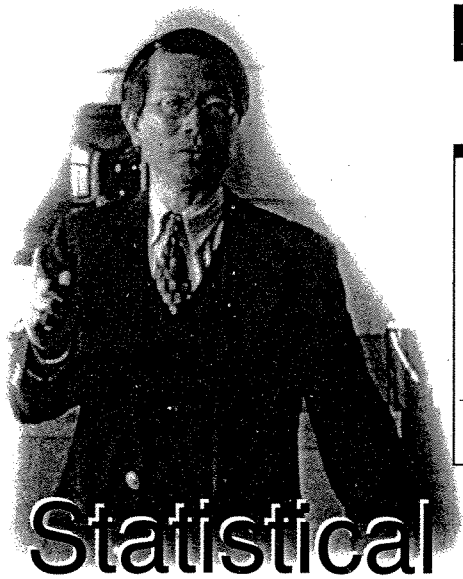
07. 2002
PP. 51-53

共进专栏：心中有数



林共进教授简介：

清华大学（台湾）学士，美国威斯康辛大学博士。现为美国宾州州立大学管理学院教授、台湾综合研究院顾问、美国统计学会院士、国际统计学社院士、西安统计学院荣誉教授。曾获杰出教学奖、杰出研究奖等10余项，著有专业论文百余篇。

林教授先后多次到国内讲学，受到好评。《中国统计》将从本期起，为林教授开设专栏，以期与读者共享他的睿智与文采。

Statistical Data Mining在台湾

（美国）林共进／文

近年来，Data Mining课题无论就教学或市场应用均受到相当的重视，其魅力无论在台湾亦或亚洲地区迅速蔓延，统计与资讯科学领域专家学者皆齐心为Data Mining之推广与理论强化而努力，而台湾即在2001年12月创立了中华资料采矿协会，期能透过学术研究与企业推展互动、理论与应用实务之激荡，全面推展Data Mining。

“厚黑学”一书中曾提及两个方法，补锅法和锯箭法。所谓补锅法指的是有位仁兄锅子破了一个小洞到店里来修补。那店家拿了锅子到后院去狠狠地把小洞敲成个大洞，然后回到店面对那人说：幸亏你碰到我这等内行人，那洞虽看很小，但经我仔细刮一刮以后，发现锅垢很厚，现在我将这大洞给补好，您从此可以安心使用！”那仁兄花了冤枉钱，还感激不已。所谓“锯箭法”指的是有位老哥腿上中了箭，跑来找外科医生。那外科医生把腿外的箭给锯掉，并宣称“医好了！”那老哥急着说“可是我还有半截箭在腿里面呢？”那外科医生不缓不急地说：“那是内科的事，不干我外科的事”。Data Mining普遍认为是Knowledge Discovery in Database(KDD)的一条重要途径。过去二、三十年有两门学科，一为Computer Soience. 二为Statistics. 这两门学科在三十年前规模和大小是差不多的!Computer Soience 走的是补锅路子，大小问题一旦找上Computer Soience，最后没有Computer Soience是不能活的，相反，Statistics走的是锯箭路子，大小问题一旦找上Statistics. 统计者总会说那不是统计问题，你把问题Well defined后再来找我。三十年过去了，Computer Soience愈补愈大，Statistics愈锯愈小，身为统计家我们得深思!

我个人在台湾大学院校或公司企业里教授统计观点于Data Mining之应用，多含括如下内容：Data Mining简介、Data Mining & e-Business(资料采拮与电子化企业)、Complexity(复杂度分析)、Classification & Clustering(分类集群方法)、Genetic Algorithms and Link Analysis(基因演算法与关联分析)、OLAP and other CS approaches(线上即时分析与资讯科学方法)、Tree-classification and kmean method(树分类与K-mean分类方法)、Modeling:Regression & Logistic Regression(建模—回归及罗吉斯回归法)、Modeling:Artificial Neural Network(建模—人工类神经网路)、Modeling: Market Basket Analysis(建模－行销篮分析)、Massive Data Set Analysis(巨型资料集分析)、Time Series Analysis(时间序列分析)、High Dimensional Plots (高维度图形)、Database and Data Warehouse(资料库与资料仓储)及Data Mining Software(资料采拮软体)；而自各内容之探究上，即双向反映了统计观点与分析工具在资料集资讯价值挖掘上之应用与分析角度，以及统计观点与相关资讯工具回向整合强化统计分析之功能需求，主要之Data Mining课题即为：

1.Classification(Supervised Leaning)分类(监督式学习)

2.Clustering(Unsupervised Learning)集群(非监督式学习)

3.Pattern Recognition 状态趋势辨别

4.Association(Correlation) 关连性

5.Modeling 建立模型

6.Estimation 估计

7.Prediction 预测

8.Description 描述

9.Visualization 视觉展现

10.Etc.

因此，谈及Data Mining导入与应用的成功关键，首先即必须对Data Mining方法论有好的了解与掌握，透过有效、正确的分析方法针对问题核心进行分析方能获取价值资讯；第二，须对资料与市场具充足丰富的知识，以准确的资料与市场认识、丰富的经验，所依循之分析观点角度才得以准确、周延，分析结果讯息以专业知识方得以正确决策，然最重要之基本关键出于对企业目标清楚的认识，Data Mining之目的即在于解决企业问题，唯有了解企业目标、掌握核心问题、提供有利资讯、发展企业策略方才是Data Mining所应扮演之角色，而方为Data mining成功的根基。

深入就Data Mining来看，Data Mining为建立在电脑资讯科学(Computer Science Root)与统计科学(Statistics Root)之基础之上，是以资讯科学技术整合统

计分析功能之需求，而运用统计分析资讯工具进行价值资讯之挖掘；拆解Data Mining一字，Data之意含意指了重新设计与维持已存在之资料库，Data之处理即含括了资料收集(Data Collection)、资料准备(Data Preparation)、资料品质(Data Quality)、资料了解(Data Understanding)、资料描述(Data Description)、资料展示(Data Visualization)与资料分析(Data Analysis)等工作，假如要从数百万笔交易记录中，分析某一商品在某一期间的交易情况，做为商品行销策略的拟定参考，此一常见却具高困难度的需求，即反应了对于资讯科技工具协助Data处理之大量需求，如"Cube"之观念与资讯工具诞生，即反应分析者所欲分析之各向角度(即为维度Dimension)与量值共同呈现之需求，就研究发现，资讯科技对于在资料准确的协助上即可提供超过五成的协助。

而就Mining而言，反映了统计分析(Analysis)之精神，即依循问题与需求，以适当之统计观点看待问题、解决问题、发掘资讯，就如假想一探讨所得与负债关系之资料如下图1所示，以各式不同统计观点角度去看待同一笔资料，图形上即可呈现不同的结果：

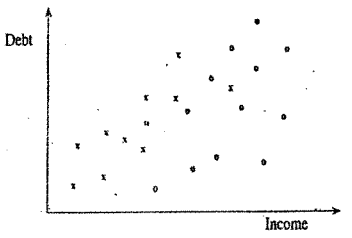
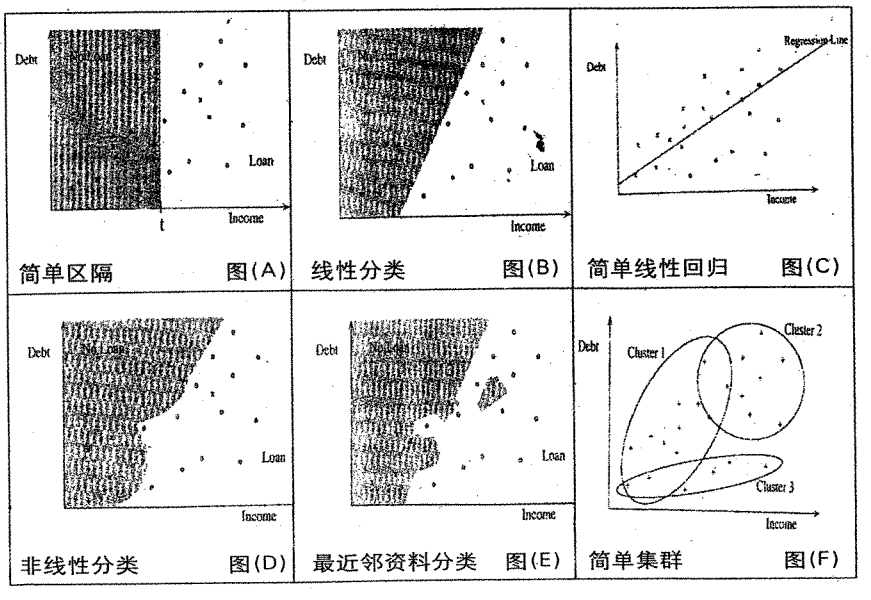


