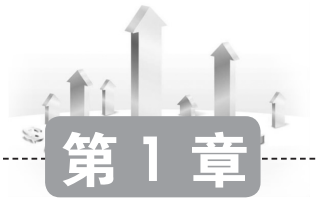


第1章 总 论

不明于计数，而欲举大事，犹无舟楫而欲经于水险也。

——管仲《管子》

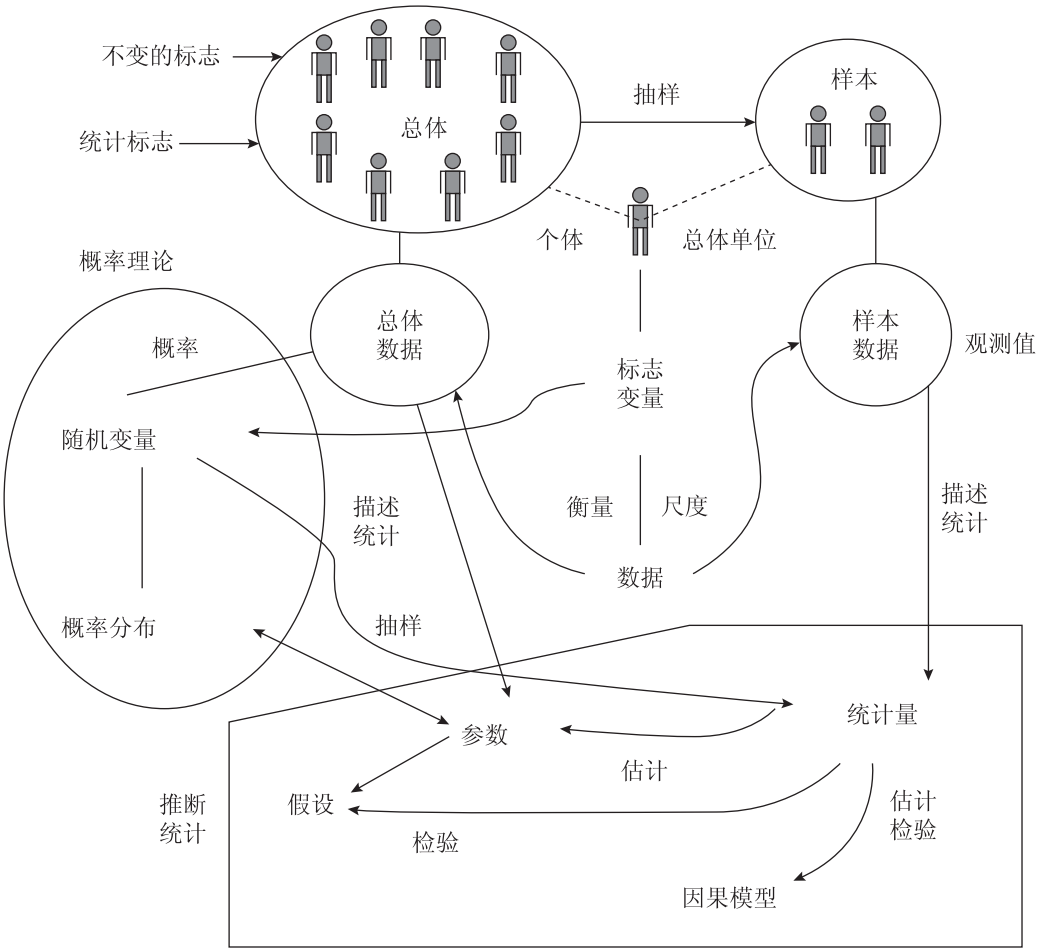
如果你不能测量，你就不能管理。

——戴明（W. E. Deming）和德鲁克（P. F. Drucker）

兵法："一曰度，二曰量，三曰数，四曰称，五曰胜；地生度，度生量，量生数，数生称，称生胜。"

——《孙子兵法·军形篇》

本章重点大纲：


不变的标志
统计标志
总体
抽样
样本
个体
总体单位
概率理论
总体
数据
样本
数据
观测值
概率
标志
变量
随机变量
描述
统计
衡量
尺度
描述
统计
数据
概率分布
抽样
统计量
参数
估计
推断
统计
假设
检验
估计
检验
因果模型

本章概念图

1.1　统计是什么

统计作为一门科学是从 17 世纪开始，起源于国情调查，有政治算数学派、国势学派以及古典概率学派，近代统计学演进到社会统计学派和数理统计学派。现代统计应用在各个领域：管理、工程、医学、农业、经济、社会、生物、气象、政治、军事等各学科。

统计一词，包括：统计工作、统计资料和统计学。本书的重点是统计学。

（1）统计工作：统计的实践，应用统计问题，统计设计、搜集、整理、分析。

（2）统计资料：统计工作获得的各种有关数据和信息，没有数据，就没有统计。

（3）统计学：统计理论、分析数据、选择分析模型、了解计算结果、信息价值。

统计应用最多的是管理，但是文、理、工、商、医、农、政、经、社、法、军等各学科，都是广义的管理。

管理大师德鲁克和质量大师戴明说："如果你不能测量，你就不能管理。"（If you can't measure it，you can't manage it）因为没有测量产生数据（data），就没有管理。有测量过的，才能管理。

上一句话反过来说："你管理过的，一定要衡量。"管理过的衡量是评估绩效，例如：营销广告的绩效、信息系统的绩效、研究发展的绩效、投资的绩效，等等。

所以，管理要根据数据，然后要有绩效衡量：是否达到目的？是否产生价值？

测量（measure）　数据（data）　统计（statistics）　管理　衡量（evaluate）　绩效

（统计工作）　（统计资料）　（统计学）　（统计工作与资料）

人们对于"统计"有好几种看法。有人认为统计是平均数、成数、表格和图形。有人认为统计是处理信息的程序和方法。有人说统计是以部分（样本）推论全体（总体）。还有人认为统计是对付与思考那些不规则发生的样本数据和不确定性的概率事件。也有人称统计是有关决策的科学。这些说法都对。

根据系统理论，一般系统有投入、处理、产出，以达到目标。从统计的投入（input：不规则发生的样本数据和不确定性的概率事件），处理（process：推论、程序和方法），产出（output：平均数、成数、图表），以及目标（objective：决策的科学）的观点来看统计。

