CongLing KaiShi DuDong TongJiXue

哈佛大学终身教授刘军说：

大数据是“原油”而不是“汽油”，不能直接使用。大数据时代，统计学依然是数据分析的灵魂。

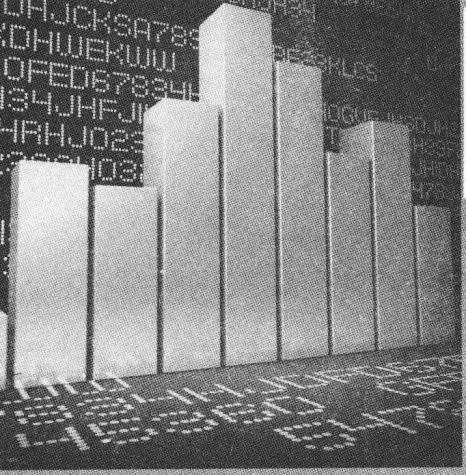
案例丰富实用，配套精选习题
大数据时代必读书

统计学应用入门手册
零基础可自学读本

从零开始读懂统计学

李慧泉◎著

立信会计出版社
LIXIN ACCOUNTING PUBLISHING HOUSE



CongLing KaiShi DuDong TongJiXue

案例丰富实用，配套精选习题
大数据时代必读书

从零开始读懂统计学

| 李慧泉◎著 |

立信会计出版社
LIXIN ACCOUNTING PUBLISHING HOUSE

图书在版编目（CIP）数据

从零开始读懂统计学/李慧泉著.--上海：立信会计出版社，2016.6

（去梯言）

ISBN 978-7-5429-4988-2

Ⅰ.①从… Ⅱ.①李… Ⅲ.①统计学 Ⅳ.①C8

中国版本图书馆CIP数据核字(2016)第091208号

策划编辑　蔡伟莉

责任编辑　蔡伟莉　于　欣

封面设计　久品轩

从零开始读懂统计学

出版发行	立信会计出版社		
地　　址	上海市中山西路2230号	邮政编码	200235
电　　话	（021）64411389	传　　真	（021）64411325
网　　址	www.lixinaph.com	电子邮箱	lxaph@sh163.net
网上书店	www.shlx.net	电　　话	（021）64411071
经　　销	各地新华书店		
印　　刷	固安县保利达印务有限公司		
开　　本	720毫米×1000毫米	1/16	
印　　张	12.5	插　　页	1
字　　数	178千字		
版　　次	2016年6月第1版		
印　　次	2016年6月第1次		
书　　号	ISBN 978-7-5429-4988-2/C		
定　　价	36.00元		

前言

大数据时代要懂点统计学

在终极的分析中，一切知识都是历史；在理性的基础上，所有的判断都是统计。

当我们提到统计学时，大多数人只会想到繁复到让人头痛的数字和图表，并且将自己归类为数据盲，很少有人会意识到，统计学其实是一种简明的生活工具，你只需要一点数学基础知识就可以入门，它可以跟数学、计量经济学有机结合，甚至可以用于分析当下的经济动态。

当然，统计学有时候还会让我们发现一些有趣的现象：孟加拉国黄油产量和标普500相关性高达0.75、全球变暖与海盗数量减少存在相关性，3月和4月出生的孩子更容易成为优秀棒球运动员……

统计学的源头其实有两个：一个是概率论，另外一个是国情学。概率论最早出现在16世纪，它的起源是一种掷骰子的赌博活动。当时欧洲流行一种掷骰子比点数的赌博，根据点数大小定输赢，这引起了一批学者的关注。学者们试图研究各种点数出现的概率，并且因此出现了一些相关的著作，其中比较有名的有卡丹诺的《机遇博弈》、尼尔·伯努利的《猜度数》等。

而到了17世纪，统计学更多地是以国情学的姿态出现的。人们应用统计学做人口统计，比如生男生女的比例问题。说到这里，我们就不得不提到约翰·格朗特，英国一个杂货店员出身的经济学家。他注意到在非瘟疫时期，一个大城市每年死亡数有统计规律，而且出生儿的性别比为1.08，即每生13个女孩就有14个男孩。他还用数据进一步说明，男性更容易在战争、公海和处以死刑中丧命，所以成年男女的数量基本相等；格朗特初步推算了不同年龄段儿童和成人的死亡比率：儿童死亡发生在4、5岁以下的比例约为1/3，发生在6岁以下的比例约为1/2，仅有7%的死亡属于自然死亡，格朗特在此基础上提出了人类的第一个生命表，并估计出伦敦16～56岁的成年男性约占总人口的34%，有7万人左右可作为战争士兵。从此，概率论和国情学逐渐融合，在这一时期，一些重要的理论被发现，二项分布和大数定律。根据二项分布建立了统计推断的最早的模型，对此分布中未知概率的研究也成为贝叶斯学派的思想起源。