图1

图1所呈现的是一组典型的例子，横座标是Iuchime(始)，纵座标是Debt(负债)"X"表示此观测值是财务状出问题，"0"表示财务状况良好。为便于说明起见，这里只列了23个Cascs.在这情况下，如何根据现有的资料找出一个分类的准则以供未来决策参考呢?几种常用的方法得到的结果如下：图(A)用的是最简单的单一变数分割。如图所示首先选取"Income"作为切割变数而切割变数之值为t时为最优切割点，换言之，整个复杂问题可以用一个简单模型来描述：如果Income高于7是一类，Income值低于7则为另一类。当变数的个数相当大时，由于计算量十分庞大，此简单方法很可能是目前使用最普遍(甚至可能是唯一可行)的方法。图(B)用的是两个变数的线性组合来切割，图(D)用的是非线性方法，所以切割线虽然连续却不是一条直线，图(E)用的是K-meaus Neigbbour法，出来的结果可以很好，但型状怪异，计算量十分庞大，数学模型十分复杂，另外图(C)用的是统计回归(regression)图(F)chetausrering。如以简单线性回归可进行不同的收入下平均负债之预测，透过简单集群分析方法可将不同收入与负债带性质相近的测量对象归属一群体加以分类，统计科学即提供了全面向的角度，使得Statistical Data Mining能透过不同的统计资讯分析工具协助挖掘出有利之各种资讯，深入问题之核心以提供对策，由此，统计观点的基础更是Data Mining价值彰显之基础关键。

而从Statistical Data Mining研究角度来看，仍有许多潜在的议题值得探究，如复杂度问题(Complexity)、巨型资料集(Massive Datasets)、单一回圈运算法(Single Loop Algorithm)、抽样(Sampling)、关连法则运算法(Association Rule Algorithm)及其他，主要，Data Mining现实仍存有许多困难，首先资料库存有数以百计种类、上亿笔资料、与达到terabyte资料量的情况均相当常见，资料库结构复杂，此外，资料库建置资料本身特性如时间序列特性亦是一资料处理上的难题，变数间的复杂关系、网路沟通处理问题的瓶颈、资料获取的系统功能限制与大量记忆体需求都是待解决的问题，事实上，对于较小的资料集上述问题可能不存在，但却相当可能是大资料的严重问题。

主要的困难有二：一是将资料量太大时，由于计算机的容量有限，无法将资料一次叫出来运算(比如：资料无法排序时，如何去找中位数median?)另一则是计算时间即使记忆体的容量够大，一个Terabye的资料用$O(n^2)$的方法来执行，在目前的计算速度也得一百年左右才能运算一次。所以许多现在的方法不适用，即使适用也不代表真能用，所以Siugee-pass Locs-Starage的方法就需要集思广意好好研究开创。针对以上几项问题，我们已有些初步的成果，就大型资料集的数量估计而言，前人提出了Minimax Trees (Pearl,1981)、Stochastic Approximation (Tierney,1983)、Remedian(Rousseeuw & Bassett,1990;Chao and Lin,1993)、Histogram Type(Hurley & Modarres,1995)等解决方法，我们则主要提出"Block"(区集)分析方法，主要将大型资料区分为数个Block，分别计算各区集的统计量，并透过加权整合以产出总统计量结果，此一区集方法透过数理证明，各估计量均具有估计不偏性之性质，且各极限分配均能趋近总统计量之分配并服从常态分配，在大资料集统计量之估计上提供了一简便、确有厚实的学理根据之结果；针对抽样观点而言，我们提出了均匀设计的观念，资料库抽样的好处是相当省时但却有不知如何抽样的问题，此与传统抽样观念是不同的，在资料分析上，我们主要先选取训练样本进行各项分析并以测试样本检视结果，在选取训练样本上，我们发现均匀设计可以在样本均匀度、代表性与学理支持



（上接第49页）

上获得较逐次历史资料与随机抽样为佳，仅在方便性、可行性与计算速度上表现为差，此一结果可提供资料库抽样上更反映整体结果之资料抽取，此两项研究结果均有效地解决了当前Data Mining发展困难的问题。[Ma,Su and Lin(2001)、Mcdermott,Liechty and Lin(2001)、Li,lin and Li(2001)、Lin and Lin(2001)、Tseng and Lin(2001)]。

Data Mining绝对不能闭门造车，专业知识是成功的主要因素。我们这里单纯的提供挖宝的工具，使挖宝工作事半功倍，但在哪里挖？能不能挖到宝则是专业知识必需配合的。

最后，对于未来的KDD发展，Data Mining仍面临几项挑战，首先，更大的资料库、更高维度资料、Over fitting、统计显著性问题、资料与知识变迁、遗失资料与资料杂音、类别间复杂的关系、状态趋势的了解、使用者互动与先验知识以及与其他系统互动问题均影响在未来资料处理、资料分析与资料解读结果，且其将面临不断进行的变化挑战与难题，然Data Mining当前所能挖掘出的价值却能有效提升企业价值、解决企业问题，创造永续发展之远景，此一新兴、热门与与众不同的方式与观点将大幅挑战传统之思维，创造未曾预期之庞大价值，因此Data Mining当是任何学术研究与公司企业体不容再等待的重要工具，Statistical Data Mining身为Data Mining持续发展的关键基础，身为统计工作者的我们又怎能置身事外呢？做的好不如做的早，你还在等什么？我不希望各位补锅，但至少不能再锯箭了。

08.2002 P.49

共进专栏：心中有数

文字与数字的学问

（美国）林共进／文

有所谓《文字的学问》，一直是人类文明重要的宝藏，扮演着举足轻重的角色；亦有所谓《数字的学问》，在人类历史的洪流中常被轻忽。早期文官体系选用人才，以咱们中国老祖宗的科举制度为例，比的是文字的学问；历史上无论是脍炙人口或是震古铄今的巨著，以及那些能名留千古的人，亦以文学大师居多；即使在今日，联合国所谓世界各国的文化水平，主要亦以《识字率》做为衡量的指标，而此识字率即意指文字的能力；相形之下，数字学问并未受到同等的重视。虽然我们口中一直强调数学为科学之母。科技的进步和数字的水平确实有密切关系，但吾等身为数字教育者当扪心自问，数字学问受重视的程度能与文字学问相提并论吗?翻开历史文献，文字墨宝处处可见，而数字学问则多半归列在旁门左道、奇门遁甲中。显然地，数字工作者在起跑点上就已落后一大截，也因此有更大的空间待吾等努力改进。

不识文字者我们称之“文盲”，而不懂数字者至今仍无一个适当的称呼，在此我姑且称之“数痴”。教育虽然愈来愈普及，但摆在眼前的事实是“文盲”日渐减少，而“数痴”却愈来愈多，对此能不多加警惕吗?我们所熟悉的看文识字，有时仍无法领会其中直接的含意，至于看数字而能了解其所蕴含意义的能力，那就更差了!随着资讯的广泛及无远弗届，做一个现代文明人，随时随地都必须与数字接触，我们必须时时刻刻“心中有数”。

统计在数学的学问中扮演着举足轻重的角色，举凡人口、税收、交通、国防、经贸，乃至食衣住行娱乐等，统计几乎无所不在、无所不能。然而，随着科技的进步与电脑的普及，大量资料的收集与储存，在日常生活中处处可见。例如：统计局的调查资料、健保局的医疗资料、出入境管理局的出入境资料、银行及信用卡公司的客户资料与交易资料、超市商店的销售资料等，林林总总不胜枚举。这些随时产生的大量资料，目前常常没有足够的能力去加以分析，探讨资料背后隐含丰富的讯息，相当的可惜。现在是一个资讯革命的时代，因为资料太多所形成资讯障碍，正考验着我们的统计专业。

我们生活上经常有一大串的数字环绕身边，这些数字试图传述某些讯息，而我们有能力去了解这些背后的含意吗?在Kimo网站上，偶然看到下面这段数字：

99：8179，7954，

76269，8406，9405，

7918934，1.91817 。

数字背后的意含为何呢?原来这是一个外甥给他舅舅的一封信。用普通话直接念，叫做：

舅舅：不要吃酒，吃酒误事，

吃了二两酒，不是动怒，就是动武，

吃酒要被酒杀死，一点酒也不要吃。

您看懂了吗?简单的说，数字会说话是不错，但我们要有能力去懂它想说的。还有许多数字的趣味表达，我将它们分列如下：

(1) $7 \div 2$

(2) $2 \leqslant X \leqslant 3$

(3) $40 \div 6$

(4) 二四六八

(5) 0000

(6) $1 \times 1=1$

(7) $1000^0=100 \times 100 \times 100$

(8) 7/8

这些数字表达式相对应的文字如下：

(1)不三不四：$7 \div 2=3.5$ 既不是三也不是四。

(2)接二连三：介于2与3之间。

(3)陆续不断：$40 \div 6=6.6666 \cdots$所以是“陆”续不断。

(4)无独有偶：没有奇数只有偶数。

(5)挂万漏一：一万(10000)少个1。

(6)一成不变：从头到尾都是1。。

(7)千方百计：一千的平方用百来计算它。

(8)七上八下：7在上8在下。

虽然这些都是数字的趣味运用，但也告诉我们：数字中往往隐藏着许多讯息，若不留心、看不懂或不了解，就会让数字失去意义，淹没许多有用的资讯。这好比老天用数字写封信给你，告诉你一些重要大事，而你却无能力去看懂它，机会当然也就让给了那些看得懂的人。“知天”是不容易，但懂得善用数字学问的人一定可以占有较大的优势。

