
- 4.15 大约 $2/3$ 。
- 4.17 方差。
- 4.19 一个高的平均得分告诉我们这种比萨饼被评为优秀。一个小的标准差表明大多数得分彼此接近,因而都很高。
- 4.21 某个样本有一些小的和一些大的观测值,则它们对均值的效应会抵消。所以,不同样本之间的均值差异不会像它们的观测值变化的那样大。

- 4 23 a. 为了考察某个观测值是否为异常。不同变量的原始得分可以通过变为标准得分来比较。
b. 学生举例。
- 4 25 从 -2 到 $+2$ 。
- 4 27 t -值。
- 4 29 某个样本的观测值有差异是因为数据的随机性,而标准差用来衡量观测值的差异有多大。
- 4 31 没有任何变差则观测值都相等。
- 4 33 收入分布常常是倾斜的,而中位数更适合倾斜分布。
- 4 35 均值不必与任何观测值相等。
- 4.37 将数据分成相等的两半。
- 4 39 a. 按照生产数据来说是准确的,由于标题没有提及排除了某些类型的工人,所以产生了误解。
b. 评分尺度没有说明什么是高什么是低。
c. 衡量那些人生产了多少可能是一件很困难的事情。很难说这些国家会怎样去比较那些类型的工人。
- 4.41 a. Braves 的均值比 Phillies 的均值少。
b. Phillies 的标准差小些,他们在每次比赛中出现相同数目的错误。Braves 错误次数的范围大一些。
c. Braves 的标准差大。
d. 由于均值为 2.0,标准差为 0.3,所以不可能有很多的观测值等于 0。
- 4.43 a. 好。在每个学术主题上得分都很高。
b. 是。得分表示平均的音乐理解力。
- 4.45 Duluth 的天气比 Hibbing 温暖。
- 4 47 否。你需要中位数、四分位数间距、可能还有直方图来显示收入的分布。
- 4 49 a. Media 均值高且标准差小,故所有增长都很大。Rose Valley 标准差大,故增长更大。
b. 在 Rose Valley, 均值往上的两个标准差等于 \$ 15000, 这大约是最大的增长了。
- 4.51 很倾斜,只有少数几个球员得到很高的薪水。
- 4 53 a. 表 A.3。

表 A.3 练习 4 53a 的统计量

变量	均值	中位数	标准差	极差
击球率	0.262	0.264	0.039	0.228
击球次数	319	290	180	665
得分次数	47.3	36.5	33.1	140
安打次数	87.6	76.0	55.6	224
本垒打次数	10.4	6.0	11.0	52
击球得分次数	44.8	35.0	34.5	150

- b. 图 A.6。
c. 击球率的均值;其余为中位数。

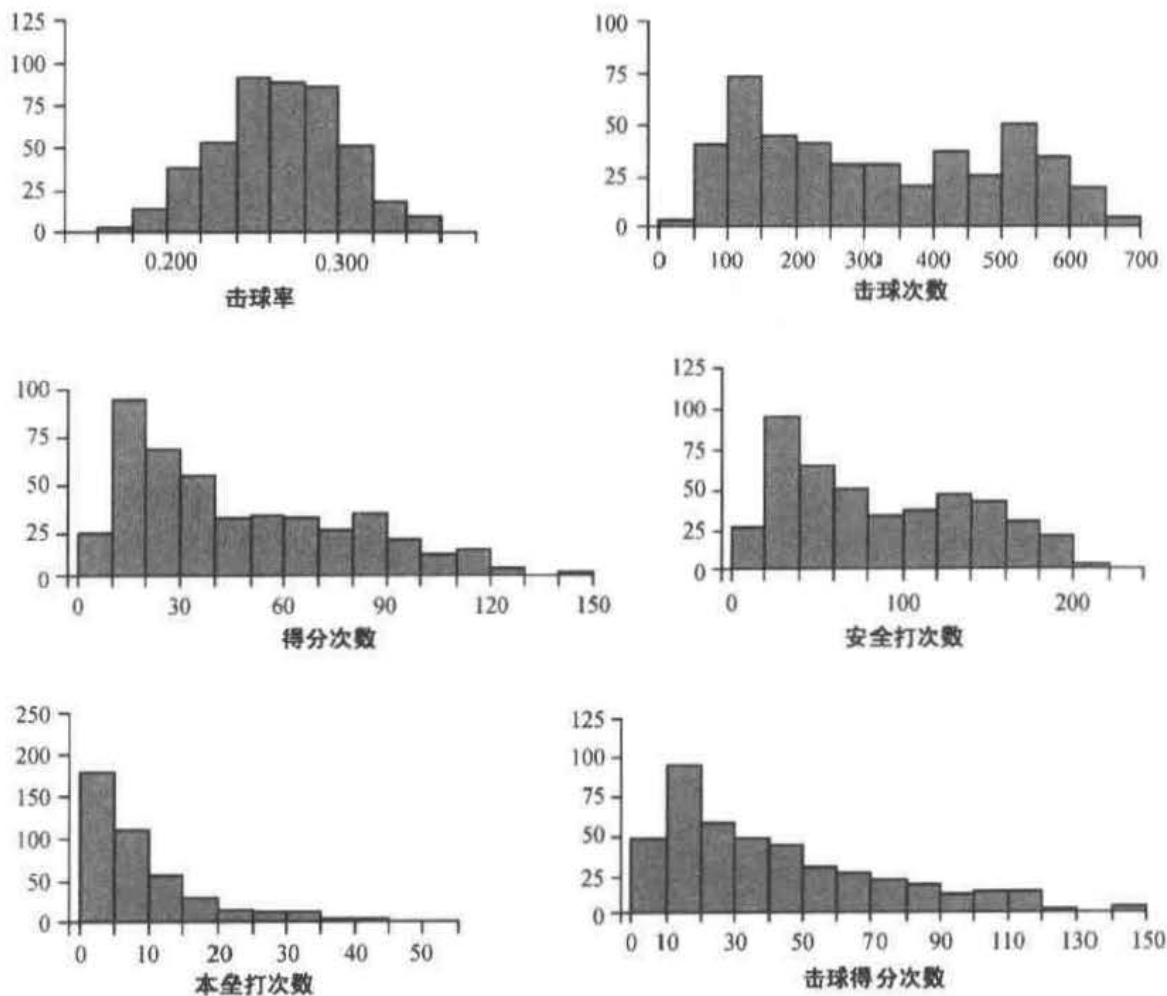


图 A.6 练习 4.53b 的频率直方图。

d. 有些分布是倾斜的, 有些有离群点。

4.55 a. 用标准差。

b. 练习 3.34 的组中 $s = 10.9$, 而新的组中 $s = 14.6$ 。

c. $14.6/10.9 = 1.3$; 一个标准差比另一个大 30%。

4.57 a. 2.18。

b. 标准得分不寻常地大, 而巧克力冰淇淋与其余甜点不同。

4.59 a. 图 A.7。

b. 1。

c. 0 表示男性, 1 表示女性。

4.61 a. 图 A.8。

b. 众数为 77, 中位数为 77, 均值为 80.4。

c. 众数给出了最经常的值, 中位数给出了既比一半的观测值小又比一半的观测值大的数值。均值给出了分布的重心。

d. 医学院的数目在这段时期内下降了。

e. 学校的学生人数增加了。

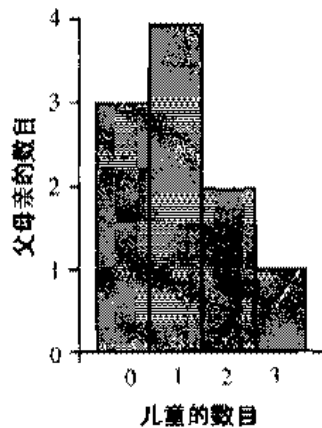


图 A.7 练习 4.59a 的直方图。

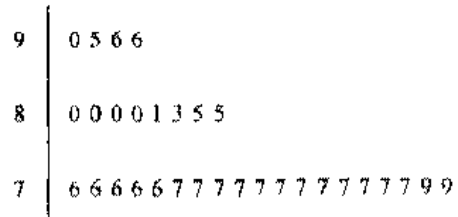


图 A.8 练习 4.61a 的枝叶图。

4.63 图 A.9c 所有五个得分值相同,都为 5。

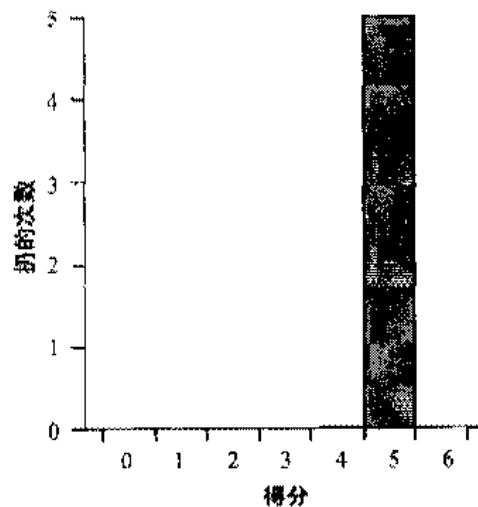


图 A.9 练习 4.63 的直方图。

- 4.65 a 在最近的一个月中至少有一次吸烟、吸食大麻、喝酒的高中学生的百分比分别为 48%, 32%, 及 38%。
b. 数目好像不符合。
- 4.67 a \$10.00。
b \$10.36。
c. \$10.50。
d \$12.13。
e 只要观测值仍然保持大于或小于(或等于)中位数,则中位数不会随着观测值大小的变化而变化。均值则受任何观测值的数值变化的影响。
- 4.69 a 均值为 21.9, 标准差为 5.2。

- b. 九。
- c. 均值为 71.3, 标准差为 11.2。
- d. 七。

4.71 新娘 27 岁, 新郎 30 岁。新郎比新娘平均大 3 岁。

5 概 率

- 5.1 似然(Likelihood), 优势(Odds), 机会(Chance)。
- 5.3 用 0 到 1 之间的一个数来描述某个事件发生的频率。概率 0 表示事件永远不可能发生而概率 1 表示事件总是发生。
- 5.5 如果某个实验有 n 个不同的且等可能的结果, 而我们想知道其中 k 个发生的概率, 则 k/n 就是这个概率。
- 5.7 $1.00 - 0.12 = 0.88$ 。
- 5.9 a. 一项主观的个人概率。
b. 这个事件是唯一的且只发生一次。
- 5.11 两种可能性。
- 5.13 a. 在有四个孩子的家庭中, 女孩的平均(期望)个数。
b. μ 。
- 5.15 σ 。
- 5.17 a. 正态、 t 、 χ^2 和 F -变量。
b. t -变量。
- 5.19 a. t -分布。
b. Gosset 先生用“学生”的笔名是因为他的老板不让雇员写科学论文。
- 5.21 自由度。
- 5.23 a. χ^2 变量不取负值。
b. 需要事先知道自由度。
- 5.25 因为如果选举区之间是平均分配的话, 则获得这样的样本数据或更为极端的数据是不太可能的, 故我们得出选举区之间不是平均分配的结论。
- 5.27 这项言论和关于爱情的一个言论类似, 而统计学家几乎从来没有发现哪件事情发生的概率会等于 1。
- 5.29 学生的课题。
- 5.31 Trudi 是最保险的, 而 Rod 则风险最大。
- 5.33 学生的课题。
- 5.35 有两个取常概率值的结果及 n 个观测到的独立事件。
- 5.37 用二项分布更好。100 个有四个孩子的家庭中有很多的样本会得出正确的结果。来自某个简单样本的数据有一个伴随着的抽样错误。
- 5.39 算出 0、1、2 和 3 个女孩的概率从而得出均值。同样的答案为 $\mu = np = 3 \times 0.49 = 1.47$ 。

