

管理应用统计学

缪柏其 主编

中国科学技术大学出版社

2002 · 合肥

内容简介

本书以管理经济理论为背景,深入浅出地介绍了描述性统计方法和数理统计中关于均值的区间估计和假设检验中常用的统计方法,如正态检验, t 检验, χ^2 检验,Wilcoxon 秩和检验,符号检验和游程检验等。也介绍了有关回归分析,时间序列分析。最后介绍了著名的统计软件包 SPSS 的使用方法。

本书可作为 MBA,MPA 学生及工商管理研究生班的教材,也可供有关管理人员阅读和参考。

图书在版编目(CIP)数据

管理统计学/缪柏其主编. —合肥:中国科学技术大学出版社,2002.9

MBA 系列教材

ISBN 7-312-01455-0

. 管... . 缪... . 经济统计学 . F222

中国版本图书馆 CIP 数据核字(2002)第 047448 号

中国科学技术大学出版社出版发行

(安徽省合肥市金寨路 96 号,邮政编码:230026)

中国科学技术大学印刷厂印刷

全国新华书店经销

开本:787mm×960mm 1/16 印张:13.75 字数:286 千

2002 年 9 月第 1 版 2002 年 9 月第 1 次印刷

印数:1—3000 册

ISBN 7-312-01455-0/C·45 定价:18.00 元

前 言

管理应用统计学是一门以经济管理理论为基础,以一般统计学为工具研究社会和经济管理的应用科学。它研究如何有效地收集、整理和解释反映社会和经济管理实践的数据,以期认识数据的规律性及内在的社会和经济涵义。管理统计学是应用统计学的一部分,现在已广泛地应用于市场调查、证券市场分析、风险分析、产品抽样调查和质量控制、投资效益评估、人力资源分析和评估、经济指标分析和预测等方面,为管理者进行正确的决策提供了科学依据。以人为本的导向使管理方法进一步由定性向定量方向发展,而这一切都离不开统计方法的使用。因此一个现代管理者必须掌握一定的统计知识。

(一) 课程的目的和要求

管理统计学是商学院的核心课程,本课程为 40 学时。课程的目的和要求是:

1. 认识到管理统计学是一门科学,也是一门艺术。不同的问题要用不同的统计方法去解决。
2. 了解如何有效地收集、整理数据,并应用基本统计原理和方法来处理和分析数据。
3. 熟悉常用的统计方法和公式,并能正确地解释。
4. 了解和能初步应用常见的统计软件包(如 SPSS, EXCEL, SAS 等)。
5. 能正确地应用统计方法来研究和处理管理和经济问题,为决策提供可靠的依据。

(二) 大纲的教学体系

管理统计学教学内容共 11 章,分为 6 个部分:

1. 描述性统计和统计量

(1) 要点和要求

这部分由前四章组成。首先要了解不是所有的数据都可以进行统计分析,我们对数据有一定的要求。因此有一个如何收集数据的问题。在此基础上用列表、图示和加减乘除等初等方法对数据进行分析,对数据表达的总体有直观的了解。具体的,第一章是叙述如何有效地收集数据;第二和第三章介绍一些常见的统计量;第四章介绍统计指数的构成。

(2) 预备知识

阅读一些统计报表、年鉴、商业和政府有关部门发表的数据。

2. 统计推断的概率论基础

(1) 内容和要求

这部分由第五章和第六章组成。为了对数据所反映的深层次的内涵进行分析,我们需

要一些概率论的知识。描述性统计和数理统计的本质差别就在于后者用了概率论的知识。我们要求学员能掌握离散分布中的二项分布和连续分布中的正态分布；了解大数定理和中心极限定理。

