高等学校统计学类系列教材

统计学：从概念到数据分析

吴喜之 编著

高等教育出版社
HIGHER EDUCATION PRESS

ISBN 978-7-04-023393-
9 787040 233933 >
定价 18.00 元

高等学校统计学类系列教材

统计学：从概念到数据分析

吴喜之　编著

高等教育出版社

内容提要

本书主要介绍了概率基础、统计的基本概念、描述性统计、估计、假设检验、回归与分类等内容，同时介绍了决策树、神经网络和随机森林等组合方法以及如何用 R、SPSS、SAS 等软件来实现相应的计算目标。

本书着重直观讨论，尽量少用公式，避免数学推导，强调统计学的基本内容及应用，使读者能够完整、准确地理解统计学的概念，学会利用统计软件进行数据分析。

本书主要是为非统计学专业的学生和读者编写，读者不需要任何概率统计基础知识。

图书在版编目(CIP)数据

统计学：从概念到数据分析／吴喜之编著. —北京：
高等教育出版社，2008.6

ISBN 978－7－04－023393－3

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(2008)第 062059 号

策划编辑 李 蕊 **责任编辑** 崔梅萍 **封面设计** 王凌波 **责任绘图** 尹 莉
版式设计 王艳红 **责任校对** 殷 然 **责任印制** 尤 静

出版发行	高等教育出版社	**购书热线**	010－58581118
社　　址	北京市西城区德外大街 4 号	**免费咨询**	800－810－0598
邮政编码	100011	**网　　址**	http://www.hep.edu.cn
总　　机	010－58581000		http://www.hep.com.cn
		网上订购	http://www.landraco.com
经　　销	蓝色畅想图书发行有限公司		http://www.landraco.com.cn
印　　刷	北京东光印刷厂	**畅想教育**	http://www.widedu.com
开　　本	787×960　1/16	**版　　次**	2008 年 6 月第 1 版
印　　张	13	**印　　次**	2008 年 6 月第 1 次印刷
字　　数	240 000	**定　　价**	18.00 元

物料号　23393－00

前　言

统计是一个对所有领域都有用的工具。一个本科生，无论其主修方向是什么，如果能够掌握一些统计数据分析方法，无疑会受益匪浅。本教材主要是为非统计专业的学生和读者编写。强调统计学最基本的内容及应用，使读者不但能够逻辑完整地准确理解统计学的概念，而且能学会如何通过计算机统计软件进行数据分析。本书包括了基本概念、描述性统计、估计、假设检验、回归与分类等统计学内容。特别要提出的是，本书还介绍了如何用计算机实现在数据挖掘中用于回归与分类的常用方法。

这是一本从入门到具体数值分析与计算的课本。我们着重直观解释统计的基本概念，尽量少用公式，避免引入只有专业统计人员才需要了解的数学推导及定理证明。

学习本书不需要读者事先学过概率论或微积分，因此，完全可以放在大学本科任何一个学期讲授。对于学过概率论或概率论与数理统计课程的读者，完全可以跳过第二章的第二节及全部第四章的内容，并且在授课时主要强调数据的描述、数值计算及结果的解释。

我们提倡启发式教学，以理解为主，杜绝死记硬背。为此，书中以“思考一下”的形式给出很多启发性的问题、注解和补充，供老师和学生讨论和思考。这些问题也可以作为习题。而在习题部分，强调数值计算和分析。希望读者用计算机来实现所有的数值计算题。我们不希望在统计课程中存在了至少五十年的“手工计算→查表→手工计算”的前计算机时代的模式再延续下去。

本书每一章后面都介绍了如何用 R、SPSS、SAS 等软件来实现相应的计算目标。不仅如此，书中对每一个计算结果的获得以及每一个重要图形的绘制，都在脚注中说明了如何用 R 软件来实现。希望这种方式对教和学都有帮助。本书例题中涉及的相关数据可登陆中国高校数学课程网（http://math. cncourse. com）下载。

可能有人担心，没有了手工计算、死记硬背的术语定义或数学推导就不易考试了。教学不是为了考试。教材也不能为迎合考试而编写。其实，考试的方式是多种多样的。开卷或闭卷、用计算机或不用计算机都可以考察学生的能力。一种行之有效的课堂闭卷考试方式为选择题。这些题目包括关于基本概念的是非题、对给定具体应用的统计方法选择题、计算机输出结果的解释题（也是选

择)等等。为有助于拉开学生成绩间的距离,题目量可以很大,而做题时间要短,部分人能够答完就行了。

本教材的内容在中国人民大学的非统计专业本科及研究生的统计学课程中得到使用。希望读者能够提出宝贵的意见。

中国人民大学　统计学院
吴喜之

目　录

第一章

引　言

§1.1　什么是科学方法?