- 5 41 1000 个孩子中有 9 个会骨折;一半的大学生会卷入事故;100000 辆 BMW 汽车中有 1 辆会被截持。
- 5 43 这些数据属于一个具有较高概率的数据集,故数据没有什么不正常的。
- 5 45 a 如果选择是真正随机的话,把所有新成员选入清扫组的概率是非常小的。由于所有组员都是新成员,很难想象选择是随机的。
b 表 A 4。看上去非常奇怪的是老成员中没有人被选中,而几乎所有的新成员都被选中。

表 A.4 练习 5 45b 的分布

成 员	选出的人 未选出的人 合计	成员		合计
		老	新	
		0	5	5
		52	2	54
		52	7	59

- 5 47 a 100000 人中有 2 人会被电击。
b 100000 个中有 2 个的频率比 0.00002 这样一个很小的数理解起来更为容易。
- 5 49 a 评价哪种慈善行动最有可能产生影响。
b 他们并不总是完全客观的。
- 5 51 数据可能来自具有相同均值的总体。
- 5 53 a $\frac{1}{6}$ 。
b $\frac{1}{6}$ 。
c $\frac{1}{6} + \frac{1}{6} = \frac{2}{6}$ 。
- 5 55 a $\frac{2}{36}$ 。
b $\left(\frac{2}{36}\right)\left(\frac{1}{35}\right) = \frac{2}{1260} = \frac{1}{630}$ 。
- 5 57 学生的课题。
- 5 59 学生的课题。
- 5 61 a $(7+1)/256 = 0.031$ 。
b 它是如此之小,以致于 0.5 可能不是正确的概率。
- 5 63 a 0.034。
b 0.114。
- 5 65 a 0.5。
b $0.5^{10} = \frac{1}{1024} = 0.001$ 。
c, 收入分布是倾斜的,均值可能很大但不具有代表性。
- 5 67 $0.33^4 = 0.012$ 。
- 5 69 a $(0.25)(0.08) = 0.02$ 。
b 直觉上,两件事情同时发生的概率应该比只有其中一个发生的概率小。

c. 打雷和闪电不是独立事件, 概率不能简单地相乘。

5.71 a. $0.07 + 0.29 = 0.36$ 。

b. $(0.07)(0.29) = 0.02$ 。

c. $(1 - 0.07)(1 - 0.29) = 0.66$ 。

5.73 $(0.05)(0.10) = 0.005$ 。

5.75 a. $2.5\% + 2.5\% = 5.0\%$ 。

b. 2.5% 。

c. 47.5% 。

5.77 某个新成员被首先选中的概率为 $\frac{7}{52}$ 。第二个人仍是新成员的概率为 $\frac{6}{51}$ 。剩下的概率变为 $\frac{5}{50}$ 、 $\frac{4}{49}$ 和 $\frac{3}{48}$ 。这些概率的乘积即为清扫小组中所有五个人都是新成员的概率。

5.79 a. 图 A.10。

b. $\mu = 0(0.05) + 1(0.19) + 2(0.31) + 3(0.28) + 4(0.14) + 5(0.04) + 6(0.001) = 2.42$ 。

c. $6(0.40) = 2.40$ 。

d. 是。

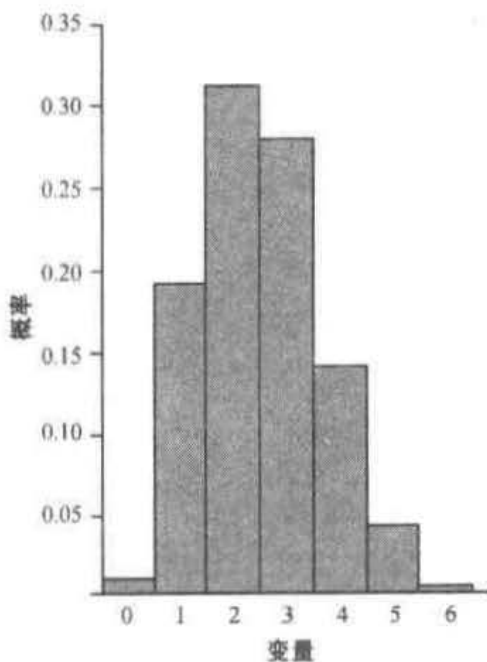


图 A.10 练习 5.79a 的直方图。

6 作出结论: 估计

6.1 了解有关总体中参数值的信息。

6.3 参数。

6.5 样本均值、百分比和标准差都被称为统计量；总体均值、百分比和标准差都被称为参数。

6.7 对来自同一总体的许多不同样本分别计算统计量，而这些统计量的均值等于总体参数。

6.9 a. $10 - 2 = 8$ 到 $10 + 2 = 12$ 。

b. 它可能是来自不同样本的许多个可能的置信区间中的一个。希望它是所有的包含总体均值的置信区间的 95% 中的一员。

6.11 样本容量扩大，置信水平降低。

6.13 我们可以从样本的数据归纳到整个总体而不用去度量总体中的每个元素。

6.15 a. 样本均值的标准差(均值的标准误差)要比样本中位数的标准差(中位数的标准误差)小。

b. 一个小的标准差告诉我们观测值之间很接近且都跟均值接近。当某个估计具有小的标准差(标准误差)，则每个统计量跟总体的参数值都很接近。

6.17 虽然我们不可能得出总体参数值，但是我们可以用样本数据来估计该参数。

6.19 a. 这些坦克必须随机地分布于各个战场上并随机地被捕获。

b. 如果最近一批坦克被送往某一前沿阵地而许多坦克在那里被捕获，这个估计可能会太高了。相似地，如果原先的坦克被派往某一前沿阵地而没有很多的坦克在那里被捕获，则这个估计太低了。

6.21 这个区间的一个高置信水平。

6.23 a. 他们可能反对或没有任何意见。

b. 样本是否是以一种恰当的方式得到的？有多少人回应了抽样调查，提问是在电话上进行还是面对面的，问题是怎样写的，问题是怎样排列的，等等。

c. 置信区间为 $56\% - 2\% = 54\%$ 到 $56\% + 2\% = 58\%$ 。用构造这个方法构造的所有置信区间中有 95% 的区间包含总体真值而另 5% 没有包含。我们不知道这个特定的区间是属于多数呢，还是那少数几个。

6.25 不真。这个从 70 到 75 的区间很可能是包含总体真值的许多区间中的一个，但也可能是没有包含真值的少数几个中的一个。

6.27 a. 一个关于学生的样本。

b. 单簧管演奏者总体。

c. 选民总体。

d. 战争死亡样本。

6.29 a. 置信区间 18 到 38 的构造方法使得在对其它样本构造区间时，得到的区间中有 95% 包含差异的真值而 5% 不包含差异的真值。我们不知道这个特定的区间是多数中的一个还是少数中的一个。

b. 由于差异的置信区间不包括 0，总体中的这两个比率似乎不能相等。

6.31 a. 用构造置信区间 6 到 18 的方法对许多其它的样本分别置信区间，其中有 95% 的区间包含差异的真值而另 5% 不包含差异的真值。我们不知道这个特定的区间是多数中的一个还是少数中的一个。

b. 由于差异的置信区间不包括 0，总体中的这两个比率似乎不会相等。

6.33 a. 用构造置信区间 5.22 到 6.70 的方法对许多其它的样本分别构造置信区间，得到的

这些区间中有 95% 包含差异的真值而有 5% 不包含差异的真值。我们不知道这个特定的区间是属于多数还是一个少数派。

b. 因为置信区间是从 5.22 到 6.70, 平均流量至少为 5.00 的说法看来是合理的。

6.35 a. 用构造置信区间 0.90 到 1.28 的方法去对许多其它的样本分别构造置信区间, 得到的这些区间中有 95% 包含差异的真值而有 5% 不包含差异的真值。我们不知道这个特定的区间是属于多数还是一个少数派。

b. 由于 1.00 ppm 包含在置信区间中, 数据有可能是来自均值为这个数的总体。

6.37 大多数置信区间都是通过计算样本误差并用样本统计量加上或减去这个样本误差得到的。

6.39 a. 不同的民意测验代表不同的样本。不同样本得出不同结果并不奇怪。

b. 样本中包含的观测数不同。

c. 38 到 46, 35 到 43, 38 到 44, 36 到 44; 图 A.11。

d. 38 到 43。

e. 因为抽样误差大约等于 $100/\sqrt{n}$, 600 人投票抽样误差为 4%, 1100 人投票抽样误差为 3%。

f. 支持 Bush(布什)的人总计为 $(0.42)(600) + (0.39)(600) + (0.41)(1100) + (0.40)(600) \approx 1170$ 。四次投票的总数大约为 2900。Bush 的比率成为 $1170/2900 \approx 40\%$, 而抽样误差为 $100/\sqrt{2900} \approx 2\%$ 。

g. 38 到 42。

h. 比较起来很相似, 差异可能是由四舍五入造成的。

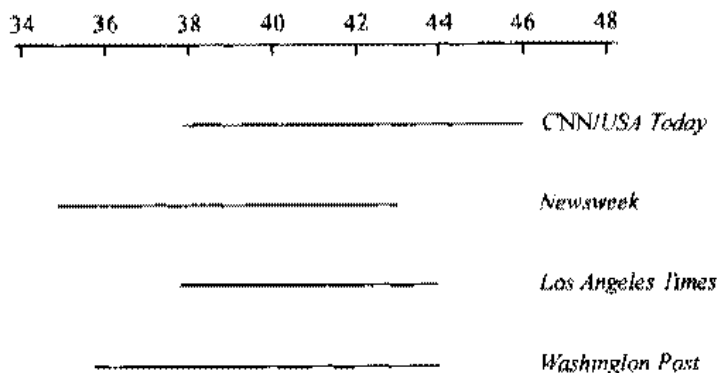


图 A.11 练习 6.39c 中的置信区间。

6.41 a. $7.0 \pm (2.01)(2.5/\sqrt{49}) = 6.3$ 到 7.7 。

b. 如果我们用构造置信区间 6.3 到 7.7 的方法对许多不同的样本分别构造置信区间, 它们中的 95% 包含差异的真值而有 5% 不包含差异的真值。我们不知道这个特定的区间是多数中的一个还是少数中的一个。

6.43 a. $12.3 \pm (2.086)(0.3) = 11.7$ 年到 12.9 年。

b. 总体均值在 11.7 到 12.9 年之间的置信水平是 95%。

- 6.45 a. $56 \pm 1.96 \sqrt{\frac{61 \times 39}{41} + \frac{5 \times 95}{807}} = 56 \pm 15 = 41$ 到 71 。
 b. 如果我们用构造置信区间 41 到 71 的方法对许多不同的样本分别构造置信区间,它们中的 95% 包含差异的真值而有 5% 不包含差异的真值。我们不知道这个特定的区间是多数中的一个还是少数中的一个。
 c. 这个区间没有包括 0,所以在年轻抽烟者的总体中这两个比率似乎有差异。
- 6.47 a. $61 \pm 100/\sqrt{531} = 57$ 到 65 ; $33 \pm 100/\sqrt{531} = 29$ 到 37 。
 b. 期望提高客户服务质量的公司数目比认为这件事已经发生的数目要大。
- 6.49 a. $0.56 \times 502 = 281$ 个“早起的人”和 $0.44 \times 502 = 221$ 个“夜猫子”。
 b. $53 \pm 100/\sqrt{281} = 53 \pm 6.0 = 57.0$ 到 59.0 ; $39 \pm 100/\sqrt{221} = 39 \pm 6.7 = 32.3$ 到 45.7 。
 c. $45 \pm 6.0 = 39.0$ 到 51.0 ; $37 \pm 6.7 = 30.3$ 到 43.7 。
 d. $74 \pm 6.0 = 68.0$ 到 80.0 ; $64 \pm 6.7 = 57.3$ 到 70.3 。
 e. 在 b 中置信区间没有重叠,表明两个组不相同;在 c 和 d 中置信区间重叠,表明两个组有重叠。
 f. 当我们用大小为 502 的样本,抽样误差变为 4.4。
 g. 用 4.4 使 d 中的答案产生差异。
- 6.51 a. $3.96 \pm 2.064(0.35) = 3.96 \pm 0.72 = 3.24$ 到 4.68 。
 b. 区间不包含 5,意味着当我们抛 10 次硬币时,5 不可能是出现正面的均值。

