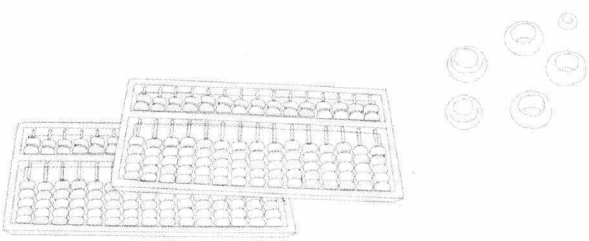
万卷方法

STATISTICS FOR PEOPLE WHO (THINK THEY) HATE STATISTICS 3rd ed

爱上统计学

中译本 第2版

(美)尼尔·J. 萨尔金德 著
史玲玲 译



重庆大学出版社

STATISTICS FOR PEOPLE WHO(THINK THEY) HATE STATISTICS, 3rd edition by Neil J. Salkind. English language edition published by Sage Publications of Thousand Oaks, London, New Delhi, Singapore and Washington D. C., © 2008 by Sage Publications, Inc.
All rights reserved, No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher. CHINESE SIMPLIFIED language edition published by CHONGQING UNIVERSITY PRESS, Copyright © 2011 by Chongqing University Press.
《爱上统计学》原书英文版由 Sage 出版公司出版。原书版权属 Sage 出版公司。
本书简体中文版专有出版权由 Sage 出版公司授予重庆大学出版社,未经出版者书面许可,不得以任何形式复制。

版贸渝核字(2010)第 16 号

图书在版编目(CIP)数据

爱上统计学/(美)萨尔金德(Salkind, N. J.)著;
史玲玲译. —2 版. —重庆:重庆大学出版社,2011.1
(万卷方法)
书名原文:Statistics for People Who(Think They)
Hate Statistics
ISBN 978-7-5624-5891-3

Ⅰ.①爱… Ⅱ.①萨…②史… Ⅲ.①统计学—研究
Ⅳ.①C8

中国版本图书馆 CIP 数据核字(2010)第 252975 号

爱上统计学
(中译本第 2 版)
萨尔金德(Salkind, N. J.) 著
史玲玲 译
策划编辑:雷少波 林佳木
责任编辑:林佳木 版式设计:林佳木
责任校对:秦巴达 责任印制:赵 晟
*
重庆大学出版社出版发行
出版人:邓晓益
社址:重庆市沙坪坝正街 174 号重庆大学(A 区)内
邮编:400030
电话:(023) 65102378 65105781
传真:(023) 65103686 65105565
网址:http://www.cqup.com.cn
邮箱:fxk@cqup.com.cn (营销中心)
全国新华书店经销
自贡新华印刷厂印刷
*
开本:787×1092 1/16 印张:17.5 字数:335千 插页:16 开 2 页
2011 年 1 月第 2 版 2011 年 1 月第 5 次印刷
印数:6 091—11 090
ISBN 978-7-5624-5891-3 定价:39.80 元

本书如有印刷、装订等质量问题,本社负责调换
版权所有,请勿擅自翻印和用本书
制作各类出版物及配套用书,违者必究

万卷方法学术委员会

学术顾问

黄希庭　西南大学心理学院教授

沈崇麟　中国社会科学院社会学所研究员

柯惠新　中国传媒大学教授

劳凯声　北京师范大学教育学院教授

张国良　上海交通大学媒体与设计学院教授

学术委员(以下按姓氏拼音排序)