数据或资料（data）是投入，信息（information）是产出。

统计学最主要的用途是：叙述已知（抽丝剥茧），与推论未知，以作为决策。统计学包括：①收集数据（定义变量、实验或调查）；②表达数据（表格、图形）；③将数据处

理为信息（百分比、平均数）；④思考概率问题；⑤产生结论预测或决策（以样本推论总体之估计、检验、及因果、关系）。

统计学的目的有以下 4 个。

（1）了解现象：描述统计是了解数据的呈现与性质，集中趋势的代表值或变异程度的离差值；时间序列和指数是了解变化因素和幅度。

（2）推测总体：统计检验和估计是推测总体。

（3）知道因果：两总体检验、方差分析、回归分析是知道因果。

（4）预测未来：时间序列是预测未来。

例题 1.1 是否新的可口可乐味道比较好？（了解现象，推测总体）

在 1985 年，可口可乐公司宣布，更改从 1886 年以来制造可乐的秘密配方。当新的可口可乐上市，《消费者报导》希望解答下列问题：是否新的可口可乐味道比较受欢迎？它跟对手百事可乐的比较又如何？《消费者报导》的研究人员找来 95 位同仁，分别尝试 3 种可乐，而杯子未注明品牌，让他们说出那一杯味道比较好。结果是：百事可乐和新可口可乐的偏好差不多；前两者比旧可口可乐多出一倍（2:1）的偏好。以上结果和新可口可乐刚上市的市场反应大相径庭，市场上反而不大能接受新可口可乐，旧可口可乐在很多地区的销售量还远大于新可口可乐。

例题 1.2 每周工作超时，中风风险大增？（知道因果）

2015 年 8 月 20 日英国权威医学期刊《刺胳针》（The Lancet）刊登伦敦大学学院流行病学基维马吉教授根据 17 份研究调查，涵盖 52 万 8908 名男女，样本追踪时间平均长达 7.2 年，考虑吸烟、饮酒和身体活动程度等因素，研究工作时数与中风风险增加概率的关系。

研究发现，比起每周工作标准工时（每周工作 35 ~ 40h）的人，那些每周工作 41 ~ 48h 的人中风风险高出 10%，每周工作 49 ~ 54h 的人中风风险更是大增 27%，而每周工作 55h 以上的人中风风险增加 33%。研究也发现，即使考虑包括年龄、性别和地位在内等风险因子，长工时也使罹患冠状动脉心脏疾病的风险提高 13%。

例题 1.3 二手烟是否对不吸烟者有害？（知道因果）

20 世纪 80 年代美国加州大学圣地亚哥分校做了一个肺功能试验。这个试验是检查 200 位不吸烟的中年人，他们通常处于不吸烟的环境；另外 200 位也不吸烟的中年人，但是 20 年来，他们经常在吸烟的环境中工作。这两组人同时和吸烟者的肺功能比较。检验结果是：吸二手烟和不吸二手烟的不吸烟者，在肺活量与吐气率两方面，无显著差异。但是在肺部的细部呼吸方面，吸二手烟者比不吸二手烟者，有显著的损害。这个研究建议：

长期暴露在二手烟的环境中，对健康是有害的，会显著降低肺部的细部呼吸的功能。

例题 1.4　民意调查可靠吗？（推测总体、预测未来）

1936 年美国总统大选，共和党的蓝登（Landon）对上民主党的罗斯福（Roosevelt）。《文学文摘》（Literary Digest）根据其读者名册、电话号码簿（当时只有 1/4 家庭有电话）、汽车注册名单、杂志读者名册和俱乐部会员名单，寄出 1000 万份问卷，回收 230 万份，预测蓝登以 57∶43 胜罗斯福。同时间，一个叫盖洛普（George Gallup）的年轻人只抽样了 5 万人，预测罗斯福会赢，被嘲笑太天真，因为样本量太少。结果，罗斯福赢了，得票率 62%！（《文学文摘》不久因此破产了，盖洛普则成为专业的民调专家）时至今日，民意调查的样本量，只要一两千个，就可以推估千万人的总体参数，数万样本可以增加的准确度很少。

1948 年，美国总统大选，共和党杜威（Dewey）与民主党杜鲁门（Truman）竞选总统。选举前，盖洛普民意调查（Gallup Poll）显示，共和党的杜威领先。1948 年 11 月 3 日，《芝加哥论坛报》刊出一个惊人的标题“杜威击败杜鲁门”。杜鲁门反而拿该报纸造势，争取同情票，结果扭转乾坤，杜鲁门赢得选举，当选总统。但是在 1948 年 11 月 3 日，是否真的是杜威领先杜鲁门，也无从查证。（这就是未知的参数）

1.2　统计学的基本概念

定义　总体（population）是要研究的数据的全体对象。

我们要研究公司的薪资所得，则全体员工就是总体。

定义　对“全部”总体进行调查，称为总体普查（population census）。

通常总体普查要花费相当大的人力、时间与金钱。有时要找到全部总体非常困难。对于质量管理的检验，有的是破坏性检验，总体普查以后，全部产品都报销。

定义　总体的基本成份，称为个体或总体单位（unit）。

个体或单位可能是人、动物或商店等。例如：学生、产品、员工、消费者等。

定义　取出总体的“部分”个体，称为抽样（sampling）。抽样出来的个体集合，称为样本（sample）。

定义　样本的数目称为样本量或样本容量（sample size）。

根据抽样的方法，总体分为有限总体和无限总体，如果总体单位的数目是有限的，每个样本抽出后不放回（不重复），且样本量占总体单位数目的比例大于10%，则为有限总体。

定义 标志是总体单位的属性和特征（characteristic）的名称。

标志有不变标志和可变标志。不变标志是总体构成的基础，例如：两个总体检验，分辨两个总体的标志，如性别、地区、处理方法等，是不变标志。可变标志是要进行统计（叙述、概率、推论）的个体的特征。例如：学生的成绩、产品的质量、员工的薪资、消费者购买的品牌、零件是否为良品、选民支持的候选人等。

标志又分为质量标志和数量标志，质量标志是定性的标志，数量标志是定量的标志。上述品牌、良品、支持的候选人是质量标志。成绩、质量、薪资是数量标志。图1-1中的方差分析检验不同教师的学生成绩的平均数是否相等。教师称为“因素”就是质量标志。不同的教师是因素的“水平”，可视为不同的总体，每个总体的教师是不变标志。“观测值”的名称（学生成绩）就是数量标志。