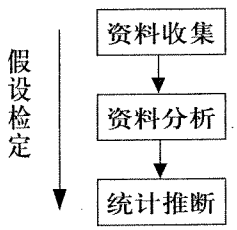
世界各国运用的语言文字虽然不同，但所使用的阿拉伯数字却是统一的。数学家们常常很得意地说：“数字是全世界通用的语言”。那我们为什么不应该在数字学问上多下功夫呢?统计学即是建立在数字上的一门学问，能帮助我们在杂乱无序的数字中找到有用的资讯，了解数字之间蕴含的意义，身为一位数字工作者，更应常以此为共勉。

共进专栏：心中有数

统计思维法

（美国）林共进／文

传统的统计思维基本上包括三个部分："资料收集"、"资料分析"以及"统计推断"。所谓"资料收集"，指的是想了解的整体（称之为母体）资料太庞大，所以透过统计方法去取得有用的样本。在统计上我们发展出"实验设计法"（Design of experiment），及"统计抽样方法"（Sampling survey）两套学问。所谓"资料分析"，指的是将已经取得的资料，加以分析、研究，甚至建立模型，主要工作内容在于Point Estimation（点估计）、Hypothesis Testing（假设检定）、Model Building（建模）以及Forecasting（预测）。从早期的叙述统计（Descriptive statistics）求得平均值、变异数等，以及EDA（Expository Data Analysis），到比较专门的回归分析（Regression Analysis）、时间序列分析（Time Series）、多元统计分析（Multivariate）、无母数分析（Nonparametrics）、可靠性分析（Reliability）等等。针对不同性质、不同假设、不同目的的资料，我们研发出许多不同的工具与方法。最后所谓"统计推断"乃经过统计分析建模之后，可以用来优化（optimization）和预测（prediction），并加以探讨此推断之可靠性如何！



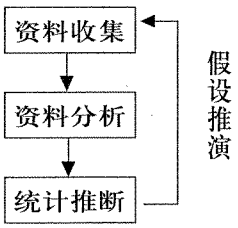
这整套架构，即从资料收集到资料分析到统计推断，长期主宰着统计思维法。基本上从科学演化来看，它是一套"假设检定"的工作。也就是说问题或假设提出之后，才开始整个架构的推动。有了假设必需判别，于是开始收集资料。有了资料收集，开始分析资料。有了成功的资料分析，便可以得到合理的统计推论（接受或拒绝假设）。这乍看十分完整的架构，却有两个重大的盲点，即资料收集的重要性及逆向思考的能力。

首先，我们的统计教育受到数学教育的影响，对统计分析特别投入。往往学生学到的便是解决一个"Given X1……

Xn,find ……" 的问题。换言之，对资料本身的好坏并不多加考虑。许多学校甚至开不出 "实验设计" 以及 "抽样理论" 的课程。反正有Data就一视同仁当作宝。

环境保护的工作同仁最痛恨的错误观念便是多年以前小学自然课本开宗明义所说的 "阳光、空气、水，取之不尽，用之不竭"。事实上，今天的环保工作花了这么多经费和人力不就是希望干净的阳光、空气、水可以取之不尽用之不竭吗？我们统计分析上所谓 "Given X_1……X_n" 便犯了同样致命的错误。资料有好坏，有可靠的、有造假的。可靠的资料不可能从天上掉下来，必需投入大批的人力物力才可以取得的。当统计局局长宣布这一季度失业率为4.5%时，那4.5% 一数很可能是透过成千上万调查人员辛苦的工作成果。调查、普查工作绝对需要大量人力、财力来保障统计数字的正确。

统计的资料收集法是科学方法，借以取得可靠有用的资料而且事半功倍。但这门学问却往往被忽略，不闻不问。反正有Data便可以分析。这话是不错，但如果Data可信度太低，推论结果很可能会误导。我们分析资料的目的是解决问题，而不是作算数习题。



其次是缺乏逆向思考。这假设检定(hypothesis testing)流程，从提出问题假设、资料收集、资料分析，到统计推断，是一套完整系列。但统计推断之后呢？我们拒绝了假设，是不是该有新的假设呢？谁来提出新的假设呢？这所谓的假设推演(hypothesis generation)在目前我们的统计思维中是空白的，更甭论教学了！

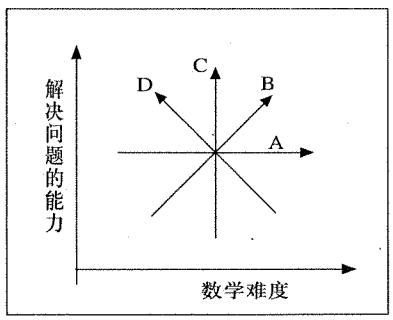
统计分析法是目前统计教育及研究的中心，其基本精神主要在于将观测值分成两个部分：模型(model)与误差(error)。即观测值＝模型＋误差，亦即 $y=f(X_1,\cdots,X_k)+\varepsilon$。曾经在五台山上众多庙宇中，见识一幅对联：

德相非空非有，应随机以恒周；
法身无去无来，住寂光而不动。

此对联不但对句工整，而且将德相与法身之境界描述的十分传神。若将此对联用在统计分析法上，那模型是固定的，"住寂光而不动"，而那误差则是 "随机以恒周"，以随机的方式来配合自然规律的运行。

吾人从古至今便在寻求 "真实"，世上当然存在又真又实的东西，但一般而言真的东西未必实，实的东西也未必真，两者的思维与期望也相当不同。真的东西是 "是非分明" 以数学为例，证明出来的便是真，证不出来的便不是真，只是推测(conjecture)而已。而且真的东西一旦证明出来，便永恒不朽，几百年前证明出来的勾股弦定理到今天还是正确的。实的东西则是 "成败论英雄"，选举每次选赢，股票每次涨，就是事实，没什么定理可言。而且实的东西也是变化无常，一百年前实的东西，一百年后不见得还是实。总的来说，学数学、物理这些理科的人求的是真，而学人文、管理(甚至工程)的人求的是实。两者追求目标不同，文化不同。也可以说，求真的人研究的对象是老天(大自然)，在有生之年变化是相当小的。而求实的人研究的对象是人以及所生活的社会环境，其变化相当大。

所以统计工作者会面对这两难。由于数字学问存在每一个阶层、每一个学科。数理统计延续数学求个真，固然是无可厚非，而应用统计，尤其是用在商学，人文则必需求实。两者其实不抵触，却往往沦于各学派叫骂之中。导致这么一个又真又实的学问，被数学家认为不够真，被使用者认为不够实。统计工作者当多多警惕之。



图一

统计受数学影响很大，数学的确是科学之母。但数学到底不是统计，统计也不应该只有数学。数学应扮演好 "妈妈" 的角色在一旁督导辅助，希望统计能独当一面快快长大，一如当年计算机从应用数学生长而茁壮。图一试图说明这样的一个观念：横座标示数学难度，越向右表示越难的数学；纵座标是 "解决实际问题的能力"，越朝上表示越有能力解决困难。A线所显示是一般数理统计发展的方向，能否解决实际问题并不重要，其努力的方向是处理数学上高难度的问题，于是发展下来越来越抽象，越来越难懂。B线所显示是一个比较平衡、比较健康的发展方向：兼顾了理论(数学)和实际的发展，一般研究计划(尤其是国科会计划)走的便是这个方向。C线是我们实务工作者，包括政府统计、现场统计的工作方向。数学在这里只是个工具，"工欲善其事，必先利其器"，有好工具的人自然解决困难上会占点便宜，但数学在这里只是工具而不是目的。具备一些数学能力是必需的，但主要工作方向是解决问题。传统的应用统计，实务统计或称现场统计均该朝此方向努力。线D则是我们未来工作的重点，如何有效的解决问题，同时把数学的负担抛去，让统计能真正走出一条自己的路来！

一个完整的流程，包括了上、中、下游。以追求科学进步或工业制程为例，上游指的是背景资料的了解、问题的提出、相关测度准则的制定。中游指的是相关资料的收集与整理分析和推断；下游指的是将所推断的结果，转到科学本质或使用者身上。

目前的统计工作干的活是吃力又不讨好的中游部份。对问题的形成与背景不多作了解，对上游工作用一个字叫 "Given" 混过去，盲目的套上一些统计公式和工具，完全不去了解其背后假设(assumption)的相容性。推断出来也不知如何诠释成相关知识，对下游工作用一句 "那是专业知识的工作" 来掩饰我们的无知。等别人把我们的报告转成有用的格式发表出去，我们统计的贡献又被埋在其中，不为人知。吾辈当警惕之！

