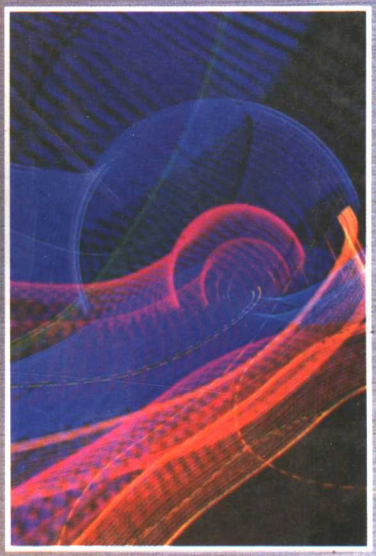
浙江省高等教育重点建设教材

YING
YONG
SHU
LI
TONG
JI



应用数理统计

陈上珠 陆传荣 编

杭州大学出版社

浙江省高等教育重点建设教材
《高等数学》编委会成员

主　编：秦禹春

副主编：陆传荣

编　委：水乃翔　王祖樾　陈上珠
姜　豪　张　皓　滕钟仁

浙江省高等教育重点建设教材
应用数理统计
陈上珠　陆传荣　编
*
杭州大学出版社出版发行
（杭州天目山路34号）
*
杭州大学出版社电脑排版部排版　浙江省上虞印刷厂印刷
850×1168毫米 1/32　10印张　250千字
1998年5月第1版　1998年5月第1次印刷
印数：0001—3000
ISBN 7-81035-477-9/O·069
定价：13.50元

前　言

什么是数理统计学?《中国大百科全书·数学卷》指出:数理统计学是一门科学,它研究怎样以有效的方式收集、整理、分析带随机性的数据,并在此基础上对所研究的问题作出统计性的推断,直至对可作出的决策提供依据或建议.通俗些说,数理统计学是有关对客观不确定现象收集和取得数据资料,并对它进行整理、分析,以对所研究的问题作出一定的结论的那些理论和方法.

这里的客观不确定现象也就是随机现象,如观察新生婴儿的性别,在出生之前是不确定的,可能是男孩也可能是女孩,生男生女是随机的(随着科学的进步,目前已经知道形成男婴或女婴的原因).又如,测定一个灯泡的寿命,其结果不会和设计的标准完全一致,在正常情形下,绝大部分高于规定标准,少量的不符合规定,除了原材料质量不一以外,在生产过程中有大量无法控制的偶然因素,如机器震动、操作上疏忽等等,使得每个产品的寿命不尽相同.这类现象在一次观察下,结果是不确定的,但在大量重复观察下,其结果又具有一定规律性.数理统计就是对随机现象进行观察和试验,从中收集和取得数据资料,然后对所获得的数据资料进行整理、分析,找出该现象反映在数量上的规律性,以期作出决策.由此可见,数理统计学有以下特点:

1.数理统计只是从随机现象外在的数量表现上去研究问题,不涉及事物的质的规定性.通俗地说,统计只能告诉你,从观察或试验结果来看,该现象如何如何,它不能回答为什么会如何如何.例如,许多统计资料表明吸烟与患肺癌之间有较大关联,这只是一种统计规律性——由外在的数量关系归纳出来的规律性.吸烟何

以引发肺癌的机制在医学上尚未研究清楚，在医学界有不同看法．由这一例子可见统计学不同于其他专门学科之间的界线．在遗传学、医学等学科中都用了很多数理统计方法，但数理统计绝不能代替这些专门学科，数理统计只是一个有用的辅助工具．因为单纯的外在的数量关系是否反映事物的实质，其本质究竟如何，必须依靠专门学科的研究才能下定论．但这也不是说，数理统计方法的作用完全是被动的．事实上，事物的本质，其根本规律性的东西常常在一些场合下有所表现．我们收集到的数据资料初看起来杂乱无章，但运用数理统计方法，就有可能透过这些纷繁的数据发现某种规律性的东西，由此作为专门研究的出发点，在科学史上有大量这样的例子，如遗传学中基因学说的提出，统计方法就起了这样的作用．所以我们说，数理统计方法在研究自然界和人类社会的规律性方面，是起着积极且主动的作用．