7 作出结论:假设检验

- 7.1 样本数据导致拒绝零假设。
- 7.3 a. 某个参数是否等于一个特定的值。
 b. 备选假设是问:参数是否等于所有没有在零假设中限定的值。
 c. H_0 和 H_a 。
- 7.5 当样本统计量与零假设中限定的值相差很大时,我们拒绝零假设。
- 7.7 a. 当零假设为真时,从总体中得到观测数据或更为极端数据的概率。
 b. 这个数据是那种当零假设为真时不太可能出现的数据。由于我们不认为这个数据是特别不可能的,唯一的解释就是数据来自另一个总体。故我们拒绝零假设。
 c. 显著水平就是考察数据之前我们选定的一个非常小的概率。而 p -值则是由此数据算出的。
- 7.9 a. 0.05。
 b. 在所有被检验为真的假设中,5%的假设会错误地被拒绝掉。
- 7.11 a. α 。
 b. α 在英文中拼为 Alpha。
- 7.13 a. $H_0: \pi = 0.5$ 。
 b. $H_0: \Pi = 50\%$ 。
- 7.15 显著水平和这两个样本的大小。

- 7.17 a. 两组的总体平均得分没有差异。
b. 室外组的平均得分比室内组的平均得分要高。
- 7.19 a. 1995 年总体平均薪水等于 \$ 43000——一个被表述为没有发生变化的假设。
b. 由于通货膨胀和可能的薪水增长,1995 年的平均薪水可能比 1985 年的高。
- 7.21 在给定零假设以后,获得现有的数据或更为极端的数据的概率为 0.50。
- 7.23 在单边检验中备择假设表述为参数大于(小于)给定值。在双边检验中备择假设表述为参数不同于某一个给定值。
- 7.25 a. 不是每个罐子正好重 18 盎司。
b. 下一个罐子的重量可能不同。
c. 随机地购买一些罐子并算出平均重量。然后用它来检验总体均值为 18 盎司的零假设。
- 7.27 a. 由于 p -值很小。
b. 全国值 53% 与样本值 45% 之间的差异并不算很大。
- 7.29 零假设常常表述为没有差异或没有变化。
- 7.31 a. 当总体参数具有零假设中限定的值时,获得我们的数据或更为极端数据的概率为 0.025。
b. 因为 p -值是这样小,所以我们拒绝零假设。
- 7.33 a. 我们拒绝平均偏好相等的零假设是因为我们不相信我们的数据来自这样一个不太可能的数据集。
b. 零假设也可能是正确的,我们的数据只不过是来自一个特别的样本罢了。获得观测数据或更为极端数据的概率为 0.001。
c. 在一个 7 分的尺度上 0.4 分的差异可能并不令人感兴趣。
- 7.35 当零假设为真时,5% 的样本会如此极端以致于我们拒绝零假设。5% 等于 20 个中的 1 个,按照 Fisher 的观点是一个很小的数。
- 7.37 某个结果从统计意义上来说可能是显著但事实上很微不足道。
- 7.39 a. 总体平均速度是否等于 5.0 呢? 某个样本平均速度为 5.96,它的 t -值为 2.49,自由度为 47。从均值等于 5.0 的总体的样本中随机抽样,10000 次里只有 49 次能得到这个或更大的 t -值。
b. 我们不认为样本来自那个总体,故拒绝零假设。总体均值大于 5.0。
- 7.41 a. 两个观测到的歧视均值 17.3 和 11.4 之间的差异在统计意义上是显著的。从均值相等的总体中观测到等于或大于这个差异的概率小于 0.001。由于 p -值小于 0.05,关于开始组织新家庭的平均考虑之间的差异在统计意义上是显著的。
b. 经历过性别歧视的女性在抑郁状况上的得分比较高。
- 7.43 我们可以检验零假设:支持总统的百分比为 50,我们也许不能拒绝它。
- 7.45 a. 由于一个样本与下一个样本之间受随机性的影响,他们会在 1.4 附近变动。
b. $t = (2.0 - 1.4) / 0.5 = 1.2$ 。
c. 它并非不寻常的大,我们的样本有可能来自一个均值为 1.4 的总体。
- 7.47 a. $H_0: \mu = 5.1$ 天。
b. $t = (7.0 - 5.1) / (2.5 / \sqrt{49}) = 5.32$,自由度为 48, $p < 0.001$ 。

c. 在总体均值为 5.1 的情况下,我们不太可能获得一个均值为 7.0 或更大的样本;拒绝零假设。工作人员更经常地生病。

7.49 a. $H_0: \Pi = 73.2\%$ 。

b. $z = (67 - 73.2) / \sqrt{67 \times 33 / 300} = 2.18$ 和 $p = 0.015$ 。这个 p -值是如此的小,我们拒绝零假设。城市雇员中单独驾驶者的减少量是显著的。

c. 减少量似乎不是很大。

7.51 $H_0: \pi = 0.5$, 其中 π 是随机选中的人为浸礼会教徒的概率。 $z = (103 - 0.5 \times 190) / \sqrt{190 \times 0.5 \times (1 - 0.5)} = 1.16$ 。因为 $z < 1.96$, 我们不能拒绝零假设。总体中浸礼会教徒和卫理公会教徒的数目可能相等。

7.53 $z = (98 - 80) / \sqrt{98 \times 2 / 50 + 80 \times 20 / 200} = 6.74$, $p < 0.0001$ 。拒绝零假设。由于零假设表述为总体中两个比率相等,这个共同值的一个估计是用支持民主党的人的总数目除以汽车拥有者的总数目, $209 / 250 = 83.6\%$ 。当我们使用这个共同的百分比时,则 $z = 3.07$, $p = 0.001$ 。

7.55 $H_0: \Pi = 50\%$ 。 $z = (134 - 200 \times 0.5) / \sqrt{200 \times 0.5 \times (1 - 0.5)} = 4.81$, $p < 0.001$; 拒绝零假设。男性乘游戏车的人数要多一些。

7.57 由 $\pi = 0.5$, $n = 10$ 的二项分布得出 $p = (120 + 45 + 10 + 1) / 1024 = 176 / 1024 = 0.17$ 。我们不能拒绝 $\pi = 0.5$ 的零假设,很可能有女孩和有男孩的概率是一样的。

7.59

$$z = \frac{77.8 - 44.4}{\sqrt{61.1(100 - 61.1)}} \sqrt{\frac{1}{36} + \frac{1}{36}} = -2.91$$

$p = 0.0018$ 。拒绝零假设;在大学,女性与男性学生之间有差异。

8 变量间的关系

8.1 一个变量的值对应于另一个变量的某些值。

8.3 在数据来源的总体中两个变量之间也许没有关系。

8.5 这种关系在统计意义上是显著的吗?

8.7 当它不是因果关系并可以用一个或多个其它变量解释时。

8.9 a. 这种关系可能是由另一个变量产生的吗?

b. 我们对这两个变量有足够多的了解来得出以下结论:温度可能是产生观测关系的另一个变量。

c. 我们很难确定产生观测到的关系的其它变量。

d. 历史上,有很多关系被认为是因果关系。例如,巫婆制造疾病。

8.11 0.17 表示一个弱的关系,0.87 表示一个强的关系。

8.13 这种关系不仅在样本中存在,而且在样本的来源总体中也存在。

8.15 错。我们不需要因果相关也可以从一个变量估计另一个变量的值。

8.17 最好是知道一个变量事实上产生了另一个变量。自变量常常是发生较早的那一个。

- 8.19 对不同的变量我们有不同的统计方法,在选择适当的方法之前,我们需要确定变量的类型。
- 8.21 a. X 是大陆而 Y 是国家的 GDP。
b. 棒球队每场比赛平均得分数和一个赛季中赢的场次。
c. 将这一周与下一周的前十名的橄榄球队排序。
- 8.23 多元统计方法。
- 8.25 一个关系可能是伪的也可能是因果的。仅仅用这两个变量的数据不能用来决定此关系是伪的还是因果的。
- 8.27 仅仅用观测数据很难证明因果关系,只有当其它所有可能影响观测关系的变量都被考虑进来后我们才可能做这项工作。但所有其它可能变量是永远不可能都被考虑进来的。
- 8.29 其它变量必须被考虑进来以考察它们是否对观测的关系有贡献。
- 8.31 吸烟与婴儿大小之间可能是伪的或因果关系。在没有其它变量的信息或进行一项实验之前,我们不能断定是哪种关系。可能还有其它变量使女性吸烟而且仍然使她们生出较小的婴儿。
- 8.33 a. 他们必须问女性这样的问题:她们感觉自己的身材如何?产生矛盾的比例可能与问题的形式有关。
b. 它可能是伪的。我们很难证明这个关系是因果的。
- 8.35 a. 表 A.5。

表 A.5 练习 8.35a 的分布

		秃头		
		是	否	合计
疾 病	心脏	214	451	665
	其它	175	597	772
	合计	389	1048	1437

- b. 665 分之 214 为 32%。
- c. 772 分之 175 为 23%。
- d. $32 - 23 = 9$ 表示秃头多出的百分比。
- e. 秃顶是自变量,我们比较自变量不同类别之间的百分比。
- f. 这些人自愿的,因为只有那些心脏病发作的人才去医院。一个更为可取的方法是跟踪一组秃顶的和另一组不秃顶的人看每组中有多少人心脏病发作。
- g. 没有。
- 8.37 a. 对品牌的选择不同。
b. 表 A.6。大多数黑人喜欢 Newports,大多数白人喜欢 Marlboros。

表 A.6 练习 8.37b 的比率分布

品 牌		种族		合计
		黑人	白人	
	Marlboro	9	71	68
	Newport	56	6	8
	Kool	9	1	1
	其它	27	22	24
	合计	101	100	101

- c. 这些比率非常不同,暗示着很强的关系。
 d. 这些选择可能与品牌的广告宣传有关。
- 8.39 a. 因为呆在家里的那一行里有更高的百分比。
 b. 我们可能对最后三年做出了错误的预计,其比例为 52%、58% 和 58%。
 c. 那时大多数女性呆在家里。
 d. 32%。
 e. 她是工作着的。
 f. 42%的时间。
 g. 我们可以对每年作一个更准确的预计。
 h. 这段时间母亲出去工作的人数增多了。
 i. 仅仅从这些数据我们不能讨论因果关系。

9 两个分类变量的 χ^2 分析

- 9.1 学生的课题。
- 9.3 一个用行和列的来显示频率的表格。
- 9.5 ϕ 。
- 9.7 χ^2 的大小与因果关系无关。它只用来得出 p -值。
- 9.9 a. 是。这两列不同。
 b. 0.62, 中等。
 c. 否, 由于 p -值很小, 我们拒绝零假设。
 d. 可能。
- 9.11 关系很弱。它可能是偶然产生的。
- 9.13 a. 是。列百分比不同。
 b. 弱。
 c. 是, 小的 p -值。
 d. 仅仅从这些数据我们不能讨论因果关系。
- 9.15 a. 秃头, 因为它先出现。
 b. 这些数据中存在关系, 但是很弱。总体中没有关系的零假设被拒绝。我们不能讨论因果关系。

- 9.17 a. 是。列百分比不同。
 b. 这个关系似乎很弱。
 c. 统计意义上显著。
- 9.19 a. 受教育越多越趋向于不回答。受教育多的孩子中更多的人有主张。
 b. 具有弱的但是统计意义上非常显著的关系。
- 9.21 a. 表 A.7。

