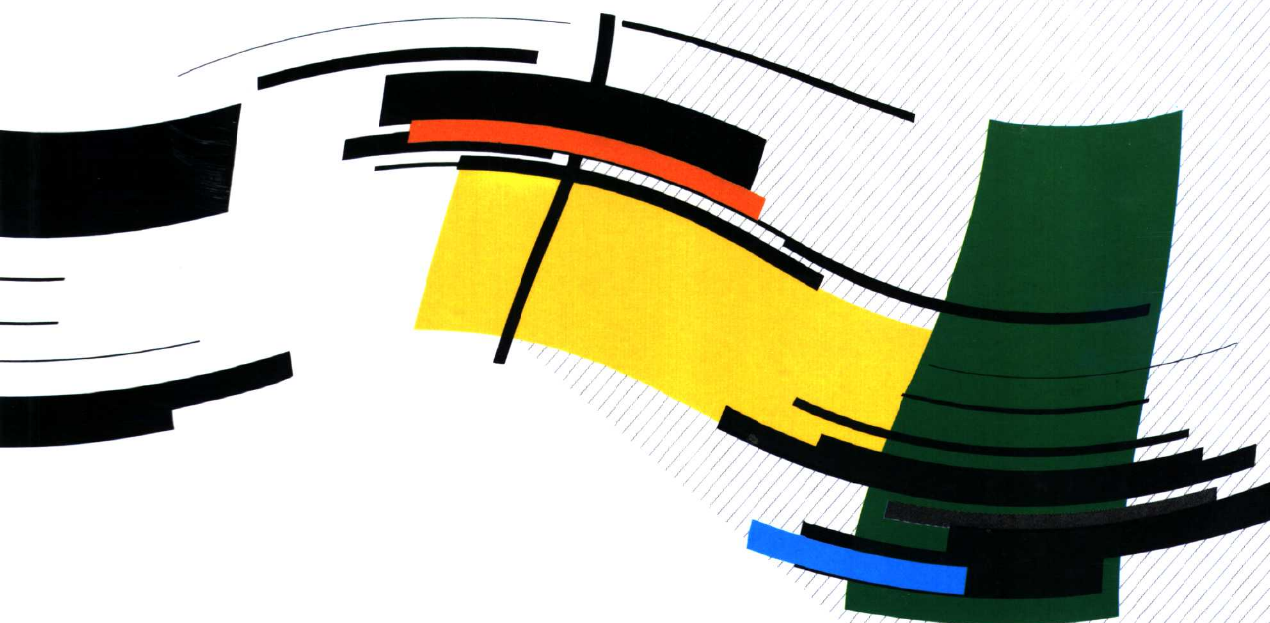
现代应用统计学

XIANDAI YINGYONG TONGJIXUE

马开玉 张耀存 陈 星 ● 编著



气象出版社

China Meteorological Press

现代应用统计学

马开玉 张耀存 陈 星 编著

气象出版社

内 容 简 介

本书简明扼要地介绍了基本的统计理论与方法，着重阐述了具体应用和实例分析计算。全书共分八章，其中第一章讲述资料的基本统计；第二、三和四章介绍了随机变量及其分布、统计推断和随机抽样；第五、六章着重阐述了常用的回归分析和时间序列分析；第七、八章介绍了极值分析和决策分析。

本书为大学本科应用型学科通用教材，也可为大专院校教学参考书。书中有较多的例题，便于读者自学。

图书在版编目(CIP)数据

现代应用统计学/马开玉，张耀存，陈星编著. —北京：气象出版社，2004.2
ISBN 7-5029-3730-7

Ⅰ.现…　Ⅱ.①马…②张…③陈…　Ⅲ.应用统计学　Ⅳ.C8

中国版本图书馆 CIP 数据核字(2004)第 006281 号

Xiandai Yingyong Tongjixue
现代应用统计学
马开玉　张耀存　陈　星　编著
责任编辑：张锐锐　陶国庆　终　审：周诗健
封面设计：世纪白马　责任技编：王丽梅　责任校对：王丽梅

