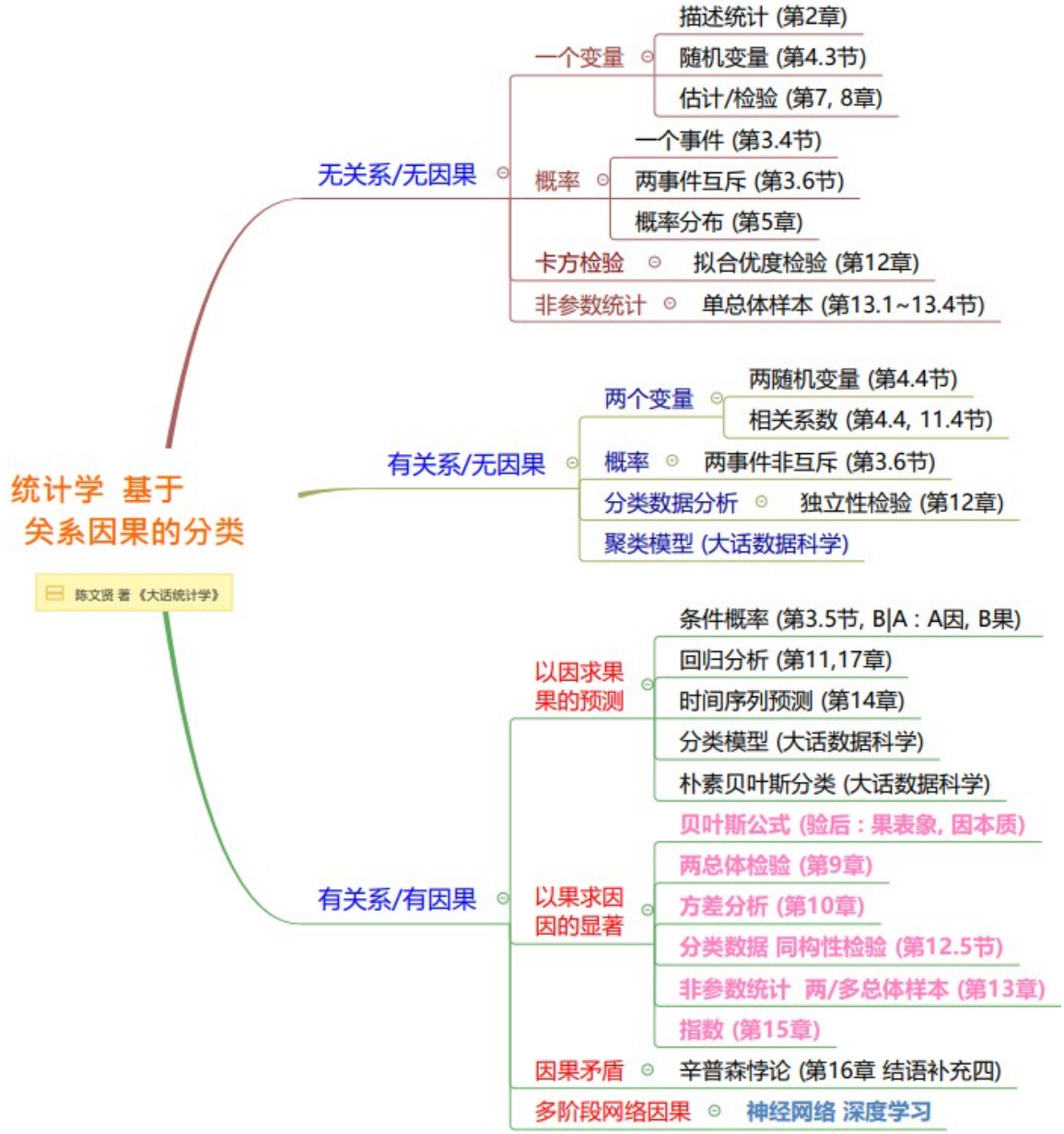
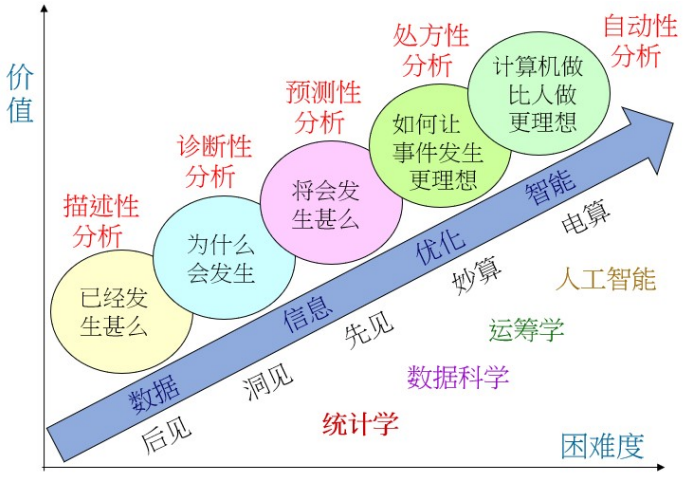
第 1 章 总论

