应用数理统计

◎ 胡良平　胡纯严　鲍晓蕾　著

SAS

统计分析系列

中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

应用数理统计

◎ 胡良平　胡纯严　鲍晓蕾　著

SAS

统计分析系列

電子工業出版社

Publishing House of Electronics Industry

北京 · BEIJING

内 容 简 介

本书的内容和写作手法介于“概率论与数理统计”与“医学统计学”或“卫生统计学”之间。可以说，本书是学习“理论统计”与“应用统计”的一座坚实的桥梁，从“待分析的数据是否值得分析”入手，阐释了“应用数理统计”与前面提及的两大类泾渭分明的统计学的区别与联系。书中从试验设计、人为定义、概率分布和抽样分布四大方面介绍了统计计算的基本原理和来龙去脉；然后紧紧抓住最小平方法和最大似然法这两大类拥有多种衍生方法的算法准则，介绍了基于这些准则构造估计方程(即求解统计模型中未知参数的过渡方程)并导出参数估计的方法。为了便于读者学习、理解和正确应用，在必要的统计推导之后，还附有许多有价值的统计应用问题与解析。

本书适合文、理、农、工、商、医、经济、法律、交通、物流等多种学科领域工作的学者、学生、教师、研究人员和医护人员学习和使用。

图书在版编目(CIP)数据

应用数理统计/胡良平，胡纯严，鲍晓蕾著. —北京：电子工业出版社，2015.6
(统计分析系列)
ISBN 978-7-121-26289-0

Ⅰ. ①应… Ⅱ. ①胡… ②胡… ③鲍… Ⅲ. ①数理统计-高等学校-教材 Ⅳ. ①O212

中国版本图书馆 CIP 数据核字(2015)第 126265 号

策划编辑：秦淑灵
责任编辑：苏颖杰
印　　刷：三河市鑫金马印装有限公司
装　　订：三河市鑫金马印装有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：787×1092　1/16　印张：14　字数：352 千字
版　　次：2015 年 6 月第 1 版
印　　次：2015 年 6 月第 1 次印刷
印　　数：3000 册　　定价：35.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010)88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：(010)88258888。

前　言

大约在5年前，电子工业出版社的编辑就热情地邀请笔者编写一本“应用数理统计”。为什么编辑会提出这样的要求呢？理由很简单，因为她清楚地知道，“概率论与数理统计”对于广大实际工作者来说，实在是太难懂了，简直就是看“天书”！而“医学统计学”或“卫生统计学”又颇像“傻瓜照相机”，使用者即使不懂其原理，也能得出一些分析结果来。然而，对于几乎都具有高智商的广大科研人员和医护人员来说，“傻瓜照相机”经常使他们处于迷茫状态。他们不甘心，发誓一定要学好统计学、用对统计学。可是，每当他们拿起“概率论与数理统计”书籍，就被其中的许多怪模怪样的数学符号、晦涩的定理和冗长深奥的公式证明搞得不寒而栗、望而却步；而每当他们拿起“医学统计学”或“卫生统计学”书籍，又被其中的“指挥棒”驱使着，像“双眼被黑布紧紧地包扎着”东碰西撞，生搬硬套，很多场合下得出了错误的结果却浑然不知原因何在。

当时，笔者欣然接受了编辑的邀请。然而，写作大纲却一改再改，迟迟不能开始正式的写作。若有人询问编写大纲究竟修改了多少遍，说实话，就连笔者自己也记不清了。这是怎么回事呢？道理非常简单：编辑发出的邀请，代表了几乎所有（文、理、工、商、农、医等）学科领域内广大读者的共同呼声。对于数学家和统计学家们来说，虽然他们能写出逻辑严谨、理论深奥的“概率论与数理统计”和“高等数理统计”书籍来，但在解决实际的科研课题（尤其是还没有真实的数据）时，却显得无能为力、举步维艰。有时，他们中的一些人只要看到“数据”就眼睛“放光”，根本不考虑数据是否值得分析，就忙着开始进行“统计建模”、“参数估计”和“假设检验”。而对于非数学和统计学专业的广大实际工作者来说，当他们在稀里糊涂套用了一阵子统计分析方法后，真正渴望了解的是统计学的精髓、真谛和原理。上述多领域的读者都对本书寄予了厚望，这就是为什么本书花了整整5年时间才得以面世的根本原因。