我们天天都在使用“科学”这个字眼,但是,有多少人认真考虑过科学的真正含义呢?

人们对世界的认识来源于他们所获得的信息(或数据).而在总结这些信息时人们头脑中会形成一些模型(也称假说或理论).这些模型会指导他们做进一步的探索,直到遇到这些模型无法解释的现象.这时,人们会改进这些模型,或者干脆建立新的模型使得新模型不仅能解释旧模型可以解释的现象,而且还能解释旧模型无法解释的现象.这就是科学的方法.而只有用科学方法进行的探索才叫科学.下面举两个人们熟知的例子.

• **天文学**:公元2世纪,托勒玫致力于传播宇宙地心说,这一思想影响了1300多年.地心说可以对当时条件下的一些天文观测提供解释.1543年,在哥白尼的《天体运行学》一书中阐明了日心说,把托勒玫的理论大大改进了.随后,开普勒发现行星运动原理.伽利略开始用望远镜进行天文观测.牛顿又建立了运动和万有引力定律.在新的观测的基础上,赖特在1750年又提出宇宙是由众多星系构成的看法.18世纪末,赫歇尔首先进行了用望远镜的巡天观测,奠定了现代恒星天文学的基础.

• **从牛顿到爱因斯坦**:牛顿建立了运动定律和万有引力定律,这些定律在当时可以解释相当一部分观测现象.然而,后来在亚原子尺度上以及在行星观测中出现了一些牛顿的惯性定律或万有引力定律无法解释的现象.这就导致了爱因斯坦狭义和广义相对论的产生.相对论是建立在光速在真空中不变的假法前提下成立的.如果人们观测到光速在真空中可变,则又会促进对相对论的修正.

上面的例子可以看出科学的一些特点.科学可以定义为对关于宇宙的所有方面的知识的认真的、系统的、合乎逻辑的研究;这些知识则是由考察最好的可

利用的证据得到的，并且这些知识总是应该在发现更有力的证据时随时予以纠正和改进. 科学也可以定义为任何知识系统，这些知识涉及物理世界及其可经受无偏见观测和系统实验的现象.

而科学方法则是目前已知的筛去谎言和错觉的最好方式. 科学方法的步骤可做如下大致的描述：

（1）观测宇宙的某些方面.

（2）发明或提出可以解释这些观测的假说或假设，它必须和观测结果是相容的.

（3）利用该假说进行预测.

（4）用实验来检验这些预测，或者做进一步观测并根据结果修正假说.

（5）重复第3、4步直到在理论和实验或观测中没有发现矛盾为止.

任何假说，如果能够说明很多现象，也可称为理论. 但任何理论都不能达到绝对的真理. 所有的科学理论都应该是可证伪的（falsifiable[①]），这意味着应该存在某种实验或可能的发现证明理论有问题. 看不见摸不着的神的存在是无法证伪的，因此宗教不是科学，而是信仰. 目前基于不能重复观测或重复实验的现象而产生的许多说法，都不是科学，最多是信仰. 没有证伪，科学是不会发展的. 从上面天文学和物理学的例子可以看出，科学的理论是在否定中发展的. 每当发现目前理论所不能解释的实验结果或观测，就产生对理论进行改进或更新的动机或需要.

注意，一个科学理论即使被发现有局限性，也不是不能应用. 比如，在一般条件下，牛顿定律还是适用的. 现在谁都知道地球是球体. 但在发现地球是球形之前，“地平说”可以解释很多现象. 即使是今天，在盖普通房子之前进行测量时，也不必要考虑地球的曲率.