第 14 页 图 1.6



第 19 页 图 1.8



1.7.6 数据集合

数据就像沙石一样，一颗细小的沙石本来没有多大的意义，可是“聚沙成塔”，数据聚集起来加以分析，则会有意义。在计算机中，数据集合起来是记录(record)，记录是个体单位的变量集合，记录集合起来是档案(file)，档案集合起来是数据库(database)。通常档案可以用一个电子表格(worksheet 或 spreadsheet)的型式来显示。如 Excel。

所以，本书所用的中文统计是建立在 Excel 上的一个加载项。

数据电子表格相当于一个矩阵，行(row)代表记录，列(column)代表变量。

表 1.2 数据集合

	变量 variables						
	姓名 name	性别 sex	生日 birth	体重 weight	身高 height	智商 IQ	考绩 rank
个体 记录 record 个案 case	朱小慧	F	750103	50.2	169	132	甲
	李大明	M	780321	76.3	173	128	优
	张一美	F	771123	47.5	158	117	乙
	林二雄	M	790815	84.7	180	122	甲
	赵三英	M	760607	70.6	165	109	丙
	陈四俊	M	770410	62.9	170	116	乙
数据尺度		定类	定距	定比	定比	定距	定序

1.10 统计与大数据 (以下是我对大数据的看法，以后会再补充修正)

大数据(big data)是近几年来热门的课题。到底是只会流行一时？还是可以持续成为重要课题？大数据是国王的新衣？还是石中剑屠龙刀？

大数据有四大特性(4 个 V)：数据庞「大」(Volume)、变化飞「快」(Velocity)、种

类繁「杂」(Variety)、真伪存「疑」(Veracity)。换句话说，大数据有四大特性「大、尔、无、当」：数据量「大」、出「尔」反尔、「无」结构化、以假「当」真。

大数据的个体纪录，也许是一句网络的查询、叙述、或对话，根本没有变量，当然没有数据尺度。但是可以将其语意转换为变量，利用统计学(描述统计)来处理。

后来，有人将最后一个 V 改为 Value 价值，这也是我要问的问题，是唬人的噱头？还是有意义的价值？

统计学的应用步骤如上一节的说明，我们再用图示如下：

(What) (Who) (Which) (Whom) (How much) (Why) (How) (So What)
问题 → 总体 → 变量 → 个体 → 数据 → 分析 → 结论 → 价值
标志 定义 抽样 收集 套模 行动 衡量

套模是套用模型，统计学有多个模型可以使用，尤其是推论统计：方差分析、回归分析等。但是要注意这些模型的假设条件，就是使用限制。所以，在分析阶段，大数据可以用到许多统计模型。

大数据的应用步骤如下：

[Who, Whom, When, Where,
(What) Which, How much (many)] (How) (Why) (So What)
问题 → 数据 → 仓储 → 变量 → 分析 → 结论 → 价值
选择 清理 转换 建模 行动 衡量

大数据要利用非结构性数据库。建模是建立模型，大数据有多个模型可以使用。

统计学的主要观念之一是，当样本容量越大，推断统计的结果越准确。所以统计学加上大数据，应该是非常有用的。如果大数据是在大海里的冲浪，统计学就是游泳池里的游泳，要冲浪一定要会游泳。