陈向明　北京大学教育学院教授

范伟达　复旦大学社会学系教授

风笑天　南京大学社会学系教授

高丙中　北京大学社会学人类学研究所教授

郭志刚　北京大学社会学系教授

蓝　石　美国 DeVry 大学教授

廖福挺　美国伊利诺大学社会学系教授

刘　军　哈尔滨工程大学社会学系教授

刘　欣　复旦大学社会学系教授

马　骏　中山大学政治与公共事务学院教授

仇立平　上海大学社会学系教授

邱泽奇　北京大学社会学系教授

孙振东　西南大学教育学院副教授

王天夫　清华大学社会学系副教授

苏彦捷　北京大学心理学系教授

夏传玲　中国社会科学院社会学所研究员

熊秉纯　加拿大多伦多大学女性研究中心研究员

张小劲　中国人民大学国际关系学院教授

孙小山　华中科技大学社会学系副教授

修养让他们采纳了我的一个小想法(那是在第一版出版以前),使其变成大家现在正在阅读的这本书,并且很成功。

但是这里必须特别感谢一些人的关注和辛勤工作。莉萨·奎瓦斯·肖(Lisa Cuevas Shaw)是每个作者梦想中的编辑——她是作者和他的工作的支持者,各种好的、坏的想法的聆听者,而且始终能够提供反馈和支持。她引导这项工作的两个版本(我的另一本题为"for people"的书)达到现在可见的成功。副主编马戈·克鲁彭(Margo Crouppen)是另一个每天都联络的人,她能看到每天需要完成的重要的工作,而且使得工作保持正规,直到出版。我欠她人情。其他需要特别提到的人是编辑助理卡伦·格林(Karen Greene)、市场经理斯特凡妮·亚当斯(Stephanie Adams)和制作编辑韦罗妮卡·斯特普尔顿(Veronica Stapleton)。特别感谢里安·列赫(Liann Lech),她有锐利的眼睛和很好的平面排版技术,使得这本书以现在这种好读的形式呈现。

Sage 出版社感谢以下书评人的洞见和建议:

Paul L. Jamison
印地安那大学伯明顿分校
Mark E. Correia
圣何塞州立大学
Jiunn—Jye (JJ) Sheu
佛罗里达大学健康教育和行为系
Autumn R. Benner
明尼苏达州立大学曼卡多分校
Thom Dunn
北科罗拉多大学
Alan Bougere, PhD, LCSW
南密西西比大学
Lisa S. Patchner, MSW, DrPH
波尔州立大学
Ron Oliver
加州州立大学富尔顿分校
Joseph P. Schwieterman
德保罗大学
Robert W. Quesal
西伊利诺伊大学

现在,关于第 3 版……

以上的内容说明了我写这本书的最初想法,但是关于新的版本介绍很少。

任何一本书都需要不断修订,《爱上统计学》也不例外。过去的六年许多人曾告诉我这本书是多么的有帮助,但是其他人也告诉我他们多么想让这本书做出修订以及为什么。在修订的过程中我尽力满足这两部分人的需求。书中部分内容仍保留,部分做出了修订。

总是有新的事物值得关注,同时也需要用不同的方式重新理解老的主题和观念。以下的清单是《爱上统计学》中新的内容。

- 在每一章的最后增加了新的练习题。不是简单的数量增加,每一个题在应用水平和兴趣(我希望大家有)方面都有很大的变化。这些练习题使用的数据集可以在 http://www.sagepub.com/salkindstudy 下载,也可以在作者主页 http://soe.ku.edu/faculty/salkind/stats_fpwhs3e 下载。
- 这些数据集仍然以两种形式存在——SPSS(流行的统计分析软件)和 Excel(许多人使用电子表格分析数据)。可以在本书附录 C 获得这些数据,也可以在 http://www.sagepub.com/salkindstudy 和作者主页 http://soe.ku.edu/faculty/salkind/stats_fpwhs3e 获得在线数据。
- "练习时间"中问题的答案放在本书的最后,不再放在每章的最后。这样更容易让问题集中排列在一起(而且学生偷看答案会困难一点)。
- 信度和效度的内容移到本书的第一部分,而不再放在后面的部分。这样做是接受了本书读者的一些建议——这部分的内容越早出现越容易理解。这是本书资料组织方面唯一重要的改变。
- 第 3 版使用 SPSS 提供的较新的版本——SPSS 15.0。通常情况下,使用 SPSS 11.0 甚至更早的版本可以完成大多数练习题,也可以打开最新版本建立的数据文件。
- 第 3 版中出现的所有错误都是我的责任,我向被这些错误困扰的老师和学生道歉。我非常感谢所有那些指正错误并使得第 3 版更好的信件、电话和电子邮件。在这一版,我们每个人都尽力修改之前的错误,并希望工作做得更好。期望能收到大家的建议、批评和意见。祝大家好运。