(2) 要点

概念：样本空间；事件；独立性；二项分布；正态分布；均值；方差；相关系数。

公式：全概率公式；贝叶斯公式；一般正态分布函数用标准正态分布函数表示的公式。

中心极限定理：正态分布随机变量的线性组合仍是正态随机变量。

3. 统计推断

(1) 内容和要求

这部分由第七章和第八章组成。当我们按一定的规则从总体中抽取了一个样本后，一个重要的问题是由样本来推断总体的性质。在管理应用统计学中我们归纳为估计问题和检验问题。在估计问题中，我们假定总体的分布类型已知，但其中的一些参数未知，由样本对这些参数进行估计。估计分为点估计和区间估计两类。关于点估计，我们最关心的是总体均值和方差的点估计。在概率论的基础上，以一定的置信度我们把总体均值和方差的真值估计在以点估计为基础的区间内。在检验问题中，我们分两种情况讨论。一种是总体类型已知，对其中的参数进行统计检验。总体均值的区间估计和参数检验依赖于正态性，在大样本情况下可以由中心极限定理来近似。另一种情况是总体的分布类型未知，我们需要对其分布进行检验，这称为非参数检验。一个重要的非参数检验统计量是 χ^2 统计量，它也常用来检验列联表中两个变量是否独立。

在这一部分中我们要求学员掌握方差已知和未知情况下总体均值和方差的点估计，均值的区间估计和检验，列联表检验和拟合优度检验。

(2) 要点

概念：点估计；置信度；精度；自由度；区间估计；假设检验；成组比较；成对比较；列联表检验；拟合优度检验；

检验统计量： u 检验（标准正态分布）； t 检验（ t 分布）； χ^2 检验（ χ^2 统计量）。

4. 相关分析和回归分析

(1) 内容和要求

这部分内容由第九章组成。在相关分析中主要介绍有序数据和比例数据相关系数的概念；在回归分析中我们着重介绍了一元线性回归系数的最小二乘估计方法以及有关的预测问题，同时我们也讨论了用变量代换可以化为一元线性回归的非线性回归问题。多元线性回归没作过多的介绍，因为除了在技术上要用矩阵外，估计和预测方法和一元线性回归思路是一样的。当然，多元线性回归中的变量选择是多元线性回归中特有的，但已超出了本书的范围。我们的要求是掌握相关系数的概念以及一元线性回归系数的估计方法和有关的预测问题。

(2) 要点

概念：相关系数、Spearman 相关系数、散点图、回归方程、回归直线、最小二乘估计、预测。

5. 时间序列分析

(1) 内容和要求

在第十章我们讨论时间序列的分析。关于时间序列的分析有两种方法，一种是描述性的，另一种是用统计知识建模的。根据对 MBA 学员的要求，我们重点介绍描述性的分析方法。用统计知识建模的时间序列我们仅仅介绍一些概念以方便大家使用有关的统计软件包。时间序列的描述性分析方法包括加法模型、乘法模型和滑动平均模型。要求学员能对时间序列进行分解，对季节因子进行修整，并能应用所学时间序列模型进行预测。

(2) 要点

概念：趋势；季节因子；循环因子；误差因子；指数加权滑动平均；预测。

模型：加法模型；乘法模型；滑动平均模型；自回归模型；自回归滑动平均混合模型；季节模型。

6. 统计软件包

(1) 内容

本章介绍如何使用 SPSS 统计软件包。在本章中我们从最基本的输入数据开始，一步步地介绍本书前十章中的统计方法如何具体地在计算机上实施。

(2) 要求

通过上机实习，初步学会使用 SPSS 统计软件包来分析和处理数据。

根据 MBA 大纲要求我们编写了该书。其中第一章至第十章由缪柏其教授编著，第十一章由蒋涛博士、彭衡博士和陈文志博士编著。陈文志博士绘制了包括附表中的所有图表。成书之前的讲义在中国科学技术大学 MBA 中心、管理科学与工程专业研究生课程班和安徽工商管理学院作为管理应用统计学教材讲授了八九次以上。在讲授过程中得到诸多同仁的关心和支持。吴耀华教授、汪道明教授、杨桂元教授、杜雪樵教授和郑坚坚副教授仔细审阅了本书的讲义，提出了不少有益的建议，并指出了许多笔误。本书的写作过程中也得到了中国科学技术大学商学院领导方兆本教授、梁 教授和商学院 MBA 中心储雪林教授和王苏鸣高级工程师等的大力支持。同时得到我妻子曹蓓和女儿的大力支持。在此谨向他们表示衷心的感谢。