2．数理统计是由部分推断整体的．我们把所观察的随机现象中客观事物全体称为总体，如我们考察某厂所生产的某种灯泡寿命，此时该厂过去、现在及未来所生产的所有该种灯泡的寿命就是这一问题的总体．总体中每一个元素称为个体．在数理统计中，我们把抽取的部分个体，称为样本．例如，抽取灯泡厂某种灯泡 n 只，测定它们的寿命，就获得样本 $x_1,\cdots,x_n$，计算其平均值 $\bar{x}=\frac{1}{n}(x_1+x_2+\cdots+x_n)$．如果停留于这一步，则所得的结果还是仅与这一“部分”有关．而数理统计还要往前走一步，认为样本能很好反映总体的性质，即以样本的平均值 $\bar{x}$ 去估计总体的平均值，并进一步讨论这样估计的误差范围及其相应的可能性有多大等．由部分推断整体，这是数理统计的特点，在这一过程中要用到许多数学工具，建立一套统计推断的理论．

3．数理统计是从事物偶然性中寻求它的必然规律的一种科学方法．统计方法的特点是大量观察，从搜集数据开始，经过整理、

分析，大量偶然因素将相互抵消而呈现出数据内在的规律性. 例如，新生儿的性别比例，从个别家庭看，生男生女是随机的，但大量地观察，会发现新生的男婴和女婴的比例基本接近，男的略高于女的，约为 106∶100. 这一规律是在人类长期遗传和发展中形成的较稳定的比值. 人类社会为保持平衡和发展，要求男女总人数大致相同. 虽然新生儿男多于女，但男孩死亡率高于女孩，越接近青年，同年龄组中男女间人数差别越小，30 岁左右的男女达到基本相等，中年后的男女，男性死亡率高于女性，老年男女，男性就少于女性了，但男女总人口数是基本平衡的.

随着现代科学技术和工农业生产飞速发展，统计方法得到愈来愈深入的广泛应用，对人类认识和改造世界产生了重大影响. 有人把统计学列为本世纪几十项最重大的成就之一. 一些国家在国情调查中不用普查方法而改用抽样的方法，取得了良好效果. 以数理统计方法分析各种社会调查的资料，已成为行政首脑决定政策的重要依据. 统计在社会与科学中各方面的应用，使统计学基本知识已普遍成为高等教育许多专业不可缺少的基础课.

在人类历史上，早就有带统计性质的工作. 我国古代就有“结绳记事”，在《二十四史》中有许多关于人口、钱粮、水文、天文、地震等资料记录. 在西方，统计(statistics)一词就是从国家(state)一词演化而来，它是指一种收集和整理国情数据资料的活动. 近代统计学的发展源于本世纪初，随着经济发展的需要及概率论的进步为统计理论与方法的产生与发展提供了充分的条件. 抽样调查、统计推断的理论与方法开始形成，这一基本方法论体系，至今已成为数理统计学教科书的基本内容. 30 年代初，前苏联数学家柯尔莫格罗夫(A. H. Колмогоров 1903～1989)把测度论和集合论引入于概率论，给出了概率论的公理化体系，由此促进了概率论飞速发展. 与此同时近代统计学在理论和应用方面也得到快速发展，数理统计的各个分支开始形成并迅猛发展，诸如：时间序列、回归分析、试

验设计与分析、非参数统计、多元分析、贝叶斯统计、统计决策、序贯分析、可靠性理论等诸多分支如雨后春笋般产生并获得广泛应用. 随着计算机的发展与应用,统计学如虎添翼般发展起来,至今已是数量分析的基础学科,研究和应用的领域越来越广,也越来越深入,已成为经济管理决策及各项科学研究的不可缺少的有效工具.

本书按经济类专业的与国际接轨的大统计课程的要求编写,以讲授基本统计思想与方法为主,适用于非数学类的文、理、工程各专业作为应用统计课教材,也可作为成人高校有关专业的应用统计课教材.

本书第一章讲授资料整理与描述,第二、第三两章属经济统计基本知识(非经济类专业,或经济类专业已另行开设经济统计的可略去不讲),第四章至第六章以最少的篇幅介绍概率论的基本知识,第七章至第十章讲授基本统计方法.本书中尽我们可能吸收了国内外部分新颖的例子和内容,也介绍了运用通用软件解回归分析等有关问题.70 学时的课程,可以从容地讲完本教材;如仅有 54 学时左右,可删去含 * 号的章节;如仅有 36 学时,可用一半多些学时讲第一章、第四章至第六章(删去 §5.5 及带 * 部分),用余下的学时讲 §7.1～§7.3,§8.1～§8.3,§9.1,§10.1 及 §10.2 的部分内容.