尼尔·J.萨尔金德
堪萨斯大学
njs@ku.edu

此为试读,需要完整PDF请访问: www.ertongbook.com

其他用图表显示数据的绝妙方法 …… 42
使用计算机图示数据 …… 44
5 冰淇淋和犯罪——计算相关系数 …… 51
相关系数到底是什么 …… 51
需要记忆的内容 …… 52
计算简单相关系数 …… 53
理解相关系数的含义 …… 58
决定性的努力:相关系数平方 …… 59
其他重要的相关系数 …… 61
使用计算机计算相关系数 …… 61
6 这就是真相——理解信度和效度 …… 67
信度和效度介绍 …… 67
关于测量尺度 …… 68
信度——再做一次直到得到正确的值 …… 70
使用计算机计算克隆巴赫系数 …… 75
效度——哦！真相是什么 …… 77
信度和效度:很亲密的堂表兄弟关系 …… 80

第Ⅲ部分 抓住那些有趣又有利的机会

7 你和假设:检验你的问题 …… 82
也许你想成为一个科学家 …… 82
零假设 …… 83
研究假设 …… 84
好假设的标准是什么 …… 87
8 你的曲线是正态的吗——概率和概率的重要性 …… 90
为什么学习概率 …… 90
正态曲线(或钟型曲线) …… 90
我们最中意的标准值:z 值 …… 94
使用计算机计算 z 值 …… 100

第Ⅳ部分 显著性差异——使用推论统计

9 显著性的显著——对你我来说意味着什么 …… 104
显著性的概念 …… 104
显著性与意义 …… 108
推论统计介绍 …… 109
显著性检验介绍 …… 112
10 两个群体的 t 检验——不同群体的均值检验 …… 115
独立样本 t 检验介绍 …… 115
计算检验统计量 …… 116
特殊效果:差异是真实的吗 …… 120
使用计算机进行 t 检验 …… 122

第Ⅴ部分 你得了解和记忆的内容

19 10个(或更多)最好的统计网址 …… 198
成堆的资源 …… 198
大量的计算器 …… 198
谁是谁以及发生了什么 …… 199
都在这里 …… 199
超级统计(HyperStat) …… 199
数据,你想要数据 …… 200
越来越多的资源 …… 201
容易,但是有趣 …… 201
在斯德哥尔摩学习统计学如何 …… 201
在线统计学教学资料 …… 201
越来越多的资料 …… 201
20 收集数据的10个原则 …… 203
附录A 30分钟SPSS教学 …… 205
附录B 数据表 …… 222
附录C 数据集 …… 232
附录D 练习题参考答案 …… 243
词汇表 …… 261

第Ⅰ部分

耶！我喜欢统计学

Yippee! I'm in Statistics

影响。

统计学——整理和分析资料并使得资料更容易理解的科学——使得研究任务可行。

通过这样的研究所得出的任何结论都是有用的，原因是我们使用统计学使得这些结果有意义。这也正是本书的目标——让你理解这些基本的工具以及这些工具的用途，当然也包括如何使用这些工具。