表 A.7 练习 9.21a 的分布

		判决		合计
		课程	监狱	
新 罪	是	6	18	24
	否	26	22	48
	合计	32	40	72

- b. 具有弱的但是统计意义上非常显著的关系。
 c. 我们不清楚样本是否为一个恰当的随机样本。
- 9.23 大多数有过失的男孩不戴眼镜而大多数没有过失的男孩戴眼镜,所以这两个变量之间有关系。这个关系为中等强,我们拒绝在所有男孩总体中没有关系的零假设。
- 9.25 a. 表 A.8。

表 A.8 练习 9.25a 的分布

		性别		合计
		男性	女性	
被选上	是	36	13	49
	否	41	7	48
	合计	77	20	97

- b. 在这些数据中有关系。它很弱且不是统计意义上显著的。你不能得出有关因果关系的任何结论。
- 9.27 在这些数据中有关系。它中度强且统计意义上显著。 p -值将决定它有多显著。仅仅从这些数据你不能得出有关因果关系的任何结论。
- 9.29 a. 表 A.9。

表 A.9 练习 9.29a 的分布

		生日		合计
		1-6 月	7-12 月	
征兵 数目	1-183	91	92	183
	183-366	91	92	183
	合计	182	184	366

- b. 这两列不同,这些数据中有关系。

c. $\chi^2 = \frac{366(73 \times 74 - 109 \times 110)^2}{183 \times 183 \times 182 \times 184}$

d. 仅仅因为偶然因素拒绝由观测表格仅由偶然产生的零假设。抽签似乎不是随机的。

9.31 a. 表 A.10。

表 A.10 练习 9.31a 的分布

		IQ		合计
		低	高	
犯罪	0,1	806	1158	1964
	2+	382	161	543
	合计	1188	1319	2507

b. 是。百分比的列不会相同。

c. $\phi = 0.24$ 。

d. $\chi^2 = 146.59$ (自由度 = 1), $p < 0.0001$ 。拒绝总体中没有关系的零假设。

e. 在低 IQ 的人中 2 次或多次犯罪的人比高 IQ 的人多。仅仅从这些数据我们不知道关系是否为因果的。

9.33 a, b. 表 A.11。

表 A.11 练习 9.33a, b 的频率和比率

		地点				
		北美	南美	合计	北美	南美
宗教信仰	天主教	190	310	500	38%	69%
	犹太教	10	10	20	2	2
	新教	120	30	150	24	7
	其它	180	100	280	36	22
	合计	500	450	950	100%	100%

c. 是。这两个比率分布不同。

d. $V = 0.33$ 。

e. 学生的课题。

f. $\chi^2 = 103.31$ (自由度 = 3), $p < 0.0001$ 。拒绝总体中没有关系的零假设。

g. $\chi^2 = 103.27$ (自由度 = 2), $p < 0.0001$ 。仍然拒绝零假设。

h. 两个 χ^2 之间的差异为 0.04, 自由度相差 1。

9.35 a. 是。好像比率列将会不同。

b. 弱。

c. $\chi^2 = 82.07$ (自由度 = 3), $p < 0.0001$ 。拒绝在所有注册选民的总体中变量之间没有关系的零假设。

d. $V = 0.14$ 。

9.37 a. 是。

b. $V = 0.34$ 。

c. $\chi^2 = 17.64$ (自由度 = 2), $p = 0.0001$ 。这些数据仅仅由偶然因素产生, 拒绝零假设。

d. 所以现在有较多的女性位于较低的等级是因为现在有更多新的女性当大学教员但她们目前还没有达到较高等级。

- 9.39 a. 是。
 b. $V = 0.30$ 。
 c. $\chi^2 = 11.69$ (自由度 = 2), $p = 0.003$ 。这种关系不可能仅仅由于偶然因素而产生。
 d. 比例上来说, 东德的女性获得最多的奖牌而苏联的女性获得最少的奖牌。
- 9.41 这些数据中有关系。 $\phi = 0.21$ 。 $\chi^2 = 4.63$ (自由度 = 1), $p = 0.031$ 。这个关系是统计意义上显著的。
- 9.43 自由度为 $(3-1)(4-1) = 6$ 。
- 9.45 表 A.12。 $\phi = 0.19$ 。 $\chi^2 = 9.45$ (自由度 = 1), $p = 0.002$ 。

表 A.12 练习 9.45 的分布

投票		轿车		合计
		Saab	Volvo	
	民主党人	49	160	209
	共和党人	1	40	41
	合计	50	200	250

- 9.47 有关系。 $V = 0.46$ 。 $\chi^2 = 181.97$ (自由度 = 1), $p < 0.0001$ 。拒绝没有关系的零假设。仅仅从这些数据你不能讨论因果关系。
- 9.49 a. 表 A.13。

表 A.13 练习 9.49a 的分布

		蛋白尿		合计
		是	否	
高血压	是	28	21	49
	否	82	286	368
	合计	110	307	417

- b. 这些数据中有关系。 $\phi = 0.25$ 。 $\chi^2 = 27.06$ (自由度 = 1), $p < 0.0001$ 。拒绝总体中没有关系的零假设。
- 9.51 这些数据中有关系。 $V = 0.26$ 。 $\chi^2 = 75.89$ (自由度 = 6), $p < 0.0001$ 。拒绝在更大的总体中没有关系的零假设。
- 9.53 a. 他们不相同。
 b. $\chi^2 = 5.61$, 自由度为 5, $p = 0.35$ 。 $V = 0.41$ 。这个关系很弱且不显著。
- 9.55 a. 这些数据中有关系。 $\phi = 0.12$ 。 $\chi^2 = 1.31$, 自由度为 1, $p = 0.29$ 。
 b. 由于按照字母次序分组, 所有没有差异。
- 9.57 a. 比率按行给出。这意味着是否修女带来了阿尔茨海默氏病作为自变量。由于语言能力先被衡量, 用语言能力作为自变量是有意义的。
 b. 这些数据中有关系。 $\phi = 0.76$ 。 $\chi^2 = 14.31$ (自由度 = 1), $p = 0.0002$ 。拒绝零假设。你不能讨论因果关系。
- 9.59 学生的课题。

10 两个数值型变量的回归分析和相关分析

- 10.1** 两个数量型变量之间关系强度的衡量表示散点图中点与回归线之间的接近程度。
- 10.3** 一个自变量。
- 10.5** a. 学生的课题。
b. 斜率和截距。
- 10.7** a. 自变量在 x -轴, 因变量在 y -轴。
b. 电视广告的次数。
c. 自变量。
d. 这些点在图中由左下角向右上角分布。
- 10.9** 孤立的点有较大效应。相关系数将增大。
- 10.11** 所有点在一条(45度)线上。
- 10.13** a. +1, 如果认为富裕的人应该付更多的税。
b. -1。随着教育发展, 文盲应该减少。
c. 0。应该没有关系。(在我们的社会中也许是正相关。)
- 10.15** a. 36.1 卡热量。
b. 这是从观测数据外推得出的。如果关系是沿同一条直线, 则一份没有脂肪的食物预计有 36.1 卡热量。
- 10.17** \hat{y}_0 。
- 10.19** a. 回归平方和及残差平方和。
b. 回归平方和是将每个观测从回归线到均值线距离的平方相加。残差平方和是将每个观测点到回归线距离的平方相加。
- 10.21** a. 所有对因变量的效应来自自变量。
b. 没有效应。
c. 加或减 1。
- 10.23** a, b. 回归给出了关系的形式, 而相关给出了关系的强度, 二者都是必须的。
- 10.25** 正。
- 10.27** 用回归或相关系数。
- 10.29** a. 当其他因素不变时, 相关系数变大。
b. 一个大的相关系数得出小的 p -值。
- 10.31** a. 当因变量是哑元而自变量是数量型变量时。
b. 购买一辆雪佛莱(Chevrolet)车或买一辆凯迪拉克(Cadillac)车是收入的函数。
- 10.33** a. 一条截距为 183 而斜率为 0 的水平线。相关系数预计为 0。
b. 这条直线斜率为负, 意味着早的月份有较大的平均征兵数目而晚的月份有较小的征兵数目。两个变量之间的关系较强。
c. 一个这么大的 t -值有很小的 p -值。拒绝变量之间关系仅仅为偶然的零假设。
- 10.35** a. 截距等于巧克力小吃的均值。斜率等于两个均值之间的差异。
b. 否, t 太小, 不是统计意义上显著的。

- 10.37 a. 上大学的人越多平均收入越高,上大学的人越少收入也越少。
b. 否。 p -值这样小以致于我们拒绝仅仅由偶然因素产生该关系的零假设。
c. 仅仅从这些数据我们不能分辨。
d. 华盛顿特区不寻常,因为作为国家的首都它有很多为政府工作的具有大学学历的人而同时它又有很多贫穷的人。
- 10.39 a. Sam 的线将会更倾斜。
b. 只是对 Sam 的数据。
c. 是,事情好像是这样的。它可能依赖于 Y_1 与 X 之间和 Y_2 与 X 之间的两个相关系数。
- 10.41 a. 药的剂量越大,有反应的动物越多。
b. 随着剂量一个单位的增长,反应的比例增长 0.13。这个关系是统计意义上显著的。拒绝在较大总体中两个变量之间不相关的零假设。
c. 你不知道关系的强度,也不知道因果关系。
- 10.43 a. 减少 3.47 秒。
b. 减少 13.525 秒。
c. 他的时间比预计的慢 3: $44.39 - 3 \times 43.82 = 0.57$ 秒。
d. $76.07 - 0.3468 \times 93 = 43.82$ 分钟或 3: 43.82 分钟。
- 10.45 a. 有强的正相关。相差一吨的两艘船平均上相差 0.00062 个船员。这个关系是统计意义上显著的,有理由相信这个关系是因果的。
b. 6.2 个船员。
c. 当吨位 = 192 时,预计船员为 10.7 人。当吨位 = 3246 时,预计船员为 29.7 人。
- 10.47 a. 图 A.12。
b. 它是负的、强的和统计意义上显著的关系。
c. 职员学科等级,不是所有学校都有对应所有职业的学科,等等。
d. 直线上方的点表示更假向于服务型的职业,而直线下方的职业是更倾向于学术型的职业。
- 10.49 a. $r = 0.40$ 。
b. $t = 2.23$, 自由度为 26, $p = 0.017$ 。显著的。
c. 赢的次数 = $45.8 + 7.0$ 平均得分的次数。
d. 平均上,每场比赛多一次得分预计多赢 7 次。
- 10.51 a. 道路距离 = $72.6 + 1.11$ 直线距离。
b. 如果距离相等,这些点会落在一条过原点的斜率为 1 的直线上。
c. $t = 1.40$, 自由度为 13, $p = 0.09$ 。你不能拒绝截距为 0 的零假设。
d. 美国有许多东-西和南-北的道路,而英国的道路则更为直接地连接城市。
- 10.53 a, b. 回归 相关
- | | |
|---|--------|
| 赢的次数 = $118 - 8.15 \text{ ERA (投手责任得分率)}$ | - 0.54 |
| 赢的次数 = $45.8 + 6.98$ 得分次数 | 0.40 |
| 赢的次数 = $71.2 + 0.056$ 本垒打次数 | 0.22 |
| 赢的次数 = $-16 + 360$ 平均击球 | 0.46 |