本书的10章内容是分别为上述提及的各方人士特别定制的，尽管笔者已使尽了浑身解数，仍感到离广大读者的迫切要求和殷切期待相差甚远。笔者衷心希望广大读者能群策群力，多多提出宝贵的意见和建议，不断充实、修改和完善这部“桥梁式”的统计学著作，使之为人类的进步和发展做出更大的贡献。

本书的问世，得益于无数先辈和广大同仁所付出的智慧和辛勤劳动，还有许多直接和间接为本书做出重要贡献的人。在此，笔者一并向他们表示衷心的感谢！

由于笔者水平有限，书中难免会出现这样或那样的不妥，甚至错误之处，恳请广大读者不吝赐教，以便再版时修正。

胡良平
于北京军事医学科学院
生物医学统计学咨询中心
2015年5月

目　录

第 1 章　待分析的数据是否值得分析

经典的数理统计是分析数据的一门学科。对于给定的一些数据，数理统计会专注于研究下述内容：如何描述其概率分布规律；如何估计总体中某些参数的数值和研究估计方法的性质；如何基于样本对特定的总体参数进行假设检验和研究假设检验方法的特性；如何揭示结果变量依赖多个原因变量(即影响因素或自变量)变化而变化的规律性；如何基于多个变量的取值情况反映多个个体或样品之间的亲疏关系；如何透过现象探索多个变量之间的本质联系；如何揭示多个变量与多个样品之间的内在关联性。然而，数理统计学家在开始实施前述的任何一类数据分析之前，都是有一个最基本的假定的，这就是“待分析的数据是值得分析的”！在实践中，人们在进行数据分析之前，无论分析者是否为数理统计学家，都应首先弄清这个“假定”是否成立。因为任何拟分析的数据不会因为处理数据的人是“数理统计学家”而变得“俯首帖耳”或“值得分析”。

1.1　哪些情形下获得的数据是不值得分析的

1.1.1　人为编造的数据是不值得分析的

当你试图运用统计分析方法处理数据时，首先要问的问题不是选择什么统计分析方法更合适，而是此数据是否值得分析。

【例 1-1】 对于表 1-1 所示资料，有人曾采用三种分析策略(参见后文中张飞给出的处理方法)得出了两个相互矛盾的结论，这一命题被称为“Symposon 悖论”。事实上，对于该命题，具有聪明头脑的人(简称张飞)与具有正常心态和思维的人(简称百姓)会给出截然不同的解读方法和答案。

表 1-1　同时按年龄和性别分层后吸烟与否与是否患肺癌的调查结果

年龄(岁)	吸烟与否	例　数			
		男性：	患	未患	女性：	患	未患
≤40	吸烟		5	5		40	50
	不吸烟		60	55		5	5
>40	吸烟		30	10		5	55
	不吸烟		30	5		5	35

【分析与解答】 张飞这样处理：分别按三种策略分析此资料，策略一，只考察吸烟与不吸烟患肺癌的概率之间的差别；策略二，在按性别分层的基础上，再采用策略一分析；策略三，先按性别分层，后按年龄分层，再采用策略一分析。得到两种自相矛盾的结论：结论一，吸烟有利于健康(策略一与策略三)；结论二，吸烟有害于健康(策略二)。