缪柏其

2002 年 4 月 18 日于中国科学技术大学

目 录

第一章 数据收集.....	1
§1 引言	1
§2 数据来源	2
§3 市场调研研究	7
§4 数据的类型	7
第二章 数据的表示.....	8
§1 离散数据的列表表示.....	8
§2 离散数据的图示法.....	10
§3 连续数据的表示.....	12
§4 散点图和曲线拟合.....	15
第三章 描述性统计量.....	17
§1 刻画数据集中程度的特征量.....	17
§2 方差和标准差	22
§3 变异系数和斜度系数.....	26
习题	28
第四章 指数.....	30
§1 指数的概念	30
§2 指数的构造	33
§3 一般零售价格指数.....	37
习题	39

第五章 随机试验、随机事件与概率.....	40
§1 事件	40
§2 概率的定义	42
§3 事件的独立性	44
§4 非独立事件与运算.....	44
第六章 随机变量及其分布.....	53
§1 离散型随机变量的分布.....	53
§2 正态分布 (NORMAL DISTRIBUTION)	56
§3 数字特征	60
§4 独立随机变量和 (差) 的均值和方差.....	62
§5 中心极限定理	63
习 题	66
第七章 参数统计推断.....	68
§1 什么是统计推断.....	68
§2 总体均值的估计.....	69
§3 百分比(数)和百分比 (数) 差的估计	76
§4 中位数的区间估计.....	79
§5 显著性检验 (SIGNIFICANCE TESTING)	80
§6 两类错误	87
§7 二样本的假设检验问题.....	89
习 题	97
第八章 非参数检验.....	98
§1 χ^2 检验	98
§2 威尔柯克逊秩和检验.....	105
§3 符号检验	108
§4 威尔柯克逊符号秩和检验.....	111

§5 游程检验 (RUNS TEST)	114
习 题	115
 第九章 回归分析和相关分析	117
§1 相关关系	117
§2 相关系数(COEFFICIENT OF CORRELATION)	117
§3 回归分析	122
§4 $\hat{a}, \hat{b}$ 的检验和预测	127
§5 另一回归直线	130
§6 可化为线性函数的非线性关系	132
§7 多元线性回归	135
习 题	136
 第十章 时间序列	137
§1 因子与模型	138
§2 趋势	139
§3 季节因素 (THE SEASONAL FACTORS)	144
§4 循环因子	147
§5 随机因素和残差(THE RANDOM FACTOR AND RESIDUAL OR FACTOR)	147
§6 预报 (PREDICTIONS)	149
§7 指数加权滑动平均	149
§8 统计时间序列模型简介	152
习 题	155
 第十一章 SPSS 统计软件使用简介	156
§1 SPSS 简介	156
§2 SPSS 的基本操作简介	156
§3 SPSS 统计分析示例	172
 附表 1	186

附表 2	196
附表 3	198
附表 4	200
附表 5	201
附表 6	202
附表 7	203
附表 8	204
附表 9	205
附表 10.....	206
附表 11.....	207
参考文献.....	208

第一章 数据收集

§1 引言

每个人都收集某种类型的数据，但商业上收集数据的量非常大，其中一些数据对商业决策是非常有用的。现实生活中虽然有许多公司在收集数据，但遗憾的是对它们没有充分加以利用。

许多商业，经济和社会问题不能简单地用“是”和“不是”来回答。它们可能需要讨论及进一步明晰，也可能需要给出一个接受或拒绝的准则(criteria)。为了考虑这些问题，需要对现有信息的完整性进行评估，然后提出还需要什么新信息。因此决策者不但需要数据，而且需要评估这些数据的质量。

什么是数据？一种描述性的定义是“可以由它作出推断的已知事情或事物”。

数据有各种不同的表达形式，如可以是数字，还可能是类别，性别，学历等等。

对决策者而言，数据的完整性是一个最大的问题。决策者需要确定，目前的数据对当前目标是够用，还是要补充什么数据。因为数据的收集是花时间、花钱的。另一个问题是数据的质量，数据有偏差或误导可以毁掉一个有效的决策过程。例如某个区行业的调查结果不能作为整个城市行业的结果。