10.2002
PP. 54-56

共进专栏：心中有数

（美国）林共进／文

处理大型资料的迷思

大型资料愈来愈普遍，此一现象并非偶然，而是必然。以下分别由宇宙发展、生物演化及人类活动等方面简单加以阐述大型资料形成的必然性。这部分没有严谨的学术证明，仅系个人观察的结果，希望能带给大家一些思考。

首先宇宙的发展本来就是愈来愈大，所以期间产生的资料当然也就愈来愈多，而这个“大”与“多”是远超乎我们想象的。根据当今比较得到认可的天文理论，现今的宇宙是在大约200亿年前诞生的。根据这个理论：宇宙一直在膨胀，这种膨胀是没有中心的，从任何一点看都能见到四周的星体远离我们而去。而且有趣的是，距离越远，退行（膨胀）速度越大，这就像一个正在充气的气球，表面上任何一点都会发现别的点正离它而去，而且距离越远，退离速度越大。宇宙的膨胀现象使我们想到，如果我们往回追溯，那么，宇宙会越来越小，就像涨大的气球放气一样，到最后就只剩下一个点了。因此科学家推论，宇宙是从点状宇宙——极小、极小的超微小宇宙，约为10的负34次方公分大小，发生大爆炸而开始膨胀的；此理论经爱因斯坦的一般相对论，及实际观测结果得到学界多数的认可。故此，目前宇宙空间尺度大约为200亿光年，大家知道，光是世界上跑得最快的东西，光速为每秒30万公里，它跑一年距离达9兆4600亿公里，可见1光年是很遥远的距离，而200亿光年即约10的23次方公里。宇宙由10的负34次方公分扩张到10的23次方公里，其间变化何等巨大！累积的讯息何等惊人！且还在不断扩大中。以我们所处的银河系为例，它的直径就有10万光年，而像太阳这样的恒星，银河系里差不多有2000亿颗，随着时间的推演，我们所知道的将愈来愈多。天文学家极力要在这庞大的资料中，找寻你我所在宇宙的过去与未来，天文学的发展已与处理超大型资料密不可分，愈研究就愈发现，从地球到苍穹（宇）、从亘古到永远（宙），实在太大了，人也实在显得太渺小了。

其次，从无生物到生物，以演化的观点来看，生物讯息也是愈来愈多。自古以来，人类对生命之探索一直很有兴趣。大家都知道母鸡生蛋，蛋孵小鸡，树木结出果子，生产种子，种子种植成树木。然而，最初的鸡或鸡蛋从何而来？最初的树木或种子从何而来？在探索生命过程中，140年前达尔文提出的“演化论”思想无疑的带来极大的震撼，引发人们对过去与未来的省思。根据达尔文的研究发现，生物是由低等演化至高等，构造由简单至复杂，所涵盖的资料讯息也是由少到多，以表一为例，体型较大、体制较复杂的高等生物（人、鸡），其DNA的总量和遗传基因总数要较低等生物（滤过性病毒、噬菌体）多得多。有些病毒的遗传基因数目较少，已被科学家发现且清楚排列出来，但像细菌这么简单的生物，遗传基因种类和数目，到目前为止还是弄不清楚，因此人类的遗传基因就更难确定。为什么难以确定？因为期间包含的资讯量太惊人了。以人类而言，人体每个细胞有46个染色体，共约65亿对

基(表一)，亦即平均每个DNA约有1.4亿对基(65亿/46)，其排列组合形成的密码变化极为惊人，因为DNA上的对基有4种，故理论上每个DNA至多可有4的1.4亿次方，亦即10的8千万次方种可能的组合方式，简直是超级天文数字，就算DNA上一小段的基因，其蕴含的资讯量亦极为惊人。照达尔文学派的理论，这是长期演化，由简至繁的结果；所以从生物演化的角度来看，生物自然界亦透露出大型资料走向的必然性，这也是生物科技在资讯技术发达的今日，能大行其道的重要原因。

【表一】 生物细胞内总基数

生物名称	核酸	基的组数(对)	遗传基因的数量(个)
人	双螺旋DNA	6.5×10^9	80,000～100,000
人的精子	双螺旋DNA	3.3×10^9	–
鸡	双螺旋DNA	2.3×10^9	–
果实蝇	双螺旋DNA	2.2×10^8	–
酵母菌	双螺旋DNA	3.7×10^7	–
大肠杆菌	环状双螺旋DNA	4.3×10^6	–
噬菌体T^2	线状双螺旋DNA	2.0×10^5	～100
噬菌体T^7	线状双螺旋DNA	38,000	～30
噬菌体φX174	环状双螺旋DNA	5,500	～8
滤过性病毒	线状双螺旋DNA	23,000	～10
噬菌体MS^2	线状双螺旋DNA	3,300	～3

说明：表内“～”符号是指大约数字。资料来源：知识解码，韦端、倪志坚，2000，晓园出版社。

最后，我们要从人类活动过程来看。根据人类科学家的计算，宇宙诞生于200亿年前，人类的祖先则在500万年前从它和黑猩猩的共同祖先分支，而有历史记载时间不超过5000年。若以24小时代表地球的年龄，那么人类历史活动长度还不到0.1秒，非常短暂，但人类却在这相对短暂时间内，快速发展，超越其他生物，一跃成为地球的主宰，这真是一项奇迹。就以刚过去的百年来看，汽车、飞机、电话的发明，使人类的起居起了革命性变化，太空梭、雷射、电脑、网际网路、核武、基因工程、复制羊都是远超过前人想像的产物，显示人类科技知识与成果，正以等比级数般的速度飞快累积，这期间各领域的资料数据库也不断扩大、扩大、再扩大，此与人类科技发达、活动频繁、分工细密与生活数据化有关，它不但是人类活动下的产物，也是提升人类生活品质的重要资讯来源。以台湾信用卡消费资料为例，20年前几乎只是少数有钱人使用的信用卡，由于所得提高，消费能力增强，再加上银行采用各种奖励措施积极抢攻信用卡市场，2001年底信用卡发卡数达4295万张，较2000年底增长32.5%，平均每人拥有1.9张信用卡，签帐笔数3.4亿笔，金额达7719亿元，占全体民间消费比重12.7%。在如此庞大消费市场诱因下，发行银行及百货商家如何从这一年3.4亿笔消费资料库中，分析消费者个人特征(如所得、职业、年龄、性别、教育程度)与消费特性(如消费金额、地点、日期、时间、商品)间之关联性，进而促销符合消费者需求之商品，提升服务品质，已是许多有心业者积极在做的事。当然，消费活动增加，消费资讯量也跟着增加，大型资料的形成也就理所当然了。

十七世纪犹太籍哲学家史宾诺莎强调理解是自由之道(他有句广为传颂的格言：“不要哭，不要笑，要理解”)。而各种大型资料库中背后隐藏的讯息是帮助我们理解的重要来源。对现在的条件了解得愈多，我们可以将这世界看得更透彻，就愈可以知道未来。人类对宇宙万物的认识虽仍有限，但知识增加的程度正以等比级数般的快速累积，如同开采一座巨大的知识金矿，从最早的用手，用血肉之躯去挖，挖1000年所得仍极有限，然后是用斧、锹、铲去挖，速度较快，挖100年的所得即超越前1000年收获，再演变至用挖土机、怪手开挖，挖10年就超越前100年的努力，今日则是以炸药、自动化机械全面大幅开挖，只要少许时间，即有重大成果。至于如何更有效率、更爆炸性的来开挖这些蕴藏超乎想像丰富的金矿，那就是接下来我们要讨论的主题了。

为了了解大量资料所内含的讯息，专家们提出了指标的观念。所谓“指标”(index)指的是将整体庞大的数据用少数(甚至少到一两个)指标来代表。常见的有平均数(mean)中位数(medium)、变异数(variance)、比率(ratio)等。比方说，两个国家比富裕用的是国民平均所得。比文化用的是认字率。当然用一个数字来涵盖所有资料很难去得到完全的面貌，只能尽力而为。而制订合理的指标是一门很深奥艰涩的学问。多年前我们在制定交通肇事率这种简单的指标便面对许多难题。所谓“交通肇事率”当然是两笔数字的比率。分子很明显的是该年度的交通事故的次数，但分母呢?如果把人口数当分母来除，结果中国大陆的交通肇事率最低，因为人一大堆但没有太多车(所以交通事故少)。如果把车辆数当分母，结果香港、日本的交通肇事率最低，因为车一大堆，但只有偶尔开，平常都搭公共运输(所以交通事故少)。有聪明人士便提出用总行驶的里程数来为分母，结果美国的交通肇事率最低，因为美国高速公路一口气可以开上百里上千里(所以分母可以很大)。如此单纯的指标都不容易，更不必说那些复杂的经济指标或社会指标了。