科学是靠证据说话的. 一个理论适用与否是靠实验或观测，靠辩论是不行的. 古希腊的伟大哲学家亚里士多德用各种理由辩论说男人和女人的牙齿数目不同. 他是好的辩才，但不是好的科学研究人员. 基于含糊不清或者不适当的前提的逻辑推理是没有多大意义的. 科学研究必须是毫无偏见的. 科学的结论应该独立于研究人员的文化背景、社会背景、种族、习惯、宗教和政治信仰等因素.

当然，也存在制造假的研究结果的现象. 但是，除非造假者的结论没有多大意义，否则总是会被人发现的. 最有名的造假案例是1989美国犹他大学的彭斯和英国南安普敦大学的弗莱什曼冷核聚变以及韩国科学家黄禹锡克隆胚胎干细胞的例子. 也有科学家犯错误的案例，其中最突出的是在伦琴发现X射线之后，

① 这个词的英文定义可以为：falsifiable. *adj*: capable of being tested (verified or falsified) by experiment or observation [syn: {confirmable}, {verifiable}].

法国教授布朗洛宣称发现 N 射线. 但无人能够重复布朗洛的实验,人们还证明他的观测有误.

此外,权力、宗教和意识形态也会对科学造成严重干扰. 拥护哥白尼的"天体运行论"的布鲁诺被罗马教廷以"异端分子和异端分子的老师"的罪名,于 1600 年 2 月 17 日烧死在罗马鲜花广场. 伽利略由于收集、分析了日心说的证据,于 1633 年被罗马天主教廷判决软禁,在软禁中度过余生结果使得地中海地区的科学传统完全停止了. 在 20 世纪 30 年代到 60 年代,苏联的全苏列宁农业科学院院长李森科出于政治与其他方面的考虑,把得到实验支持的孟德尔和摩尔根遗传学斥为资产阶级的异端邪说,并在斯大林的支持下对苏联的研究基因的学者实行迫害. 李森科事件是政治权威取代科学权威裁决科学争论的典型案例. 这件事也对中国遗传学界产生了恶劣影响.

思考一下

1. 那些关于耳朵识字、透过封闭的瓶子取物等报道的事件是科学的吗?
2. 举出一些科学、信仰和魔术的实例. 它们之间有什么区别?
3. 有人说他们见到了不明飞行物,这能够证明它们存在吗?
4. 谈谈你对电视上报道的一些问题的看法.
5. 举例说明利用科学方法得到的结论应该是可重复的.
6. 举例并讨论如果信仰和偏见混入了大前提,逻辑过程再正确,不可能得到科学的结论.
7. 为了某种目的,一些人伪造研究结果,这些最终会被发现吗? 举例说明.
8. 有人说科学也是信仰,所信仰的是"宇宙存在规律". 请大家讨论这个"信仰"是真的宗教式的信仰,还是可以用证据来说明的.

§1.2 统计是什么?

人们总是在现实世界中收集各种证据,试图找出一些规律或模型来描述所研究的对象. 物理学、化学、生物学、地理学、天文学都是这样做的,统计学也是一样. 不同的是,那些自然科学本身的规律是比较确定的. 叙述简洁的牛顿定律就是一个很好的例子. 此外,在一定条件下化学反应中的结果也是完全确定的,许多天体的运动轨道也是基本可以确定的. 而世界上有许多事物是无法用这些确定性的理论来描述的. 比如,一个企业家去年增加投资可能利润有所增加,而今年增加投资就可能赔本. 再如保险公司希望减少汽车保险中的风险,这就需要找出具有哪些特征的人群具有高风险. 这些问题绝对不能从逻辑推理来得到解决方案,必须通过分析相关的数据才能总结出规律.

拿汽车保险为例. 如国外保险业数据所显示的,年轻人、开红色车的人、开跑车的人均容易出车祸. 从物理学角度,鲜明的红色应该更醒目,性能好的跑车应该更安全,眼明手快的年轻人更不易出事. 但保险公司的规律是统计学家从事故数据总结出来的. 现在,心理学家可以从这一类人有炫耀倾向的心理因素来解释. 统计学仅仅是从数据中找出规律而已. 但是在统计学家找出什么类型的人易出事故的规律之前,心理学家往往不会往这方面去想.