目　录

第一章　资料的搜集、整理与描述

§1.1　资料的搜集

进行统计分析和研究，首先就是要根据研究的目的和要求去搜集数据资料. 而资料的齐全和准确与否，直接影响到能否进行高质量的统计分析. 如果搜集到的数据不准确或残缺不全，那么根据这种数据进行整理和分析的结果就不能如实反映客观事物的真相，甚至还会得出相反的结论. 所以说，资料的搜集对于整个统计研究是十分重要的.

收集资料的方法有两种，即观察和试验. 所谓**观察**是客观事物发生的过程与结果的记录. 一般说来，观察者只是对感兴趣的事物客观地记录所发生的结果，而不能或不去企图改变被观察的事物，观察者是处在被动的地位，如天文观察、对社会经济现象的调查等. 所谓**试验数据**是指试验者处在主动的地位，可以在一定范围内自由地控制某些因素，以考察其他因素的作用. 例如，化工生产中，原料的配比、化学过程的温度、时间等都对产出率有影响，试验者可以控制某些因素以观察另一些因素对产出率的影响，以及它们之间交互作用对产出率的影响.

资料的来源也分为两类：一类是来源于出版物的资料，另一类是统计调查资料. 有些资料的搜集是无法进行调查的，尤其对于社会科学领域中社会宏观分析，如各国“现代化”进程的比较等等. 对前几年经济状况也无法重新观察或调查，尤其对国外的资料就更难以进行实地调查，必须应用第二手资料，如从国家统计部门编

制的统计年鉴和联合国统计局编制的统计资料中取得，或在别的研究中曾经加以搜集并公开发表的资料中取得，这些数据资料统称为来源于出版物的资料．当没有现成资料可以利用时，由研究者亲自去进行观察或调查所获取的资料，称之为原始资料或统计调查资料．

在搜集原始资料、组织统计调查时，应根据不同的情况采用不同的调查方式．调查方式可分为全面调查和非全面调查两种方式．

全面调查即普查，费用较高，而且所需的调查时间较长，但可以获得每一个研究对象的数据，信息量比较大．如果只需要了解研究对象的总体特征，而不是每一个对象的情况，则采用非全面调查．非全面调查主要指抽样调查，可以节省人力、物力，减少调查时间，提高调查质量（为什么要强调抽样调查，在第六章将作进一步论述．）

为了使调查工作能有组织有计划地进行，以达到预期的目的，在统计调查工作之前，必须做好各种准备，事先设计一个切实可行的统计调查方案．一个完善的统计调查方案，应包括以下基本内容：

1．确定调查目的．确定调查目的就是明确调查要解决什么具体问题，因为它涉及到确定调查对象、调查内容及调查时间与经费等一系列问题．

2．确定调查对象．调查对象的选择首先要服从于调查的目的，而且必须考虑调查的可行性，即要使调查者有充裕时间来与被调查者接触，才有希望达到调查的预期目的；其次应取得被调查者的合作．此外，还须明确调查对象的范围，一旦调查对象的范围界限被确定，就构成了统计总体．

3．编制调查大纲．即实际进行调查时的行动纲领，它包括调查的目的、对象、方式及组织分工、步骤和调查项目．重点是调查项目，它的含义要具体明确，切忌似是而非、含糊不清，避免调查者或

被调查者按照各自的理解进行解释或回答问题.此外,还需对调查项目的重要程度进行排队归类、分清主次,既考虑横向的项目,也要考虑纵向的项目,列出必不可少的调查项目.

4. 设计调查表格和问卷.在设计调查表格之前,最好先对所要调查的课题有个初步了解,然后动手设计.设计时力求做到:繁简得当,概念明确,附有必要的填表说明.

5. 调查时间、调查人员的培训及调查经费的筹措.应明确规定调查时间,这样才能使资料及时汇总,保证数据有意义.为保证调查的质量,必须对调查人员的基本素质进行专门的训练,统一口径范围.为保证调查能顺利地进行,还必须落实经费的来源.如果是抽样调查,还应规定样本的容量和抽选的方法(其基本方法在§6.1中介绍).总之,统计调查是统计人员必备的一项非常重要的基本功,要有很好的技巧,而这些技巧主要不是从书本上学来的,而是从实践中摸索出来的.