统计学主要建立在三个基础：1.总体和样本，2. 变量和数据，3. 误差和变异。

为什么要抽样？统计学与大数据的差别之一 1. 总体普查不可行，大选前的民意调查，总体太大或无法找齐总体；2. 破坏性抽样，质量管理；3. 实务考虑，时间、成本的限制；4. 统计方法的进步，样本量的可靠度增加。

例题 1.7：说明表 1.2 的总体和变量

解答: 总体是 全部泰坦尼克号的旅客和组员。

变量是 性别、身分、死活。

习题

1. 说明下列数据变量的衡量尺度：

(1) 身份证编号 (2) 学生的学号 (3) 出生年月日 (4) 年龄
(5) 性别 (6) 籍贯 (7) 住址 (8) 婚姻状态
(9) 电话号码 (10) 入学年度 (11) 身高 (12) 总成绩

(13) 智商 (14) 鞋子的号码 (15) 血型 (16) 每月房租

(17) 对老师的评鉴 (18) 是否身心障碍

2. 说明上列数据变量的分类。

3. 调查 10 个以上学生的上述数值数据，并且用「中文统计」，储存数据。

4. 举出一个例子，说明下列名词：

(1) 总体 (2) 样本 (3) 变量 (4) 参数 (5) 统计值

5. 为了了解大学生抽烟的比例，抽样北中南各一所大学，每个大学抽样 100 人，调查抽烟人数。

(1) 总体是什么？ (2) 样本是什么？ (3) 变量是什么？ (4) 参数是什么？

(5) 统计值是什么？ (6) 可能的非抽样误差是什么？

6. 我们在 1500 家餐厅放置问卷，调查小费的比率，统计结果是 15%，请讨论这 15%是参数还是统计值？

7. 以下数据是 4 家超级商店的 A 牌饮料的价格：11 元、12 元、9 元、8 元。

下列叙述何者是描述统计，何者是推断统计：

(1) 这 4 家超级商店售价最低是 8 元。

(2) 有 2 家超级商店售价超过 10 元。

(3) 所有超级商店平均售价是 10 元。

(4) 这 4 家超级商店平均售价是 10 元。

(5) 所有超级商店的售价是在 8 元到 12 元。

8. 某大学学生有 12000 人，上学期平均每人购买 9.3 本教科书，平均花费 3520 元。现在抽样一班学生，25 人，平均每人购买 8.5 本教科书，平均花费 3850 元。

