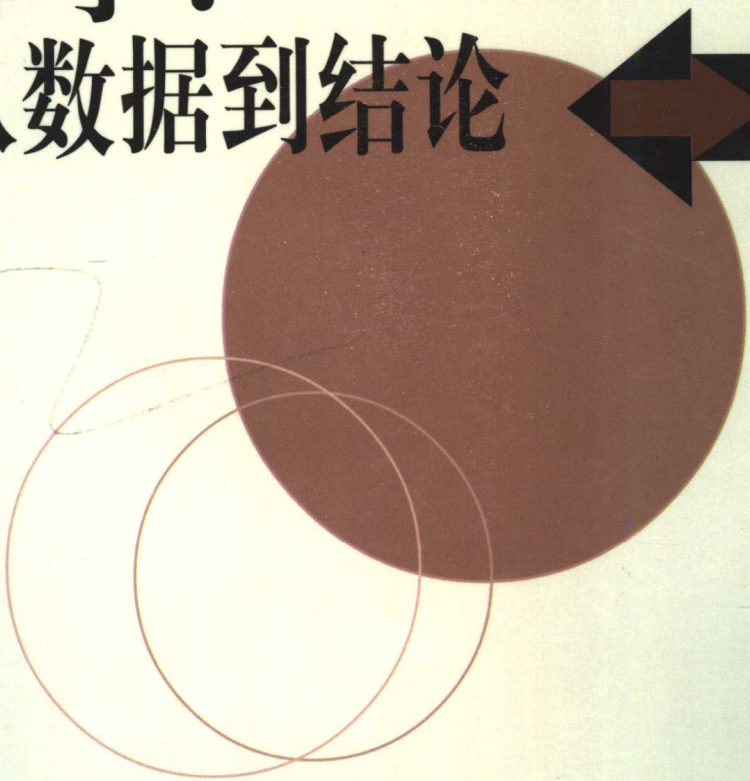
全国统计教材编审委员会

“十五”规划教材

STATISTICS: FROM DATA TO CONCLUSIONS

吴喜之·编著

统计学：从数据到结论



中国统计出版社
China Statistics Press

全国统计教材编审委员会
“十五”规划教材

统计学：从数据到结论

Statistics: From Data to Conclusions

吴喜之 编著

中国统计出版社
China Statistics Press

(京)新登字041号

图书在版编目(CIP)数据

统计学:从数据到结论/吴喜之 编著.
-北京:中国统计出版社,2004.6
全国统计教材编审委员会"十五"规划教材
ISBN 7-5037-4383-2

Ⅰ.统…
Ⅱ.吴…
Ⅲ.统计学-高等学校-教材
Ⅳ.C8

中国版本图书馆 CIP 数据核字(2004)第045963号

统计学:从数据到结论

作　　者/吴喜之
责任编辑/杨映霜
装帧设计/艺编广告
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街75号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲6号
电　　话/(010)63459084、63266600-22500(发行部)
印　　刷/科伦克三莱印务(北京)有限公司
经　　销/新华书店
开　　本/787×1092mm　1/18
字　　数280千字
印　　张/16.625
印　　数/1—6000册
版　　别/2004年8月第1版
版　　次/2004年8月北京第1次印刷
书　　号/ISBN 7-5037-4383-2/C·2048
定　　价/35.00元

本书附配套数据盘CD-ROM一张。

出版说明

“十五”期间是我国加大教育改革力度，全面推进素质教育，教育体制、教育方法发生重大变革的时期。教材建设必须紧跟教育改革的步伐，建设适应社会主义市场经济和现代化建设需要的高质量教材。为了适应这种新形势的需要，全国统计教材编审委员会制定了《2001—2005年全国统计教材建设规划》（以下简称《规划》），并根据《规划》的要求，主要采取招标的方式组织全国有关院校的专家、学者编写了这批统计学“十五”规划教材。