在《爱上统计学》的第 I 部分介绍统计学学习的内容，以及为什么值得花费精力来掌握这个领域关键的基本术语和思路。这都是学习本书其他部分的准备。

果用到计数，人们就会知道太阳在一季升起多少次，度过整个冬天需要多少食物，以及一个人拥有多少资源。

这只是开始。一旦数字成为语言的组成部分，似乎接下来的步骤就是将数字和结果结合在一起。在 17 世纪早期就开始收集与人口相关的数据。以此为基础，科学家（大多数是数学家，此外是物理学家和生物学家）需要发展特定的工具来回答特定的问题。例如，弗朗西斯·高尔顿（Francis Galton）（顺便说一下，他是达尔文的表兄弟，生卒年代是 1822—1911 年），他对人类智力的性质非常感兴趣。为回答家庭成员智力一致性的基本问题，他使用了特定的统计工具——相关系数（首先是数学家发展的），接着他把相关系数广泛地应用于行为科学和社会科学。你会在第五章对这个工具有全面的了解。

实际上，你将学到的大多数基本统计过程最初应用于农业、天文学甚至政治学领域。在人类行为领域的应用则比较晚。

在过去的 100 年中，在发明新方法应用旧观念方面取得了极大的进步。最简单的用于比较两个群体的均值差异的检验方法在 20 世纪初取得首次进展。在此基础上建立的技术十年之后才提出，而且得到了极大的完善。随着个人电脑和类似 SPSS 的软件（见附录 A）的应用，任何人想研究这些有意义问题都可以使用复杂的统计技术。

影响力深远的个人电脑的应用有利有弊。有利的方面是大多数统计分析不再需要使用巨大、昂贵的中央处理机。而价值不超过 1 000 美元的个人电脑就能满足 95% 的人 95% 的需求。另一方面，很少有大学生（如已经通过这门课的你的同学）会使用他们已有的旧数据，并且认为应用复杂的 SPSS 软件分析这些数据可以得到可靠、可信和有意义的结果——这是不正确的。你的老师可能会说“垃圾进，垃圾出”——如果你开始就没有使用可靠、可信的数据，那么分析这样的数据所得到的结果就既不可信也不可靠。

现在，不同领域——从司法、地球物理学到心理学——的统计分析人员发现他们基本上使用相同的技术来回答不同的问题。当然在资料收集方面有重要的差别，但是大体上来说，随着数据（大量的数据）收集所进行的分析（多个的分析），即使说有所谓的差别也是非常类似的。那么伦理问题呢？这个课上会给你提供一定的工具来理解统计学如何应用于几乎所有的学科，帮你轻松搞定那些 3 ~ 4 个学分的课程。

如果你想更多地了解统计学历史，而且想按历史发展来了解，可以从一个很好的地方开始，就是圣安赛姆学院的网络，网址是 http://www.anselm.edu/hompage/jpitocch/biostatshist.html；或者 http://www.stat.ucla.edu/history/（加利福尼亚大学洛杉矶分校）。

统计学：是什么（或不是什么）

《爱上统计学》是一本关于基础统计学，以及在不同的情况下如何应用统计学分析和理解数据的书。

就一般意义而言，统计学是描述一系列可用于描述、整理和解释资料

本工具。

5. 如果你计划获得教育学、人类学、经济学、护理学、社会学或其他社会、行为或生物科学领域的任何一个学科的硕士学位，统计学课程是你前进的基础。

6. 最后，你可以夸口说你完成了人人都认为是相当于建立和运行核反应堆的高难度课程。

使用这本书的十种方式（同时也在学统计学咯！）

耶，好像手里的统计学书总不好用，但是这一本是不同的。这本书是针对学生写的，但不是降低标准，而是信息充分，也尽可能展现基础内容。本书也没有假定在课程开始之前应该具备什么知识，只是进程安排较缓、步骤较小，可以让学生按自己的节奏安排。

大家都认为统计学是很难掌握的课程。我们也这么认为，因为统计学的一部分的确充满挑战。另一方面，无数的学生已经掌握了这门课程，你也可以。在开始我们的第一个主题之前，先看一下导论这一章的要点。

1. 你不笨。这是真的。如果你是笨蛋，你不可能在学业方面走这么远。因此，对待统计学就像对待其他新的课程那样吧。听讲座，学习基本内容，做书上的练习或课堂练习，那么你就会学得很好。火箭研究者精通统计学，不过你不需要像火箭研究者那样研究统计学。

