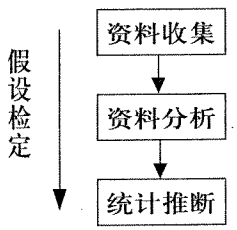
共进专栏：心中有数

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！



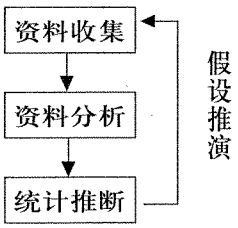
这整套架构，即从资料收集到资料分析到统计推断，长期主宰着统计思维法。基本上从科学演化来看，它是一套"假设检定"的工作。也就是说问题或假设提出之后，才开始整个架构的推动。有了假设必需判别，于是开始收集资料。有了资料收集，开始分析资料。有了成功的资料分析，便可以得到合理的统计推论（接受或拒绝假设）。这乍看十分完整的架构，却有两个重大的盲点，即资料收集的重要性及逆向思考的能力。

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！

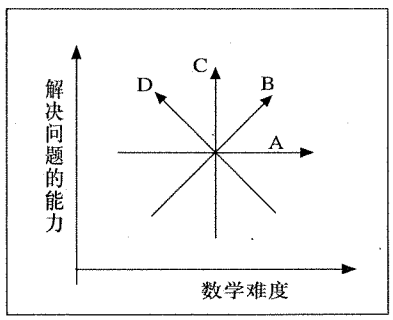
统计分析法是目前统计教育及研究的中心，其基本精神主要在于将观测值分成两个部分：模型(model)与误差(error)。即观测值＝模型＋误差，亦即 $y=f(X_1,\cdots,X_k)+\varepsilon$。曾经在五台山上众多庙宇中，见识一幅对联：

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！