数理统计理论、方法、应用和软件计算

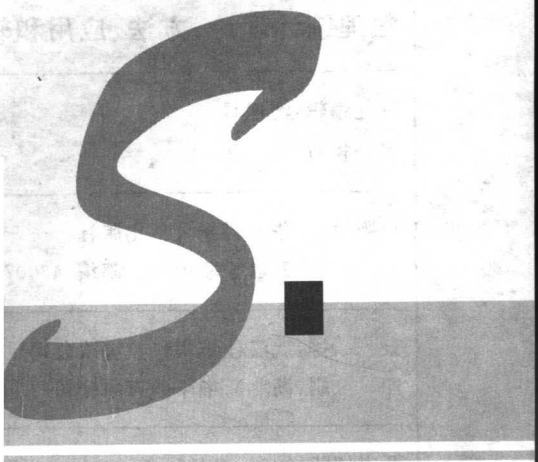
刘顺忠 编著

华中科技大学出版社
E-mail: hustpp@wuhan.cngb.com

- 强调理论、方法、实例和软件计算相统一
- 注重基本方法的使用技巧，有助于提高实际应用能力
- 通俗易懂，特别适合非数学专业人员学习
- 依据例题，逐步操作，轻松学习SPSS13.0

数理统计理论、方法、应用和软件计算

刘顺忠 编著



华中科技大学出版社

图书在版编目(CIP)数据

数理统计理论、方法、应用和软件计算/刘顺忠　编著
武汉:华中科技大学出版社,2005 年 9 月
ISBN 7-5609-3445-5

Ⅰ.数…
Ⅱ.刘…
Ⅲ.数理统计-理论-应用
Ⅳ.O212

数理统计理论、方法、应用和软件计算　　刘顺忠　编著

责任编辑:沈旭日　　封面设计:刘　卉
责任校对:刘　飞　　责任监印:熊庆玉

出版发行:华中科技大学出版社
武昌喻家山　邮编:430074　电话:(027)87557437

录　　排:武汉佳年华科技有限公司
印　　刷:湖北新华印务有限公司

开本:787×960　1/16　　印张:13.25　　字数:225 000
版次:2005 年 9 月第 1 版　　印次:2005 年 9 月第 1 次印刷　　定价:18.80 元
ISBN 7-5609-3445-5/O・361

内 容 提 要

本书将数理统计的原理、方法、计算、实例和SPSS软件计算相结合，以数理统计原理和方法为出发点，将手工计算过程和SPSS软件计算过程相结合，使读者能够从数学角度深刻理解SPSS软件的数据处理方法、过程和处理结果。本书介绍了大量的非参数统计方法，这些方法是当前社会科学实证研究的锐利武器，这也是本书的独特之处。

全书共有11章和两个附录。前2章介绍数理统计基本原理，从第3章到第11章介绍了实际工作中经常使用的数理统计方法及其SPSS软件计算方法。分别为：数理统计的基本概念、参数估计和假设检验、相关分析、回归分析、定性数据统计分析、非参数检验、方差分析、聚类分析、典型相关分析、主成分分析、因子分析。

本书可作为各大专院校非数学专业如经济学、统计学、市场营销学、医学、心理学、人文地理学、社会学、管理学等专业学生学习数理统计及SPSS软件的教材，也可供统计人员、科研人员以及广大自然科学工作者做科研定量分析时参考。同时，还可作为高等院校财经类专业研究生和本科生掌握SPSS统计分析方法和软件使用方法的工具书。

序　言

数据量庞大和数据关系越来越复杂是目前统计应用中遇到的主要问题，只有快速准确地处理庞大的数据量和复杂的数据关系，才能够有效解决实际工作和科研中遇到的问题。电子计算机的广泛普及和软件的不断完善，为统计分析中的数据处理和计算提供了强有力的工具，促进了统计方法在各个领域中的广泛应用。目前有关统计分析的软件主要有SPSS(Statistics Package for Social Science，社会科学统计软件包)和SAS(Statistical Analysis System，统计分析软件系统)，这些软件大大方便了统计数据的处理和分析，特别是SPSS软件由于使用Windows的窗口操作方式进行数据的处理和分析，避免了繁琐的编程工作，是深受非数学专业人员欢迎的统计分析软件。统计软件是专用软件，不同于不需要专业知识的普通通用软件(如Word)。只有掌握了统计分析理论，才能够选择相应的软件模块对数据进行处理分析，并对软件输出结果进行科学解释，从而为科学决策提供有力的数据支持。