2. 你怎么知道统计学很难学？统计学很难学？既是也不是。如果你是从上过这门课的朋友那儿听来的，而他们没有努力学习也没有学好，那么肯定愿意告诉你统计学是多么难学，甚至会说，统计学即使不是对整个生活的灾难，对整个学期来说也是灾难。不过我们不要忘记——我们总是倾听抱怨者的抱怨。因此，我们建议你应该以这种态度开始这门课程，也就是，等着看统计学是否难学，并且依据自己的经验做出判断。更好的选择是找几个上过这门课的人讨论一下，获得他们的总的看法。不要依据让人泄气的人的经验做出判断。

3. 不要逃课——按顺序学习各章的内容。《爱上统计学》的每一章都是下一章的基础。我希望在课堂上学习了所有内容后，你就可以回顾整本书，并把这本书作为参考书。如果你需要确定表中的特定值，可以查阅附录 B。或者，你需要记起如何计算标准差，可以回顾第三章的内容。但是现在要按照本书的顺序学习每一章。当然也可以不按照顺序，先了解后面的学习内容。但是在学习后面章节之前要掌握前面的章节。

4. 形成学习小组。这是确保通过这门课程的最基本的方式之一。在一个学期的开始，要和朋友商定一起学习。如果没有朋友选择同一课程，那么就要结交新朋友，或者邀请和你一样看起来很高兴学习统计学的学生一起学习。如果你学得比他人好，一起学习可以帮助他人，反之，他人学得比你好你就可以从他人那里受益。每个星期安排特定的时间聚在一起一个小时，复习每一章后的练习题，或者相互提问。或者依据需要安排更多时间。与他人一起学习是帮助你理解和掌握课程内容的有用的方式。

5. 向老师提问或者向朋友提问。如果你不了解课堂上所讲授的内容，

此为试读，需要完整PDF请访问：www.ertongbook.com

浏览《爱上统计学》全书你会发现许多小梯子符号，就像在这儿看到的。这表明这里有许多步骤，沿着这些步骤可以指导你通过特定的过程。这些步骤已通过检验，也被任何一个证明这些步骤的机构证明。

绑着蝴蝶结的手指是很可爱的符号，但它的主要目的是用于强调你所读到主题的重点。在学习过程中，强调这些重点是由于这些重点通常是这个主题的关键。

《爱上统计学》的大多数章节提供了一个或多个特殊的统计过程和同步计算的详细信息。计算机符号用于表明每章“使用计算机……”的部分。

本书的许多章节都包含如何使用 SPSS 15.0 版本来完成同样的统计过程的指导，因此你可以手动计算，也可以应用现在的任何一款非常有效的统计分析软件包来计算。

附录 A 是对 SPSS 的介绍。若准备使用 SPSS，通览这个附录是现在要完成的任务。如果你有更早的 SPSS 版本（或者是 Max 版本），这些资料仍然非常有用。实际上最近的 SPSS 视窗版本和 Max 版本在形式和功能方面几乎一致。

可以使用附录 C 的数据集来完成本书练习题。在每一章的“使用计算机……”部分，都会发现数据集的索引（例如“第二章数据集 1”）。每一个数据集都可在附录 C 中找到，如果你按部就班地学习本课程就可以使用这些数据来成功地完成“使用计算机……”部分。可以手动建立数据库，也可以从 Sage 出版社的网址 http://www.sagepub.com/salkindstudy 或者作者的网址 http://www.soe.ku.edu/faculty/salkind/stats_fpwhs3e/下载。数据文件以 SPSS 和 Excel 两种形式存在。