这批教材力求以第三次全国教育工作会议作出的《中共中央、国务院关于深化改革全面推进素质教育的决定》为指导思想，在充分总结“九五”期间统计教材建设经验的基础上，认真贯彻大胆探索和创新的原则，努力使统计教材具有前瞻性和实用性。选题中不仅包含了一些国内统计研究和教材建设上的空白领域，也包含了统计研究的最新成果。为了配合教师教学、方便地使用这批教材，我们还特地编制了专供教师使用的电子课件，这些课件将在国家统计局统计教育中心网站（http://edu.stats.gov.cn）上挂出，以供需要的教师下载。另外，对于部分教材，我们还将编辑出版与之相配套的习题集，以方便教师和学生在教学中使用，也使这批教材在编辑出版上形成一个比较完整的体系。我们相信，这批教材的出版和发行，对于推动我国统计教育改革，加快我国统计教材体系和教材内容更新、改造的步伐，都将起到积极的促进作用，同时，对我国统计教材建设也将起到较好的示范与导向作用。

限于水平和经验，这批教材的编审、出版工作还会有缺点和不足，诚恳欢迎教材的使用单位、广大教师和同学们提出批评和建议。

全国统计教材编审委员会

2004年1月

前 言

有什么在本书中等待着你们去发现，去探讨，去欣赏呢？当然不是数学公式和定理定义的堆砌，也不是和枯燥的公文报表相关的政府工作的培训。等待着你们的，是一门充满了哲学韵味的认识世界的学问。

不知读者们是否意识到，统计已经渗入到人们的社会、生活、工作等各个领域。每天新闻媒介报道的各个方面都离不开各种统计数据和各种分析与预测。人们可能对于这些统计内容已经习以为常，也可能会有一些好奇或神秘感。由于国情不同，统计的地位与人们对统计的看法也不同。在发达国家，一般民众觉得统计学和数学类似，是一门高不可攀但极易找到满意工作的学问。在中国，又有一些人认为统计就是处理政府报表的职业。但自从中国向世界开放之后，越来越明确的一点是，没有什么学科或领域能够真正离开统计。

以应用为目标学习统计，究竟是为什么？是为了流利地背诵一大堆定义、概念和抽象的名词和术语吗？是为了学习如何进行推导和证明一些复杂的定理和公式吗？这些问题不仅学生会思考，更重要的是统计教师要思考。本书的目的是希望读者在学习之后，能够知道实际中哪些是统计问题，最好能够自己解决一部分统计问题，即使不能解决也知道能够在哪里查到答案和向谁请教。知识固然重要，更重要的是通过学习获得解决和处理问题的能力。

学习并不总是一个令人生畏或至少成为某种负担的过程。人们学会走路、说话、骑车、下棋、打球等大都是在一种乐趣中进行的。为什么涉及到日常生活的每一个方面的统计就不能和看侦探小说那么引人入胜呢？其实任何一门科学，都有其趣味性；而只有把科学研究当成游戏的人才会真正成为大师。本书并不想使读者都成为统计学家，而仅仅想让读者如同学会使用电脑、手机，学会辩论、上网或讨价还

价那样愉快地认识或理解在人生中无法躲开的统计。

本书由浅入深地把统计最基本和最有用的部分在这么一本不厚的教科书中完整地介绍给读者；而且让读者可以边学习，边着手用统计软件处理数据。篇幅大、语言啰嗦的教材对读者是个负担，不但浪费了资源，也抓不住要领。因此，作者力图惜墨如金，既节省篇幅，又要把该解释的全部说清。希望读者慢慢咀嚼，不必图快。

很少有一本统计教材包括像本书那么多的统计内容。我觉得，这些内容本来并不深奥，只是其貌似复杂的数学工具把它搞成阳春白雪，再加上强调数学推导的教学方式，使得统计显得高不可攀。本书要还这些统计应用以其本来面目。使得统计变成人人都能够基本上理解和掌握的有用工具。多数使用计算机的人都不是计算机专业毕业的，多数开汽车的都不会修汽车，但这对他们的使用毫无妨碍。难道不会推导或背诵与统计有关的数学公式就不能应用统计这个工具了吗？