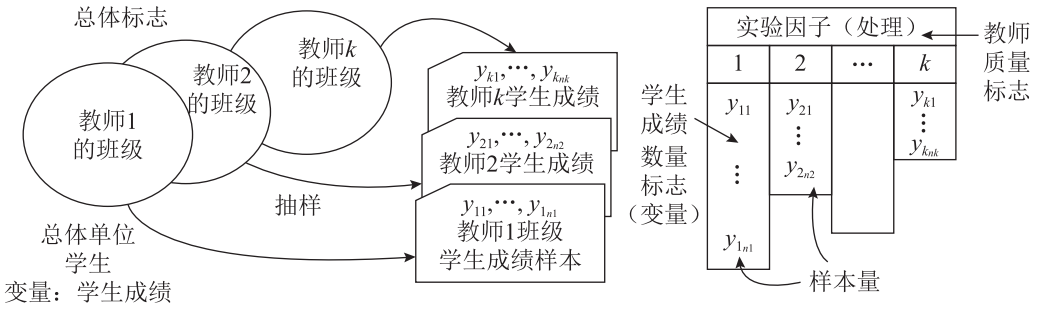


图1-1 质量标志和数量标志

标志有标志名称，如教师、学生成绩；有标志值，如教师1、教师2、学生成绩值。

定义 统计指标（indicator）是说明总体数量特征的名称及数值，例如国内生产总值、总人口数等。

在第2章描述统计，量数（measure）是总体的数量特征也是指标，例如集中趋势量数。

指标又分为：总量指标、相对指标、平均指标和变异指标。完整的指标应具备：时间限制、空间限制、指标名称、统计数值、计量单位等5个构成要素。（下一页说明）

定义 变量（variable）是可变标志和统计指标，前者是总体单位（个体）的变量，后者是总体的变量。

本书的变量，多数是指个体的变量。例如：总体是某一班级（不变标志）的学生，变量是学生的性别、分数、身高等。

变量有数量（quantitative）变量和质量或品质（qualitative）变量。（请见 1.7.2 节）

定义　数据（data）是变量的观测值或计算值，包括：总体的数据和个体的数据。总体的数据经计算产生参数，个体的数据经计算产生统计量。

定义　总体变量数据的衡量值，描述总体特征的数值，称为参数或母数（parameter）。统计的参数有：平均数（均值）、方差、标准差、比例等。

定义　样本变量数据的衡量值，描述样本特征的数值，称为统计量（statistic）。

参数或统计值是：总体或样本的变量数据的一个衡量值，是变量的公式。

如果总体普查不可行，则参数是未知的（固定）常数。

统计学主要名词关联如图 1-2 所示。

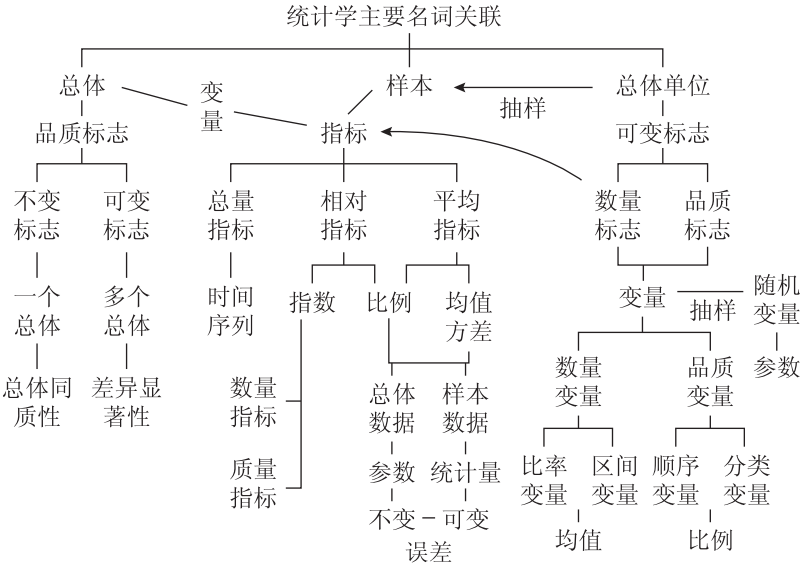


图 1-2　统计学主要名词关联

例题 1.5　大学生菜英文

2012 年，中国台湾报考托业（TOEIC）英文测验的考生的平均总成绩是 539 分，其中大学生的平均总成绩是 504 分，高中生的平均总成绩是 582 分。中国大陆报考托业测验的考生的平均总成绩是 747 分。中国台湾的大学生输给中国台湾的高中生 78 分，中国台湾的大学生输给中国大陆的学生 208 分。为什么中国台湾的大学生是菜英文（菜是差劲不好的意思）？

因为，中国台湾的大学将托业测验列为毕业门槛，大学生一定要考，报考人数有 12.8 万人，这是普查中国台湾大学生的英文。高中生报考是为甄试进大学，英文程度好的才会报考，报考人数只有 3.4 万人。全球每年报考托业的人数达 700 万人，中国大陆的报考人

数不知道。中国台湾报考托业的高中生不能代表所有高中生总体，中国大陆报考托业的考生是否代表所有中国大陆大学生的总体？

例题 1.5 说明完整的指标应具备：时间限制（2012 年）、空间限制（中国台湾的大学生）、指标名称（平均总成绩）、统计数值（504）、计量单位（分数）等 5 个构成要素。其实还有一个构成要素就是容量限制：总体容量和样本容量（每年大学生的人数和报考人数）。

常用的参数和统计量的符号如表 1-1 所示，更完整的参数和统计量的符号，如表 16-2 所示。

表 1-1　总体与样本的符号

指标名称	总体 参数符号	样本 统计量符号	计量单位
容量（数目）	N	n	无单位：整数
（算术）平均数	μ	$\bar{x}$	定距尺度：x 的单位
几何平均数	G	G	定比尺度的百分比
调和平均数	H	H	相对单位：速度，单价
比例	π	p	无单位：百分比
方差	σ^2	s^2	同 x 的单位的平方
标准差	σ	s	同 x 的单位
中位数	M_ρ	M_ρ	同 x 的单位
协方差	$Cov(X, Y) = \sigma_{XY}$	$Cov(x, y) = q_{XY}$	x 单位 × y 单位
相关系数	ρ	r	无单位
指数	P，Q	P，Q	无单位

1.3　统计学的分类

统计学的内容，分成两大类：描述统计（descriptive statistics）与推断统计（inference statistics）。描述统计是探讨总体数据的性质或样本数据的性质，是将数据加以组织分析，并且用图形或数值（指标）表达一些现象，描述某些关心的主题，例如：集中代表值，离散程度，分布型态。指数和时间序列归类在描述统计，主要是将历史数据整理成一个信息（变动比率、趋势值、季节值等指标）。至于时间序列的回归预测，则应该属于推断统计的范围。