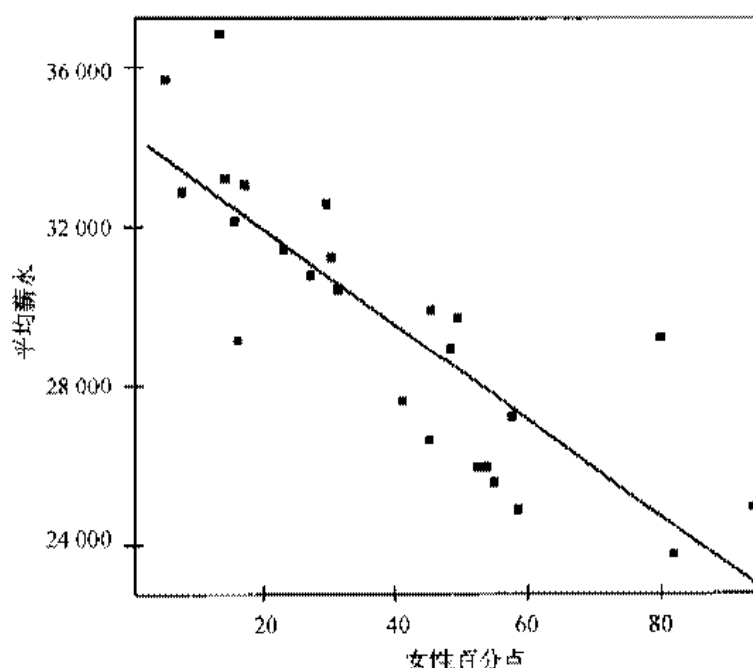


图 A.12 练习 10.47a 的散点图和回归直线。

改善(降低)ERA 一分赢的次数将增加 8.5 次。每场比赛多一个得分将导致多赢 7 次。在这个赛季中多 100 个本垒打的得分将导致多赢 5.6 次。在击球上平均提高 0.1 将导致多赢 3.6 次。

10.55 a. 击球率为均值;其它为中位数。

b. 安全打的次数 = $-9.0 + 0.30$ 论到击球的次数 $r = 0.98$

得分的次数 = $-2.2 + 0.57$ 安全打的次数 $r = 0.95$

得分的次数 = $-7.1 + 0.17$ 论到击球的次数 $r = 0.93$

本垒打的次数 = $-2.7 + 0.29$ 击球得分的得分次数 $r = 0.91$

本垒打的次数 = $-21.1 + 120.1$ 击球百分率 $r = 0.42$

这些是四个最高和最低的相关系数的回归方程。

10.57 散点图看起来是线性的,正相关的。回归方程为:

$$\text{首领年龄} = -51.2 + 2.7 \text{ 成员的平均年龄}$$

对于成员平均年龄相差一年的两个组,首领的平均年龄相差 2.7 年。关系的强度是 $r = 0.70$ 。对仅仅由偶然因素产生这种关系的零假设的检验得出 $t = 2.62$, 自由度为 7, $p = 0.03$ 。这个 p -值是边沿显著的,主要是因为样本很小。

10.59 a. 学生的课题。

b. 学生的课题。

c. 学生的课题。

d. 给出一个对每个人的大概拟合。

- e. 预计其毕业的 GPA 为 $0.6 + 0.74 \times 1.9 = 2.0$; 看起来好像 Elmer 可以毕业
f. 这个方程将不变, 而相关性将变弱。

10.61 a. 图 A.13。

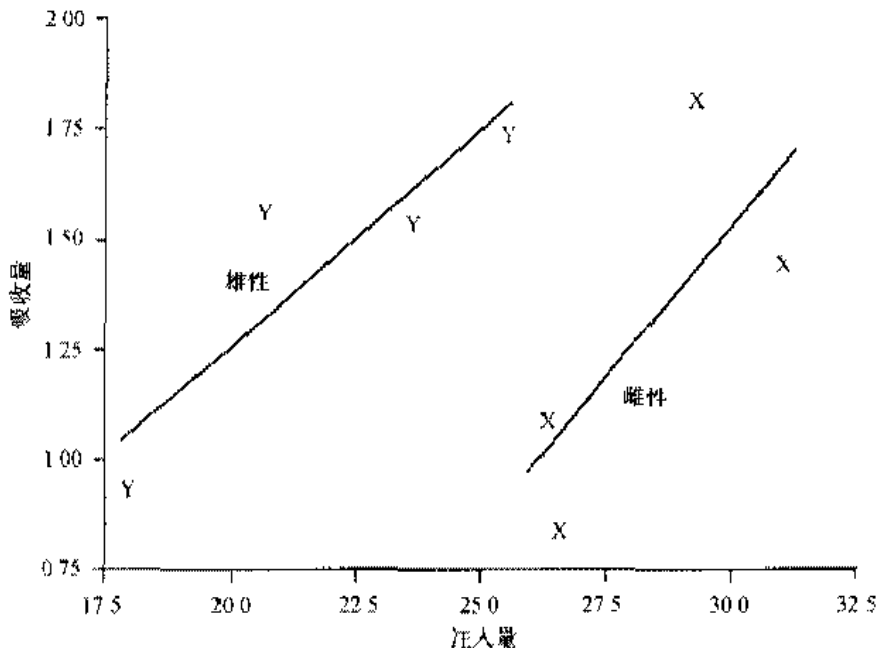


图 A.13 练习 10.61a 的散点图和回归直线。

- b. 两种关系都是正相关。雄性关系的斜率比雌性关系的斜率小。相关性相同。
c. 雄性: 吸收量 = $-2.70 + 0.14 \times \text{注入量}$
雌性: 吸收量 = $-0.67 + 0.10 \times \text{注入量}$
d. 雌性老鼠的直线比雄性老鼠的直线陡, 因而雌性老鼠的截距比雄性老鼠小。
e. 预测值为 $-2.70 + 0.14 \times 25 = 0.80$ 和 $-0.67 + 0.10 \times 25 = 1.83$, 所以差异变为 $1.83 - 0.80 = 1.03$ 。
f. 即使这些均值之间差异不大, 对于一个特定的注入量雄性老鼠具有更高的吸收值。

10.63 a. 因为各个州有不同的病床数目, 以及其它的一些原因, 医疗补助计划在州与州之间不一样。

b. 图 A.14。

- c. 三个南部州再加上 Indians 和 North Dakota 两个州所具有的人均床位较由获得医疗补助计划的人员百分比解释的人均床位更多。床位较少的州都在北部。North Dakota 具有最大的残差, 因此是最不同寻常的州。
d. 每 100000 人床位数 = $286 + 9.4 \times \text{获得医疗补助计划的百分点}$, $r = 0.44$, $t = 1.38$, 自由度为 8, p -值 = 0.10。获得医疗补助计划相差一个百分点的两个州在每 100000 人床位数上相差 9.4 个。这个关系中等强。 p -值的大小表明, 这个关系可能仅仅是由偶然因素产生的。

10.65 a. 图 A.15。

- b. 随着剂量的增长, 每窝幼崽的数目在下降, 畸型的百分比在增长, 而且胎儿重量在

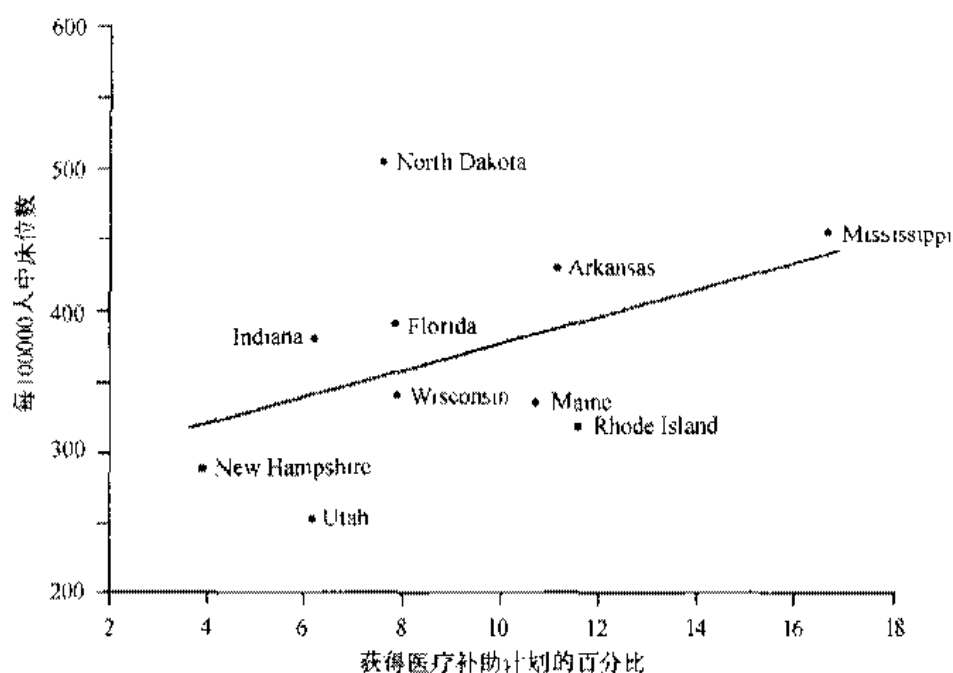


图 A.14 练习 10.63b 的散点图。

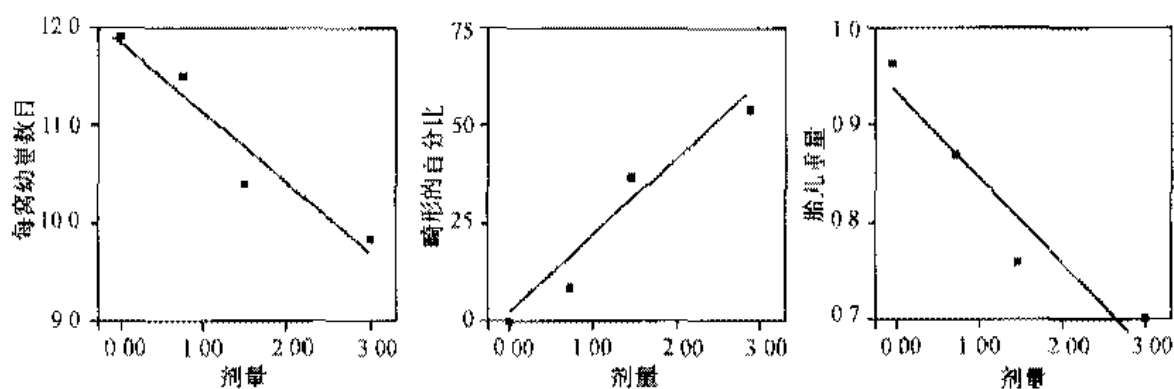


图 A.15 三个变量对剂量的散点图。

下降。

c 每窝幼崽数目 $= 11.85 - 0.75 \text{ 剂量}$ $r = -0.96$ $p = 0.02$

畸形的百分比 $= 0.3 + 19.9 \text{ 剂量}$ $r = 0.97$ $p = 0.02$

胎儿重量 $= 0.95 - 0.09 \text{ 剂量}$ $r = -0.96$ $p = 0.02$

这个关系是强的和统计意义上显著的。

d 否。它们是以不同的单位测量的。

e. 这三个中的每一项分析都是基于四个观测。对于个体数据每个分析将有 100 个观测值。在每个均值点附近都有变差,而且残差变量将会更大。但是对于更大的样本, p -值可能会小些。