在我们为决策而准备数据前，首先要明确问题是什么？难点是什么？

在收集数据时我们要考虑：

1. 什么是我们考虑的相关总体(population)?
2. 数据从何处而来？
3. 在分析所收集的数据时要进一步分析一下我们提问了多少人，他们是如何被挑选出来的？
4. 如何从这些回答者中收集信息？
5. 谁没有回答提问？
6. 收集的数据类型是什么？

在数据收集中一个重要的概念是总体，即我们要分析的某一特定人群中某个指标取值的全体。因为数据收集是花钱、花时间、有代价的，因此首先要明确向哪些人群收集数据。例如，要调查为什么人们买外国产的化妆品却没有与购买国产化妆品的人接触，则你不可能区别如下事实：某些购买者没有意识到他们购买的化妆品是国外产的。

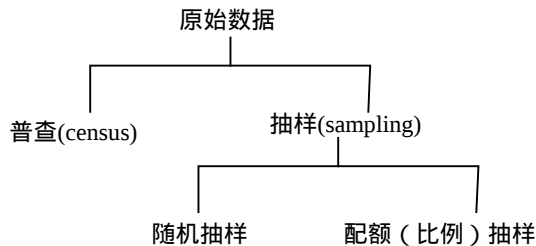
确定有关人群后,下一问题是知道这些人是谁(比如姓名,地址),称为抽样框(sampling frame),然后明确向这些人调查内容。

有些时候可以用政府调查报告中的数据,如统计年鉴。

§ 2 数据来源

数据可分为两大来源:原始数据(primary data) 和二手数据(secondary data)。

原始数据来源可以表示如下:



二手数据:已有的数据(别人调查所得,包括政府数据),如“年鉴”,别的公司调查结果等。

原始数据:自己经过调查所获得的数据。它的特点是为指定目的而收集。

二手数据来源:行业或地区组织内部;国内外有关资料。二手数据也称为副产品统计(by-product statistics)数据。如失业人员集合是通过下岗登记人数总数和找到工作人数之差得到。

二手数据大部分来自官方统计,如

年鉴 (*Annual Abstract of Statistics*)

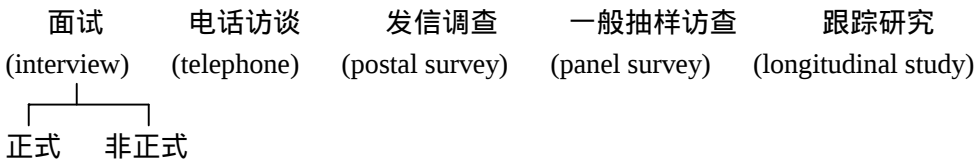
月度统计报表 (*The Monthly Digest of Statistics*)

财经统计 (*Financial Statistics*)

经济趋势 (*Economic Trends*)

.....

原始数据得到的方式如下:



普查:仅局限于总体较小时。如对每个住户调查以估计有空调的户数,则因普查太费钱而不太可能实施。但若你代表一工厂,而该工厂仅把产品卖给一小批批发商,要批发商

讲述空调使用情况时，则可用普查。

样本大小(sample size)的选择：当总体太大时不能全面调查，这时要用抽样方法，关键是样本要有代表性，即全面反映该总体观点，所以抽样前要做好抽样方案：确定要调查多少人，调查什么样的人等。抽样代表性是指每个人都可能被抽到，且抽到的可能性相同。这样的抽样称为随机抽样(random sampling)。

例1 抽查时，不抽身高 ≥ 1.80 米，则不能反映1.80米以上人群的观点，因此这种抽样是非随机的。如何做到抽样有很好的随机性？通常用如下手段：首先把总体中要调查的每个人编上号。然后用计算机，随机数表，摇奖机等方法得一组随机数。把随机数对应的人群构成一个样本。

当总体很大时，很难做到随机抽样，这时可用一种很好的非随机抽样——配额抽样(quota sampling)。

例2 如某商场希望根据年龄和性别来调查对营业时间的意见，计划调查周边地区的1000个人。已知年龄和性别分布如下：

年龄(岁)	比例%
15~20	16
20~30	26
30~50	30
>50	28

性别	比例%
男	51
女	49

若准备抽1000人，则配额抽样方案为：

年龄(岁)	男性	女性
15~20	82	78
20~30	133	127
30~50	153	147
>50	143	137

也可以作更复杂的抽样方案(随机和配额抽样的复合)，如分层抽样，分地区抽样等。在分层抽样方案中，每层人数比例已知。如预测美国总统选举结果，已知民主党和共和党的百分比为 P_1 P_2 ，当决定抽取1000人作调查时，分层抽样方案为抽民主党 $1000P_1$ ，共和党 $1000P_2$ ，然后在每个党派中再作随机抽样。