百姓这样处理：首先提出一个问题，此资料是否值得分析？经过审读，发现它不值得分析，因为它是造假的产物！何以见得？此资料中有两个“疑点”，由此可推测其纯属人为编

造的“调查资料”。疑点之一，调查了≤40岁男性吸烟者10人，居然有一半人患了肺癌；调查了≤40岁女性不吸烟者10人，居然也有一半人患了肺癌。这样的调查结果只有在肺癌病房中才有可能获得，即肺癌病房中有5位患者，另5位是照顾病人的家属或护工。疑点之二，随机调查400位对象，再按“年龄”、“性别”、“吸烟与否”和“是否患肺癌”将其划分成16个小组，每个小组中人数的个位数不是0就是5(由基本常识可知，一个正整数的个位数有0、1、2、…、9共10种可能性)，出现这种奇特现象的概率是极低的，大约为 $P=\left(\frac{2}{10}\right)^{16}=(0.2)^{16}=6.5536\times10^{-12}$。以如此低的概率出现的奇特现象，在实际调查资料中几乎是不可能出现的！

另外，即便表1-1中未出现上述两个疑点，而且，调查资料是在科学完善调查设计方案指导下获得的，张飞的三个分析策略都是不正确的，它们均属于“用单因素分析方法取代多因素分析方法”的错误。正确的分析方法是直接分析三个原因(“年龄”、“性别”、“吸烟与否”)对“是否患肺癌”这个“二值结果变量”的影响是否具有统计学意义。具体地说，可以选用对数线性模型分析，也可以选用多重logistic回归模型分析。若分析目的仅仅是希望考察吸烟与不吸烟患肺癌的概率之间的差别是否具有统计学意义，还可以选用CMH校正的χ^2检验，此法是把性别和年龄视为两个分层因素，对每一层(研究层内吸烟与不吸烟患肺癌概率之间的差别)做一些基本计算，然后再进行合并计算，最后得出汇总分析的结果和结论(即在消除分层因素影响的前提下，分析吸烟与不吸烟患肺癌概率之间的差别是否具有统计学意义)。

事实上，基于表1-1的调查设计，即使在样本含量、样本的代表性和资料的准确性等方面都做得很好，统计分析方法选用得也正确，其结论也不一定就很科学。因为导致一个人是否患肺癌，还有很多可能的影响因素未加以考察，如是否有肺癌的家族史、生活环境(如空气污染情况、饮水的质量、饮食构成、饮食习惯、生活习惯)、锻炼身体情况、职业、劳动强度、家庭经济状况、家庭生活是否幸福美满、人际关系是否良好、心态是否平稳、睡眠质量和每晚平均深睡眠时间长短等。这一切也都与一个人“是否患肺癌”有很大关系，它们被统称为“重要的非试验因素”。在任何一项调查或试验研究中，很关键的一个问题是，是否“找准找全”了对研究目的和主要评价指标有影响的“重要非试验因素”。不仅如此，还必须在尽可能大的样本含量且对总体有极好代表性的受试对象身上准确地获得试验因素、全部重要非试验因素和主要评价指标的数值。所有这一切，都与“数据是否值得分析”密切相关。

由此可知，统计学与纯数学(这里特指数理统计)是有本质区别的！纯数学所研究的数据是有很多隐含的假定的，在这些隐含假定(即所给定的数据是值得分析的)成立的条件下，直接研究数据之间内在的规律性。而统计学则不然，在使用者分析数据之前，必须对那些“隐含假定”逐一进行考证。例如，这些数据是否来自于研究目的相符合的总体(即数据是否具有同质性、样本对于总体的代表性如何)？它们是一个指标的不同取值还是多个指标的不同取值(即是一元资料还是多元资料)？它们是定量的还是定性的(即资料的性质是什么)？它们受到多少个因素的影响(即数据是受单因素影响还是受多因素影响的结果、是什么设计类型下收集的资料)？它们是怎样被收集到的(即是随意或人为选取的还是随机获得的)？所提供的数据对要揭示的问题是否足够多(即样本含量是否充足)？不同组数据之间是否具有可比性(即是否有合理的对照组、组间所受影响是否均衡)？在实施研究或进行试验或调查过程中的质量控制是否严格(即难以控制的众多非试验因素的影响是否被控制在尽可能低的水平上；至少应有