10.67 这个散点图显示了一种线性的正相关。

$$\text{脂肪百分比} = 3.2 + 0.55 \text{ 年龄} \quad r = 0.79$$

由此得出 $t = 5.19$, 自由度为 16, $p = 0.0001$, 所以这个关系是统计意义上显著的。你不知道这个关系是否为因果的。

10.69 a. 散点图显示了一种关系。

$$\text{死亡率} = -21.8 + 2.4 \text{ 温度} \quad r = 0.87$$

由此得出 $t = 6.76$, 自由度为 14, $p < 0.0001$ 。这个关系是显著的。

b. 仅仅从这些数据我们只能猜测是因果关系。

10.71 学生的课题。

10.73 a. 斜率为 1, 截距为 0。

b. 斜率为 1, 截距为 5。

c. 斜率为 1.1, 截距为 0。

d. 新郎 = $-1.6 + 1.13$ 新娘。

e. 新郎的年龄一般比新娘大, 尤其是在大龄夫妇中。

f. 在散点图当中每对夫妇可被识别, 而两个枝叶图却可以分别地显示两个分布。

11 ANOVA: 一个分类变量和一个数量变量的方差分析

11.1 a. 犯罪数目除以人口总数。

b. 大一些的州因为大, 可能会有更多的犯罪。

c. 我们可以考察单个的州并衡量它是否与其它州不同。但是这样一来就没有自变量去解释为何攻击会不同了。我们猜想地区会对犯罪有影响, 所以我们把地区当作自变量。

11.3 a. 性别是自变量而犯罪率是因变量。

b. 分类型和数量型。

11.5 学生的课题。

11.7 各个州的比率都相同。

11.9 问题 3。

11.11 a. 五。

b. $50 - 6 = 44$ 。

11.13 学生的课题。

11.15 a. 我们感兴趣的变量已经被选出来作为自变量。所有其它变量都放进残差变量中。

b. 当一个因素被测量时, 观测值和真值之间的差异会变为测量的误差。

11.17 如果在每一数对中都没有差异, 则两个值是同样的。画这样一个点, 它将位于过原点的 45 度线上。

11.19 a. $56.6 - 48.6 = 8.0$ 。

b. 是。

11.21 a. 地区。

- b. 除地区之外的所有自变量。
 c. 如果 R^2 是被自变量解释的部分, 则 $1 - R^2$ 是被残差变量解释的部分。
 d. 总的平方和。
 e. $R^2 = 1.00$ 。
 f. 它们将全部相等。
- 11.23 在变量之间是一种弱的关系。
- 11.25 a. 大学和得分之间关系的强度是 $R = \sqrt{0.40} = 0.63$ 。在大学之间有显著的差异。
 b. 哪所大学与众不同, 哪所却不这样?
- 11.27 a. 在这些数据中国家变量与旅行数日之间有关系。这个关系是统计意义上显著的。
 b. 这个关系的强度。
- 11.29 a. 雌性 109.6 mm, 雄性 113.4 mm。
 b. $109.6 - 113.4 = -4.8$ 得出 $t = -3.484$, 自由度 = 18, $p = 0.0013$ 。(表 A.14)。

表 A.14 练习 11.29b 的方差分析

来源	平方和	自由度	均方	F-比	p 值
性别	115.2	1	115.2	12.14	0.0026
残差	170.8	18	9.489		
总计	286.0	19			

注意由于 F 的自由度为 1 左右, $(-3.484)^2 = 12.14$ 。这意味着 F 的 p -值等于 $t < -3.484$ 或 $t > 3.484$ 的概率。

- 11.31 a. 图 A.16。从散点图我们发现在冻酸奶与冰牛奶之间差异很小或没有差异, 而冷冻小吃就没有这么受欢迎了。
 b. 64.2, 64.0, 27.2。
 c. $R = \sqrt{6364/9395}$, 一种强的关系。 $F = 3182.1/126.3 = 25.20$ 自由度为 2 和 24, 得出 $p = 0.000001$ 。拒绝均值相等的零假设。
- 11.33 a. 当一个南部城市具有一个低/高的得分, 相应的北部城市得分也为低/高。
 b. 平均上, 北部城市比相应的南部城市高 4.8 分。
 c. 对于一个配对检验, $t = 1.47$, 自由度为 9, 而 $p = 0.09$ 。平均差异不是统计上异于 0 的。
- 11.35 a. 运动的种类。
 b. 要预测的身高。
 c. $t = (5.48 - 8.00) / \sqrt{(0.32^2 + 0.50^2)} = -4.24$, 自由度为 41, $p = 0.00006$ 。
 d. p 的确切值告诉我们结果是非常显著的。如果我们只知道 $p < 0.05$, 则 p 可能为小于 0.05 的任何值。如果 p 接近那个值, 则结果仅仅是边界显著的。如果 p -值比它小很多, 则结果是很显著的。
- 11.37 a. 表 A.15。

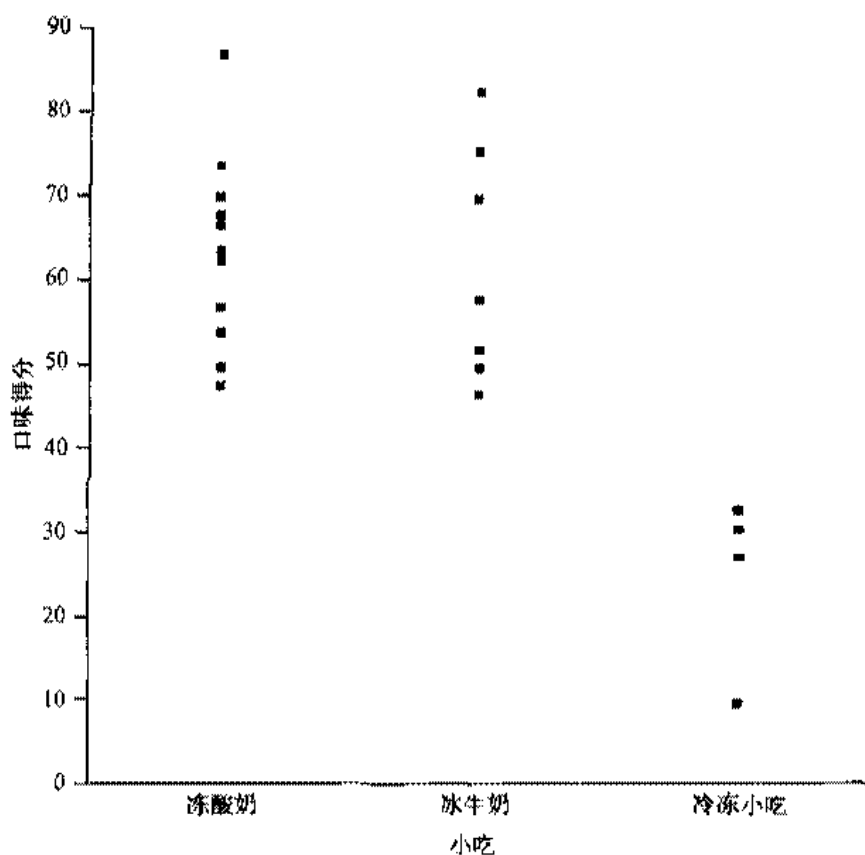


图 A.16 练习 11.31a 的图。

表 A.15 练习 11.37a 的方差分析

来源	自由度	效应	比率	均方	F-比	p-值
气候带	7	88866	0.41	12695	3.93	0.002
残差	40	129165	0.59	3229		
总计	47	218031	1.00			

b. 这个关系的强度为 $R = \sqrt{0.41} = 0.64$ 。由于 p -值很小, 这个关系是统计意义上显著的, 且气候带之间有差异。

c. 我们可以成对地比较这些气候带看看哪些互相不同。

11.39 a. 有 9 个正的差异和 1 个负的差异。用样本大小为 $n = 10$ 和概率为 $\pi = 0.5$ 的二项分布得出 $p = (10 + 1)/1024 = 0.001$ 。这个 p -值很小, 我们拒绝 $\pi = 0.5$ 的零假设。这样, 正和负的差异就不具有相同的可能性。

b. 它们都给出了统计意义上显著的结果, 符号检验的 p -值在此比 t -检验的 p -值小。这就产生了一个问题, 是否差异的分布足够达到做 t 检验的正态程度。

11.41 这四个均值是全日制托儿所为 6.69, 私人家庭照管为 5.34, 雇佣保姆为 7.36, 亲戚照管为 5.06。由于均值不同, 在这些数据的支出与护理种类之间有关系。 $R = 0.99$, 这种关系很强。因为 $F = 173.96$, 自由度为 3 和 12, $p < 0.0001$, 我们拒绝四个总体均值相等的零假设。

- 11.43 对两个均值之间差异的 t -检验得出 $t = 4.02$, 自由度为 37, $p = 0.0001$ 。在两组之间有统计意义上显著的差异。
- 11.45 a. 差异的均值为 67。检验均值为 0 的零假设得出 $t = 1.87$, 自由度为 6, $p = 0.055$, 为边界显著。
b. 对于不配对数据 $t = 0.78$, 自由度为 12, $p = 0.34$ 。

12 两个顺序变量的秩方法

- 12.1 a. 变量值是可以从多到少排序的变量, 但我们又不知道一个值与另一个相差多少。
b. 学生的课题。
- 12.3 a. 衡量两个以词作为其值的顺序变量之间关系强度的系数。
b. -1 到 $+1$ 。
c. r_s 。
d. 大多数观测值位于表格从左下到右上的主对角线上, 而 γ 将为正。
- 12.5 γ 需要变为 z 。用统计表 1 或统计软件算出 p -值。
- 12.7 学生的课题。
- 12.9 a. 顺序变量。
b. 用全班的排名(秩)我们只知道某个学生比另一个学生好。用 GPA 我们可以知道当以成绩衡量表现时, 某个学生比另一个学生好多少。(即使在计算平均学分绩时成绩被当作数值变量处理, 它们本身也可以是顺序变量。)
- 12.11 次序(秩)变量。
- 12.13 社会地位最高的人比社会地位最低的人有更多的小孩。这个关系是弱的且统计意义上不显著的。
- 12.15 a. 把 γ 变成 z -变量并算出 p -值。
b. 如果样本不是太小, p -值将会小从而拒绝零假设。
c. 在所有葡萄总体中变量之间没有关系。
- 12.17 a. 在某个赛季表现好的球队在其它赛季表现也好。而在某个赛季表现差的球队在其它赛季表现也差。
b. 更小。
- 12.19 a. 没有哪个城市会赶上另一个, 故次序也不会改变。
b. 没有哪个城市会赶上另一个, 故次序也不会改变。
c. 大多数城市具有相同的等级, 但少数城市会比其它城市具有更高等级。
d. 拒绝零假设。这个排序不是仅仅由偶然因素产生的。
- 12.21 a. 表 A.16。由于百分比分布不同, 在这些数据的两个变量之间有关系。

表 A.16 练习 12.21a 的比率分布

		对于堕胎的立场			综合
		反对	混同	支持	
重要性	最重要之一	38%	12%	11%	14%
	重要	36%	46%	44%	44%
	不大/很不重要	25%	42%	45%	42%
	合计	99%	100%	100%	100%

b. $G = \frac{407336 - 595409}{407336 + 595409} = -0.19$ 。

c. $z = \frac{-0.19}{\sqrt{4(3+1)(3+1)/9(2421)(3-1)(3-1)}} = -7.01$ 。这个 p -值很小,我们拒绝总体中没有关系的零假设。

d. 仅仅从这些数据我们不能区分这个关系是否为因果的。

12.23 学生的课题。

12.25 a. 表 A.17。

表 A.17 练习 12.25a 的年数排序

年数	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
结婚	1	2	3	4	5	6	9	8	7	10	13	14	15	12	11	16	18	17	19
离婚	1	2	3	4	5	6	7	8	9	10	11	16	13	12	14	15	17	18	19

b. 图 A.17。

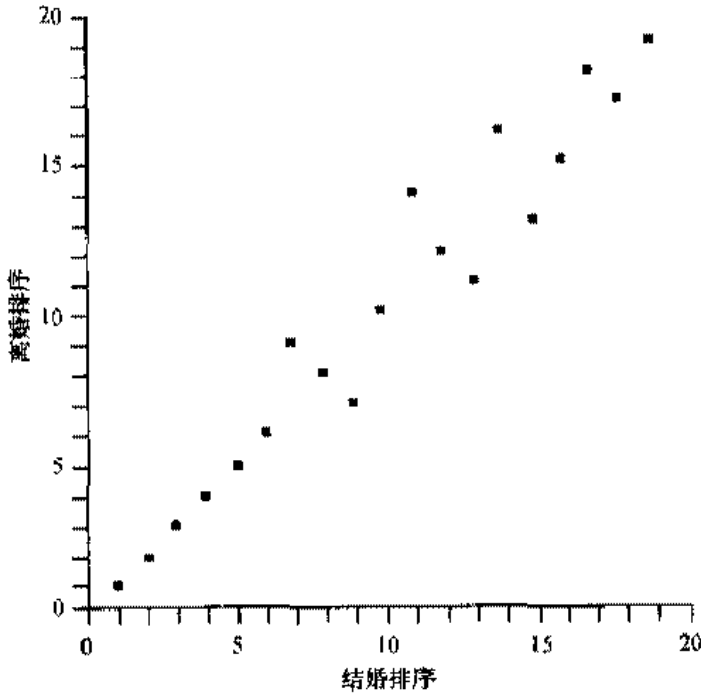


图 A.17 练习 12.25b 的散点图。

- c. 数据显示了很强的线性关系。原始数据中没有向左弯曲的点。
 - d. $r_s = 0.97, t = 17.03$, 自由度为 17, $p < 0.0001$ 。这个关系非常强且是统计意义上显著的。我们不知道它是否为因果的。
- 12.27
- a. $r_s = 1.00$ 。一个变量的秩与另一个变量的秩完全相同。
 - b. 概率很小。大多数女性没有乳腺癌。
 - c. 总体增大了, 更多的女性面对可能性。
- 12.29
- a. 图 A.18。