因此,在无法用自然科学的定律来解释的情况下,在研究许多说不清楚原因的现象时,统计学可以通过研究数据来找出规律甚至答案. 统计学进行推断的基础就是数据. 因此可以像不列颠百科全书那样,把统计学定义为**收集、分析、展示和解释数据的科学**①. 这个定义是国际上普遍接受的. 这里所说的数据就是科学中的事实和证据. 数据不仅限于数字,它也可能是图像或者是文字. 实际上,任何信息都可以称为数据.

统计研究的许多对象被认为具有**随机性(randomness)**. 随机的事物有规律吗? 当然有. 比如掷一个色子,得到几点是随机的,事先无法确切地预测. 但只要没有人在色子上做手脚,得到各点的机会大致上是一样的. 一个人能够活到多少岁是由无数很难说清的因素确定的,但整个人群的预期寿命是大体确定的. 这些都是随机中有规律的例子. 从某种意义上说,统计是研究随机现象的科学,它要在随机现象中寻找规律. 实际上,这里所说的"随机现象"可能并不是随机的,只是人们习惯于把弄不清机制的现象归到随机现象而已.

思考一下

1. 请举出一些有规律的随机事物的例子.
2. 有人认为,整个世界原本都是确定性的,只是因为科学不够发达,或人类能力的局限性,使得我们无法找到确定性模型来描述. 当然,目前没有足够证据或手段来肯定或否定这个观点. 你怎么想呢?
3. 也有人认为,一切量都有随机性? 你又怎么看呢?
4. 就你目前所了解的,比较统计和其他自然科学之间的区别.

统计学和数学一样,是所有学科的工具.

统计学与各个学科的数据都打交道,统计学实际上已经应用于所有领域. 作为例子,它们包括②: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,

① Encyclopedia Britannica, 2006.

② Encyclopedia of Statistical Sciences, 1981.

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，气象改善，博彩，等等. 现在，任何领域的研究结果，如果没有根据数据所作出的结论，是很难被认可的.

目前，随着科技的进步，我们面对着所谓的“信息爆炸”. 在网络、遥感、金融、电讯、地理、商业、旅游，军事以及生物医学等各个领域不断产生大量的数据. 现在从各个领域中产生的数据量远远超过了人们分析和处理它们的能力. 如何能够把数据中的重要信息迅速有效地提取出来是非常重要的. 数据挖掘、人工智能、机器学习等领域的出现对统计学、计算机科学及各个相关领域提出了更高的要求，带来了机会和挑战.

那么，统计实践的一个全过程究竟是由什么组成的呢？首先，做任何事情都需要有一个目标. 比如一个企业想要知道他们的某项产品受欢迎的程度、市场占有率、知名度以及什么因素影响人们对该产品的看法等，需要做的第一件事就是收集数据，而收集什么样的数据则是首先要关注的. 至少，要了解多少人购买该产品或对该产品有兴趣、人们为什么认可该产品以及这些人的文化程度、工作性质、性别和年龄等特征. 人们也许还需要了解竞争对手的各种信息. 实际上，即使是有经验的人，也可能忽略一些需要调查的内容. 在确定调查内容和所要提的问题之后，就要考虑如何设计问卷. 问卷的设计对于调查结果至关重要，问卷的质量直接影响调查的结果. 一个不合格的问卷既浪费资源、又不会得到预期的结果. 确定了问题，下面就要考虑如何收集数据了. 比如，一个重要问题是在什么人群中调查，因为每一个产品都有其适应的人群. 对缺水地区推销洗衣机，对低收入阶层推销奢侈品都是荒唐的. 在确定调查对象和范围之后，还要确定调查多少人. 调查人越多，结果就越可靠，但也更耗费资源. 用什么方式调查也是必须确定的问题. 如果用问卷调查，选择面对面调查、电话调查、网上调查和邮寄问卷，可能会产生不同的结果. 调查之后，就要从数据中找到规律. 这些规律可能是由一些数学公式表示的模型，也可能是由一些算法所界定的. 这些模型可能是已知的，也可能是改进的或者是根据数据新构造的. 根据这些模型，人们可以理解数据所表达的含义，也可以对未来进行预测.