探讨总体与样本之间性质的另一方向是推断统计：利用样本数据，加上推论或归纳，得到总体未知参数的估计或检验。推断统计是利用概率分布的理论以及估计、检验、预测等方法，利用抽样的有限数据，来归纳或推断总体的一般性质。在工程、医学、管理等领

域，或在日常生活，我们都无法掌握完全的信息，来知道事实的全部真相。我们都是利用有限和不完全的信息在做决策与推论。所以推断统计是，以有限的信息，来了解与处理，周遭的不确定事件，或者是已经存在但未知的事实（总体参数）。

在描述统计和推断统计之间，扮演串场角色的是概率理论。概率理论是：总体参数已知，利用演绎或仿真，得出样本或事件（总体的部分集合）的概率，再利用随机变量，定义概率分布函数，于是得到抽样统计量的概率分布，然后再将其应用到推断统计。

统计学分类的关系说明如下（第 8，9 章的统计量对应第 2 章的指标公式）：

总体　标志、变量　数据、尺度　指标、参数、表格、图形（第 2 ~ 4 章描述统计）

总体 + 已知参数　概率、随机变量、概率分布、抽样　统计量（第 5 ~ 8 章概率理论）

样本数据　统计量　估计、检验　总体参数或因果关系（第 9 ~ 15 章推断统计）

另外，统计学又可分为理论统计和应用统计。理论统计是概率理论和数理统计，数理统计学则是讨论应用统计背后的理论基础的学科。应用统计分为描述统计和推断统计，是应用在各学科的统计学，例如商业统计、生物统计、农业统计、社会统计等。

图 1-3 是统计学的内容分类和中文统计菜单（选单）。

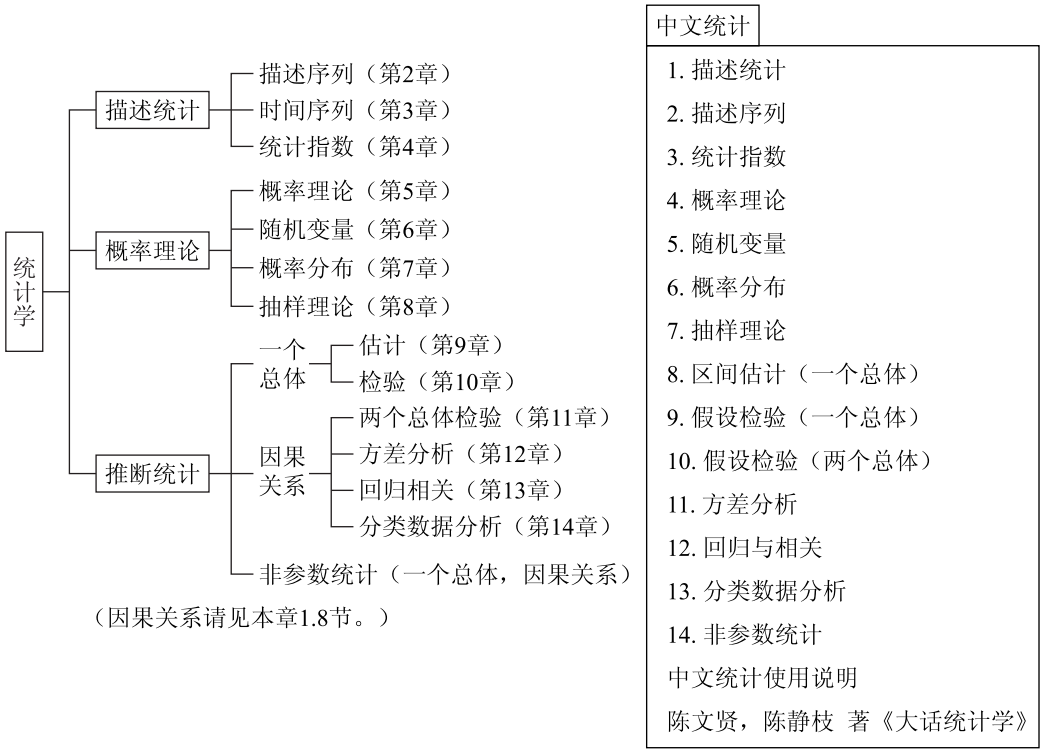


图 1-3　统计学的内容分类和中文统计菜单（选单）

1.4 抽样误差

推断统计主要建立在3个基础之上：①总体参数与样本统计量；②抽样随机性（概率理论）；③抽样误差。总体参数是推断统计的目的（What，要做什么），概率理论是原因（Why，为什么），抽样误差是方法（How，如何做）。这是5W1H（What，Why，Who，When，Where，How）中，最重要的3个：Know What，Know Why，Know How。

因为抽样数据只是总体数据的一部分，所以抽样数据计算出来的统计值（例如：样本平均数），与总体数据计算出来的参数值（例如：总体平均），会不相等，其差异是误差。

误差是“样本数据统计值与总体的参数值”的差（绝对值），误差分为抽样误差和非抽样误差。

定义 抽样误差（sampling error）是因为抽样的“随机性”所造成的误差。

抽样误差是因为不同组的样本而有不同的误差，是因随机性（random）而产生的误差。衡量抽样误差和下列名词有关：样本标准误、残差、置信度、显著性水平（允许误差的概率）、第一类错误（p 值）。影响抽样误差的有：抽样的样本量（样本量越多，抽样误差越小）、抽样的方法（分层抽样或实验设计等方法）、选择的统计量（请见第9章）。推断统计最主要是使变异（误差的平方）减小，使检验结果显著。

抽样误差是推断统计的基础。

（1）描述统计的平均数：每个变量值和平均数（均值）之差（误差）的平方和最小。

（2）大数法则：样本容量越大，抽样均值和总体均值的误差（标准误差）越小。

（3）区间估计：置信度是控制“置信区间不包含总体参数”的误差。

（4）估计之标准误差：抽样误差，越小越好，只有这样估计才会越准确。

（5）假设检验的显著：第一类错误（p 值）不超过显著性水平。

（6）检验的检验值：统计值与参数值之差除以标准误差，所以，标准误差越小，检验值就越大，才有检验的显著结果。

（7）回归分析的最小二乘法：每个变量值和回归预测值之差（残差）的平方和最小。

（8）方差分析：总（误差）平方和 = 组间（差异）平方和 + 组内（误差）平方和，检验各组的均值是否不相等（显著），要看各组间差异越大，各组内越小。

（9）分类数据分析：以样本值和理论值之差（误差）的平方和，检验一个变量的概率，或两个变量的独立性。

抽样误差 { 估计的标准误差 / 检验的第一类错误 / 方差分析的变异来源 / 回归分析的最小二乘法 / 分类数据的误差平方 } 推断统计的公式（统计量）

定义　非抽样误差（non－sampling error）是在抽样过程中，由于人为错误而造成的误差。

非抽样误差是因“人”（研究者或受测者）而产生的误差。非抽样误差包括以下几种。

（1）选择样本抽样框（sampling frame）的错误，样本不能代表总体。抽样框是抽样个体的名册，用来抽选样本的个体，如：电话簿名册、毕业纪念册、会员名单等。