出版发行：气象出版社

出版社地址：北京市海淀区中关村南大街 46 号　　邮政编码：100081
出版社电话：68407112　　传真号码：62176428
电子邮箱：CMP01@263.NET
出版社网址：HTTP://CMP.CMA.GOV.CN/

印　刷：北京昌平环球印刷厂
开　本：787mm×960mm　1/16　　版　次：2004 年 2 月第 1 版
印　张：15.0　　印　次：2004 年 2 月第 1 次印刷
字　数：302 千字　　定　价：24.00 元

前　言

统计是人类认识社会和自然的有效工具，在各学科领域和各行各业中有着广泛的应用。在实际生活中，尽管具体从事统计工作的人是少数，但对于那些或是从事某一学科领域的研究，或是从事某一部门或单位管理的人来说都要与统计数据打交道，甚至在个人或家庭生活和理财中也常常要用到一些统计知识。

现有的统计学教材，或是理论性较强，不利于应用性学科学生的学习和掌握，或是专业性较强，不利于其它学科用作教材。为使统计学教材更适应教育改革和市场经济的发展，更适合应用型学科学生的学习和掌握，本教材吸收了近年来国内教学改革和教材编写的最新成果，紧密结合生产实践和经济建设中的新问题和新资料，系统介绍了基本的统计学理论与方法，着重阐述了具体应用和实例分析计算，以帮助学生理解和掌握，培养学生运用统计理论与方法分析问题和解决问题的能力。

《现代应用统计学》一书不仅涵盖了一般统计学中常见的概率论与统计推断基础知识、回归分析和时间序列分析等内容，还增加了一般统计学中不做介绍而实际中又需要应用的一些章节，如逐步回归分析、极值统计方法和统计决策理论等章节。因此，本书不仅内容广泛、理论性强、应用面宽，而且还是一本实用性较强的统计学教材。

本书由马开玉执笔编写，张耀存和陈星对全书进行了编审。全书共分八章，其中第一章讲述资料的基本统计；第二章介绍随机变量及其分布函数；第三章介绍统计推断理论；第四章介绍随机抽样方法；第五章和第六章着重阐述常用的回归分析和时间序列分析；第七章介绍极值统计方法及其应用；第八章介绍统计决策理论及其实际应用。本书不仅可作为大学本科应用型学科的通用教材，也可作为大专院校的教学参考书。特别是书中附有较多例

题，便于读者自学和掌握统计学的基本理论和应用方法。

《现代应用统计学》一书由南京大学继续教育学院"985"教改项目和南京大学大气科学系联合资助出版。

在本书出版之际，对王晓如、蒋龙海和李汝谅三位教授表示真诚的谢意，感谢他们对本书的编写要求和内容提出的许多宝贵意见，以及编写过程中给我们提供的帮助。杨菲菲和汪志洋两位同学为本书制作了所有插图，在此也表示衷心的感谢。

书中不妥之处请读者批评指正。

编　者

2003年12月

目　　录

第一章　资料的基本统计

§1.1　统计学的几个基本概念

统计学也像其它各门学科一样，在叙述本门学科的理论和方法时，采用一些专门的名词、术语，这些名词术语也是从研究对象所具有的特性中总结概括出来的，并且是研究工作中经常应用的。其中的大部分将在后面各有关章节的内容中予以叙述，这里介绍的是应用统计学中常用的几个基本概念。

1.1.1　变量

各种自然现象和社会现象都是可以进行观测、试验和调查研究的，每次观测、试验和调查研究的结果可用数量表示，也可以文字叙述，在统计学中称这些结果为变量。变量有两种，一种是定量的，是以数值大小来表示的，称为定量变量，如某河流 6～8 月的最高水位，某企业工人的工资等等；一种是定性的，是以某些特征来表示的，称为定性变量，如某地 6～8 月的旱涝程度，工人的性别、工种和文化程度等等。显然定性变量也可以转化为定量变量，例如工人的性别可以用 0,1 表示，0 表示女，1 表示男；文化程度也可用 0,1,2,3,…分别代表文盲、小学、初中、高中……。