指标的制定固然不容易，在一般的资料中，指标的计算倒不是太大的问题。但当资料量相当庞大时，指标的计算和统计的分析便面对不同的困难。拿最简单的平均值和中位数为例。我们都知道平均值便是将所有观测值加起来除以其总个数。而中位数便是所有观测值排序，从小排到大，那排在正中央的数便是中位数。这可能是最基本而简单的统计指标。当资料太庞大时，我们面对什么样的困难呢?首先是我们的记忆体不够大，一般的PC(personal Computer)可能有个40GB(1GB=1024MB)，而当我们的资料需要的memory是几百个GB甚至几个TB(1TB=1024GB)时，我们没有能力把所有资料同时叫出来求和，我们只能逐步(sequentially)的做。其次是计算时间，即便我们的memory够大，执行计算的时间可能相当大。以排序而言，目前最快速的排序方法也是OrderO(nlogn)，在资料相当庞大时，则可能需要几年的时间来运作，那么如何不排序又能求得中位数便成了问题。

资料的复杂性基本有两个项目，一个是资料本身的复杂性，一个是使用分析方法的复杂性。资料本身的复杂性我们以其所占的记忆体(memory)大小来区别。如表二所示，当资料量只有10^2项以下，我们称它为“细小型”(tiny)资料，这大约可以记载在一张笔记纸上。当资料在10^4项以下，我们称为小型(small)资料，这大约可以记载在几张纸上。当资料量在10^6项以下，我们称它为中型(medium)资料，纸张已不足以记载，我们得用电脑磁碟片(floppyDisk)来储存。当资料在10^8以下，我

们称它为大型(Large)资料，磁碟片已不够用，我们必须用电脑的主记忆体(HardDisk)来储存。当资料量在10^{10}以下，我们称它为巨大型(Huge)资料，需要数个HardDisk来储存，资料量在10^{12}以及10^{15}我们分别称它为庞大型(Massive)和超大型(supermassive)资料，这时必需用特殊的电脑设备来储存。

【表二】资料大小的分类

资料类别	资料大小(以Bytes计算)	储存方式
细小型	10^2	一张笔记纸
小型	10^4	几张笔记纸
中型	10^6	电脑磁碟片
大型	10^8	电脑的主记忆体
巨大型	10^{10}	数个电脑的主记忆体
庞大型	10^{12}	特殊磁带
超大型	10^{15}	特殊磁带

分析方法的复杂性则以其计算的复杂来区别，我们利用数学上的Order(bigO)来作判别。简单的说，从头到尾把资料走一回便可以得到答案的方法是O(n)。比方说求平均值便是O(n)的演算法。又好比传统clustering的方法必需先求得所有点之间的距离才可以分析。这是C(n，2)的演算法，也就是$O(n^2)$。当然Order越小，算起来越快。如表三显示，常用的分析方法可以简单列成下列几组。划一个scatterplot是$O(n^{1/2})$；计算平均值，变异数是O(n)；求FastFourierTransformation或快速排序(QuickSort)是O(nlogn)；解回归分析，对矩阵分解均是O(nc)；多数clustering方法是$O(n^2)$；多元分析找outlier则是$O(a^n)$。

【表三】分析方法的复杂性(AlgorithmicComplexity)

$O(n^{1/2})$	Plot a Scatterplot
O(n)	Calculate Means,Variances,Kernel Density Estimates
O(n log(n))	Calculate Fast Fourier Transforms
O(n c)	Calculate Singular Value Decomposition of an r x c Matrix;Solve a Multiple Linear Regression
$O(n^2)$	Solve most Clustering Algorithms
$O(a^n)$	Detect Multivariate Outliers

将资料复杂性(表二)和分析方法复杂性(表三)综合在一起我们便可以得到表四。对OrderO(n)而言，针对tiny、small、medium、large和huge分别是10^2、10^4、10^6、10^8以及10^{10}，而对$O(n^2)$则是10^4、10^8、10^{12}、10^{16}、10^{20}，对O(nlogn)则是2×10^2，4×10^4，6×10^6，8×10^8，10×10^{10}。现今假使我桌上有一部Pentium III的PC(目前比较快而普遍的机器)，此类机器每秒钟可执行10^7个运算，所以表四可以除以10^7而得到表五。表五所呈现的是每一个情况的计算时间。我们可以看出当资料量是tiny时，即使计算复杂度为$O(n^2)$，大约千分之一秒可以计算完毕。但同样的计算用在巨大型资料上则需要317000年，换句话说在有生之年是看不到计算结果的。当然工作站(work station)或更先进的高科技计算机可以大幅提高这速度，但要达到可以接受的速度(比方说10多分钟甚至小时)则还需要一段漫长的研究。

所以两个研究重点值得进一步开发：

(一)所有的分析方法(Methodology)与算则(Algorithm)必需标明其order让使用者知道该不该使用，这好比药品食物必需标示其成份让使用者知道该不该使用，其道理是一样的。

【表四】综合资料大小与分析方法的复杂性

n	$n^{1/2}$	n	nlog(n)	$n^{3/2}$	n^2
tiny	10	10^2	2×10^2	10^3	10^4
small	10^2	10^4	4×10^4	10^6	10^8
medium	10^3	10^6	6×10^6	10^9	10^{12}
large	10^4	10^8	8×10^8	10^{12}	10^{16}
huge	10^5	10^{10}	10^{11}	10^{15}	10^{20}.

【表五】表四各种组合之运算时间(PentiumIII 10megaflop)

n	$n^{1/2}$	n	nlog(n)	$n^{3/2}$	n^2
tiny	10^{-6} seconds	10^{-5} seconds	2×10^{-5} seconds	0.0001 seconds	0.001 seconds
small	10^{-5} seconds	0.001 seconds	0.004 seconds	0.1 seconds	10 seconds
medium	0.0001 seconds	0.1 seconds	0.6 seconds	1.67 minutes	1.16 days
large	0.001 seconds	10 seconds	1.3 minutes	1.16 days	31.7 years
huge	0.01 seconds	16.7 minutes	2.78 hours	3.17 years	317,000 years

(二)多数较复杂的方法(O(nlogn)以上)并不适用于巨大型资料，必需研发其替代方案，即所谓的Single-Loop方法。

这两个研究方向在未来几年中将有进一步的结果可期。

大型资料同时也带来另外两个冲击，其一，资料所以如此的大主要原因是因为它是个时间序列(time series)，而时间序列方法在DataMining几乎是空白。其二，资料量这么大因为它什么都收集，换言之，一般而言它常常是population(母体)而不是我们统计课本内所学的样本(sample)。如何分析母体数据对我们而言是个新挑战。如果我们有能力去计算，那么算出来的便是population mean，population median而不是估计值(estimate)。

所以传统的统计思维在这里需要修正，传统的统计思维从"资料收集"，"资料分析"到"统计推论"自成一个完整的体系。但对大型资料而言，资料为什么会如此大，因为它常常是所有的资料，所谓母体(populaton)，而不是我们一般经由抽样技术而得到的样本(sample)，所以没有所谓"资料收集"的问题。在"资料分析"和"统计推论"方面，传统的思维是假设一个可能的机率分布(Distribution)然后透过收集而来的样本来估计机率函数中的参数(parameter)，而种种估计的方法便成了许许多多研究的重点。一旦参数值敲定，从机率分布函数便可有效地统计推断。对大型资料而言，整个母体资料都有了，为什么还需要去假设分布函数呢?是否应该从手中大笔资料直接去决定分布函数，直接作统计推论呢?技术层面上还需进一步开发但这观念应有所改变。

过去我们一直以为"数大便是美"，数据不嫌多。但当科技已进步到大小通吃，资料量几乎可以无限上网时，我们面对完完全全不同的困难。这些困难是等待——被突破，——被解决。

11. 2002
PP. 46-49

共进专栏：心中有数

- 我们统计工作主要的精神便是从纷繁复杂的数据之中摸索出有用的讯息，并赋予其生命意义。
- 在教授统计分析课程时，我会特别强调对数据背景的了解，这和统计分析方法几乎是同等重要。
- 一个成功的案例应该有上、中、下游。上游收集背景资料以提出有意义的问题；中游利用统计方法来解读、推敲答案；下游给予统计结果赋于生命，提供有意义的结论以方便领导决策。

（美国）林共进／文

从铁达尼号沉船资料谈起

数字工作者经常犯的一个毛病，就是常把实际数据当作代数问题来处理。反正所有数据丢入电脑软件包中，跑出一堆有用没用的统计量，再胡乱解释一通。需知道"数字本身是没有意义的"。我们统计工作主要的精神便是从纷繁复杂的数据之中摸索出有用的讯息，并赋予其生命意义。所以数据本身的背景资料是非常重要的。