（2）选择抽样方法的误差，选择抽样的方法包括：便利抽样，以最方便的方法选择样本，如街头调查、利用学生作实验；自发性响应样本，样本以自动应答的方式取得，如电视电台的叩应（call－in）或报纸、杂志、博客、BBS 的来应（write－in），其回答的样本都是有心人；还有例题 1.5 报考托业的中国台湾高中生，以上都不能代表总体。

（3）取得数据的误差：问卷设计得不好，问题敏感，受访者不愿答或故意答错，回收率低的误差（未回应的误差，邮寄问卷回收率低，大多数会有此问题）。

（4）量测误差：记录数据的误差（记载错误或笔误）、计算数据的误差（输入错误或计算错误）等。

非抽样误差要在实验与调查的设计上考虑，注意抽样对象是否有代表性，尽量避免这项误差。注意问卷的设计。增加样本量，并不能减少非抽样误差，如例题 1.4。

抽样误差是得到样本数据之“后”的差异。非抽样误差是得到样本数据之“前”的错误。推断统计学是考虑“抽样误差”。统计工作和统计资料，要考虑“非抽样误差”，如图 1-4 所示。

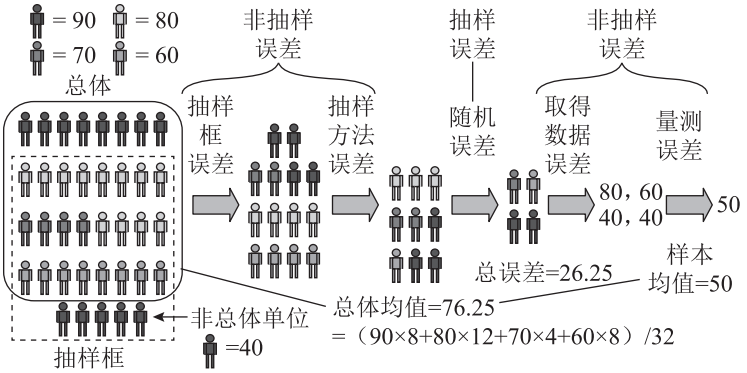


图 1-4　抽样误差与非抽样误差

统计学是“数据进，信息出”（Data in，information out）；如果有非抽样误差，则是“垃圾进，垃圾出”（Garbage in，garbage out）；如果用错统计方法，就是“数据进，垃圾出”（Data in，garbage out）。

赵民德（中文参考书目［7］）：“（统计学）方法的一个最大特征是：统计学家深切地体会到误差的存在，并积极地面对可能的误差，而使得经过这套方法所导出的结论，其因误差而产生的暧昧减少。统计学的方法并不能无中生有，但它的确致力于尽量滤去误差，而得到传统方法所不能得到的结论。误差如水，真相若石。水落，所以石出。如果水中原本无石，水落当然也仍然无石。统计的方法之所以大行其道，是因为误差是近代生活的一部分。一切人类所搜集的数量数据中，其不包含误差者百不得一。社会越进步，则所需要搜集与分析的（包含着误差的）数据也越多。而我们越尝试地去以无误差的概念去推演结论，所得的结论里便越充满误差。而统计方法，因为能正视误差的存在，反而可以得到更合理的结论。……统计方法：围绕着包含了误差的数字，所作的种种精巧的努力。”

以上所说的“误差”是“抽样误差”。本书第 16 章结语，介绍了“误差”的名词与关联性。

1.5 统计数据的收集

要得到样本数据，可以通过实验与调查（包括观察）两种方式。

定义 实验（experiment）是对样本，加以控制（control）分组，再进行测量或观察。

例如：医学实验，将病人分成两组或两组以上，利用不同的药物治疗（控制），再观察其病情。实验数据通常经由方差分析或双总体统计检验，做统计分析。

利用实验进行统计分析，要注意下列几点。

（1）分组时要利用随机性，以尽量消除其他非控制因素。例如：要记录病人吃药后的病情，除了不同药物治疗（控制分组）以外，还有病人的年龄、血型等其他因素，会影响病情。所以随机性是指将病人随机分布到不同的药物治疗分组。

（2）实验分组做统计分析的主要目的是，比较不同组的结果是否有不同。为了使其结果比较客观有意义，可使受试者不知道自己在那一组，例如：消费者不知道自己用的是哪一品牌。如果结果是主观地评定，不是客观地衡量，那么评分者最好也不知道受试者是哪一组，这种方式称作双盲（double blind）。例如：病人不知道自己吃的是实验药或维他命片；同时，医生或护士也不知道病人是哪一组。总之，双重隐瞒是消除实验中可能的个人感情因素，以避免影响实验结果。双盲实验通常在实验对象为人类时使用，目的是避免实

验的对象或进行实验的人员的主观偏向影响实验的结果，通常双盲实验得出的结果会更为严谨。在双盲实验中，实验的对象及研究人员并不知道哪些对象属于对照组，哪些属于实验组。只有在所有资料都收集及分析过之后，研究人员才会知道实验对象所属组别，即为“解盲”。解盲结果，若主要疗效指标未呈现统计学上“显著”意义，则“解盲失败”。