提问(asking questions)

认定调查总体及决定用问卷方式以得到所需信息后，下一步是确定问什么，如何问。

- 问卷设计(questionnaire design)

问卷结构特点：问答由一个问题顺势转入下一个问题；由一个主题转到另一主题。不要有跳跃而导致回答的无方向性。

建议：从一般问题到特殊问题。

盖洛普(Gallup)组织总结提问的目的有如下 5 种：

1. 找出回答者是否觉察到这一结果，

例. 你是否知道合肥和徐州之间要修条高速公路的计划？知道/不知道

2. 获得关于结果的一般感觉

例. 你是否同意合肥和徐州之间要修条高速公路？请在下列选择中选择一个：

强烈同意、同意、无所谓、不同意、强烈不同意

3. 获得该结果指定部分的答案

例. 你认为高速公路会对当地环境产生影响吗？有/没有

对是否影响环境，有时可分得再细一点，如指出是噪音还是风景。

4. 获得反映回答者观点的理由

例. 如果反对，那你反对的理由是：

(a) 已有一条可用的主干道

(b) 合肥和徐州之间运量不足

(c) 高速公路会损坏美丽的乡村风景

(d) 道路会破坏历史文物

(e) 其他，请指定.....

这一条是先知道他（她）的观点，然后给机会以让他（她）说明理由。把条件叙述“如果反对”看作一种预先设定的过滤。此外，这一条也可提成“为什么你反对修这条高速公路”。这种提法称为开放式的。这两种设定互有优缺点。预先设定的容易回答；开放式的结论需要回答者有较高的文化修养。

5. 找出持有这些观点的强烈程度

例. 你准备用下列哪一种行动来支持你的观点？

(a) 给地区人大代表写信

(b) 给本单位的全国人大代表写信

(c) 给新闻单位写信

(d) 在政府召开的公众听证会上发言

(e) 在网上发表呼吁文章进行调查

建设高速公路将会毁坏一些历史文物，因此可以提这样的问题：这些历史文物的重要性有多大？请在下列小格中打√

(应不惜一切代价保存) →

1	2	3	4	5	6	7
---	---	---	---	---	---	---

 → (不重要)

在许多情况下，不仅要知道某件事是好还是坏，还要知道有多好或多坏，因此应该在问卷上给出一系列可能的态度和行动。

可以用 1, 2, ... (称为刻度) 来表示，一般取 4 - 5 个刻度。

● 提问的方式

有开放式提问和预设答案提问两种方式。

开放式提问的优缺点：

缺点：导致偏爱社区中有文化和教育程度高的人，因他们能很快地组织和表达他们的思想和理念。

优点：有助于人们回答自己想要表达的观点。

预先设定答案提问的优缺点：

优点：容易回答。

缺点：不能准确和全面反映被调查者的意见。

例如，你的提问是“你同意在×国布置核武器吗？”，预先设定的答案如下：

同 意 ☐

不同意 ☐

不知道 ☐

某些回答者可能说“是的，但仅仅是某一种核武器”，或“不，但没有可以代替它们的”，或“是的，只要它们的发射受到严格控制”。

对这类问题的解决办法是试图扩大预设答案的范围。如果不行，只能提成开放式的。

- 提问的措辞（wording questions）

提问的措辞在诱导有代表性回答中是非常重要的。一个有偏差的或诱导性提问将偏离给定的回答。在问卷设计中发现的偏差主要有：

1. 把两个或两个以上问题表达为一个。例如在调查一个小区建公用车库还是建自有车库时，你的问题设计为“由于自有车库使用方便和比较清洁，因此你赞成修自有车库”。使用方便和比较清洁是两个不同的问题，有人愿意修自有车库是因为使用方便，但不一定清洁。