可靠的方法确保所有非试验因素在试验因素的不同水平组中处于基本平衡状态)？只有弄清了上述诸多疑问之后，当答案是“此资料值得分析”时，才能选择具体的统计分析方法(首先是进行探索性统计分析，以了解资料的基本情况，然后再进行正式的统计分析，以便达到合理选择统计分析方法，正确实现研究目的)对资料予以处理。

然而，上述看似烦琐的“循证”过程并非是正确应用统计学的全部“法宝”，它仅仅适用于有了数据之后。在更多情况下，是在仅仅有了研究课题和确定了研究目标之后，就如何能够“多快好省而又科学严谨”地实现既定的研究目标提供“强有力的保证和具有可操作性的做法”，才是统计学的精华之所在！也就是说，统计学的精华是如何制订出科学完善、严谨高效、经济可靠的设计方案。在此方案的指导下，努力做好试验或调查过程中的“质量控制”，才有望获得值得进行统计分析的研究数据。

应当清楚，这里所说的“数据”是广义的，正确的表述应当是“科研资料”。有时，科研资料是纯文本的，更多情况下，科研资料包括原因变量、结果变量、变量的具体取值和变量的专业含义。也就是说，变量、变量的取值(含单位)和变量的专业含义构成了科研资料的三要素。

怎样才能确保所制订出来的设计方案质量高、可操作性强呢？这离不开正确的统计思想的指导，离不开“透过现象看本质”的具体措施保驾护航、离不开“基本常识”的鼎力相助，离不开“各科专业知识”这个坚如磐石的后盾。

【例 1-2】 下面有两组试验记录，试审查此类资料是否值得分析。第一组试验记录见表 1-2，第二组试验记录见表 1-3。

表 1-2　某项试验的试验记录

某药浓度(?)	?	?	?
0	72	97.042	6987.024
25	72	93.431	6727.032
50	72	98.056	7060.032
100	72	96.056	6916.032
200	72	130.694	9409.968
300	72	139.403	10037.020

注：表中“?”处为空白，读者不知道其下方的数据代表什么含义，也不知道它们的单位分别是什么。

表 1-3　某项试验的试验记录

动物编号	性　别	体重(g)	X	Y	Z
1	1	230	3.73	300	16
4	1	225	3.73	600	24～36
5	1	231	3.73	510	16～24
6	0	211	3.73	540	16
12	0	220	3.73	570	24～36

注：X 代表气溶胶浓度(mg/L)；Y 代表翻正反射消失持续时间(min)；Z 代表恢复时间(h)；Z 的取值为“16”的还有 7 只动物，即共有 9 只动物都在 16h 恢复。

【分析与解答】 在表 1-2 中，第 3 列上有 2 个数据的小数部分都是“056”，而第 4 列上有 3 个数据的小数部分都是“032”，这种巧合令人难以置信！在表 1-3 中，有 9 只动物的恢复时间都是 16h，还有 3 只动物的恢复时间的跨度很大，前者过于精确，后者过于不准，都说明结果不真实！如此不真实的试验结果是不值得分析的！

1.1.2 产生于质量控制不严的试验的数据是不值得分析的

有些大型科研课题，需要划分成大约 10 个分课题，每个分课题下面可能还需要再分成 5 个子课题。这样一来，子课题的总数目有 30～50 个。其中，有相当多的观测指标是相同的或在专业上是有联系的，为了探讨这些指标之间的相互关系和依赖关系，人们常常需要把它们合并在一起进行统计分析。很显然，完成这些子课题研究的研究者们的责任心、技术水平等可能相差很多，若事先没有统一的操作规程(SOP)和严格的技术培训，所收集到的试验数据必然是参差不齐甚至杂乱无章的。