§1.2 资料的整理

通过观察或试验得到的原始数据,一般是杂乱无章的,往往看不出其中的规律性,需要加以整理.统计资料的整理工作质量如何,直接影响整个统计工作的效果.因此,在整理之前,必须对原始数据进行认真审核,逐一检查原始记录是否按规定的要求填写完全、正确,查明有过失错误的数据应予舍去,发现有计算或记录错误的数据应予纠正.但不能轻易剔除数值异常的数据,因为这些数据可能反映了重大变化的影响,应进一步查明其原因.

如果资料的数量较多,还需要对资料进行科学分类、分组编制分布数列,这是资料整理的核心问题.本节重点介绍对数据资料进行整理的一种重要方法 —— 次数(频数)分布及其图示.

一、资料的分类

如果资料的数量较少，只需按照一定的顺序加以排列就形成一个统计数列. 如果资料的数量较多，就要将总体中各单位按某种标志分为若干组，在各组内将数据排成数列，这就是所谓统计资料分类. 对于统计数列，按其内容可分为三种：空间数列、时间数列和变量数列. 空间数列（也称为地区数列）是按照不同的空间（如国家、地区等）排列的数列，如全国人口数按省、市分组等；时间数列也称动态数列，是按照时间顺序排列的数列，可以按年排列，也可以按季、月排列；变量数列则是由数据的变量差异所形成的数列，但变量又可按其性质上的特征，分为属性变量（定性变量）和数字变量（定量变量）两类. 属性变量是指无法用数字来计量的变量，如人口按性别、职业、宗教信仰分类；数字变量是指可以用数量来划分的，如人口按年龄分类等.

二、次数分布

资料整理的第一步是根据研究任务的要求对原始资料进行分组或分类. 在统计分组基础上，将总体中所有单位按一定标志的分组列出数据观察值在各组中出现的次数，就形成次数分布（或分配）数列，简称分布数列.

按分组标志特征的不同，分布数列可分为品质标志分布数列和数量标志分布数列.

按品质标志（非数字）分组编制的分布数列简称**品质数列**. 例如，某公司把购买该公司信用卡的顾客按其职业类型进行分组，可以编成品质数列，如表 1-1 所示.

按数量标志分组编制的分布数列简称为**变量数列**. 下面结合一个例子说明形成次数分布的过程，即变量数列的编制方法. 表 1-2 中的数据资料是对 30 个新生进行综合测评后的平均等级分.

表 1-1　顾客按职业类型的分布

职业类型	顾客人数
管理人员	38 835
技术人员	31 262
服 务 员	14 011
职　　员	12 797
工　　人	1 577
其　　他	22 273
合　　计	120 755

表 1-2　对 30 个新生测评的平均等级分

2.0	3.1	1.9	2.5	1.9	2.7
2.3	2.6	3.1	2.5	2.1	2.8
2.9	3.0	2.7	2.5	2.4	2.2
2.7	2.5	2.4	3.0	3.4	2.7
2.6	2.8	2.5	2.7	2.9	2.1

1. 确定组距. 组距的大小要适度，要能正确反映总体的分布特征及其规律. 组距与组数成反比例，组距越大组数就越少(组数 = 全距 ÷ 组距). 组数过少，容易把不同质的单位归在一个组内；组数过多，又容易把同质的单位分散在不同的组内，两者都不符合分组的要求. 至于是采用等距分组还是不等距分组，要根据现象的特点、研究的目的及所搜集到的资料分布是否均匀来确定. 如果资料分布比较均匀，就可采用等距组. 一般组数以不少于 7 组或不超过 15 组为原则. 以表 1-2 数据资料为例，因为资料数量不多，以分 8 组较为适宜. 组距通常是将资料中最大值与最小值之差除以组数，然后进行适当调整使之凑成便于运用的整数，即

$$\text{组距} = \frac{\text{最大值} - \text{最小值}}{\text{组数}}.$$

例如，表 1-2 中组距为$\frac{(3.4-1.9)}{8}=0.1875\approx 0.2$.

2. 确定组限. 组的上限和下限统称为**组限**. 确定组限的基本原则是：按这样的组限分组后，要能使全体数据资料“不重不漏”. “不重”就是任一数值只能分在一组中，不能同时分在两组中；“不漏”就是任一数值都不能遗漏. 当变量值都是整数时，它们之间有明显的界限，因此可用肯定性的数值表示组的上下限，组限非常清楚. 例如，人数分组，其组限可表示为

100 人以下，
100 ～ 499 人，
500 ～ 999 人，
1000 人以上.