具有数理统计专业基础的读者在学习SPSS软件过程中经常遇到这样的问题：首先，如何组织和加工数据，才能利用SPSS软件进行数据处理和分析？其次，怎样操作SPSS软件对数据进行处理和分析以及如何理解和合理解释分析结果？……随着计算机的普及和数据处理的需要，目前越来越多的科研工作者、高校学生、市场营销人员和管理工作者使用统计软件对统计数据进行处理和分析，他们不具备数理统计专业基础，在用SPSS软件对数据处理过程中不但会面对统计专业人员面临的SPSS软件使用问题，更为重要的是需要学习和理解各种数理统计方法的原理及其应用。

我在市场调研公司和广告公司工作期间，以及在攻读博士期间都深感统计分析在实际工作和科研中的重要作用。从2000年开始，我就开始了本书的编写工作，为了适合非数学专业人员的特点，本书在讲述数理统计理论方法的同时，侧重从这些方法的计算过程出发讲解数理统计分析方法的应用。本书对各种数理统计方法的计算过程都给出了手工计算实例，不但便于读者理解数理统计方法，也方便读者理解SPSS软件的计算过程。经过给研究生和高年级本科生授课，使初稿不断完善，形成了现在的这本书。

本书将数理统计的原理、方法、计算、实例和SPSS软件计算相结合，以数理统计原理和方法为出发点，将手工计算过程和SPSS软件计算过程相结合。本书各章结构基本相同，都是数理统计原理介绍，计算方法，手工实例计算和统计软件实例计算。

本书的前2章介绍数理统计的基本原理，第3章到第11章介绍了实际工作中经常使用的数理统计方法及其SPSS软件计算方法，书籍的附录介绍了数理统计的基本概念、线性代数基本概念和SPSS软件的基本操作。

本书可作为各大专院校非数学专业如经济学、统计学、市场营销学、医学、心理学、人文地理学、社会学、管理学等专业学生学习数理统计及SPSS软件教科书，也可供统计人员、科研人员以及广大自然科学工作者做科研定量分析时参考。同时，还可作为高等院校财经类专业研究生和本科生掌握SPSS统计分析方法和软件使用的工具书。由于水平所限，书中的问题与错误在所难免，恳请读者批评指正。

刘顺忠

2005年5月于武昌南湖

目　录

第1章

数理统计的基本概念

由于偶然性和不确定性，数据的随机性常常是无法避免的。数理统计是对带有随机性影响的数据进行有效搜集、整理、分析和推断，将杂乱无章的各种数据转化为简洁有序和容易解释的信息，并对观察到的现象作出推断或预测，为决策提供依据的数学方法。掌握数理统计的基本概念是理解和使用各种数理统计分析方法的前提，下面首先介绍数理统计分析中常用的基本概念。

1.1 抽样方法和测量水平

1.1.1 总体和样本

在数理统计中，把研究对象的全体称为总体，而把组成总体的每一个单元称为个体。例如，若研究某城市人口的年龄结构，则该城市全部人口就构成了总体，而其中每个城市公民就是个体；若研究某纺织厂一天生产的20000件服装的合格率，则这20000件服装就组成一个总体，而每一件服装就是一个个体。

在实际研究中，我们关心的常常是研究对象的某个数量指标X(如人口寿命，产品是否合格)，它是一个随机变量。因此，总体通常是指某个随机变量取值的全体，每个个体就是其中的一个实数。因此，可以把总体和数量指标X可能取值的全体组成的集合等同起来，随机变量X的分布就是总体的分布。