【例 1-3】 笔者曾参与评审过一种新药的临床试验项目，该项目由 5 个临床试验中心(通常是 5 所规模相当的医院)共同完成。其中，有一个主要安全性评价指标是某种不良事件发生率(%)，一所医院的试验结果为 22.3%，另一所医院的试验结果为 78.2%，其他医院的试验结果介于这两个数据之间。同一个重要的评价指标在不同医院之间相差如此之大，表明此项临床试验研究的质量控制很差，其整个临床试验数据的准确性非常值得怀疑。因此，质量控制做得不好的任何科研课题(最常见的是多中心临床试验研究课题)所产生的数据都是不值得分析的。

1.1.3 经过错误的方法加工整理后的数据是不值得分析的

【例 1-4】 有些医生常需要采用两种或多种方法对疑似患有某种疾病的患者或患者的标本同时进行检测，以便诊断患者是否确实患了预先推断的某种疾病。某医生收集了同时用 B 超和 CT 检查的 94 例某病患者的检测资料，整理成表 1-4 所示的形式。收集和分析此资料的目的是希望回答可否用 B 超取代 CT。基于表 1-4 的资料，这位医生选择什么统计分析方法可以实现其分析目的呢?

表 1-4 B 超及 CT 检查结果

检查方法	例数			
	轻度	中度	重度	正常
B 超	18	3	3	70
CT	38	7	3	46

【分析与解答】 与这位医生的分析目的对应的统计分析方法属于“一致性检验”或称“Kappa 检验”。然而，该医生将两种仪器检测 94 名患者的结果按表 1-4 的形式整理后，就无法实施一致性检验了。因为表 1-4 属于“结果变量为多值有序变量的 2×4 列联表资料”，本质上是“单因素两水平设计一元定性资料”，就是人们习惯表述的“两个独立样本资料”。针对此资料，只能采用秩和检验，其分析结果只能回答“两种仪器检测结果之间的差异是否具有统计学意义”。事实上，实际收集资料是这样操作的：用两种仪器检测每位患者，每种仪器检测结果都可能出现 4 种结果(正常、轻度异常、中度异常、重度异常)之一，故对每位患者而言，两种仪器就有 16 种可能的组合结果之一出现，即应按“配对扩大形式”来列表，见表 1-5。

基于表 1-5(注意：表中的 f_{ij} 处必须是具体的频数)采用 Kappa 检验，就可回答两种仪器检测结果之间是否具有一致性。

表 1-5　B 超及 CT 检查结果比较(例数)(表 1-4 的正确表达形式)

B 超检查结果	例数				
	CT 检查结果:正常	轻　度	中　度	重　度	合　计
正常	f_{11}	f_{12}	f_{13}	f_{14}	70
轻度	f_{21}	f_{22}	f_{23}	f_{24}	18
中度	f_{31}	f_{32}	f_{33}	f_{34}	3
重度	f_{41}	f_{42}	f_{43}	f_{44}	3
合计	46	38	7	3	94

注：表中的 f_{ij} 为第 i 行第 j 列上的例数，由于原始资料整理错误而无法获知。

1.1.4　不符合特定统计分析方法要求的数据是不值得分析的

【例 1-5】　表 1-6 是某市工业部门 13 个行业 8 项指标的数据，拟采用主成分分析方法确定 8 项指标的样本主成分(综合变量)，若要求损失信息不超过 15%。现分别基于相关系数矩阵、协方差矩阵进行计算，所用程序及结果如下。