对于连续型变量，其变量值有小数，组限不能用肯定性的数值表示，只能用前一组的下限与后一组的上限重叠的方法表示. 例如，学生考分的分组：50 ～ 60，60 ～ 70，70 ～ 80，80 ～ 90，90 ～ 100，原则是“上组限不在内”，即每组的上组限数值不包括在本组内，如考分是 80 分时，它应落在 80 ～ 90 分之内. 为避免出现这种重叠组限的情况，也可采用 50 ～ 59，60 ～ 69，70 ～ 79，…，有时，为了使组限清楚，选择比变量值的小数多一位小数的处理方法. 例如，对表 1-2 中的数据，带一位小数，最小值是 1.9，我们选择 1.85 为下限，这样组限就用肯定性的数值：1.85，2.05，2.25，2.45，…，表示，使变量值之间有明显的界限.

3. 编制次数分布表. 用手工整理资料编制次数分布表时，先编制划记表. 现以表 1-2 的资料为例，用划记法来编制次数分布表，如表 1-3.

4. 累积次数分布. 有时需要观察某一数值以上或某一数值以下次数之和，即累积次数，如表 1-4.

表 1-3　次数分布表

组	组　限	登记法 Ⅰ	登记法 Ⅱ	次　数
1	1.85 ～ 2.05	下	𝍸	3
2	2.05 ～ 2.25	下	𝍸	3
3	2.25 ～ 2.45	下	𝍸	3
4	2.45 ～ 2.65	正下	𝍸𝍸	7
5	2.65 ～ 2.85	正下	𝍸𝍸	7
6	2.85 ～ 3.05	下	𝍸	4
7	3.05 ～ 3.25	下	𝍸	2
8	3.25 ～ 3.45	一	𝍸	1

表 1-4　累积次数分布表

分　组	次　数	较小累积	较大累积
1.85 ～ 2.05	3	3	30
2.05 ～ 2.25	3	6	27
2.25 ～ 2.45	3	9	24
2.45 ～ 2.65	7	16	21
2.65 ～ 2.85	7	23	14
2.85 ～ 3.05	4	27	7
3.05 ～ 3.25	2	29	3
3.25 ～ 3.45	1	30	1
合　计	30		

较小累积，是以最小组的次数为起始点逐项累计各组次数. 例如，第 3 组的较小累积次数是 9，说明计分在 2.45 以下的人数合计是 9，占全部人数的 $\frac{9}{30}=30\%$.

较大累积，则是从最大组的次数开始，逐项累积各项的次数，表示该组下限以上的次数的合计. 例如，第 6 组的较大累积次数是 7，说明计分在 2.85 以上的人数合计为 7，占全部人数的 7/30 ≈

23.3%.

三、次数分布的图示

为了使次数分布更直观，采用图形来表示次数分布.常用的分布图有直方图、折线图、茎叶图等等.

1. 直方图.在平面直角坐标系中，将次数分配中分组标志作为横轴，对等距分组时，将各组次数作为纵轴，一个组对应一个矩形条(组距为宽度、次数为高度)，就可以绘出直方图.由表 1-2 的次数分布所绘制的直方图如图 1-1 所示.对不等距分组时，要先计算出各组的频数密度，然后以组距为宽，以频数密度为高画直方图.频数密度 = 频数 ÷ 组距.

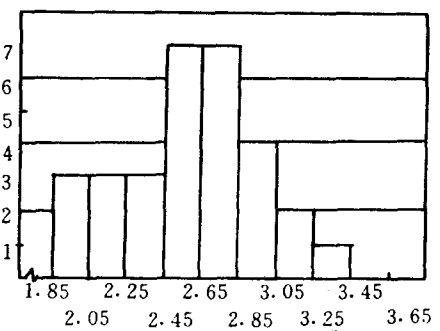


图 1-1　次数分布的直方图

2. 折线图.在直方图的基础上，将直方图各矩形顶端中点用直线连接而成.起点是在距左边最低组半个组距处的横轴上，终点在距右边最高组半个组距处的横轴上，如图 1-2.

3. 茎叶图.茎叶图是一种在现场统计中用来描述、分析和图示资料的较好方法.它既简单直观，又能较全面反映数据的规律性.它是将分组和画直方图两步工作合成一步.