在教授统计分析课程时，我会特别强调对数据背景的了解，这和统计分析方法几乎是同等重要。毕竟我们是在解决问题，而不是在作代数习题。铁达尼号(Titanic)沉船资料常常是我的第一个案例分析。学生必须上网去查询所有可能得到的背景资料，然后提出问题。最后数据才呈上台面，使用统计方法来解答这些问题。当然统计并非万事皆通，有些问题可以解答，有些则不行。以铁达尼号为例，学生搜集的背景资料包括：造船历史背景、船体资料、乘客背景、航线与沉船经过、以及事后发展。有了这些资料，这才进一步着手了解数据、根据数据，正式开始统计分析。其基本架构详述如下。

前言

铁达尼号搭乘的乘客详细资料当中分别针对每个乘客的属性作记载，我们在其中看到了一些特别有趣的现象，我们针对这笔数据提出这样的问题：乘客资料的各项属性(年龄、性别、所在舱等)中，何者是影响生存与否最重要的因素，是否有特定的属性下生还机率明显不同的情况，我们认为这应该也是历史学家极欲了解的资讯。

历史背景

铁达尼是十九世纪人类科技与工艺的登峰造极之作，同进也是为了满足当时人类的自信与奢华而打造的昂贵产物；她可以说是一个图腾，一个因工程科学的进步而给人类带来幻想与希望的巨大海上图腾，如同巴黎圣母院、埃及金字塔以及中国万里长城那样的伟大与不朽。在她长眠大西洋底近一世纪后的今日，人们依旧歌颂着她所谱写出的传奇乐章，遥想铁达尼号给予当时各社会阶层乘客所带来的美梦与幻想，牢记着她带入海底昂贵的历史教训，缅怀着长伴她左右超过一千五百名的罹难乘客不朽的精神。

英国的白星航运公司的超级豪华邮轮，由汤玛士.安德鲁处理设计，她的船长882英尺9英寸，宽度为92.5 英尺，吃水线以上至甲板的高度为60.5 英尺。排水量6万6千吨，三个推进器，最大航速

24～25节。有20个救生艇，可容纳1178人。总重46329吨(净重21831吨)，动力3000匹马力，有16个防水隔舱，双层船底，因此被认为是不沉之船。船上装潢极具堂皇，可称为融汇了当时科技及精湛艺术结晶的"海上浮宫"。1911年5月31日下水，经过数月的试航后，1912年4月10日作首次航行，船长为E.J. 史密夫，搭载了当时社会上不同阶层的乘客共二千多人，于是日正午离开了南汉普顿港，预定渡过北大西洋，直达美国纽约。

铁达尼号属英国白星海运公司所有，当时为了与对手吉娜海运公司的摩尼速尼亚轮竞争，花了750亿美元，由William Pirrie's Belfast firm设计建造完成。铁达尼号于1912年建造完成，成

为全世界最大的游轮。至少以其头等舱来看，也可称得上是最豪华的游轮了。它的头等舱能容纳905名旅客，二等舱能搭乘564名，而三等舱则有1124名旅客的容量，另外加上900名船上工作人员。不过这次首航旅程上，乘客并未载满。铁达尼号与现在的运输工具大小的比较大略如下。

船体内部的资料

长度 882尺8寸/268公尺(约等于3个草地足球场长度)

甲板 9层(包括最下层甲板)分为A，B，C，D，E，F，G及G下的锅炉房

横梁 92.5 尺/28 公尺

高度 入水深60.5 尺，水面至烟囱175尺(相等于11层楼高)

乘客 329(头等)，285(二等)，710(三等)

全体机员 899人

引擎 2个直接启动转化引擎，4个气缸：30000hp 75rpm，1个低压涡轮机：16000hp 165rpm

螺旋桨 3个：16尺(中间)；23尺6寸(左右)

总重量 46328吨

实际排水量 24900吨

排水量 每日14000加仑

锅炉 29个

燃料 每日825吨煤

气压 215P.S.I.

密室 16个，伸展至F甲板

总运载量 3547人(全船 fully loaded)

救生艇 20艘—16艘木制及4艘折叠(可运载人数1178人)

救生设施 3560件救生衣及49个救生圈。

全速 24海里

乘客背景

1912年，铁达尼号建造代表着一个航向新世界的梦想，2208位来自社会各阶级的各流仕绅乃至于无名小卒登上这艘号称永远不会沉没的巨轮，一趟由英国横跨大西洋的旅程，前往人们当时心目中的新大陆—美国。

铁达尼号建立在一个社会阶级分明的时代。从铁达尼号的船舱等级可以发现，其将等级分为三个层级，即头等舱、二等舱、三等舱。头等舱是最高级的舱别，搭乘的大多为当时的贵族阶级、富翁，带着他们的仆人一同登上船，其中不乏有时常往来大西洋两岸的商贾和他们的管家们一同参与这趟旅程；也包含了许多著名的美国名流，富人，因此他们通常会有随从跟着，例如他们的仆役，护士、女婢，甚至司机等等。

至于二、三等舱的乘客，背景大多来自不同的国家以及各种不同的社会阶级，对他们的大多数而言，这趟旅程几乎都是他们一生中第一次搭船长途旅程的经验，甚至是第一次的旅行经验，这些人大部分都向往着美国这块新大陆，怀着希望和梦想，期待能够在那块土地上，再创一个新的人生；另外有些人则单纯是属于返乡的旅客，一些前往欧洲旅游而有幸在回程搭乘这趟具有历史性的旅程。

航线与沉没点

铁达尼的处女航从英国的南汉普顿先航向法国与爱尔兰，目的地是美国的纽约。预定单程是七天，她最后靠岸是在爱尔兰女王街港。四月十四日深夜，铁达尼号以每小时二十三里高速航行于大西洋上。深夜十一点四十分，当铁达尼航行于北纬四十一度四十六分、西经五十度十四分，在新凡兰岛南部400英里，忽然撞上游离冰山。冰山像锋利的钢刀在客轮左舷撕开一个大裂口，5个防水隔舱顿时破裂，一片混乱。到了4月15日凌晨2点20分，这艘巨型客轮悲哀地结束了短促的生命，在距离纽约港东北2575公里的冰海沉没，1500多名乘客葬身在海底。

沉船经过

铁达尼号这次的擦撞，不是水线以上的船身，而是船头下右舷的底舱部分，擦撞并没有撞出破洞，而是船身擦撞处的几块钢板凹陷，板端铆钉崩脱而向外张开，形成了长达百公尺的一道口子，占全船长的三分之一，涵盖了六舱，前五舱都有水密舱，一舱又一舱灌满，虽有五万匹马力的海上巨无霸，也随着海水涌入而下沉，十一点四十分擦撞，凌晨两点十八分全船沉没，它只在海面上支持了两小时四十二分。

这座海上皇宫首航之日，乘客满载，连船员共达两千一百五十四人。而船上仅仅只有十四艘救生艇和四艘可折小艇，救生艇每艘可载四十人，可折小艇每艘可载七十五人，总共能载八百六十人(实际上只救了六百五十一人，有些救生艇上的人数才四成)，其他的一千两百九十四人(实际为一千五百零三人)，就注定了要眼睁睁随船下沉淹死！船从中间断成了两截，船头沉的很快，然后船尾似乎有站起来一分钟，便悄然沉下海底。

附近的Cunard liner Carpathia，得知铁达尼的困境，立即转向铁达尼，以19海里全速前进，她费时4小时行驶58英里漂着冰山的水域。铁达尼号上2200多人中仅有705人生还。那些搭上救生艇

的生还者在早上时被Cunard liner Carpathia 号搭救。

在1912年4月14日号，发生整个铁达尼号的撞冰山事件，以下我们就时间的经过来叙述整个事件：

11：40PM 发现冰山，船员通知舰桥左转以避开右侧冰山的撞击，经过了37秒以后整艘船才开始向左边转向，然而已来不及避开与冰山的碰撞，冰山水面下的部分撞击铁达尼号并在与其右侧的船壳上发生总计长248英尺的摩擦，造成接下来在这个部分的船壳出现破洞与裂痕，并开始在前方的四个船舱进水，Titanic的主要设计者Edward Wilding估计在发生撞击后的40分钟内，约16000立方英尺的海水灌进铁达尼号的船舱，并判断铁达尼号即将沉没，当Smith船长了解船的进水状况之后，随即发出求救信号，Frankfurt与Carpathia这两艘位于附近的船马上回应的求救的讯号，但是即使距离最近的Carpathia也必须在4小时后才能赶抵现场，以铁达尼号的进水速度已经来不及赶到现场救援，Smith船长了解在救生艇只有一半乘客的乘载量的状况下，船上半数的人在华氏30度的海水中终将罹难。

00:05AM 船员开始放下各艘救生艇。

00:25AM 乘客开始纷纷登上救生艇。

00：45AM 第一艘救生艇Lifeboat7出发，然而就在此时，在北方的海平面上出现一道船舰的光线，Californian正从铁达尼号附近经过，然而尽管船员施放求救的信号火箭，Californian终究没有看到这个信号并逐渐驶离。