从一个总体X中，随机抽取n个个体$X_1,X_2,\cdots,X_n$，其中$x_i(i=1,2,\cdots,n)$是第i个个体抽样观察结果，称$x_1,x_2,\cdots,x_n$为总体X的一组样本观察值。对于某一次抽样结果来说，它是完全确定的一组数，但是，由于抽样的随机性，所以它又是随每次抽样而改变的，这样每个x_i都可以看作某一个随机变量$X_i(i=1,2,\cdots,n)$所取的观察值。将$X_1,X_2,\cdots,X_n$称为容量为n的样本，$x_1,x_2,\cdots,x_n$就是样本$X_1,X_2,\cdots,X_n$的一组观察值，称为样本值。

比如，在研究某城市人口年龄结构时，从全部城市公民X(总体)中随机抽取

1000 人，记这1000 人为 $X_1, X_2, \cdots, X_{1000}$，则 $X_1, X_2, \cdots, X_{1000}$ 构成样本。若样本 $X_1, X_2, \cdots, X_{1000}$ 的取值记为 $x_1, x_2, \cdots, x_{1000}$，则 $x_1, x_2, \cdots, x_{1000}$ 就是样本 $X_1, X_2, \cdots, X_{1000}$ 的一组观察值。

1.1.2 抽样方法

若要获得总体中的样本，就必须采取一定抽取样本的方法。抽样方法主要有两种：概率抽样和非概率抽样。

1. 概率抽样

所谓概率抽样就是根据已知的概率来选取样本，无须在抽样过程中对样本进行判断和抽选。概率抽样主要有以下 4 种。

简单随机抽样是最基本的抽样形式，它是完全随机地选取样本。该方法要求有一个完美的抽样框，或每一个个体的详尽名单。如在调查某高校学生月消费的研究中，按照随机数表选取给定学号的学生进行抽样调查，就是简单随机抽样。

分层抽样是先将总体分成不同的层，然后在每一层内进行简单随机抽样。分层抽样可以防止简单随机抽样造成的样本构成同总体构成不成比例的现象。如在研究城市人口年收入状况时，先将一个城市按照不同城区进行分组，然后在各个组中采取简单随机抽样的方法进行抽样，这就是分层随机抽样。

整群抽样是将一组被调查者视作一个抽样单位而不是个体的抽样方法。如在市场研究中，选取的是某个小区的指定楼号的所有居民进行调查，这就是整群抽样。

等距抽样又称系统抽样，是在抽样框中每隔一定距离抽取一个样本。如在研究某大学新生家庭收入情况时，按照学号每隔 100 号抽取一个学生进行调研，就是等距抽样。

2. 非概率抽样

非概率抽样不是完全按照随机原则，而是按照研究人员主观判断或研究方便的原则选取样本。非概率抽样主要有以下 4 种。

方便抽样又称偶遇抽样，是根据调查者的方便来抽取样本的抽样方法。如在市场调查中经常使用的街头拦截访问就是方便抽样。

判断抽样又称目的抽样，它是凭研究人员的主观意愿、经验和知识，从总体中抽取典型代表性样本的方法，主要是根据研究目的选取具有代表性的样本或极端样本。如企业质量管理人员根据以往的经验，从所有会出现次品的产品中抽取样本，这就是判断抽样。

配额抽样是非概率抽样中最流行的一种。它首先将总体按照一定标志分为若

干组，然后在每组中用方便抽样或判断抽样来选取给定数量的样本。如在研究某种新型移动通信服务时，将目前手机用户按照月话费额进行分组，然后在每组中根据方便抽样原则抽取给定数量的样本，这就是配额抽样。

雪球抽样是以滚雪球的方式来抽取样本，即在少量样本的基础上不断扩充以获得更多样本的信息。这种抽样方式运用的前提是总体各样本之间存在一定的联系，可以在不甚了解总体的情况下对总体或总体部分样本情况进行把握。如在研究知识密集型服务业创新机理过程中，由于知识密集型服务业显著特征就是相互之间联系密切，因此可以以少量知识密集型服务业企业作为原始“雪球”，采取“滚雪球”的抽样方式不断扩充对知识密集型服务业企业的抽样，这就是雪球抽样。