一般来说，统计过程大体上可以总结成下列步骤：

1. 明确目标，并根据目标确定需要收集的变量，也就是收集什么类型的数据. 还要确定收集数据的方法.

2. 收集数据.

3. 选取或者改进已知的模型，或者基于数据构造新的模型. 这个建模步骤和前面步骤 1、2 中的收集数据的步骤可能要重复多次.

4. 利用模型对数据进行描述和分析.

5. 得出所需要的和实际问题相关的结论. 如果结果不能解释,可能还要从步骤 1 重新开始.

一个统计学家或一个实际领域的工作者不一定从头到尾经历所有这些步骤. 他们可能仅参与其中的部分工作. 实际的统计过程也可能有更少或更多的步骤. 这些步骤从时间上和资源耗费方面并不均等. 希望大家能够从实践中总结自己的经验和作法.

在为其他领域服务时,统计工作者的重要原则是:在提供你的统计结论时,必须同时提供你的结论可能犯错误的概率①,而让熟悉该领域的工作者对于实际问题来做决策. 这是因为只有他们才了解什么是他们的利益之所在. **在提供统计结论的同时避而不谈该结论可能造成的风险是不负责任的**. 统计结论可能犯错误的概率本身不能代替与实际问题相关的风险. 比如,某个探险行动使你有 10% 的可能失去生命,你可能不会去参加. 但某项收益颇丰的投资有 10% 的可能失败,你可能就会去加入. 同样的概率,人们可能会采取不同的决策. 这说明概率仅仅给决策者一个参考,而在决策者心中还有与概率相关的损失或收益的评价体系.

思考一下

1. 文学看来和统计没有关系. 但是,在英国,人们用统计来研究哪些作品是莎士比亚写的,哪些不是;中国也有人对《红楼梦》各章回的作者做过统计分析. 你能想出他们是如何做的吗?
2. 如果你想知道校园中学生对食堂满意和不满意的比例,以及不同的回答与回答者的性别、年级、专业、经济状况等特征是否有关,及如何有关,你的调查计划将会是怎样的呢? 如果你想分析不满的原因,问卷中需要包含哪些问题?
3. 如果你投资一万元,有 10% 的可能血本无归,但有 90% 的可能得到加倍的偿还. 你会去投资吗? 如果你的一万元投资有 10% 的可能获得 50 倍的偿还,但有 90% 的可能血本无归. 你会去投资吗? 你的决策是完全根据概率吗? 请解释.

§1.3 学习统计需要的基础知识和技能

知识总是多一点好,但人们的生命有限,不可能什么都学. 我觉得,除了学习

① 虽然我们在中学已经了解了概率这个概念,我们在第二章还是会涉及这个概念,在第三章还要给出和概率有关的分布概念.

本领域的知识之外，最重要的两项是数学和计算机，再加上想象力、通常的逻辑推理和常识判断的能力.

好的数学基础无疑对于逻辑思维和学习新知识都有帮助，但本书并不需要超过你们已经学过的知识. 对简单概率论知识的需要可能多于微积分. 此外如果掌握一些向量和矩阵的知识，则对某些统计方法就更容易理解了. 重要的是，在需要某些新知识时必须具备自学能力. 本书尽量做到基础知识自给自足，以免学生受查资料之苦.

我们对于计算机的需要较多. 实际上，本书从头到尾都要用计算机. 我相信所有读者都会用计算机做一些事情. 但是我们需要的是利用统计软件来描述和分析数据，这可能并不是每个读者都做过的. 统计软件入门其实并不难. 下面对一些统计(计算)软件做些介绍.

在没有视窗及鼠标的时代，所有的计算软件都需要写程序. 最初的语言是用机器语言写的，称为初级语言，后来又有了诸如 FORTRAN、BASIC 等和手写公式近似的所谓高级语言. 这些高级语言在执行时需要由编译器换成机器语言. 后来出现了专门针对统计计算或工程计算的软件. 针对本书的每个例题，我们都介绍如何用三种统计软件来计算：R、SPSS、SAS. 下面主要介绍这三种软件.