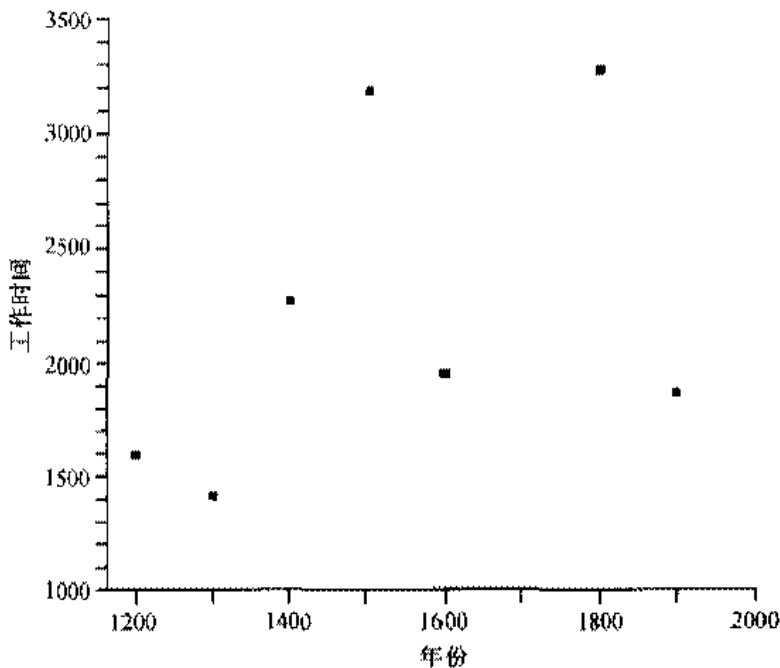


图 A.18 练习 12.29a 的散点图。

- b. 小时数增加了一会儿, 然后又减少了。由于数据可能不是线性的, 我们不能用通常的回归和相关分析。
- c. 表 A.18。

表 A.18 练习 12.29c 的年份和工作小时散排序

年份排序	每年工作小时散排序
1	2
2	1
3	5
4	6
5	4
6	7
7	3

- d. 这个关系仅仅由偶然因素产生。算出 r_s , 把这个值变为 t , 再算出相应的 p -值。
- e. $r_s = 0.50, t = 1.29$, 自由度为 5, $p = 0.13$ 。我们不能拒绝零假设。
- f. 这些变量都是数值变量, 可以用某种形式的相关和回归分析。

12.31 a. 图 A.19。

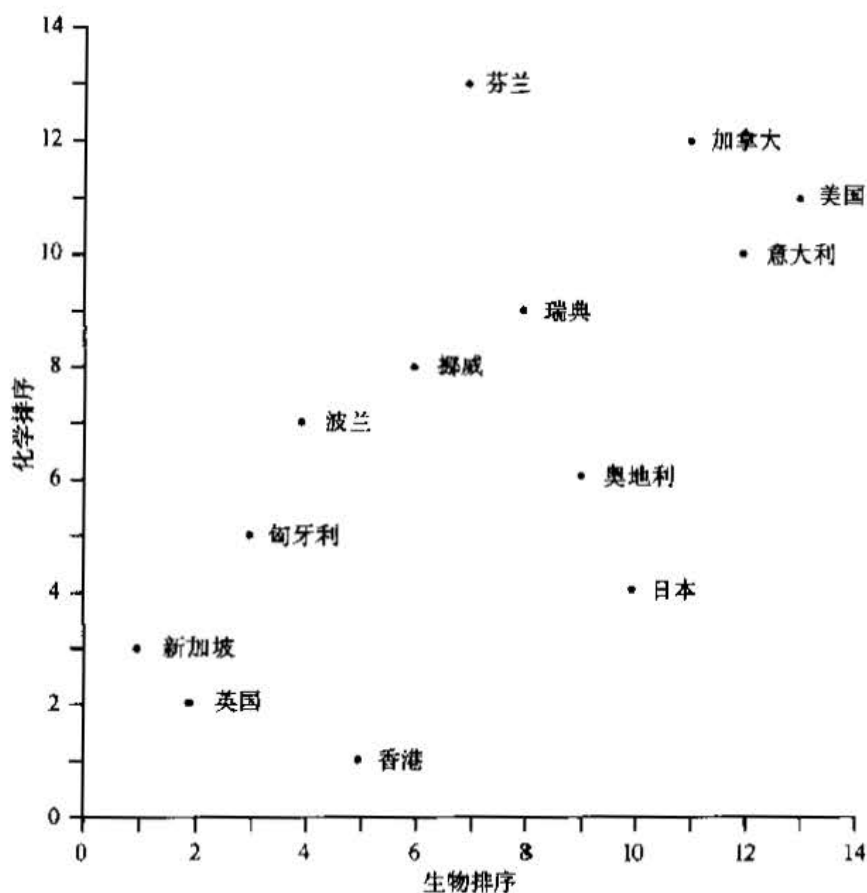


图 A.19 练习 12.31a 的散点图。

- b. 在某个学科上好的国家在其它学科上也好,而在某个学科上差的国家在其它学科上也差。香港、日本、澳大利亚、意大利和美国在化学上比生物上好。
- c. $r_s = 0.65$ 。
- d. $t = 2.82$, 自由度为 11, $p = 0.008$ 。这个关系不可能仅仅由偶然因素产生。
- e. 我们不能讨论有关因果关系。

12.33 a. 图 A.20。

- b. 从一周到下一周散点图显示的变化很小。
- c. $r_s = 0.98$ 。
- d. $t = 15.19$, 自由度为 10, 故 $p < 0.0001$ 。这不像是一个仅仅由偶然因素产生的结果。

13 多元分析

13.1 a. 用多个自变量来研究因变量。

- b. 对于一个自变量,残差变量包含其它所有变量的效应。通过把所有自变量同时分

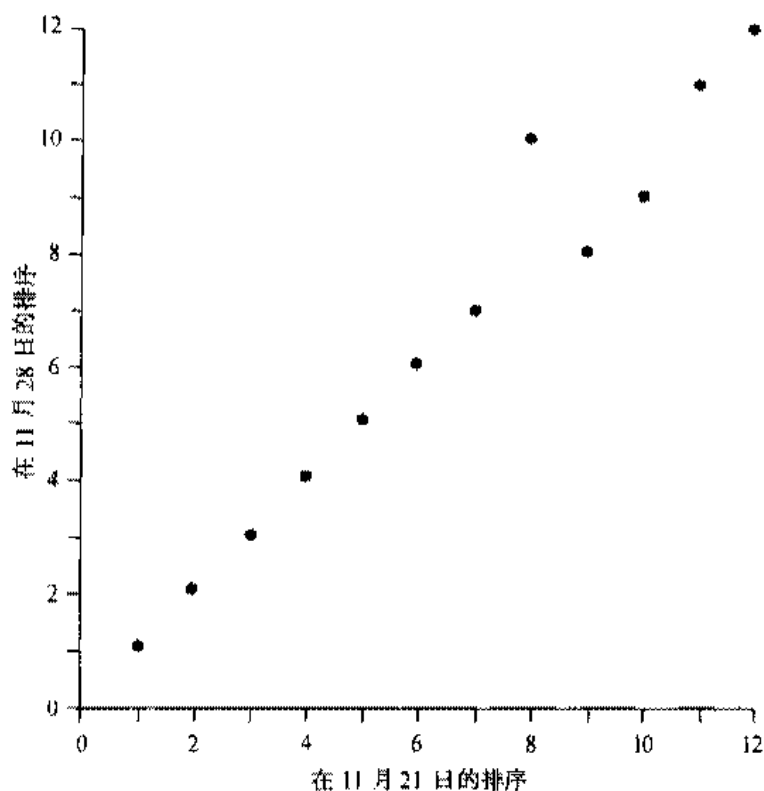


图 A.20 练习 12.33a 的散点图。

离出来,我们减少了残差变量的效应。

- 13.3 使控制变量保持常数,并通过考察由控制变量定义的各个分组中因变量与自变量之间的关系。
- 13.5 a. 我们将数据按控制变量分组,然后对每组计算 ϕ 并对所有 ϕ 取平均而得出的系数。
b. 将数据分成四组。在每组中算出 ϕ 并对所有 ϕ 取平均。
c. 在平均的过程中我们考虑了各组观测值的数目。
- 13.7 共线性。
- 13.9 a. 由具有两个类别的分类型变量产生的一个变量,一个类别的所有观测值被赋予一个数值而另一个类别的所有观测值被赋予另一个数值。
b. 投票,0 表示民主党人,1 表示共和党人。
c. 一个分类型变量可以被包含在回归分析中。
d. 这个分类型变量只能有两个类别。
- 13.11 假设检验或置信区间估计。
- 13.13 分析中用到的两个分类型变量之外的所有变量的联合效应。
- 13.15 a. 两个变量各自效应之外的联合效应。
b. 吸烟(是/否)和饮酒(是/否)作为两个自变量而期望寿命作为因变量。
- 13.17 我们永远也不知道究竟是否囊括了所有适当的控制变量。
- 13.19 a. 将数据分成民主党人和共和党人两个组并在每组中做分别的分析。
b. 对来自个别分析的两个 ϕ 取平均。

- c. 由于偏 ϕ 不等于 0, 原来的关系仍可能是因果的。
- 13.21 当控制政党时, ϕ 没有改变, 故原来的关系可能是因果的。
- 13.23 甜食种类与口味之间具有强的显著的关系。巧克力/香草变量和交互变量与口味之间具有弱的非显著的关系。
- 13.25 a. 正的。
b. 正的。
c. 大的。
d. 负的。
e. 负的。
f. 大的。
- 13.27 a. 这天是否为上学日或周末日, 因为儿童在周末玩得更多。
b. 我们可能发现在上学日之间和周末日之间没有差异。这样, 原来的关系就不是因果的。
- 13.29 回归系数的单位是每 X 的 Y 。如果 Y 是每加仑的英里数而 X 是马力, 则系数的单位是每加仑英里/马力。当 X 是重量时, 则系数的单位是每加仑英里/千磅。因为两个系数的单位不同, 他们不能被比较。
- 13.31 控制变量的选择是一个非统计的问题, 这个选择必须来自那些对所研究的对象本身有了解的人们。
- 13.33 如果拒绝零假设所有总体中的三个回归系数为 0 的想法。三个数为 0 的反面是至少有一个不为 0。
- 13.35 a. 对每种轿车, 不同性能的得分之间变差很大, 故残差变量效应较大。Mercedes 平均得分最大, VW 的均值最低。Volvo 的残差变量效应最小。
b. 车型与得分之间的关系强度是一个弱的 $R = \sqrt{140/4117} = 0.18$ 。因为 p -值大, 轿车之间没有统计意义上显著的差异。
c. 性能得分的差异不够大。
- 13.37 重量对每加仑英里数可能仍然有因果效应, 而马力与每加仑里程数之间是伪关系。
- 13.39 a. 图 A.21。
b. 练习 13.23 中的表 13.15 显示了种类之间的显著差异, 但是它不能区别哪种类型高哪种类型低。交互效应没有告诉我们冰牛奶与其它两类不同。
- 13.41 a. \$2500, 因为这是受大学教育年数的系数。
b. 两个妇女的差异将相同。
c. 对于两个其它变量都相同而受大学教育相差一年的两个人, 大学教育的年数产生了 \$2500 的差异。而对于两个其它变量都相同而服务相差一年的两个人, 产生了 \$400 的差异。受大学教育的年数从 0 到 4, 这个变量对薪水最大的贡献是对一个受了 4 年大学教育的人来说为 \$10000。对一个工作了 40 年的人来说, 服务年数这个变量对总薪水的贡献为 $\$400 \times 40 = \16000 。
- 13.43 a. 学生的课题。
b. 吸收量 = $-1.79 + 0.11$ 注入量 + 0.83 性别。
c. 分别分析时, 系数为 0.10 和 0.14。在多元分析中这些系数的加权平均值为 0.11。