2. 问题中含有含糊或不熟悉的单词。例

你通常到哪儿去购物？问题是几次算通常？

感冒后你吃盐酸马啉脲胍片吗？

盐酸马啉脲胍片是医学用的专业词汇。通常老百姓只知道感冒灵。

3. 用软化难度或方向的词开始的提问。如对一个女孩提如下问题：

我希望你不要在意我提出这个问题，你多大了？

回答者会立即警惕起来，向你发出警告（不回答）。再举一个例：

像大多数人那样，你是否认为在北大西洋公约组织里应该有英国的声音？

是/不是

这类引导性问题有两种可能反应：

(a) 导致同意这种论述，以作出和大多数人一样的回答

(b) 单纯为了不同意而不同意。

在两种情况下，回答者都不一定能反应出自己所持的观点。

4. 包含条件或假设条款。例

如果你有 6 个小孩，你认为你的生活会有多少改变？

一般情况下，有 6 个孩子的情况很难发生，因此回答者无法回答你的问题。

5. 包含对回答者一个或多个指令的提问。例

如果你拿到周薪后，在预留下所有日常支出的钱以及小孩上学费用后你还有多少钱可以消费或存起来？

一般机关事业单位人员拿月薪或年薪，水电费也是一个月或两个月收一次费，而孩子上学费用一般是半年一次。因此被访者无法回答你的问题。

一个完整的问卷需要有一个合理的流向，当问题用来对回答者分类时，比如回答“是”从问题 1 到问题 8，回答“不”从问题 9 到问题 15，则整个问卷上的路径必须用满足上述流向的指令来组成。

● 调查中的无反应

调查中经常会发生无反应的情况。主要原因是从抽样框中选出的人，由于如下原因无反应：

- (1) 已搬走，
- (2) 不善于用言词表达或体弱多病，
- (3) 打电话时人不在（如果被调查人不在，一般建议至多重复 6 次电话），
- (4) 调查期间人不在，
- (5) 拒绝合作（大约 5%）。

例 2 1936 年美国总统选举前，*Literary Digest* 杂志预测共和党候选人兰登（London）和民主党候选人罗斯福（Roosevelt）的得票比例是 53 43，但实际选举结果是罗斯福赢得了这场选举，他与兰登得票比例是 62 38。*Literary Digest* 杂志自第一次世界大战以来一直进行总统预测，而且每一次都预测准了。这次该杂志发出了 1000 万张调查问卷。回收到 240 万张，预测是根据 240 万人的民意调查回单作出的。出现预测错误的原因主要有两个，一是选取样本的方式不是随机的，它主要是从电话号码簿和各种俱乐部会员名册上去选取的，这些人大多是有钱人，他们支持共和党候选人兰登，第二是有 760 万张问卷无反映，在没有寄回的人中支持共和党和支持民主党人的比例一般与寄回来的 240 万答卷中的支持比例不一样，例如该杂志向芝加哥地区三分之一的登记选民发了调查问卷，有 20% 的被询问者作了回答，其中一半以上支持兰登，但实际选举的结果是，在芝加哥地区是 2 1 有利于罗斯福。这种由于无反应（等于改变了原先随机抽取的样本）而造成的偏差叫“无反应偏差”。因此在调查中我们要以各种方式减少这种偏差。

§ 3 市场调查研究

市场调查研究(market research)是顾客的声音,它对工业,商业和政府有极大的影响力。市场调查研究又是一种手段,通过它可以影响货物和服务的发展和交易,以及社会政策的形成。它用各种技术手段来提供有关消费者的爱好、态度、嗜好和需求,以帮助公司了解顾客的各种要求。国家和地方政府则依靠市场调查研究提供的数据,制定从地区运输到社会服务等各种政策。市场调查研究可以对我们经济的良好运行起很大的作用,可帮助我们正确认识国内居民的需要。市场调查研究可以分为两大类:

1. 研究新产品或市场需求。目的是研究是否有市场或市场份额有多少,如有,研究对提供的产品或服务有什么需求。

2. 对已有产品或服务进行研究。目的是识别在市场中的份额,观察如何变动。

两类市场研究的基本研究技术有类似的方面,但必须以不同的方式来进行。例如对已有的产品和服务,许多被调查者已经熟悉了它们,因此我们可以假定一个确定的先验知识。而对新产品和新的服务,则必须询问最一般的市场调查问题以确定各种可能性,即先要做可行性研究。