R 软件：这是一个免费的，由志愿者管理的软件. 其编程语言与商业软件 S - plus所基于的 S 语言一样，很方便. 此外，R 网站提供不断更新的统计学家编写的涉及各种最新方法的包含数据和程序的统计软件包. 因此，其软件包和函数的数量及更新速度远远超过其他软件. 这个软件受到世界各国统计师生的欢迎，是用户量增加最快的统计软件. 它的多数计算过程和代码都是公开的，不像“傻瓜”软件的“黑盒子”式模块. 这些函数还可以被用户按需要改写. 它的语言结构和 C + +、FORTRAN、MATLAB、Pascal、BASIC 等很相似，容易举一反三. 虽然它没有点鼠标式的“傻瓜化”，但它是最容易入门的软件. 关于 R 系统的主要资源可以在 R 的网站找到①，它的网站还提供各种数据挖掘的软件包.

SPSS：这是一个很受欢迎的统计软件. 它容易操作，输出漂亮，功能齐全，价格合理. 它也有自己的程序语言，但基本上已经“傻瓜化”. 它对于非专业统计工作者是很好的选择. 这似乎是最“傻瓜化”的软件之一. 虽然它也有编程语言，但一般用户主要使用对话框用鼠标选项来实现计算. 它是模块化软件，人们无法知道模块的代码. 它不包括数据挖掘的方法.

SAS：这是功能很多的软件. 尽管价格相当不菲(每年要交租金)，许多公司，特别是美国各制药公司都在使用，这多半因为其众多的功能. 尽管现在已经部分

---

① R 网址为：http://www.R-project.org. 其资源包括不断更新的系统本身、各种可加的最新写成的关于不同统计方法的软件包以及手册、说明等.

“傻瓜化”,但仍然需要一定的培训才可以掌握. 可以编程;但也是无法知道代码的模块化软件. 它对于基本统计课程不是很方便. 较新版的 SAS 包括数据挖掘的模块.

其他通用的统计软件包括 Minitab、Statistica、S-plus(和 R 一样,同样使用 S 语言,并且已经开始“傻瓜化”)、Gauss(类似 R)、MATLAB(工科用得较多,和 R 很像)、SYSTAT 等软件. 还有一些专门针对某一两种统计方法的软件,这里不做介绍. 上述软件基本上是一边计算一边编译,因此速度不如把程序编译之后再计算的 FORTRAN(也有许多数学和统计函数库)和 C 等语言快. 当然,学习 FORTRAN和 C 则比上述软件稍微多花一些时间,但完全可以弥补某些大计算量运算所耗费的时间. 许多人把 Excel 也当成统计软件. 其实,对于稍微复杂的统计问题,Excel 不方便,更不灵活,需要 VBA 的知识. 最新的 Excel 也包含了一些复杂的数据分析功能.

根据经验,学习统计软件的最快和最好方法是在学习统计方法的过程中学会使用统计软件. 当有无法解决的问题时,可以在各种网站或者介绍软件的书上找到答案.

思考一下

1. 你用过何种计算软件? 你对这些软件评价如何?
2. 下载 R 软件,并且根据帮助中的手册(PDF)的“An Introduction to R”的一些命令做些简单的计算. 当然,以后在学习本书的过程中,你会学到更多的使用方法.
3. 许多人认为 R 软件是最基本的数学和统计计算软件,可以在两小时内学会. 你可以试试.
4. “傻瓜软件”可以代替统计课程吗? 其实在对话框中点鼠标选项本身就是学问. 使软件输出结果容易. 但输出合理的、可以解释的结果就那么容易吗?
5. 有人认为一周就可以掌握 FORTRAN 的基本运算编程. 你可以试试.

§1.4 习题

1. 观察你看到的各种论点、观点和推理方式. 说明哪些是科学,哪些是魔术,哪些是信仰.

2. 虽然还没有接触到统计的具体内容,你肯定有许多关于统计的看法或疑问. 请把这些问题列出来,在以后学习中注意找到答案.

3. 举出随机现象有规律的例子.

4. 搜寻到 R 网站(http://www.r-project.org/),并在 CRAN(The Comprehensive R