随机抽样是以概率论为基础，容易进行估计、检验和控制抽样误差，因此以概率论为基础的数理统计分析适合处理随机抽样得到的数据。从原则上讲，数理统计分析的方法不适合处理非随机抽样得到的数据，但是，由于非随机抽样简单易行，适合对数据进行探索性分析。

1.1.3　测量水平

当我们对样本进行测量时，也就是赋予给定变量以具体数值。变量测量水平主要分为以下4类，不同类别的测量水平所容许的运算是不同的。

1. 分类量表

分类量表又称为称名量表。它是使用一些数字或其他符号来识别不同事务所属类别的一种方法。这些数字或符号就构成一个称名量表，它是测量水平最低的量表。

比如，在对性别进行分类时，男性用1表示，女性用2表示，这时，数字1和2就构成一个称名量表。又比如，在对学历进行分类时，学士用B表示，硕士用S表示，博士用D表示，这时，符号B,S,D也构成了一个分类量表。

分类量表运算是把一个给定的集合分成一个个互不相交的子集合，它们所包含的关系是等价关系，即任何一个子集成员在被标度的特征方面是相同的。比如在使用数字1和2对性别进行标度时，所有被标度为数字1的都是男性，所有被标度为数字2的都是女性。

2. 顺序量表

顺序量表又称评秩量表。如果在分类量表中，所有子集之间都保持一个顺序关系，则该量表构成一个顺序量表。

比如，按照年龄对人进行划分，则人可以划分为儿童(A)、少年(B)、青年(C)、中年(D)和老年(E)5个子集。每个子集如果按照年龄排序，则存在一个偏序关系，

即A＜B＜C＜D＜E。这样，不同年龄的人之间就存在一个顺序关系，因此也就构成了一个顺序量表。

顺序量表包含偏序关系和等价关系，但是，还不能进行算术运算。

3. 间隔量表

当一个量表具有顺序量表的特征，并且量表中任意两个数之间的距离为已知时，其测量水平就高于顺序量表，称为间隔量表。间隔量表的特征是有公共的、不变的测量单位，它将一个实数赋给样本中的每个个体。在间隔量表中，零点和单位是任意的。

比如温度的测量，通常采用两种不同的量表——摄氏量表和华氏量表。在测量温度时，测量单位是任意的；这两种量表具有不同的零点和单位，却包含了相同的信息量，一个量表中温度差的比值和另一个量表中温度差的比值是相同的。

在摄氏温度量表和华氏温度量表中，表1-1所示的温度是相同的。

表1-1　摄氏和华氏温度对照表

量表	相同温度			
摄氏/°C	0	10	30	100
华氏/°F	32	50	86	212

在摄氏温度量表中$\frac{30-10}{10-0}=2$，在华氏温度量表中$\frac{86-50}{50-32}=2$，在使用不同间隔量表时其温度差的比值是相同的，与零点和单位没有关系，由于间隔量表是将实数赋给样本观测值，所以该量表可以进行算术运算。

4. 比例量表

当间隔量表以真正的零点作为原点时，就称为比例量表。比例量表是测量水平最高的量表。

比如，在测量物体重量时，可以使用公斤和市斤，两个量表都具有零点。任意两个重量间的比值与测量单位无关。

比如，在测量两个物体重量时，公斤量表分别为10和50，市斤量表分别为20和100，则50/10＝5，100/20＝5，两个重量间比值与测量单位无关，比例量表可以进行算术运算。

1.2　统计量及其计算

通过试验或观察取得的样本值往往呈现为一堆杂乱无章的数据。因此，在利用

样本数据推断总体特征时，往往不能够直接利用样本数据，而需要对它们进行一定的加工和整理，通过计算出来的一些量来对总体进行推断，否则无法有效利用样本数据中的信息。

【例1-1】 某系随机抽取30名学生的英语考试成绩，以调查该系学生英语考试情况。现得到下列数据，试对该系学生英语学习成绩和相互之间的差距状况进行大致分析。

73	61	67	71	90	88
66	91	75	80	71	61
80	61	68	66	87	60
64	80	83	80	86	75
44	75	60	98	65	56