实验要注意：随机性分组与双重隐瞒。

定义　调查（survey）是对总体或样本，不加以控制（control）分组，直接进行访问或观察。

例如：市场问卷调查、电话访问、座谈或个人访问等都是调查。调查方式有：自我测验、访问、电话等，必要时先做试测。

1.6　变量与数据的衡量尺度

数据衡量是将变量给予一个实数值（观测值），但是因为变量的性质不同，所以有不同的衡量尺度。以下我们介绍 4 种衡量尺度，如图 1-5 所示。

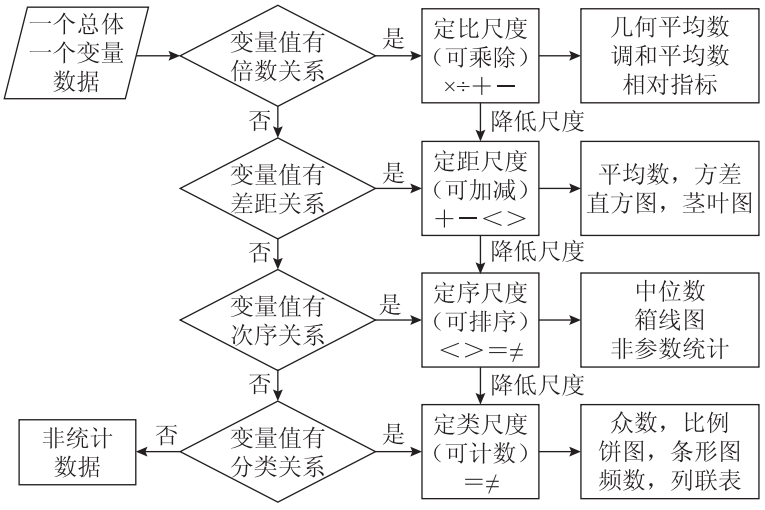


图 1-5　数据衡量尺度与统计方法

1.6.1　定比尺度（ratio scale）或比率尺度

数据之间有“次序”、“大小”及“比率”的关系。例如：长度尺度、质量尺度、绝对温度、金钱、面积、体积、时间间隔等。同时可以定义一个原点，即零值“0”，零值表

示“没有”，没有长度、质量、温度、钱等。

定比尺度的数据做分析时，可以经过数学运算（+、-、×、÷）及转换（X_i^2 或 $\log X_i$）等。

定比尺度的数据常用的代表值是平均数。在统计推论中，是估计及检验其算术平均数。用定比尺度的数据制作统计图时，有时会故意改变尺度比率，或故意忽略原点，或用二度空间尺度面积表示，但可能造成误解。

1.6.2 定距尺度（interval scale）或区间尺度

数据之间有“次序”及“大小”的关系，而没有“比率”的关系。主要是衡量一个数据大于另一个数据多少，而并非其倍数关系。

定距尺度的原点“0”并非代表“无”（无温度，无智商，无尺寸）。例如：①温度尺度（℃或℉）；②智商分数；③衣服或鞋子的号码。

定距尺度的数据有基本的测量单位，可以计算其数值的加减。定距尺度的数据常用的代表值是平均数。在统计推论中，也是估计及检验其算术平均数。

定距尺度的数据也可以用定序尺度（非参数）的统计方法。定距尺度降低为定序尺度：n 个数据顺序排列，从小到大，给予 1 到 n 的秩（等级），若数值相同，其秩（等级）作平均。

1.6.3 定序尺度（ordinal scale）或顺序尺度

数据之间只有“次序”关系，其数值大小并不重要，不能用加法。例如：①考绩等级；②门牌号码；③问卷问题同意程度；④学历。

定序尺度的数据，可以排序后计算，例如计算中位数，无参数统计计算其秩（等级），检验中位数。定序尺度降低为定类尺度：数据按照各秩（等级）出现的频数，不管其顺序秩（等级）。或者合并秩（等级），计算其频数，例如：成绩（A,B,C,D）计算及格不及格，计算（或估计检验）及格的比例。

1.6.4 定类尺度（nominal scale）或分类尺度（categorical scale）

数据之间没有任何“次序”，“大小”及“比率”的关系，只有“分类”关系。例如：①性别数据；②颜色数据；③电话号码、邮政编码、球员编号、职业别、地区别。

定类尺度的数据常用的代表值是众数。在统计推论中，定类尺度变量是计算总体比例，或在回归分析中当作虚拟变量（dummy variable）。

定比尺度和定距尺度的主要参数是平均数；定序尺度和定类尺度的主要参数是比例，如图 1-5 所示。

1.7　数据的类型

数据以不同的性质分类，有下列分类方法。

1.7.1　连续数据与离散数据

连续数据（continuous data）：数据是可以有小数或分数。定比尺度和定距尺度通常是连续数据。计量值数据相当于连续数据。

离散数据（discrete data）：数据是整数，没有小数或分数。定序尺度和定类尺度通常是离散数据；定比尺度也可能是离散数据，例如：生产的个数。计数值数据相当于离散数据。

1.7.2　定量数据与定性数据

定量（数量）数据（quantitative data）：利用客观标准衡量而得到的数据。例如：产品寿命数据，长度数据。有的书将定量数据定义为数字数据，以数量表示的数据。

定性（质量）数据（qualitative data）：利用主观判断而得到的数据。例如：考试数据，同意的程度。有的书将定性数据定义为文字数据，描述特性或性质的数据。

1.7.3　初级数据与次级数据

初级数据（primary data）：数据是由直接观察、调查或实验而得到的原始数据，未经他人的整理或分析，这种数据一定符合搜集数据者的研究目的。通常是内部数据（internal data）。

次级数据或称二手数据（secondary data）：数据是已经他人的整理或分析，变成频数分布表或某种统计结果的数据。次级数据通常取自政府机构、数据公司、广告公司等。引用次级数据要注意研究目的是否相符、来源是否可靠以及时效性。描述统计中的分组数据，可以说是次级数据的描述统计。通常是外部数据（external data）。例题 1.2 可以说是次级数据的研究。

1.7.4　横断数据与纵向数据

数据来源根据是否与时间相关，通常可分成横断（面）数据（cross - sectional data）及纵向数据或追踪调查数据（longitudinal data or panel data）。横断数据是静态数据，收集一个时间点的数据，在“同一时间”的单总体、多总体或多变量的数据。纵向数据是动态

数据，是经过一段时间，收集“不同时间点”的数据，指数的数据和时间序列数据是纵向数据。只做一次式的调查，是横断数据；实验虽然要经过一段时间，但是如果只记录最后结果的数据，那么也是横断数据。

1.7.5 数据集合

记录或个案是个体单位的变量集合，记录和变量可以用一个电子表格（worksheet 或 spreadsheet）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。

1.8 因果关系

在统计学中，可以利用两组数据（“两个变量”或“两个总体”），分析其因果关系或相关性。两总体的平均数或比例检验，其“因”是两总体的分类变量，例如“性别”或“地区”；其“果”是平均数或比例的变量，例如“成绩”或“候选人得票率”。所以，因果关系的假设检验是：同年龄的“男生和女生”的智商或成绩平均数，是否相等或有显著差异，即“性别”是否影响“智商或成绩”；地区是否影响候选人得票率；吸烟影响健康是确定的因果关系；出生月份决定未来的职业、健康与命运，你相信这个推断吗？

