全国统计教材编审委员会“十一五”规划教材

统计学：从数据到结论

第三版

★ 吴喜之 编著

中国统计出版社
China Statistics Press

全国统计教材编审委员会“十一五”规划教材

统计学：从数据到结论

第三版

★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.—3 版.
—北京:中国统计出版社,2009.9
ISBN 978—7—5037—5801—0

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(2009)第 166058 号

统计学:从数据到结论(第三版)

作　　者/吴喜之
责任编辑/吕　军
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 57 号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲 6 号
网　　址/www.stats.gov.cn/tjshujia
电　　话/邮购(010)63376907　书店(010)68783172
印　　刷/北京市顺义兴华印刷厂
经　　销/新华书店
开　　本/787×1092mm　1/18
字　　数/260 千字
印　　张/21
印　　数/1—5000 册
版　　别/2009 年 9 月第 1 版
版　　次/2009 年 9 月第 1 次印刷
书　　号/ISBN 978—7—5037—5801—0/C·2277
定　　价/39.00 元

前　言

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。这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觉得习以为常，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。这门课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。本教材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的，多数开汽车的都不会修汽车，但这对他们毫无妨碍。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。这种安排使得本教材能够适用于各种不同水平、不同要求的读者群体。本教材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。为了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（第二版之后加了 SAS 格式）。

软件方面，本书主要使用 SPSS 和部分地应用 Excel（第二版之后加了 SAS 和 R 的应用）。这里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。

这个教材的全部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，这个课程前 15 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六章，或最多前九章的内容，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教材内。

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

2003 年 6 月

第二版说明

本书的第一版发行不到一年，已经作为参考书或教科书在许多学校使用。各个学校的师生对本书提出许多宝贵的意见，并且指出了很多错误和不妥之处。他们的支持和鼓励，促使了本书的第二版的诞生。

和第一版相比较，第二版对许多内容完全重新写过，还进行了一些调整，同时加强了对概念和方法的解释，使得该书更加容易理解。第二版还对例题和习题都做了很多修订和增减。此外，还增加了一些内容；除了基于SPSS的操作和输出结果的分析和解释之外，还增加了SAS软件的使用，特别是增加了关于如何通过免费的，功能强大的，需要自己动手写程序的R软件来理解概念及处理数据。R软件的代码公开及透明的优势是一些“黑匣子”式的傻瓜软件所无法比拟的。R软件是使用S语言来编程的(和S-plus的编程语言一样)，在其问世的不到十年的时间，已经成为国外统计研究生的首选软件。它有强大的网上支持系统。多数最新的统计计算方法，在进入商业软件之前，就已经以R语言的形式在R网站上免费提供了。使用本书的师生最好也使用R语言。在掌握R软件之后，对其他统计方向的学习和研究都会有很大的帮助，甚至会有一种到了自由天地的感觉。在R软件的帮助中有完整的入门材料和各种命令的意义和使用例子，因此本书没有必要自己编写关于R语言的附录来增加篇幅。

本书选择的与内容有关的SPSS软件选项和SAS软件语句(或选项)的原则是容易理解和掌握，当然，由于编者知识有限，对于有些方法，没有找到(因此也无法提供)

最合适、最方便或者最新的软件选项或模块，希望读者提出建议，使得再版时予以弥补。

由于使用软件比查表更加方便和可靠，有人说，你自己都不查表，为什么要教学生去查表呢？的确，编者除了在最初等的统计课教学过程中曾经涉及少数统计表之外，从来都是使用软件。“己所不欲，勿施于人”，本书不提供任何统计分布表。希望有条件的读者尽量使用计算机，而不去查表。实际上，如果没有计算机的支持，很难对有一定规模的数据在任何统计方向进行较深入的分析。

许多人，比如各层管理人员，并不一定都进行第一线的实际数据计算，但为了理解手中关于本单位及有关方面信息的意义，为了更好地进行明白的决策，他们必须理解各种统计推断结果的意义。对这些人，不一定要求能够熟练使用软件，更不需要理解数学推导，但他们必须明白各种统计概念和方法以及输出结果的意义。相信本书对他们也会有所帮助。

作为教科书，本书内容对于每周两学时的课程似乎太多。我觉得，什么讲或者什么不讲应该根据学生的需要由老师自己安排。实际上，对于任何课程，最好是由任课教师来决定讲哪些内容以及如何讲。因为他们最了解他们所面对的学生。教科书编者的思维方式不见得和老师的一致，而老师最好按照自己的理解来讲述。一个好的教科书，应该给教师以较大的余地和自由。

希望读者继续对本书予以宝贵的支持和批评指正。

吴喜之

2006 年 3 月

第三版说明

这一版是在第二版的基础上做了一些改进，对内容安排也进行了一些调整，并纠正了一些各种原因造成的错误和不妥之处。没有广大读者的鼓励、帮助和建议，没有中国统计出版社的支持，本书的第三版是无法面世的，为此，特在这里对所有方面表示感谢。

吴喜之

2009 年 7 月

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. 当你买了一台电脑时,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？

10. 如何根据税务部门过去的税收记录来预测下一年的税收收入，供政府部门制定预算时参考？

11. 如何根据某地区的寿命记录来确定人寿保险的既有竞争力，又有利可图的定价？

其实，这些都是统计应用的例子。这样的例子太多了，无法一一列举。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若干个应用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。当然书中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。统计是一门科学，它以现实世界待解决的问题为目标，这一点，和诸如物理学等其它科学一样。科学研究的方法是：观测世界或进行试验，得到数据，提出可以解释这些观测的假说或理论，试图尽可能地接近现实世界的规律，当出现理论或假说无法解释的现象(数据)时，就有可能需要原有理论进行修正或者代之以新理论。统计学的假说或理论通常称为模型。按照不列颠百科全书关于统计的定义，**统计学(statistics)是"收集、分析、展示和解释数据的科学。**[①]"与物理学的假说类似，统计学的模型仅仅是对现实的近似，没有任何模型是"正确"的，也无法证明任何模型是正确的。只能够说，在某些可能有争议的准则之下，某些模型比另外一些要更合适一些。在

① **statistics.** (2008). Encyclopædia Britannica. *Encyclopædia Britannica 2007 Ultimate Reference Suite*. Chicago: Encyclopædia Britannica.

数学逻辑中存在的确定性在统计中完全不成立。针对于不同学科问题而发展的统计学中的数学完全不成为一个完整封闭的体系，也没有必要成为一个数学体系。能否解决实际问题是评价统计方法的最终标准。

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。

思考一下：

1. 你周围经常会有辩论，是不是这些辩论都是以科学的方法来进行的？
2. 科学之外就是信仰，举例说明科学和信仰之间的区别。
3. 举出一个你认为是统计应用的例子。

§1.2 现实中的随机性和规律性，概率和机会

从中学起，大家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、而且经常吃荤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，我国公民的平均预期寿命却是非常稳定的，而且随着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预期年龄，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