(1) 总体是什么？ (2) 样本是什么？ (3) 变量是什么？ (4) 参数是什么？ (5) 统计值是什么？

(6) 如果 25 人，平均每人购买 8.5 本教科书，平均花费 12650 元。你的看法是什么？

9. 下列问题，何者是描述统计，何者是概率问题，何者是推断统计，如何进行统计学的应用步骤：

(1) 昨天股价加权指数涨 100 点

(2) 今天每个顾客平均购买 150 元

(3) 某视昨天八点档收视率为 30%

(4) 好彩头节目「海底捞月」，得大奖的可能性：红绿黄蓝四色球，加上三白球，每次取一球放回，一分钟内，四色球都出现即得大奖。

10. 庙宇的筊杯是一面平、一面凸，掷筊(跋杯)是用两个筊杯，如果出现一平一凸，称为圣杯，表示神明认同。如果我们宣告：出现圣杯的概率是 50%。请问：

(1) 如何统计学的应用步骤，测试这个宣告？

(2) 总体是什么？ (3) 样本是什么？ (4) 变量是什么？ (5) 参数是什么？

(6) 统计值是什么？ (6) 如何使用统计推论来做决策？

11. 请问以下做法为描述统计或推断统计？
    (1) 请找出抽样一个样本中有 75 个生产产品的重量？
    (2) 请计算由宅急配所运送的 100 箱物品的平均重量？
    (3) 请估计下次台北市市长选举的结果？
    (4) 选择去年出生 100 位的新生儿，以此来估计去年新生儿的平均重量。
    (5) 随机选择 100 罐罐装玉米的平均重量，决定是否包装上的 250 公克是否是真的？
12. 请问以下问卷调查，所收集的答案值应是：比率、区间、顺序、名目尺度？
    (1) 请问您是否为台湾公民？
    (2) 请问您的婚姻状况？
    (3) 请问您一天花费在饮食的费用？
    (4) 请问您对美食区餐厅，从最好到最差的排列？
    (5) 请问您上星期花在做功课的时间？
    (6) 请问您认为学校教学设备的总评价：非常好、好、普通、不好、很糟糕。
13. 理文科技学院的学务长想了解在学学生平均上学的通车距离，于是就从全校学生中随机抽样 150 学生，调查他们上学的通车距离，经计算后得到平均的通车距离为 15 公里。请就这个调查回答以下问题：
    (1) 请问本调查中的总体为何？
    (2) 请问本调查中的样本为何？
    (3) 请问本调查中的母数为何？
    (4) 请问本调查中的统计值为何？
    (5) 请问 15 公里为母数或是统计值或都不是？
14. 维维玩具制造商的厂长想知道所生产出来的玩具的不良率，于是就从全厂生产的玩具中随机抽样 500 个玩具检查其状况，调查生产出来的玩具的不良率，经计算后得到的不良率为 0.8%。请就这个调查回答以下问题：
    (1) 请问本调查中的总体为何？
    (2) 请问本调查中的样本为何？
    (3) 请问本调查中的母数为何？
    (4) 请问本调查中的统计值为何？
    (5) 请问 0.8%为母数或是统计值或都不是？
15. 某律师想参与 A 市的议员选举，于是就至 A 市的第一选区注册，他发现此选区总共的选民数为 50,000 人，为了解其选上议员的概率，她从第一选区所有选民中随机抽样 500 位选民，调查这些选民的投票意愿，经统计后会投票给她的比率为 51%。请就这个调查回答以下问题：
    (1) 请问本调查中的总体为何？
    (2) 请问本调查中的样本为何？

(3) 请问本调查中的母数为何？

(4) 请问本调查中的统计值为何？

(5) 请问 51%为母数或是统计值或都不是？

16. 立乐旅行社在旅游行程结束后，都会发问卷调查参与游客的意见，部份问题如下，请问这些问题的答案值应是比率、区间、顺序、名目尺度？

(1) 请问在本次行程之前，你曾经参加过几次本旅行社所举办的旅游行程？

(2) 请问本次行程的天数是否适当？

(3) 请问你认为本次行程的旅馆住宿那项特色是最吸引人：地点、旅馆设施、房间大小、客房服务或是价格。

(4) 请问本行程中每天你可以接受的最大车程时数为？

(5) 请问你认为本次行程的总评价：非常好、好、普通、不好、很糟糕。

17. 请问以下做法为描述统计或推断统计？

(1) 请找出抽样一个样本中有 75 个生产产品的重量？

(2) 请计算由宅急配所运送的 100 箱物品的平均重量？

(3) 请估计下次台北市市长选举的结果？

(4) 选择去年出生 100 位的新生儿，以此来估计去年新生儿的平均重量。

(5) 随机选择 100 罐罐装玉米的平均重量，决定是否包装上的 250 公克是否是真的？

18. 请问以下问卷调查，所收集的答案值应是：比率、区间、顺序、名目尺度？

(1) 请问您是否为中国公民？

(2) 请问您的婚姻状况？

(3) 请问您一天花费在饮食的费用？

(4) 请问您对美食区餐厅，从最好到最差的排列？

(5) 请问您上星期花在做功课的时间？

(6) 请问您认为学校教学设备的总评价：非常好、好、普通、不好、很糟糕。

19. 「单身易得神经病」(2010 年 4 月 8 日 台湾卫生署长说)，还是「神经病易成单身」？这是因果问题，统计学无法解答。但是利用统计学做研究，或接受统计结论的人，应该要问：总体是怎么定义？鳏寡独居是单身？有没有包括年龄区域(多少岁以上才算单身)？神经病的变量如何衡量？如何搜集样本数据？参数是什么？信赖度或误差有多少？