表 1-6　某市工业部门 13 个行业 8 项指标的数据

编　号	行　业	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	冶金	90342	52455	101091	19272	82.00	16.100	197435	0.172
2	电力	4903	1973	2035	10313	34.20	7.100	592077	0.003
3	煤炭	6735	21139	3767	1780	36.10	8.200	726396	0.003
4	化学	49454	36241	81557	22504	98.10	25.900	348226	0.985
5	机械	139190	203505	215898	10609	93.20	12.600	139572	0.628
6	建材	12215	16219	10351	6382	62.50	8.700	145818	0.066
7	森工	2372	6572	8103	12329	184.40	22.200	20921	0.152
8	食品	11062	23078	54935	23804	370.40	41.000	65486	0.263
9	纺织	17111	23907	52108	21796	221.50	21.500	63806	0.276
10	缝纫	1206	3930	6126	15586	330.40	29.500	1840	0.437
11	皮革	2150	5704	6200	10870	184.20	12.000	8913	0.274
12	造纸	5251	6155	10383	16875	146.40	27.500	78796	0.151
13	文教艺术用品	14341	13203	19396	14691	94.60	17.800	6354	1.574

注：X_1 表示年末固定资产净值(万元)；X_2 表示职工人数(人)；X_3 表示工业总产值(万元)；X_4 表示全员劳动生产率(元/人·年)；X_5 表示百元固定原资产值实现产值(元)；X_6 表示资产利税率；X_7 表示标准燃料消费量；X_8 表示能源利用效果。

以下是算法 1(基于相关系数矩阵)进行主成分分析的第一部分计算结果：

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.10491252	0.20747090	0.3881	0.3881
2	2.89744162	1.96722608	0.3622	0.7503
3	0.93021555	0.28809329	0.1163	0.8666
4	0.64212226	0.33803813	0.0803	0.9468
5	0.30408413	0.21748637	0.0380	0.9848
6	0.08659776	0.05441338	0.0108	0.9957
7	0.03218438	0.02974261	0.0040	0.9997
8	0.00244178	.	0.0003	1.0000

从这部分分析结果可以看出，取前3个主成分就可以解释全部原始数据86.66%的信息。

以下是算法2(基于协方差阵)进行主成分分析的第一部分计算结果：

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	5.44387E10	4.63867E10	0.8668	0.8668
2	8052070898	7827548141	0.1282	0.9950
3	224522758	136007774	0.0036	0.9986
4	88514983.7	86405779.1	0.0014	1.0000
5	2109204.6	2105915.97	0.0000	1.0000
6	3288.62949	3274.59278	0.0000	1.0000
7	14.0367172	13.9062234	0.0000	1.0000
8	0.13049382	.	0.0000	1.0000

从这部分分析结果可以看出，只需取1个主成分，就可以解释原始数据86.68%的信息。算法2的分析结果看上去似乎更为合理，更好理解。

【分析与解答】 一般而言，根据上述算法2中的这个第1主成分的值，可以对13个行业进行排序和分类。而且，由于这个第1主成分主要包含的是X_7的信息，也就是说，本资料可以根据X_7的信息对13个行业进行排序和分类。

表面上，基于算法2分析本资料后得到了一个令人兴奋的结果！其实不然，在上面的分析和解释过程中，已经犯了一个不可饶恕的错误！

显然，由算法2所得的结果，与算法1所得的结果很不一样。算法1是基于相关系数矩阵进行的主成分分析；算法2是基于方差协方差矩阵进行的主成分分析。按照统计学上的基本常识，这两种算法所得结果不应相差很大，而对本资料而言，其结果却相差十分悬殊。问题出在资料不符合主成分分析的要求上！

严格地说，主成分分析适合用在单组设计多元定量资料上。所谓单组设计，即受试对象或样品来自于同一个总体，他们(或它们)对于研究目的和全部评价指标而言应具有相同的性质。此时，可以考察不同定量指标之间内在的相互和依赖关系，甚至包括它们与隐变量(即主成分变量)之间的关系。而在本例中，13个行业之间存在着本质区别，X_7一个指标就可明显区分出哪些行业是以“燃料消费多”为标志的行业。用这一个指标就可对这13个行业进行排序，这个结果似乎令人感到十分欣慰，但这是十分片面的，甚至是相当错误的，因为它仅能反映哪些行业燃料消耗量大、哪些行业燃料消耗量小，而并不能反映其他任何方面！