在现代，统计学的发展成为各个知识点的交融，我们可以说统计可以运用于各个领域：经济中计量经济学、医学统计、数据挖掘、生物统计、农业统计、公共卫生、零售等。一句话，只要出现数据的行业，都需要统计学，而随着大数据时代的到来、随着各行各业的发展，越来越多的行业都将开始

需要数据分析这一个职业。

大数据时代已经来了，不管我们是否从事统计、数据分析，都应该了解一些基本的统计学知识，这样才不会在纷乱的数据中迷失自我判断！

导读

为了高效地、一步步理解“统计学”

❶本书为什么由两部分构成

这是一本统计学的入门书。可以大胆地说，这是一本“内容再削减就不能称为统计学”，最浅显易懂的“超级入门书”。

本书由两部分构成。第1部分从最基础的知识开始，力求在最短的时间达到理解“检验”和“区间估计”等统计学最重要的目标。

阅读第1部分，可以让我们在短时间内对“学习统计学要达到的目的以及如何实现”有整体上的了解。

那些正在为无论到哪学习统计学都“无法理解”而抱头苦恼的人，或是无论阅读多少入门书却总是遇到相同难题的人，可以试着浏览本书的第1部分。这里一定有你想要理解却总是理解不了的内容。平日忙碌的读者，当你读到书中的某处时，你一定会感叹“原来统计学是这样的啊”，并认为本书物有所值。

第2部分是对第1部分内容的深化，解说关于母群体的推论统计方法。第2部分的目标是最高效地达到使用 t 分布进行小样本检验、区间估计的

程度。尽管只要理解这些，就能掌握统计学的要点，但很多学习者在此之前就已经备受挫折。

导致此情况最常见的原因就在于数据处理和概率这两部分。这两者几乎以同样的计算来定义，但原理该如何区分却极其难以理解——学习者大概就是因为这一点而陷入迷茫。

本书的第2部分，将包括数据处理与概率之间区别在内的，易使初学者陷入混乱的概念和枝节剪掉（在保证学术正确性的基础上），而选择统计的估计的本质结构，使读者能够直接地理解。也就是说，第2部分在某种意义上是对达到统计学重要目标的全力冲刺。

2 什么是统计学——描述统计和推论统计

大体而言，统计学由描述统计和推论统计两部分构成。

所谓描述统计，概括地说就是从取得的数据中抽取其特征的技术，起源可以说相当古老。比如，将人口调查作为一种数据来看的话，诞生了“摩西十诫”的摩西时代和罗马帝国时代就已经有了统计。汉朝时代的中国和大化革新时代的日本也有为了征税进行的人口调查和土地调查。

而描述统计学的确切起源在17世纪。

德国学者海尔曼·康令的《国势论》、英国军人约翰·格兰特的《关于死亡表的自然与政治的观察》以及威廉·配第的《政治算术》、爱德蒙·哈雷的《死亡率推算》等就是描述统计的先驱之作。在这些著作中，我们可以看出作者们从有关出生率和死亡率的数据中明确地抽取出了特征，这正是站在描述统计学的立场上使用的研究方法。

此后，作为清晰抽取数据特征的工具，人们又开发出了频数分布表、

直方图等图表方法，还有（各种）平均值、标准差等统计量方法。而现代人正在利用这些方法，对社会和经济状况进行把握，对气象和海洋等环境加以调查。

与此相对的推论统计，将统计学手法与概率理论相融合，对“无法整体把握的大的对象”或“还未发生而未来会发生的事情”进行推测。这是20世纪确立的方法论，从“部分推测整体”的意义上来说，即使称其为前所未有的全新科学也不为过。

就从我们身边来讲，选举速报可以算作是典型的推论统计的成果。在开票率仍在百分数阶段就可以进行“确定当选”的报道，这就是推论统计的功劳。此外，在全球变暖、股票、金融商品和保险商品的定价等问题的预测上，推论统计也是一种不可或缺的工具。

❸本书最重视标准差（*S.D.*）

本书第1部分的前半部分在解说描述统计时，选取了“标准差”为要点说明其意义。所谓标准差，是表示“数据在平均值周边分散程度”的统计量。笔者认为“标准差是统计学最重要的工具”，但很多统计学教科书只笼统地说明了其定义和计算方法。这使得学习者无法切身体会究竟“什么是标准差”。

而如果不能充分领悟标准差，在之后利用正态分布、卡方分布和 t 分布等展开推论统计时，就不能顺利地理解这些究竟是在做什么。这就是很多人学习统计学受挫的原因。

因此，本书从简单内容入手，从各个角度对标准差进行了解说，并引以自信地讲，这种解说在书中所占篇幅之大是其他教科书所无法企及的。说得具体一些，本书不只是在单纯地提示定义，而是利用杂乱的公交车