变量按其值是否连续可以分为连续型变量和离散型变量两种。有些变量只能取一些离散值，例如一年中出现某种天气现象的日数只能取 0,1,2,3,…,365 这样一些值；再如某企业的职工人数、机器的台数等等的取值。这种变量称为离散型变量。连续型变量的取值不是互相分离的，而是连续地充满一个区间，如降水量、气温、气压及人的身高等等。

变量按所受影响因素的不同又可分为确定性变量和不确定性变量。确定性变量是由某种起决定性作用的因素影响着它的起伏变化。不确定性变量的影响因素很多，因而可能出现多种可能的结果，表现出一定的波动性与随机性，因此不确定性变量也称随机变量，在第二章我们将详细介绍。

1.1.2　总体与个体

1.1.2.1　总体

总体是统计所研究的对象，是由客观存在的、具有相同性质的许多事物所组成的全

体,亦称母体。总体可以是各种社会现象和自然现象。例如以某市所有工业企业为研究对象,则全市的工业企业就构成一个总体。在这个总体中每个工业企业都具有共同的性质,即它们都是进行工业生产活动的基层单位。又如以某河流 6～8 月最高水位为研究对象时,则该河流历年 6～8 月的最高水位就构成一个总体,在这个总体中每一个数值都是该河流某年 6～8 月观测到的水位最高值。

1.1.2.2 个体

个体是构成总体的每一个别事物,是构成总体的基本单位(单元)。如某市的工业企业构成一个总体,则该市每一个工业企业就是这一总体中的个体。又如某河流历年 6～8 月最高水位构成一个总体,则在这个总体中每一个最高水位都是它的个体。

总体的所有个体在主要属性上必须是相同的,这是形成一个总体的必要条件,称为总体的同质性。一个总体所包含的个体数可以是无限的,称为无限总体;也可以是有限的,称为有限总体。

总体与个体的概念不是固定不变的,随着研究目的不同,它们可以转化。同一个研究对象,在一种情况下是个体,在另一种情况下则变成了总体。上例中某市工业企业构成总体,每个工业企业就是一个体。如果要研究某个工业企业职工工资状况,则该企业全部职工就成为总体,每个职工就是这个总体中的个体。

1.1.3 样本

由总体中一部分个体组成的集合称为样本或子样。样本内所含个体的个数称为样本容量,表示样本数量的大小。例如我们可以把一定自然环境下某一天气现象看成一个变量,把无限长时期内这一大气现象的全体看成一个总体,而把已有的观测资料看作样本,观测记录的总数就是样本容量。又如某地区要估计 300 000 个大小相同的地块上水稻的产量,用测框法调查了 100 个地块的水稻产量,这里水稻产量是变量,300 000 个地块的产量是总体,100 个地块的产量是样本,而 100 则是样本容量。

在统计研究中,样本是一个很重要的概念。往往由于人力、物力的原因或者研究对象本身获取的困难,无法得到总体,因而通过对样本的统计、分析和研究推断总体的基本特征和变化规律。

§1.2 平均数、众数与中位数

1.2.1 平均数

统计平均数反映了研究对象的一般水平和集中位置,它是变量的所有不同取值的一种概括和抽象,体现了个别与一般的统一。按不同的计算方法可以分为算术平均数、

调和平均数、几何平均数和平方平均数等。下面分别予以介绍。

1.2.1.1　**算术平均数**

算术平均数是统计工作中应用最广泛的一种平均数。这是因为它的计算方法同许多研究对象中的个别现象和总体现象之间存在的客观数量关系相符合。例如每亩[①] 地的粮食产量之和恰好等于全部播种面积的总产量，企业中每个职工的工资之和恰好等于全部职工的工资总额等等。

算术平均数的计算公式是

$$\bar{x}_A = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.2.1}$$

式中 $x_1, x_2, x_3, \cdots, x_n$ 代表变量 x 的取值，n 表示取值的次数或样本容量。

如果将 $x_1, x_2, x_3, \cdots, x_n$ 分成 m 组，每组的代表值或平均值为 $\bar{x}_i (i = 1, 2, \cdots, m)$，每组中取值的次数称为**频数**，分别为 $n_1, n_2, \cdots, n_m$，则(1.2.1) 可以写成

$$\bar{x}_A = \frac{\bar{x}_1 n_1 + \bar{x}_2 n_2 + \cdots + \bar{x}_m n_m}{n_1 + n_2 + \cdots + n_m} = \frac{1}{n}\sum_{i=1}^{m} \bar{x}_i n_i = \sum_{i=1}^{m} \bar{x}_i f_i \tag{1.2.2}$$