【解】 显然，面对这些参差不齐的数据，很难得出什么印象。但是，只要对这些数据稍事加工，便能进行分析，并得出结论。设这30名学生的英语成绩分别为x_1，x_2，$\cdots$，x_{30}，则

$$\bar{x}=\frac{1}{30}\sum_{i=1}^{30}x_i=72.73 \tag{1-1}$$

$$S=\sqrt{\frac{1}{30}\sum_{i=1}^{30}(x_i-\bar{x})^2}=12.24 \tag{1-2}$$

由$\bar{x}$得出该系学生英语平均成绩为中等，由S得出该系学生英语成绩差距不大的结论。因此，对样本数据进行加工和处理是十分重要的。

对样本数据进行加工和处理主要就是构造统计量。如果用数学的语言来描述，则统计量就是一个不含未知参数的样本已知函数。设样本为X_1，X_2，$\cdots$，X_n，则统计量通常记为

$$T=T(X_1,X_2,\cdots,X_n) \tag{1-3}$$

设X_1，X_2，$\cdots$，X_n为总体X的样本，其观测值为x_1，x_2，$\cdots$，x_n，则常用统计量如下：

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{1-4}$$

$\bar{x}$称为样本均值。

$$S=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{1-5}$$

S称为样本标准差，S^2称为样本方差。

众数就是一组样本中出现次数最多的变量值，记为M_0。对统计量的描述还有最大值

$$\mathrm{MAX}=\max(x_1,x_2,\cdots,x_n) \tag{1-6}$$

最小值

$$\mathrm{MIN}=\min(x_1,x_2,\cdots,x_n) \tag{1-7}$$

样本极差

$$D^*=\mathrm{MAX}-\mathrm{MIN} \tag{1-8}$$

已知 $X_1,X_2,\cdots,X_n$ 的取值 $x_1,x_2,\cdots,x_n$，假设 $x_1<x_2<,\cdots,<x_n$，则

$$M^*=\begin{cases} x_{\frac{n+1}{2}}, & n\text{ 为奇数} \\ \dfrac{1}{2}\left(x_{\frac{n}{2}}+x_{\frac{n}{2}+1}\right), & n\text{ 为偶数} \end{cases} \tag{1-9}$$

M^* 称为样本中位数，中位数是将样本从中间等分为两部分。

与中位数类似的还有四分位数，四分位数也称为四分位点，它是通过3个点将全部样本数据分为四部分，其中每部分包含25%的样本，处于分位点上的样本值就是四分位数。很显然，中间的四分位数就是中位数，通常所说的四分位数是指第一个四分位数(下四分位数)和第3个分位数(上四分位数)。

设下四分位数位为 Q_L，中间四分位数为 Q_M，上四分位数为 Q_U，则 Q_L 的位置是 $\frac{N+1}{4}$，Q_M 的位置是 $\frac{2(N+1)}{4}$，Q_U 的位置是 $\frac{3(N+1)}{4}$，显然 $Q_M=M^*$。

同中位数计算相同，当四分位数不在某一个数值上且不为整数时，可以根据四分位数的位置，按照比例分摊四分位数两侧数值的差值。

根据例1-1的数据计算上述各统计量，首先对例1-1中30个成绩数据从小到大排序，结果如表1-2所示。

表1-2　成绩数据升序排序表

序号	分数	序号	分数	序号	分数
1	44	11	66	21	80
2	56	12	67	22	80
3	60	13	68	23	80
4	60	14	71	24	83
5	61	15	71	25	86
6	61	16	73	26	87
7	61	17	75	27	88
8	64	18	75	28	90
9	65	19	75	29	91
10	66	20	80	30	98

根据表1-2中的数据，可计算出该班分数样本数据的各个统计量的值：