§ 4 数据的类型

根据数据的来源,我们可以把数据分成:

1. 定性数据

又称分类(名义)数据(categorical or nominal data)或分类标志,例如调查人群有工人、教师、商店服务员和公司老板,我们用数字 1、2、3 和 4 分别表示这 4 个人群。这儿的数字称为分类数据,他们之间没有大小先后之分,仅仅用来表示不同的职业。

2. 定量数据

(1) 有序数据。它们之间有大小之分,但没有大多少的概念。例如可以把人们对某一事件的态度进行量化,用 1,2,...来表示对某一事件的态度。

(2) 间隔数据。间隔数据是有序数据,它们之间的差是有意义的。例如温度是一个典型的间隔数据。 0°C 不表示没有热量, 40°C 和 30°C 的温差在数量上等于 80°C 和 70°C 之间的温差,但它们代表的热量是不一样的。

(3) 比例数据。测量的最高水平是比例数据。这类数据有一个起点,记为 0。例如距离和时间,0 是有意义的,可以看作距离和时间的起点。两倍的距离和两倍的时间也是有意义的。

定量数据可进一步分为离散数据和连续数据,这在统计上是有本质区别的,即离散数据是可以一个一个地数的,而连续数据往往用一个区间或一条直线上的数值来表示。

第二章 数据的表示

§1 离散数据的列表表示

设有一组数据 $x_1, x_2, \dots, x_n$ ，为找出这些数据的变化规律，要对数据进行合理的加工。

例1 观察每 10 分钟进入某超市的人数，144 个数据如下：

23, 23, 35, 7, 29, 3, 15, 20, 30, 18, 23, 32, 34, 40, 17, 25, 30, 33, 51, 37, 43, 52, 67, 34, 20, 26, 46, 68, 79, 82, 57, 61, 96, 75, 59, 64, 77, 99, 87, 48, 58, 95, 96, 68, 46, 73, 57, 39, 45, 28, 24, 35, 43, 25, 27, 30, 42, 30, 22, 18, 21, 26, 21, 8, 19, 15, 24, 12, 8, 8, 6, 8, 20, 8, 25, 29, 26, 36, 26, 28, 36, 22, 41, 37, 30, 50, 28, 35, 24, 36, 50, 44, 48, 38, 47, 55, 30, 44, 50, 40, 47, 83, 75, 64, 66, 75, 83, 90, 87, 59, 63, 78, 75, 86, 86, 77, 64, 70, 65, 69, 56, 55, 42, 47, 33, 36, 34, 29, 33, 25, 16, 30, 22, 18, 9, 14, 16, 20, 26, 10, 18, 9, 7, 8。

如何加工？首先把数据从小到大排列起来，当数据太多时，可以用枝叶法初步排队，即把第一位相同的数据放在一组，本例数据枝叶表示为

枝	叶	频数
0	7 3 8 8 8 6 8 8 9 9 7 8	12
1	5 8 7 8 9 5 2 6 8 4 6 0 8	13
2	3 3 9 0 3 5 0 6 8 4 5 7 2 1 6 1 4 0 5 9 6 6 8 2 8 4 9 5 2 0 6	31
3	5 0 2 4 0 3 7 4 9 5 0 0 6 6 7 0 5 6 8 0 3 6 4 3 0	25
4	0 3 6 8 6 5 3 2 1 4 8 7 4 0 7 2 7	17
5	1 2 7 9 8 7 0 0 5 0 9 6 5	13
6	7 8 1 4 8 4 6 3 4 5 9	11
7	9 5 7 3 5 5 8 5 7 0	10
8	2 7 3 3 7 6 6	7
9	6 9 5 6 0	5

第二步

1) 分组找出最大最小值 $X_{(n)}$ 和 $X_{(1)}$ ， $X_{(n)} - X_{(1)}$ 称为极差 (range)，常用 R 表示。

2) 确定组数和组距，在决定组数 k 时，必须考虑到数据整理的目的，一方面在于简化数据，以利于显示规律性，另一方面又要适当保持数据的细节，以免失之过粗，但组数太