式中

$$n = n_1 + n_2 + \cdots + n_m$$

$$f_i = \frac{n_i}{n}$$

f_i 表示 x 在第 i 组取值的频率。(1.2.2) 式计算的平均数称为加权算术平均数。

例 1.1：根据 2000 年某公司 50 名职工调查资料，每人每月生活费支出如下(单位：元)，求该公司每月人均生活费。

228	242	242	245	248	251	251	251	254	255
256	258	258	259	259	260	260	260	260	261
262	263	264	264	265	265	266	266	266	267
275	277	279	280	282	285	285	286	288	292
265	267	268	269	262	263	275	266	278	292

① 1 亩 $= 666.\dot{6} m^2$，下同。

解：按(1.2.1)式计算，得人均生活费为

$$\bar{x}_A = \frac{1}{50} \times (228 + 242 + 242 + \cdots + 292)$$
$$= \frac{13240}{50}$$
$$= 264.8(\text{元})$$

按(1.2.2)式列表(表 1.2.1)计算，得人均生活费为

$$\bar{x}_A = \frac{1}{50} \times (288.00 + 977.00 + \cdots + 292.00)$$
$$= 4.56 + 19.54 + \cdots + 5.84$$
$$= 264.8(\text{元})$$

如果每组中的代表值不是该组的平均数，则用分组方法(1.2.2)式计算的平均数往往没有(1.2.1)式计算的结果准确。

表 1.2.1 某公司职工月生活费支出

组序	按每人月生活费支出分组(元)	组平均($\bar{x}_i$)	频数(n_i)	频率(f_i)	$\bar{x}_i n_i$(元)	$\bar{x}_i f_i$(元)
1	240 以下	228.00	1	0.02	228.00	4.56
2	240～250	244.25	4	0.08	977.00	19.54
3	250～260	255.20	10	0.20	2552.00	51.04
4	260～270	264.45	20	0.40	5289.00	105.78
5	270～280	274.50	8	0.16	2196.00	43.92
6	280～290	284.33	6	0.12	1705.98	34.12
7	290 以上	292.00	1	0.02	292.00	5.84
合计			50	1.00	13239.98	264.80

1.2.1.2 调和平均数

调和平均数是与算术平均数不同的另一种性质的平均数，它是根据各变量值的倒数计算的，故又称为倒数平均数。

在资料未分组时，调和平均数的计算式为

$$\bar{x}_H = \frac{1}{\frac{\sum_{i=1}^{n} \frac{1}{x_i}}{n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{1.2.3}$$

这里 $\frac{1}{x_i}$ 为变量 x 的逆变量 $\frac{1}{x}$ 的取值，x 称为正变量。

在资料分组的情况下，调和平均数的计算式为

$$\bar{x}_H=\frac{1}{\dfrac{\sum\limits_{i=1}^{m}\dfrac{1}{x_i}\cdot n_i}{\sum\limits_{i=1}^{m}n_i}}=\frac{\sum\limits_{i=1}^{m}n_i}{\sum\limits_{i=1}^{m}\dfrac{1}{x_i}\cdot n_i}$$