所以，进行主成分分析时，务必要强调受试对象或样品应具有同质性，不要盲目地滥用主成分分析。若分析本资料的目的是为了基于给定的8项定量指标进行综合考虑，给13个行业进行排序，建议采用“综合评价”方法，如秩和比法。

综上所述，什么样的数据值得进行统计分析呢？数据必须满足以下四个前提条件：其一，数据是真实的，不是造假的，也不是随意添加或删除某些数据而形成的；其二，所取数据的试验(或调查)设计无严重错误；其三，数据在收集和整理过程中没有出现任何严重的失误或偏倚；其四，数据应满足拟选用的统计分析方法的要求(或前提条件)。

以上四点都是十分重要的，对事先确定的研究目的来说：

第一点关系到研究者或数据分析者的科研道德和科研作风问题。研究者或数据分析者无

论出于什么目的，造假及随意添加或删除数据都是绝对不允许的！

第二点关系到对总体的定义是否清楚，所抽取的样本对于总体的代表性是否很好，样本中个体的同质性是否好，样本是否足够大，所考察的影响因素是否“找准找全”，所观测的指标是否涵盖了与研究目的对应的全部主要指标，指标的测定时间是否合适和测定结果是否准确，课题设计的质量成为所收集的数据是否值得分析的第一道防线。

第三点关系到试验(或调查)过程中可能会受到来自环境、试验条件(仪器设备、试剂等)、研究者和受试者的心理等因素或条件的改变对观测结果造成的歪曲实际的影响，“质量控制”的严格程度成为所收集的数据是否值得分析的第二道防线。

第四点应当注意的是，应根据设计类型、比较类型(如新药临床试验中，还有非劣效性检验、等效性检验、优效性检验，通常的假设检验被称为差异性检验)和统计分析目的，选择最合适的统计分析方法，而不应当盲目套用。例如，有些人针对某病患者与正常人的多项定量指标的数值，进行变量聚类分析和样品聚类分析，这是很不合适的。若研究目的是考察全部定量指标的平均值(即均值向量)在两组之间的差别是否具有统计学意义，就应当选用成组设计定量资料多元方差分析处理资料；若研究目的是希望构造出函数式，以便对新个体究竟属于某病患者还是属于正常人进行判定，就适合选用判别分析。还有人面对多项二值变量及其取值时，采用主成分分析和探索性因子分析，完全忽视了这些多元统计分析方法要求拟分析的资料应当是多元定量资料的前提条件。考察拟分析的科研数据与拟选用的统计分析方法是否吻合成为所收集的数据是否值得分析的第三道防线。

1.1.5　盲目解释基于误用统计分析方法所得到的分析结果是不可取的

有些研究者习惯于脱离实际或无中生有地解释多元统计分析的计算结果。有些多元资料并不适合所选用的某些多元统计分析方法，得出的计算结果是解释不清的。然而，分析者却凭自己对多元统计分析方法的一知半解，盲目解释计算结果，不仅牵强附会，更严重的是对未来的研究工作或政策制订会产生歪曲的指导。

【例 1-6】 原文题目为《过敏性紫癜患儿血清白三烯 B4 白介素 -5 的测定及其临床意义》，测得患儿 31 例，正常儿 27 例的血清 IL-5、LTB4、CRP 的结果如表 1-7 所示。

表 1-7　两组血清 IL-5、LTB_4、CRP 的检验结果($\bar{x} \pm s$)

组　别	例　数	IL-5(pg/ml)	LTB_4	CRP
正常组	27	12.7 ±3.2	17.6 ±5.7	4.75 ±2.85
患儿组急性期	31	53.8 ±4.2	95.3 ±12.0	36.10 ±11.78
恢复早期	31	37.8 ±3.9	45.7 ±10.1	18.35 ±6.43