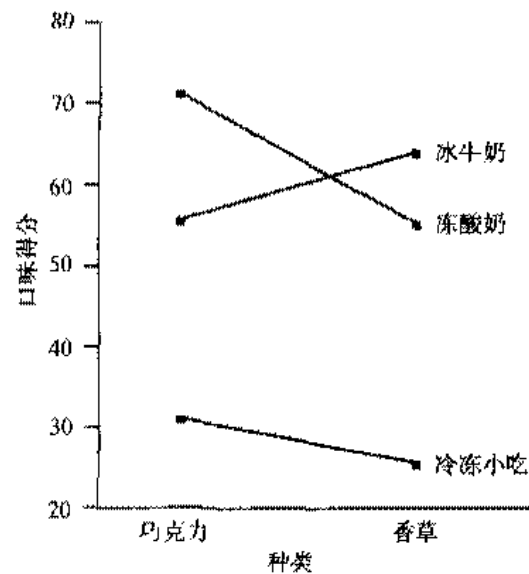


图 A.21 练习 13.39a 的图。

d. 性别 = 0 时得出

$$\text{吸收量} = -1.79 + 0.11 \text{ 注入量} + 0.83 \times 0 = -1.79 + 0.11 \text{ 注入量}$$

性别 = 1 时得出

$$\text{吸收量} = -1.79 + 0.11 \text{ 注入量} + 0.83 \times 1 = -0.96 + 0.11 \text{ 注入量}$$

e. 这两条直线是平行, 而练习 10.61 的直线是不平行的。

f. 0.83。

g. 控制注入量意味着保持注入量为常数。对于注入量的一个特定值, 雄老鼠的预计得分比雌老鼠的预计得分高 0.83。这样, 性别的效应为 0.83。

13.45 a. 表 A.19。

表 A.19 练习 13.45a 的被告和被受害者的种族

		被告的种族		合计
		黑人	白人	
受害者 种族	黑人	1438	64	1502
	白人	228	745	973
	合计	1666	809	2475

b. $\phi = 0.72$ 。

c. 控制判决时种族的偏 ϕ 系数 = 0.70。

d. 被害者种族与被告种族之间的关系可能是因果的。

e. 我们可以考察控制被告的种族时判决与被害者种族之间的关系, 及控制被害者种族时判决与被告种族之间的关系。

13.47 简单回归分析得出

$$\text{热量} = 287 + 19.8 \text{ 脂肪} \quad r^2 = 0.85$$

多元回归分析得出

$$\text{热量} = 688 + 22.2 \text{ 脂肪} - 17.2 \text{ 胆固醇} + 0.03 \text{ 钠} \quad R^2 = 0.97$$

简单分析得出这些食品稍大的回归系数。多元分析显示钠可以去掉。

14 日常生活中的统计

14.1 学生的课题。

14.3 学生的课题。

14.5 学生的课题。

14.7 学生的课题。

索引

A

Alternative hypothesis 备择假设 169, 183

Analysis of variance 方差分析 303

one-way 单因子 312, 372

two-way 双因子 370, 371

Analysis of variance table 方差分析表

one-way 单因子 312, 320

regression 回归 372

two-way 双因子 377

Average 平均

mean 均值 88 ~ 90

median 中位数 86 ~ 88

mode 众数 84 ~ 86

B

Bar graph 条形图 49

two variables 两变量 219

Beta β 271

Bias 偏差 150

Bimodal 两众数 56, 84

Binomial distribution 二项分布 121

hypothesis testing 假设检验 182

Boxplot 盒形图 52

C

Categorical sum of squares 分类平方和 307

Categorical variable 分类变量 48, 204

Causality 因果关系 205

Census 普查 21

Chartjunk 图垃圾 63

Chi-square variable χ^2 变量 11

Class size 组大小 34

Collinearity 共线性 364

Confidence interval 置信区间 152

for mean 对均值 159

for percentage 对百分比 158

for difference of two means 对两均值之差 159

for difference of two proportions 对两比例之差

159

for regression coefficient 对回归系数 271

versus hypothesis testing 与假设检验对比 180

Confidence level 置信水平 152

Constant 常数 11

Contingency table 列联表 217

Control 控制 30, 361

Control group 控制组 30

Convenience sample 方便样本 23

Correlation analysis 相关分析 251

Correlation coefficient 相关系数 257

Critical values 临界值 176

D

data 数据

experimental 实验的 29

observational 观测的 21

Data analysis 数据分析 3

Data density 数据密度 64

Data file 数据文件 35

Decision analysis 决策分析 133

Degree of freedom 自由度 126

analysis of variance 方差分析 319

chi-square χ^2 225, 226

t t - 171, 172

two-way analysis of variance 两因子方差分析

370, 371

Dependent variable 因变量 203

Distributions 分布

binomial 二项 121

chi-square χ^2 128

F F - 129

standard normal 标准正态 124

Poisson Poisson 121

t t - 126

Draft lottery 征兵抽签 8,9

Dummy variable 虚拟变量 274

in multiple regression 在多重回归中 368

E

Error type I 第一类错误 169

Error type II 第二类错误 170

Errors in sampling 抽样误差 26

Error variable 误差变量 307

Estimation 估计 148

interval 区间的 152

point 点的 149

Expected frequency 期望频率 233

Experiments 实验 30

Explained amount of variation 对变化量的解释程度
308

F

F -variable F -变量 11

F -ratio F -比 320

Fisher, Sir Ronald 罗纳德·费希尔爵士 33

Four questions 四个问题 198

G

Galton, Francis 弗兰西斯·加尔顿 252

Gamma γ 337

Graphical excellence 图优性 62

Graph 图 45

H

Hawthorne effect Hawthorne(豪森)效应 31

Histogram 直方图 54

Hypergeometric distribution 超几何分布 123

Hypothesis testing 假设检验 148

analysis of variance 方差分析 312

contingency table 列联表 224 ~ 227

correlation coefficient 相关系数 271

difference of two means 两均值之差 185

difference of two proportions 两比例之差 187

Gamma γ 337

mean 均值 185

multiple R 多重 R 365

partial regression coefficient 偏回归系数 367

proportion 比例 177

rank correlation 秩相关 341

simple regression coefficient 简单回归系数 271,
272

versus confidence interval 与置信区间对比
180

I

Independent variable 自变量 203

Inference 推断 148

Interaction sum of squares 交互作用平方和 377

Interaction variable 交互作用变量 377

Interquartile range 四分位数极差 92

Interval estimate 区间估计 152

Interval variable 区间变量 51

L

Least squares 最小平方,最小二乘 263

Linear relationship 线性关系 257

Linear correlation coefficient 线性相关系数

Literary Digest 文摘 7

Logistic regression Logistic 回归 276

M

Matched pair analysis 配对分析 315

Mean 均值 88 ~ 91

for probability distribution 对概率分布 123

Mean squares 均方 312

Median 中位数 86

Metric variable 度量变量 51

Mode 众数 84

Multiple correlation coefficient 多重相关系数 366

Multiple regression 多重回归 358

Multivariate analyses 多元分析 357,362

N

Nonresponse error 未响应误差 27

Normal distribution 正态分布 124

Null hypothesis 零假设 168,176,179,183

O

- Odds 优势 118
 Odds and probability 优势和概率 136
 One-sided test 单边检验 176

P

- p value p -值 170, 173
 one-sided and two-sided 单边和双边 171
 Paired data 配对数据 315
 Parameters 参数 11
 Partial correlation coefficient 偏相关系数 365
 Partial phi 偏 ϕ 358
 Partial regression coefficient 值回归系数 363
 Pearson correlation coefficient Pearson 相关系数 257
 Pearson, Karl 卡尔·皮尔逊 257
 Percentiles 百分位数 87
 phi ϕ
 Pie chart 圆饼图 48
 Point estimate 点估计 149
 Poisson distribution 泊松分布 121
 Population 总体 21, 148
 Prediction 预测 202
 Probability 概率 9
 finding 找寻 115
 computing with 计算 117
 Pythagoras triangle 毕达哥拉三角(勾股定理) 309

R

- r r 257
 R R
 in analysis of variance 方差分析中 308
 in multiple regression 多重回归中 367
 r^2 260
 interpretation of 解释 256
 R^2 R^2
 in analysis of variance 方差分析中 308
 in multiple regression 多重回归中 367
 Randomness 随机性 23
 Range 极差 91
 Rank correlation r_s 秩相关 r_s 340
 Rank variable 顺序变量, 秩变量 333
 Ratio variable 比例变量 51

- Regression analysis 回归分析 251
 Regression coefficient 回归系数 263
 Regression equation 回归方程 262
 Regression line 回归直线 261
 Regression sum of squares 回归平方和 267
 Relationship 关系 198
 Residual sum of squares 残差平方和
 in analysis of variance 方差分析中 307
 in regression 回归中 267
 Residual value 残差值 310
 Residual variable 残差变量 265, 310
 Response errors 响应误差 28
 Response rate 响应率 25

S

- Sample 样本 23
 convenience 方便 23
 random 随机 23
 simple random 简单随机 24
 Sampling error 抽样误差 26
 Scatterplot 散点图 57, 254
 Sign test 符号检验 316
 Significance level 显著水平 175
 Simple regression 简单回归 252
 Skewed distribution 倾斜分布, 非对称分布 55
 Slope 斜率 263
 Spearman rank correlation Spearman 秩相关 342
 Spurious relationship 伪关系 199
 Standard deviation 标准差 92
 Standard error 标准误差 96
 for mean 对均值 96
 for regression coefficient 对回归系数 281
 Standard normal distribution 标准正态分布 124
 Standard score 标准得分 97, 171
 Statistics 统计 3, 148, 198
 Stemplot 茎叶图 53
 Strength of relationship 关系强度 201
 two categorical variables 两个分类变量 204
 two metric variables 两个数量变量 204
 Subjective probability 主观概率 117
 Summary number 汇总数 86
 Surveys 抽样调查 399 ~ 401
 Symmetric distribution 对称分布 55

T

t-variable *t*-变量 11

Tables 表 45

Test of significance 显著性检验 148

Time series plot 时间序列图 58

Total sum of squares 总平方和 267

in analysis of variance 方差分析中 307

in regression 回归中 267

Two-sided test 双边检验 176

Two-way analysis of variance 双因子方差分析 370, 371

Type I error 第一类错误 169

Type II error 第二类错误 170

U

Unbiased estimate 无偏估计 150

V

V *V* 228

Value 值 10

Variable 变量 10

categorical 分类型 204

empirical 经验 11

metric 数量型 204

rank 秩, 顺序 204

theoretical 理论的 11

Variance 方差 96

from probability distribution 从概率分布计算
140

Variation 差异

interquartile range 四分位数极差 92

range 极差 91

standard deviation 标准差 92, 96

standard error 标准误差 96

variance 方差 96

Z

z-variable *z*-值 11