难度指数

1 非常难	☺
2 比较难	☺☺
3 一般	☺☺☺
4 比较容易	☺☺☺☺
5 非常容易	☺☺☺☺☺

词汇

正文中一些重要的术语都汇总在本书最后的词汇表中进行了简单的解释。

- 最后，n 是从中计算均值的样本的规模。

下面是计算均值的步骤：
1—以一列或多列的形式列出所有数值。这些数值就是那些 X。
2—计算所有数值的总和或总计。
3—总和或总计除以数值的个数。

例如，你需要计算三个不同商店的消费者的数量的平均数，你可以计算均值得到这个值。

场　所	年顾客数量
兰哈姆公园商店	2 150
威廉斯堡商店	1 534
下城商店	3 564

每个商店顾客数量的平均数或均值是 2 416。公式 2.2 表明如何应用公式 2.1 计算这个值。

$$\bar{X} = \frac{\sum X}{n} = \frac{2\ 150 + 1\ 534 + 3\ 564}{3} = 2\ 416 \qquad (2.2)$$

瞧，我已经告诉你很容易了。小菜一碟！

需要记忆的内容

均值有时也用字母 M 表示，也叫做典型平均数或中心值。如果你在看其他的统计学读物或者研究报告，会看到类似 $M = 45.87$ 这样的表达式，这就可能意味着均值等于 45.87。

- 在上面的公式中小写字母 n 表示从中计算均值的样本的规模。大写字母 N（像这样的）表示总体规模。在一些书中和一些杂志的文章中没有对这两者做出区别。
- 样本均值是非常准确地反映总体均值的集中趋势量数。
- 均值就像跷跷板的支点。均值就是中心点，也就是均值一边的所有数值等于均值另一边的所有数值。
- 最后，不论好坏，均值对极值很敏感。极值会使得均值向一方或另一方倾斜，也使得均值对数据组的代表性减弱，同时作为集中趋势量数的有效性减弱。当然这依赖于计算均值的数值。后面会有更多的讨论。

均值也是指**算术平均数**（arithmetic mean），当然你也会看到其他类型的均值，如调和均值。调和均值用于特殊的情况，你现在不需要关注。如果你想表现得更专业，就需要知道算术平均数（也就是我们讨论到现在的均值）定义的基点是均值的偏差的总和是 0

25 567 美元

13 234 美元

那么中位数和中间的两个数相同。在这个案例中中位数是 25 567 美元。

我们有一组数据——7 个病人运动伤害的康复天数。数据具体如下：

43

34

32

12

51

6

27

和之前所做的一样，我们将这些数排序（51，43，34，27，12，6），接着选择中间的数作为中位数，在这组数据中是 32。因此，康复天数的中位数就是 32。

TECH TALK

如果你了解中位数，也应该知道百分位点（percentile points）。百分位点用于定义数据集或数据分布中等于或者小于一个特定数据值的个体的百分数。例如，你的成绩是“处在 75 百分位点”，这意味着成绩分布中你的成绩刚好是或者超过 75% 的其他人的成绩。大家也知道中位数是 50 百分位点，因为数据分布中 50% 的个体在这一点之下。其他的百分位点也很有用，如 25 百分位点，通常用 Q_1 表述，75 百分位点，通常用 Q_3 表示。那么 Q_2 呢？当然是中位数。

现在给出一些问题的答案，这些问题可能在开始讨论中位数时就出现在你的头脑中了。为什么使用中位数而不是均值？一个非常好的原因是中位数对极值不敏感而均值却不是。

如果你的数据集中有一个或多个极值，中位数相对其他集中趋势量数来说能更好地代表数据集的中心值。是的，甚至比均值更好。

我们所说的极值是什么？非常容易就可以想到，一个极值就是与其所属数据组非常不同的值。例如，考虑之前的案例中已经用过的收入清单，在这里再次列出：

135 456 美元

54 365 美元

37 668 美元

32 456 美元

25 500 美元

这组数据中数值 135 456 美元与其他 4 个数之间的差异较大。我们可以认为这个数值就是极值。

最好的说明中位数作为集中趋势量数是多么有用的方法是计算包含

一个或多个极值的同一数据集的均值和中位数，然后比较哪一个值能更好地代表数据集。现在就来计算和比较。

上面有 5 个收入数值的数据组的均值是 5 个数值的总和除以 5，结果是 57 089 美元。而这个数据组的中位数是37 668美元。哪一个值能更好地代表这个数据组？数值 37 668 美元更加明显地位于数据组的中间，但是我们习惯于认为平均数具有代表性或者占据中间位置。事实上，均值 57 089 美元在第4 高位数值（54 365 美元）之上，不是这个数据分布的中间值，也不具有代表性。