1：15AM 随着进水情形更趋严重，乘客纷纷争相登上救生艇。

1：40AM 在大部分救生艇都已驶离的情况下，船上的乘客只好朝甲板上较安全的地方移动。

2：05AM 最后一艘救生艇驶离铁达尼号。

2：17AM 铁达尼发出SOS call，也是她最后一通求救讯号。

2：18AM 船上的灯光熄灭。

2：20AM 船上锅炉的部分由于进水逐渐沉入水中，造成前半部船身几乎垂直于水面，最终沉入水中，至此铁达尼号终于沉没。

事后发展

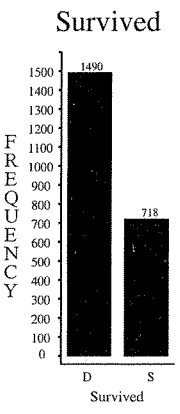
科学家推测，可能是冰山撞凹了约3公尺×9公尺的钢板，使铆钉因此而迸脱，海水便由接缝处灌入。但原因似乎并不如此单纯。比较对钢板边缘的接缝处和船沉没前船员所看到的受损状况纪录，令人觉得难以理解。而且船不到3个小时就沉没，速度也太快了。1985年海洋地质学家拜勒德在约3844公尺深的海底，找到了TITANIC，令人惊讶的是，在海底找到的铁达尼号已断成两截，船头和船尾的距离相当远，各种碎片散置各处，经研究结果发现TITANIC的钢板太脆了，以至于碰到冰山后，钢板没有弯曲，却断裂了，因当时的人没有脆性断裂的概念，通常只要通过抗张强度的测验就可以了，因而造成此一重大悲剧的发生。

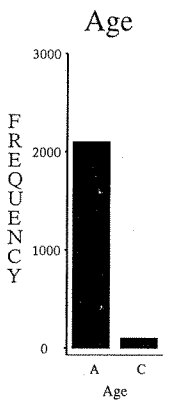
生死	年龄	性别	舱等
D	A	F	1
D	A	F	1
D	A	M	1
D	A	M	1
S	A	M	1
S	A	M	1
S	A	F	1
S	A	F	1
S	A	F	1
S	A	F	1
S	A	F	1
S	A	F	1

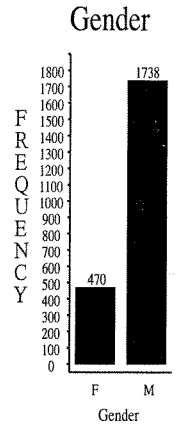
资料描述

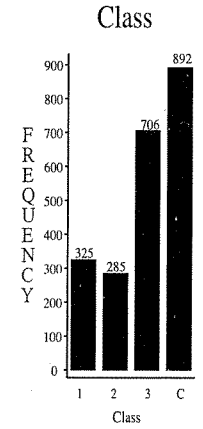
本组资料具有生死、年龄、性别、舱等四个属性(栏位)共计2208笔，下面列出部分的资料：其中符号说明如下：生死栏：S表示存活，D表示死亡；年龄栏：A表示成人，C表示小孩；性别栏：F表示女性，M表示男性；舱等栏：1表示头等舱，2表示二等舱，3表示三等舱，C表示船员。

单一变数直条图









统计分析方法

首先我们采用简单的单一变数直条图，可以看出各栏位大略之分布情形。大略而言：生死比率约2：1；成人小孩比率约20：1；男女比率约3.5：1；三等舱人数为头等舱与二等舱之人数总和。

生死与年龄之交互关系表格

	大人	小孩	总数
存活	661	57	718
死亡	1438	52	1490
总数	2099	109	2208

生死与性别之交互关系表格

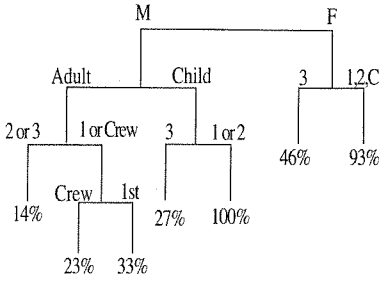
	男性	女性	总数
存活	374	344	718
死亡	1364	126	1490
总数	1738	470	2208

生死与舱等之交互关系表格

舱等	1	2	3	C	总数
存活	203	118	178	219	718
死亡	122	167	528	673	1490
总数	325	285	706	892	2208

接着我们利用交叉分析表来分析其他栏位与生死栏之关系。这是常见的表格。如何看这交互关系呢?我们拿生死与性别之表格为例。在全部2208人之中有718人生存149人死亡比率约1：2。如果生死栏与性别无关，那么不管男性或女性其生死比率都应该在1：2左右。事实上在470位女性之中有344人生存126人死亡，比率为2.7：1。这代表着生死栏与性别栏有着密切的关系。

另外比较先进的方式则是采用树状分类法。如下所示，决定生死最重要变数是性别。所以树状分类将所有数据一分为二：男性在左，女性在右。在右方的女性资料之中，决定生死最重要的变数是舱等：三等舱在左方生存率只有46.6%。而头等舱，二等舱以及女性的船员则有93%的高生存率。男性方面最重要变数是年纪。其中决定男性儿童的最主要因素是舱等：头等舱与二等舱的男性儿童全部生还(100%)，而三等舱的男性儿童则只有27%的生存率。总而言之，女性与孩童的确有比较高的生存率。这与当时的急救措施应该有一定的关系(所以可以大胆假设当时的确是妇孺优先登上救生艇)。另一值得注意的是三等舱。三等舱的生存率似乎不如其他舱等高(女性只有46%，男性儿童则只有27%)。所以也可以合理的推断当时船舱裂洞进入处很可能是在三等舱。简单的统计工具，便可以看出当时的沉船轮廓。当然还有许多先进的统计方法可以使用，在这里仅就几样简单的统计工具作个初步说明。



相对铁打尼号沉船资料分析，另一个极端的例子是"大陆情书"。这一个例子是标准的八股文。报告本身的架构，一次搞定。尔后每一回换汤不换药，照本宣科。就我个人了解，国内许多单位均采用之。我换个比较温和的题目叫"大陆情书"，但用在一般统计报告上，倒也十分传神。此类统计报告，没头没脑，对数字背景资料不求甚解。原文如下：

亲爱的齐：

我们的感情，在组织的亲切关怀下、在领导的过问下，一年来正沿着健康的道路蓬勃发展。这主要表现在：

(1) 我们共通信121封，平均3.01天一封。其中你给我的信51封，占42.1%；我给你的信70封，占57.9%。每封信平均1502字，最长的达5215字，最短的也有624字。

(2) 约会共98次，平均3.7天一次。其中你主动约我38次，占38.7%；我主动约你60次，占61.3%。每次约会平均3.8小时，最长达6.4小时，最短的也有1.6小时。

(3) 我到你家看望你父母38次，平均每9.4天一次，你到我家看望我父母36次，平均10天一次。

以上充分证明一年来的交往我们已形成了恋爱的共识，我们爱情的主流是互相了解、互相关心、互相帮助，是平等互利的。当然，任何事物都是一分为二的，缺点的存在是不可避免的。我们二人虽然都是积极的，但从以上的数据看，发展还不太平衡，积极性还存在一定的差距，这是前进中的缺点。相信在新的一年里，我们一定会发扬成绩、克服缺点、携手前进，开创我们爱情的新局面。

因此，我提出三点意见供你参考：1.要围绕一个爱字；2.要狠抓一个亲字，3.要落实一个合字。

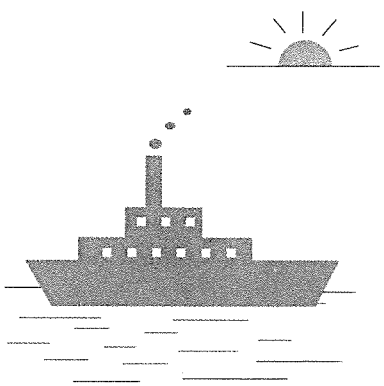
让我们弘扬团结拼搏的精神，共同振兴我们的爱情，争取达到一个新高度，登上一个新台阶。本着我们的婚事我们办，办好婚事为我们的精神，共创辉煌！

你的小惠

综合而言，一个成功的案例应该有上、中、下游。上游收集背景资料以提出有意义的问题，中游利用统计方法来解读、推敲答案，下游给予统计结果赋于生命、提供有意义的结论以方便领导决策。很遗憾，一般统计教育只重视那既辛苦又得不到重视的中游工作。拿到数据，对资料背景不作深入了解便草草下手。统计方法任意套用，当作代数习题解答。错误的方法或使用不当，造成下游工作毫无着力点，得到统计结果只会写八股文——"大陆情书"只是套用轻松的一则美丽错误。数字工作者千万不要把统计当代数习题。面对现实、实事求是、让我们开拓另一个统计新格局。看看下面这段法官与犯人的对话，您作何感想?

法官："你为什么要印假钞?"