本书每一章的主要部分是用日常语言来引进和解释一些概念，如果可能就通过例子来说明。如果不涉及应用，这部分就足够了。涉及应用的各章后面的小结中，有一部分是说明如何通过统计软件来处理本章的数值例子；这会给多数想要自己动手分析数据的读者以方便。小结的最后还展示了与概念及计算有关的一些数学公式，使那些精力充沛的读者能更深刻地理解内容。这种安排使得本书能够适用于各种不同水平、不同要求的读者群体。本书不仅可供没有学过概率论和数理统计的非统计专业的本科生和研究生使用，也可以供统计专业的本科生作为理解统计本来含义的教材使用（以代替不能满足需要的“描述统计学”等类课程）；它还可以为各领域的广大实际工作者作为应用各种统计方法的参考书。为读者可以使用各种软件来进行分析，本书所涉及的所有电子版数据都有文本格式、SPSS格式及部分EXCEL格式。

在软件方面，本书主要使用SPSS和部分地应用Excel。我们不可能介绍所有软件，也不可能介绍某个软件的所有细节。经验证明，只要把某个方法在某个软件的基本选项指出，学生就可以通过自己的经验（最多借助于帮助）来得到所需要的结果。在课本中罗列使用软件时的出现的各种对话(选项)框的做法对本课程完全没有必要。

在前计算机时代，几乎所有的统计教科书都给出了各种与分布有

关的表格。但随着计算机的普及，所有统计软件（无论是商业的还是免费的）都给出了和各种分布有关的各种函数，把人们从繁琐而又不精确的查表中解放出来。目前很多国外的统计教科书都不再提供既占用篇幅又比较粗糙的分布表。本书不准备提供任何和分布有关的表格。本书第四章会介绍如何使用软件来进行与概率分布有关的计算。

本书的绝大部分内容曾作为非统计专业硕士和博士的课程分别在北京大学光华管理学院及中国人民大学讲授过，受到普遍欢迎。实践证明，本书前16章的内容完全能够轻轻松松地在一个学期（每周三个学时）中全部讲完。一些热心而又好奇的非统计背景的人士也曾读过本教材的全部内容，没有任何理解上的问题。当然，根据不同的教学对象和需要，有些章节可以完全不讲或少讲。

本书前面的章节，是对统计基本概念的介绍。而后面的部分则是更有针对性的一些统计模型和方法。一般传统统计学的课程包括前6章，或最多前9章的内容；而第十章到第十四章一般属于多元统计分析的课程内容；第十五章一般属于时间序列课程包含的内容；第十六章一般属于非参数统计课程的内容；第十七章介绍了生存分析；第十八章对指数进行了必要的介绍。目前大多数流行的统计应用都已包含在本书内。

本书的编写是在国家统计局教育中心的建议和鼓励下产生，并得到其大力支持。本书还受到北京大学、中国人民大学以及各兄弟院校老师和学生的鼓励和帮助。中国统计出版社一直关心着本书的写作和出版。SPSS北京办事处的专家也一直积极对写作过程中出现的有关计算问题予以帮助。特别要指出的是敬爱的汪仁官老师又一次为我所写的统计教材进行了非常认真的审校，使我重新感受到做学生的幸福；中国统计界的老前辈茆诗松老师也热心地对本书提出了许多宝贵而又中肯的建议。他们的审校和建议使本书避免了许多错误和不妥之处。没有这些支持和帮助，本书是不可能面世的。谨在此对所有各方面表示衷心的感谢。

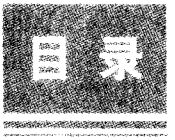
吴喜之

2004年6月

目录

目录

目录



目录

目录

第一章

一些基本概念

§1.1 统计是什么?