时刻表和冲浪者等比喻，还有股票指标等指数来使读者形象地理解其意义。而作为附加效果，读者还能理解在判断金融商品优良性上有重要作用的波动率和夏普比率。在21世纪高度发展的金融社会中，这些知识是非常有用的。

④本书几乎不用“概率”

像前言中描述过的，为了将统计学应用于预测，必须在描述统计的方法上加上概率理论。描述统计学中学习过的平均值，在这里以随机变量中期望值的名称再次登场，而数据的标准差在随机变量中也以相同的标准差再次出现。虽然计算方法相同，但被当作不同的概念对待，就很容易使学习者产生混乱（实际上笔者在最初学习的时候也遇到过此问题）。

这种混乱在学习推论统计的进程中会变成大问题，最终导致学习者完全搞不清自己的学习内容是什么。

而之所以会出现这种混乱，原因在于统计和概率之间的微妙差异。统计是观测所得数据的集合，是“对于过去发生的事情的描述”。而概率，是“对于未来将发生的事情的描述”。所以，以“现在”为基准来看，两者意义完全不同。而若是从时间轴的往复来看，则可消除这种差异。

之所以这样说，是因为“未来发生的事情”在经过那一时点后，就变成“已经发生的数据”，而“过去发生的事情”追溯到那一时点之前，就成为“未来发生的事情”。对于这种微妙的既相同又有差异的统计和概率，使用平均值和标准差等相同的计算时，产生混乱也是在所难免的。而且，在推论统计的方法（本书第9章）中，“作为已经过去的事情而取得

的数据恰巧在未来出现”，这看上去像是进行了预测。因此，越是喜欢深思熟虑的人，越会产生“完全不明白到底是在做什么”的迷茫心境。

所以，为了避免这种混乱，本书大胆尝试了“尽量不使用概率”的解释方法。

实际上，即使你只是随意浏览本书也能马上理解其内容，而像其他统计学书籍中必然会出现的组合公式nCk或$P(X=x)$等随机变量的符号，在这里全都不会出现。本书将“数据组中的数据x占全部数据的p%”和“观测数据组中的1个数据时，其为x的概率是p%”两者同一视之。虽然这种做法无视了推论统计学者辛勤建构的理论框架，有些令人心痛，但却是可以避免很多初学者产生混乱的不可或缺的捷径，一般读者应该也不会感到难以接受。

❺以“95%预测命中区间”来说明

但是，只有一处是必须拘泥于“过去和未来的区别”的，这就是作为检验、区间估计的基础思维方式的部分。

笔者在这里展现了其他书中完全没有提到的，自己独有的思维方法，并创造出了“95%预测命中区间”这一词语来表现它。这是笔者关于推论统计学的独家解释，而笔者也可能会因此成为统计学专家批评的对象。但作为使用概率论进行理论决策的专家，笔者愿意在此将错就错（哲学的意义上），并坚持正是这种解释才可以向众多初学者传达推论统计思路精髓的信念。从这个意义上说，这种解说是会给本书带来最大争议的同时，也是本书最大卖点。

⑥几乎不用数学符号和数学公式

本书大胆地削减了概率部分，所以没有使用高中以上的数学知识的必要。其他的统计学教科书，无论如何强调“入门级”，如何强调“简单”，只要触及概率，就无法排除高中以上的数学知识。组合符号、求和符号和随机变量的期望值自不必说，而更难的微积分符号和计算也必然都会出现。

而本书不仅不使用组合符号、求和符号和随机变量的期望值，还完全排除了微积分。这里只会用到初中数学的知识，大概就只涉及一元一次不等式和开方计算。

当然，这样的简化会对全面理解统计学造成障碍。但尽管如此，笔者还是选择了这种方法，这样做既是出于“即使没有数学符号和数学公式，也能传达统计学思维方式的本质部分”这一考虑；更是出于对有“数学过敏症”而无法理解统计学的初学者来说，如果能够理解统计学“不含混合物的本质”，那么也就能理解其他书籍中包括数学在内的统计学全部内容这一意图。

而且，本书尽量使用语言来描述统计学公式。因为不擅长数学符号而回避数理统计学内容，就像因为不懂音符就不听音乐一样，实在是太可惜了。大家应该都赞成“音乐的本质和音符是两回事”这一观点吧。同样，笔者也想呼吁大家接受“统计学的本质和数学符号是两回事”这一观点。

⑦填空式的简单练习题便于自学

想要熟练掌握统计学，必须亲自去做练习题，所以本书每章后均设有练习题。这些练习题是对这一章内容的简单复习，而且格式也都经过

精心设计，只要按照顺序填空，自然就能解答出来，所以请把它们全都做完。

希望所有拿到这本书的读者都能读完它，可以进入统计学的大门。那么，现在就让我们开始吧！

目录