原作者对此资料做了多种统计分析，其中包括直线相关分析，给出了相关系数的计算结果，见表 1-8，该计算结果可信吗？

表 1-8　血清 IL-5、LTB_4、CRP 的相关系数 $r(n=89)$

	IL-5	LTB_4	CRP
IL-5	—	0.772	0.715
LTB_4	0.772	—	0.735
CRP	0.715	0.735	—

【分析与解答】 原作者给出的表1-8中的计算结果是不可信的！因为所做的线性相关分析是错误的。道理很简单，由表1-7可知，任何两个定量变量的89对数据实际上来自58名儿童，“89”是“例次”而不是“例数”，因为有31名儿童被重复测量了两次。即使都是不同的个体，把他们的数据放在一起进行线性相关分析也是错误的，因为他们是不同质的。正常儿童与处于不同时期的患病儿童，在与所研究疾病有关的指标上的取值大小是不同的，而进行线性相关与回归分析有两个隐含的前提条件：其一，数据所测自的受试者必须具有同质性；其二，所研究的变量或指标在专业上应有联系(例如，若研究人的身高与转氨酶之间的相关关系或回归关系就缺乏专业依据了)。

1.1.6 缺失值过多的数据是不值得分析的

有些课题在实施过程中出现了不可控的严重问题，导致所收集的科研数据严重缺失。例如，在新药临床试验研究中，可能由于未采用盲法分配受试对象或盲法执行得不够严格，导致有相当比例的受试者在试验过程中要求更换治疗方案或干脆脱离此项临床试验研究，这些受试者的试验数据就不完整，较多变量或较多时间点上的变量表现为缺失值。当具有缺失值的受试者数目的比例占总例数的20%以上时，应宣布此临床试验失败了。又例如，在一个调查研究课题中，若发出的问卷调查表的数目为10万份，而收回的调查问卷为1万份，其中各项内容填写有效的问卷为3千份。若仅依据这3千份有效调查问卷，即使采用各种正确的统计分析方法予以处理，得出的结论很可能是无用的，甚至有可能是有严重错误的。因为有效问卷的回收率仅为3%，这些被调查者很可能在全部10万名被调查者中对调查目的或其中某项项目具有很强的倾向性。换句话说，有效问卷回收率低的调查数据是不值得进行统计分析的。

1.2 怎样保证数据是值得分析的

1.2.1 必须制订出科学完善的科研设计方案

由1.1节介绍的内容可知，数据是否值得分析的关键有两条，其一，考察科研设计是否正确，只有在无严重错误的科研设计方案指导下收集的数据才具备值得分析的首要前提；其二，在科研工作实施(试验或调查或从文献中提取信息)过程中，考察质量控制是否严格，是否在可能出现问题的所有环节上都事先想到了，重在防范一切可能影响结果准确可靠性的非试验因素的干扰和影响，一旦出现也有切实可行的对策使其影响降到最低程度。

什么是科研设计呢？在开始一项科学研究之前，需要提出研究目标，为了很好地实现事先制订的研究目标所做的一切考虑和安排，称为该研究项目或研究课题的科研设计方案，包括课题框架设计方案和课题技术设计方案两大类。对于任何一个科研课题而言，要想圆满地完成事先确定的研究目标，要考虑的问题或方面往往是很多的，稍有不慎，一旦落下了某些重要方面或在某些方面考虑不周，就可能导致整个课题研究前功尽弃。因此，对于任何一个科研课题来说，事先制订出科学完善的科研设计方案显得尤为重要。

科研设计的涵盖面非常广，其基本内容概括起来如图1-1和图1-2所示。图1-1是从结构上来划分的；而图1-2是从功能上来划分的，它实际上就是图1-1中“课题技术设计方案”的一种变形。