就是由于这个原因，特定的社会和经济指标（大多数与收入相关）的分析使用中位数作为集中趋势量数，例如“美国家庭平均收入的中位数是……，”而不是使用均值来概括收入。总是存在太多的极值改变或者明显地扭曲一个数据组或者数据分布的中心点。

你已经知道均值有时用大写字母 M 表示而不是 $\overline{X}$，中位数也有表示符号。我喜欢使用字母 M，但是一些人会将 M 和均值混淆，因此他们使用 *Med* 或 *Mdn* 表示中位数。不要让这些拌住你的脚——只要记住什么是中位数以及中位数代表什么，你就不会存在适应这些不同符号的困难。

需要记忆的内容

这些有关中位数的内容十分重要也很有趣，需要记忆。

- 均值是一系列数值的中间点，而中位数是一系列个体的中间点。
- 因为中位数关注的是有多少个体而不是这些个体的数值，极值（有时也叫做奇异值）就不会产生影响。

计算众数

我们要学习的第三个也是最后一个集中趋势量数是众数，也是最笼统、最不精确的集中趋势量数，但是在理解特定的数据集的特征中扮演着非常重要的角色。众数（mode）就是出现次数最多的数值。没有计算众数的公式。

按照下面的步骤计算众数：

1—列出一个数据分布中的所有数值，但是每一个数值只列出一次。

2—计算每个数值出现的次数。

3—出现次数最多的数值就是众数。

例如，调查 300 个人的政党背景会形成如下的数据分布结果。

政党背景	次数或频数
民主党	90
共和党	70
无党派人士	140

众数是出现次数最多的数值,在上面的例子中是无党派人士。这就是数据分布的众数。

如果我们看到的是100个多项选择题的众数反应的答案,你会发现相对其他选项人们更多的选择选项A。数据形式可能如下。

选项	A	B	C	D
次数	57	20	12	11

这项包含100个多项选择题的测试中每个题有四个备选答案(A,B,C和D),A被选择57次。这就是众数反应。

你知道计算众数时最容易、最经常出现的错误是什么吗?就是选择某个分类选项出现的次数而不是分类选项的标签本身。对一些人来说很容易就可以得出众数是140而不是无党派人士,为什么?因为他们看到的是数值出现的次数,而不是最经常出现的那个数值!这是一个稍不留神就会犯下的错误,因此让你计算众数时一定要注意。

双峰分布的苹果派

如果数据分布中每一个数值的出现次数都相同,那么就没有众数。但是如果不止一个数值的出现频数相同,那么这个数据分布是多峰分布。数据集可能是双峰分布(有两个众数)的,如下面头发颜色构成的数据组所示。

头发颜色	次数或频数
红色	7
金色	12
黑色	45
棕色	45

在上面的案例中,数据是双峰分布,因为黑色头发和棕色头发出现的频数相同。如果众数相当接近但不是完全相同,数据分布也是双峰分布,如45个人的头发是黑色,44个人的头发是棕色。问题就成为一种类别在多大的程度上与其他类别相区分?你的数据是否是三峰分布的?那就是说出现了三个频数相同的数值。一般来说是不可能的,特别是在处理大的数据集过程中,但出现的可能性是存在的。

何时用什么

好吧,我们已经定义了三种不同的集中趋势量数,而且每一种都给出了简单明白的案例。但是还有一个最重要的问题没有回答,即"何时使用哪一种?"

此为试读,需要完整PDF请访问: www.ertongbook.com