你想过下面的问题吗?

1. 当你买了一台电视时,被告知三年内可以免费保修。那么,厂家凭什么这样说?说多了,厂家会损失;说少了,会失去竞争力,也是损失。到底这个保修期是怎样决定的呢?
2. 在同一年级中,同样统计学的课程可能由一些不同教师讲授。教师讲课方式当然不一样;考试题目也不一定相同。那么如何比较不同班级的统计学成绩呢?
3. 大学排名是一个非常敏感的问题。不同的机构得出不同的结果;各自都说自己是客观、公正和有道理的。到底如何理解这些不同的结果呢?
4. 任何公司都有一个信用问题。如果这些公司试图得到贷款时并没有不还贷的不良记录。如何根据它们的财务和商业资料来判断一个公司的信用等级呢?
5. 我国东部和西部的概念是一个比较笼统的概念。如何能够根据某些标准或需要,选择一些指标来把各省,或各市县甚至村进行分类呢?

6. 疾病传播时，如何能够通过被感染者入院前后的各种经历得到一个疾病传染方式的模型呢？
7. 如何通过问卷调查来得到性别、年龄、职业、收入等各种因素与公众对某项事物（比如商品或政策）的态度的关系呢？
8. 一个从来没有研究过红楼梦的统计学家如何根据比较写作习惯得出红楼梦从哪一段开始就不是曹雪芹的手笔了呢？
9. 如何才能够客观地得到某个电视节目的收视率，以确定插播的广告价格是否合理呢？

其实，这些都是统计应用的例子。这样的例子太多了。因为统计学可以应用于几乎所有的领域，包括精算，农业，动物学，人类学，考古学，审计学，晶体学，人口统计学，牙医学，生态学，经济计量学，教育学，选举预测和策划，工程，流行病学，金融，水产渔业研究，遗传学，地理学，地质学，历史研究，人类遗传学，水文学，工业，法律，语言学，文学，劳动力计划，管理科学，市场营销学，医学诊断，气象学，军事科学，核材料安全管理，眼科学，制药学，物理学，政治学，心理学，心理物理学，质量控制，宗教研究，社会学，调查抽样，分类学，和气象改善，博彩等。当然，大家用不着也不可能理解所有的统计应用。只要能够解决自己身边的统计问题就足够了。

在解决上面所提到的9个问题时所需使用的大多数统计分析方法将会在本书后面章节中陆续介绍。当然我们的例子并不一定就刚好是上面问题中的具体例子，但至少所使用的分析方法是类似的。

上面的例子并没有明确说出什么是统计。其实很简单。上面的所有例子都要通过各种直接或间接的手段来收集**数据（data）**；都要利用一些方法来整理和分析数据；最后通过分析得到结论。一句话，**统计学（statistics）是用以收集数据，分析数据和由数据得出结论的一组概念、原则和方法。**比如要得到某电视节目的收视率，可能首先要在该节目播出时，利用电话对看电视的人进行采访，同时问他们在观看什么节目。在得到了被采访的看电视的总人数，和其中观看该节目的人数之后，就有可能得到这部分观众中，观看该节目的比例，即粗糙的收视率了。之后还要经过统计分析，评估这个收视率的可信度和代表性等等。显然，这是一个收集数据，然后通过分析数据得到结论的简单例子。

§1.2　现实中的随机性和规律性，概率和机会

从中学起，我们就知道自然科学的许多定律，例如物理中的牛顿三定律，物质不灭定律以及化学中的各种定律等等。但是在许多领域，很难用如此确定的公式或论述来描述一些现象。比如，人的寿命是很难预先确定的。一个吸烟、喝酒、不锻炼、而且喜好油荤食物的人可能比一个很少得病、生活习惯良好的人活得长。因此，可以说，活得长短有一定的**随机性(randomness)**。这种随机性可能和人的经历、基因、习惯等无数不易说清的因素都有关系。但是从总体来说，我国公民的预期寿命却是非常稳定的，而且由于生活水平的提高在逐步增长；比如1996年的平均预期寿命为70.80岁而2000年为71.40岁。这就是规律性。一个人可能活过这个平均年龄，也可能活不到这个年龄，这是随机的。但是总体来说，预期寿命的稳定性，却说明了随机之中有规律性。这种规律就是统计规律。