被告无辜地说："因为我不会印真钞呀!"



12.2002
PP.46-47

共进专栏：心中有数

统计未来与时俱进的

（美国）林共进／文

许多重要的改变总是在一念之间，总是突破传统，总是跌破眼镜。传统思维历经多年考验，有它一定的准确性和稳定性。但突破和稳定是有冲突的。基本严厉的训练有助于稳定的操作以及保持现状的缓度进步，但大改革是需要打破传统的。许许多多的例子值得我们借鉴学习。

Macintosh 的发迹在于发展个人用电脑。当时的电脑公司，以IBM一个公司独大一面。IBM公司总裁（CEO）认为电脑就是给大公司使用，个人电脑是笑话一则，不会有市场。结果等到Macintosh坐大了，IBM才猛然反省直追。而今个人电脑几乎是人人不可缺少的东西，几乎是人手一部。

想当年最早的字典，不也是只放了一些难字吗？当时的想法也很有道理：简单的字大家都懂，何必放在字典里头？从“难字字典”发展到现今的“实用字典”曾经是一段漫长的历程。

Coke Cola 的故事也相同。当年可乐（Cola）内因含有咖啡因（cocaine）所以只能在酒店里头卖。有人突发奇想一一把可乐装在小瓶子内，如此便可以把可乐送到每个人家中使用。Coke Cola 老板把这想法当作笑话一则：他以为可乐就是要在酒店内喝，不会有人愿意在家里喝可乐。所以象征性地以一美元的权利转移金让此人贩卖瓶装可乐。结果Coke公司却因而大行其道，成了美国文化的一部分。

信用卡的发行当然从 American Express 开始。当然是采用会员制：会员缴会费，所有账目到月底一起算。之后 Visa 和 Master Card 公司更创下信用卡的使用高峰。但信用卡真正的关键转变则是在转移其发行权给各银行。如此虽然减少利润，但便可大大降低信用卡公司本身的风险。此一小小个动作却造成金融革命：各大公司，如 Sony，GM，AT&T... 等等，甚至连大学都可以结合财团信用卡。

不要轻估每一个小小的念头。许多后续影响是我们现阶段不能理解想象的。注意自己的每一小步：一个小念头可以飞黄腾达，一个小岔可以身败名裂。念之！！

知识是累积来的，科学的进步因而是呈指数成长（而非线性成长）。过去10年的成长很可能是相当于之前100年的总和。飞机的发展便是一个简单的例子。资讯科学在过去几年，由于计算机的迅速成长，无论在数据收集、资料储藏、资料库的建立等等，都有空前突破的成果。相形之下，对数据科学家、统计工作者也有着一定深远的影响。时代潮流走到这一步，数据工作者倘使还不能觉悟，与时共进，将被时代给淘汰。我首先举几个简单的例子。

锣鼓：锣鼓的发明，不同于其他乐器。由于敲打起来特别响亮，其使用目的便成为提高人们的注意力。所以热闹的庆典、舞龙舞狮轰轰烈烈的大排场、到早期召开村民大会…等等，锣鼓便扮演着重要的角色，其特点便是大声。可是等到低音喇叭发明以后，锣鼓再是怎么大声也比不上演唱会的低音喇叭响亮。这时候，除非锣鼓能够作个转型，否则

终究会被淘汰，送入历史博物馆内陈列。

绘画：绘画的源起，当然是为了绘的像；不管人物、山水、花鸟到速描。美术科老师教导我们，甚至评价我们画得好不好，都以画得像不像为主要指标。可是等到照相机发明问世以后，绘画画得再怎么像也不可能比照出来的相片像。这时候除非绘画能够作个转型，否则终究会被淘汰。所以绘画渐渐重视那感观、意境、心灵感受，而不一味要求像不像。

邮政：邮政的兴起，主要在于送信。人们借着单文片字来保持联系，交换心得，联络情感。这从古代便一直是人类生命重要的一环。从早期的驿站到今日的邮局都是大事业。可是等到电话、电报发明，人们可以直接对话交换意见，不管相隔多远。接着电子邮件、传真机相继问世兴盛行，人们可以瞬息间传达信件、话语、甚至影像。这时除非邮政能作个转型，否则终究会被时代给淘汰。

我们关心的统计呢？是不是该到转型的时刻了？如果答案是肯定的，那又该如何转型，让我们的统计未来与时济进呢？

统计工作主要是一门数据的科学。在过去数据取之不易，数据量不大的年代，统计工作相当辛苦；但也发展出一套套相当牢靠的方法。如何从抽取有效样本，花少量样本来具体得解读总体？如何预测未来趋势建立决策模？…等等。基本的分析架构大致上着重于：

（1）探讨母体与样本之关系：因为我们手上有的只是代表性的少数样本；

（2）假设可能的概率分配模型；

（3）利用少量样本来估计概率模型内的参数；

（4）一旦参数敲定，整个模型会可以用来作决策与预测。

但时代变了，随着计算机的盛行，资料收集不再是遥不可及。资料储藏轻而易举，再大的资料基本都还有能力可以储存下来。于是我们手上有的不再是"样本"，而往往是母体资料。分析母体资料和分析样本来作推断是不同思维。样本推算出来的统计量是个"估计值"，而母体算出来的便是"真值"。所以如何从母体资料直接下手，导出分配模型，这分配函数不是用假设，而是直接从整体资料归纳出来的。在思维和方法上是大大不同。有了分配模型，我们便可以用来作决策与预测。所以我们必需发展一套处理大型数据的思维与方法。当然，传统方法还是有它一定的可靠性。在新思维尚未成熟之前，恐怕还得用上一阵子。这里仅提供一点前瞻性的看法。

表一　　统计角色的调整

操作执行	战略指导
设计与收集数据以及分析	指导与选择适当的统计方法
教导传统统计方法	设计整体训练课程
与现场工作人员、操作员沟通打交道	与经理、上层领导沟通打交道
顾问统计方法上的疑难杂症	顾问大型整合计划发展
专业走向（可能被忽略）	广泛到受到重视（随时随地都被需要）

另外统计在实务上扮演的角色也应该略有调整。一个比较成功的例子便是在GE (General Electrics, 美国奇异公司)：在过去几年大力推广6 σ (Six Sigma)质量管理方法得到成效后，统计家在GE显然脱胎换骨成为另一种新的地位。全体GE人员或多或少都必须接受统计思维训练。而统计工作者的地位也有所调整。表一列出在Six-Sigma计划前后统计工作的比较。

综合而言，统计工作者的定位有几个大方向的调整，值得我们借鉴与推展：

· 统计工具是大家的，而不是统计工作者的。愈多人懂统计方法，对统计愈有利，统计工作者的地位愈提升。二十多年前，计算机工作人员只会替别人写电脑程式，而现今每个人都可以写一点电脑程式时，计算机人员的地位并没有降低，反而大大提升。

· 统计用在决策性和服务性的项目将愈来愈重要，甚至可能超过在制造性的质量管理上和政府经济性和社会性的工作。

· 统计思维将愈来愈普遍，这可能比统计方法来的更重要。

· 统计工作者将不再扮演着服务业的角色：输入资料、分析一些简单图表、等等。这些工作将来每个人透过简单而普遍的软体（如微软的Excel），自己都能作。相反的，统计工作者提升到顾问的工作，教导其他人如何解读这些图表结果，如何正确选择正确的统计方法，如何作有效的决策和预测。

全国最富的人请了全国最好的建筑师来给自己建造陵墓。三年之后，大富翁问建筑师："全部工程结束了吗？？"

"差不多了"

"还差什么呢？？"

"只差你了。"

如果您同意这些看法的话，剩下就只差您了：您的收入！您的宣导！您的改变！

后　记

感谢您过去六个月的支持与爱护"心中有数"专栏得以顺利刊登。如前所言，许多重要的改变，总是在一念之间：几年前有幸与中国统计出版社谢鸿光社长会晤。在谢社长鼓励之下，决定写篇文章谈一些统计工作的心得与看法。没想到一下笔便停不下来，一口气写了六篇。这中间得感谢谢社长的支持，以及张玉妹小姐给予各方面的协助。同时国家统计局教育中心的王吉利主任与研究所的文兼武所长也都给予相当多的宝贵意见。林林总总点滴在心，只盼对统计工作尽一点心力。本栏所提的一些看法，有些部分比较成熟，有些部分则尚在萌芽阶段。倘使读者有不同看法，也欢迎来函指导。若能抛砖引玉，也不失美意一桩。

编后语：

"共进专栏"到本期即全部结束了。我们虽是编者，但也和读者一样，有着很强的意犹未尽之感。林共进教授贵为知名华裔统计学家，不仅学识深远、文笔精湛，且轩昂卓荦，诚信有佳，虽远在大洋彼岸，每期稿件，传递及时，字正迹清，从未贻误，令我等倍感钦佩。在此，与读者一并向林教授表示感谢，并欢迎读者就读后感受说来写来分享。