分类数据分析的卡方检验，其“因”是两类以上的定类变量，其“果”也是定类变量，例如：如表 1-2 所示，1912 年，泰坦尼克号撞上冰山而沉没，乘客和组员共 2223 人，死亡 1517 人，其中不同“性别”（因）的死亡率（果）是否有显著差异？不同“身份（旅客等级或组员）”（因）的“死亡率”（果），是否有显著差异？第 5 章的条件概率与第 14 章的分类数据分析将对此给予回答。

表 1-2 泰坦尼克号生死录

因 果	头等舱		二等舱		三等舱		组员		总和	
	男	女	男	女	男	女	男	女	男	女
存活	54	145	15	104	69	105	194	20	332	374
	199		119		174		214		706	
死亡	119	11	142	24	417	119	682	3	1360	157
	130		166		536		685		1517	
总和	173	156	167	128	486	224	876	23	1692	531
	329		285		710		899		2223	

通常，第 11 ~ 14 章的原假设是“没有因果关系”，检验结果“拒绝原假设”表示有“显著差异”，所以“有显著差异”表示“有因果关系”。

回归分析就是两个变量的因果关系，检验自变量 X（因）对因变量 Y（果）的直线关系是否显著。例如：广告预算对销售额的影响是否显著，信息科技的支出对企业的获利绩效的影响是否显著。

不同数据尺度检验因果关系的统计方法也有不同。表 1-3 是从第 11 章开始到第 15 章，不同尺度的因果或相关的统计方法。

表 1-3　不同尺度的因果关系的统计方法

因 / 果	定类尺度		定序尺度	定距尺度
	2 分类	≥2 分类		
定类尺度	两个总体比例检验 游程检验（run test）	列联表 分类数据分析		判别分析 Logistic 回归
定序尺度	符号检验（sign test） Wilcoxon 符号秩检验 Wilcoxn 秩和检验	KW 检验 Friedman 检验	Spearman 检验	Spearman 检验
定距尺度	两个总体平均数检验 两个总体变异数检验	方差分析	时间序列指数	散点图、回归分析、相关系数

1.9　统计的应用步骤

《孙子兵法・军形篇》中写道：“兵法：一曰度，二曰量，三曰数，四曰称，五曰胜；地生度，度生量，量生数，数生称，称生胜。”地：分析地形险易情况。度：判断战区战线区域。量：计划部署战场容量。数：决定所需人力数量。称：权衡比较双方优劣。胜：未战已经胜券在握。统计的应用步骤（如图 1-6 所示），和兵法不谋而合。

统计工作：

（1）（地）了解问题，定义总体、变量。总体是什么？有几个总体？（分类总体的标志是什么？）有什么变量？（要衡量总体的什么性质？）有几个变量？是否有两个以上变量的相关或因果关系？

（2）（度）认定变量值的数据尺度，决定指标、参数。数据的尺度是什么？什么指标？什么参数？描述统计或推断统计？

统计资料：

（3）（量）决定实验、调查、观察或二手数据。实验是抽样，调查决定普查或抽样。

设计实验步骤或调查方式。选择抽样方法，决定样本容量。

（4）（数）收集数据，决定数据特性（符合假定条件如正态）、统计量、统计模型。辨认（identify）统计模型，检查假定条件，统计模型的假定条件有：数据尺度、正态分配、抽样独立性、方差条件等。

统计学：

（5）（称）数据分析，普查是描述统计，选择表达的方式。抽样是推断统计，选择统计分析模型。表达方式有：表格、图形或代表值等。计算（compute）结果。

（6）（胜）得到信息、报告结论，或导出决策。

解释（interpret）结果，得到信息、报告结论、实施决策、衡量决策的结果。

一般统计学教科书的例题或习题的解答步骤，通常已有数据，只要做第4，5，6步骤。

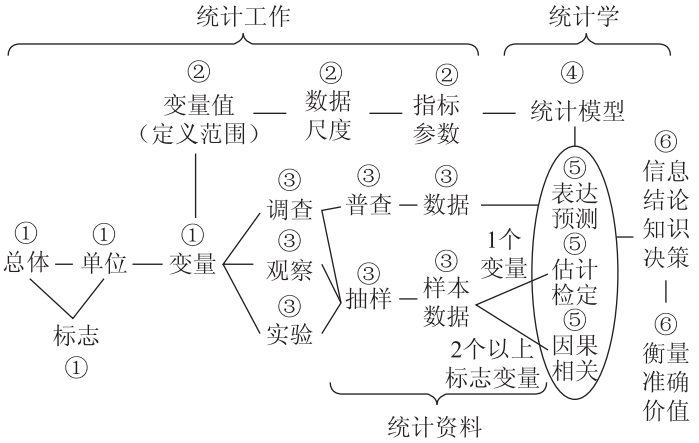


图 1-6　统计的应用步骤（数字代表上述步骤）

例题 1.6　可乐独卖权利的决策（参考英文书目［14］Keller 2009）

T大学有30 000学生，要和P可乐饮料公司签定独卖合约，在校园内只能卖P可乐，学校每年收100万元回馈金，加上P可乐全年销售金额的30%。目前在T大学有3种以上的可乐销售，P可乐每周平均销售20 000罐，但是不知道其他可乐的销售量。一年40周（寒暑假没学生），所以P可乐在T大学的年销售量为800 000罐。如果P可乐每罐售价￥3，每罐成本￥1。P可乐公司有2周的考虑期间，请问该公司如何做决定？

解答：目前（没有独卖）P可乐每年获利 $20000 \times （￥3 - ￥1） \times 40 = ￥1\ 600\ 000$

假设 π = P可乐在T大学的市场份额（市占率）

则每年T大学可乐独卖的销售数量为 $X = 800\ 000 \div \pi$（罐）

P可乐独卖每年的获利 $X \times ￥3 \times 0.7 - X \times ￥1 - ￥1\ 000\ 000 = ￥1.1X - ￥1\ 000\ 000$