你可能经常听到**概率(probability)**这个名词。最常见的是在天气预报中提到的降水概率。大家都明白，如果降水概率是百分之九十，那就很可能下雨；但如果是百分之十，就不大可能下雨。因此，从某种意义说来，概率描述了某件事情发生的机会。显然，这种概率不可能超过百分之百，也不可能少于百分之零。**换言之，概率是在0和1之间(也可能是0或1)的一个数，说明某事件发生的机会有多大。**

有些概率是无法精确推断的。比如你对别人说你下一个周末去公园的概率是百分之八十。但你无法精确说出为什么是百分之八十而不是百分之八十四或百分之七十八。其实你想说的是你很可能去，但又没有完全肯定。实际上，到了周末，你或者去，或者不去；不可能有分身术把百分之八十的你放到公园，而其余的放在别处。有些概率是可以知道的。比如掷骰子。只要没有人在骰子上做手脚，你得到6点的概率应该是六分之一。得到其他点的概率也是一样。这反映了掷骰子的规律性。但掷出骰子之后所得到的结果还只可能是六个数目之一。这体现了随机性。如果你掷1000次骰子，那么，大约有六分之一的可能会得到6；这也说明随机结果也具有规律；而且有可能通过试验等方法来推测其规律。

§1.3 变量和数据

做任何事情都要有对象。比如一个班上注册的学生有200人,这是一个固定的数目,称为**常数(constant)**或者常量。但是,如果猜测今天这个班有多少人会来上课,那就没准了。这有随机性。可能有请病假或事假的,也可能有逃课的。这样,就要来上课的人数是个**变量(variable)**。另外对于某项政策同意与否的回答,也有"同意"、"不同意"或者"不知道"三种可能值;这也是变量,只不过不是数量而已。当变量按照随机规律所取的值是数量时该变量称为**定量变量**或**数量变量(quantitative variable)**;因为是随机的,也称为**随机变量(random variable)**。像性别,观点之类的取非数量值的变量就称为**定性变量**或**属性变量**或**分类变量(qualitative variable, categorical variable)**。这些定性变量也可以由定量变量来描述,比如男性和女性的数目,同意某政策人数的比例等等。定性变量只有用数量来描述时,才有可能建立数学模型,才可能使用计算机来分析。

有了变量的概念,什么是数据呢?拿掷骰子来说,掷骰子会得到什么值,是个随机变量;而每次取得1至6点中任意某点数的概率在理论上都是六分之一(如果骰子没有作假)。这依赖于在掷骰子背后的理论或假定;而在实际掷骰子过程中,如果掷100次,会得到100个由1至6点组成的数字串;再掷100次,又得到一个数字串,和前一次的结果多半不一样。这些试验结果就是数据。所以说数据是关于变量的观测值。

通过数据可以验证有关的理论或假定。比如通过很多次掷骰子验证得到每个点的概率是不是1/6。对于顾客是否喜欢某种饮品的调查也类似,但这里不象掷骰子那样事先可以大致猜测顾客喜欢与否的概率。在随机问了1000人之后,可能有364人说喜欢,而480人说不喜欢,其余的人可能不回答,或说不知道,或从来没有喝过这种饮料。当然,它仅仅反映了1000个被问到的人的观点;但这对于估计整个消费群体的观点还是有用的。从这些数据可以估计出喜欢这种饮料的大约占36.4%左右。后面还要介绍得到数据的一些途径和方法。

§1.4 变量之间的关系

现实世界的问题都是相互联系的。不讨论变量之间的关系,就无从谈起任