独卖优势 A = 独卖每年的获利 - 没有独卖每年的获利 = ￥1.1X - ￥2 600 000

$A = 1.1 \times (800\,000 \div \pi) - 2\,600\,000 = 880\,000/\pi - 2\,600\,000$

P 可乐在 T 大学的市场份额与独卖优势如表 1-4 所示，市场份额越高，独卖优势越少。

表 1-4　可乐在 T 大学的市场份额与独卖优势

市场份额 π	0.2	0.25	0.3	0.35	0.40	0.45	0.5
独卖优势 A（万）	180	92	33	-8.6	-40	-64	-84

当 P 可乐的市场份额为 33.85%，独卖优势为 0，为独卖的损益两平点。P 可乐的市场份额越低，和 T 大学签定独卖合约越有利。因此，P 可乐决定用一周的时间，进行统计推断，推断市场份额。统计学的应用步骤如下。

（1）总体 = 30000 学生，变量 = 每个学生每周购买可乐的品牌及数量，

参数 = T 大学每年可乐的销售数量或 P 可乐的在 T 大学的市场份额。

（2）变量有定类尺度（可乐品牌）及计数值数据（可乐数量）。

（3）决定调查，抽样 500 个学生，记录每人一周买可乐的品牌及数量。

（4）每个样本数据如：{CCP} {PP} {CCCT} {PC} {T} 等，P 表示 P 可乐、C 表示 C 可乐、T 表示 T 可乐。计算 500 个学生中，P、C、R 的个别总和，及全部可乐的总和。

（5）“P 的总和”除以“全部总和”即为 P 可乐的市场份额的估计值，这是点估计，根据这个市场份额的点估计，就可以决定是否和 T 大学签订独卖合约。如果要检验下列假设，则可能要多抽样几次 500 个学生。

原假设 $H_0: \pi \geqslant 33.85\%$，　　不同意独卖合约。

备择假设 $H_1: \pi < 33.85\%$，　同意独卖合约。

（6）得到决策：是否同意独卖合约。

（7）进一步考虑因素：这个推论的假定条件是 T 大学全部可乐的销售量等于 P 可乐独卖的销售量。实际上其他牌可乐的忠诚度，使得独卖不见得将所有市场份额，都转为 P 可乐。例如 C 可乐爱好者，在独卖后，可能不会买 P 可乐。因此，“独卖后”P 可乐的销售数量，不等于 P 可乐销售数量除以“没有独卖的市场份额”。应该将每年独卖后的销售数量打折。

（8）问题：如果能够推导 π 的估计量或统计量的概率分布或方差（标准差），那么才可以进行第 5 步的假设检验推论。决策法则：如果 π 的点估计值大于 33.85%，则不同意独卖合约。

最后，将统计的应用步骤，再整理如下：（套模是套用模型）

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

1.10 本章流程图

本章流程图如图 1-7 所示。

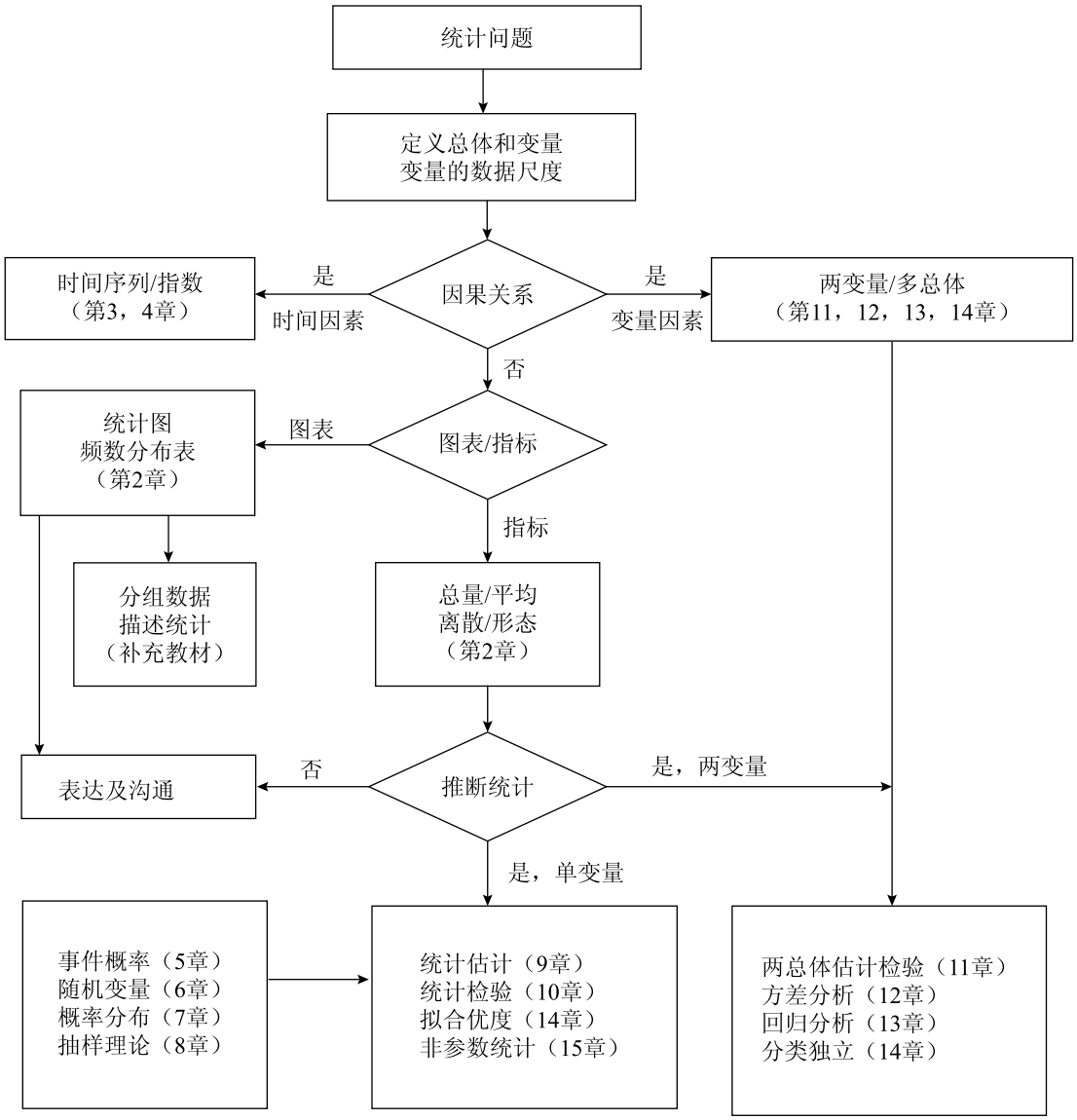


图 1-7　第 1 章流程图