$$=\frac{1}{\sum\limits_{i=1}^{m}\dfrac{1}{x_i}\cdot f_i}\tag{1.2.4}$$

上式计算的平均数称为加权调和平均数。

调和平均数在经济统计学中常常应用，如劳动生产率、商品流转速度、资金周转速度、货币购买力等都可以用正变量和逆变量表示。以劳动生产率为例，它的正变量和逆变量是：

$$劳动生产率=\frac{产品产量}{劳动消耗量}\qquad（正变量）$$

或

$$劳动生产率=\frac{劳动消耗量}{产品产量}\qquad（逆变量）$$

劳动生产率的正变量与逆变量是互为倒数的关系，即

$$\frac{产品产量}{劳动消耗量}=\frac{1}{\dfrac{劳动消耗量}{产品产量}}$$

在统计研究中，计算类似劳动生产率这些变量的平均水平，当用正变量计算时，用算述平均法，用逆变量计算时，则用调和平均法。

例 1.2：某电器设备厂甲、乙、丙 3 个工人制造某种元件的劳动生产率，如表 1.2.2 所示，试用算术平均法和调和平均法计算 3 人的平均劳动生产率。

解：根据表中 3 个人的劳动生产率第 1 列，因为已知条件是每人每天做元件数量，所以采用算术平均法计算平均劳动生产率，即 3 人平均每人每天做元件。

$$\bar{x}_H=\frac{\sum\limits_{i=1}^{n}x_i}{n}=\frac{0.575}{3}=0.19(个)$$

若根据表中劳动生产率第 2 列，因为已知条件是做 1 个元件所需要时间，所以采用调和平均法计算平均劳动生产率，即 3 人共同做 3 个元件平均需要时间。

$$\bar{x}_H=\frac{n}{\sum\limits_{i=1}^{n}\dfrac{1}{x_i}}=\frac{3}{\dfrac{1}{4}+\dfrac{1}{5}+\dfrac{1}{8}}=\frac{3}{0.575}=5.22(天)$$

表 1.2.2　3 个工人制作元件的劳动生产率

工人	每天做元件数量 x_i（个）	做 1 个元件所需时间 x_i（天）
甲	0.250	4
乙	0.200	5
丙	0.125	8

例 1.3：现某电器设备厂有 9 个工人，分为 3 组制造某种元件，劳动生产情况如表 1.2.3所示，求平均劳动生产率。

表 1.2.3　劳动生产情况表

组序	每人做 1 个元件需时间 x_i（天）	工人数 n_i
1	4	3
2	5	4
3	8	2

解：据表 1.2.3，已知每 1 个元件所需时间和每组人数，因此采用加权调和平均数计算：

$$\bar{x}_H = \frac{\sum_{i=1}^{3} n_i}{\sum_{i=1}^{3} \frac{n_i}{x_i}} = \frac{3+4+2}{\frac{3}{4}+\frac{4}{5}+\frac{2}{8}} = \frac{9}{1.8} = 5(\text{天})$$

即每人做 1 个元件平均需要 5 天，或者是每人每天平均做元件 0.2 个。

1.2.1.3　**几何平均数**

几何平均数是 n 个变量值连乘积的 n 次方根，主要应用于平均发展速度、平均增长速度和银行利率等方面。如资料是未分组的，计算式为

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \cdots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{1.2.5}$$

式中符号 $\prod$ 表示连乘。

如果资料是分组的，则计算式为

$$\bar{x}_G = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot x_3^{n_3} \cdot \cdots \cdot x_m^{n_m}} = \sqrt[n]{\prod_{i=1}^{m} x_i^{n_i}} \tag{1.2.6}$$

式中 $n = n_1 + n_2 + n_3 + \cdots + n_m$。

例 1.4：某公司 1996～2000 年每年销售收入分别是前 1 年的 105%，105.5%，107.5%，106.5%，108%，求该公司在这 5 年中平均每年销售收入的增长速度。

解：由于公司各年的销售收入的增长是在前 1 年基础上的，因此平均增长速度应

采用几何平均法计算，即求得

$$\bar{x}_G = \sqrt[5]{1.05 \times 1.055 \times 1.075 \times 1.065 \times 1.08} = \sqrt[5]{1.3696941}$$
$$= 1.065$$

后再减去 1，得该公司 5 年平均销售收入的增长速度为 6.5%。

例 1.5：某先生在银行存款 15 年，年利率按复利计算，其中 2 年为 10%，5 年为 12%，3 年为 11%，4 年为 9%，1 年为 8%，求平均年利率，若该先生存款 15000 元，共得本利多少？

解：首先计算平均年利率。将各年利率加 1 后，得到各年的本利率，即变量 x，用(1.2.6)式，得

$$\bar{x}_G = \sqrt[15]{1.10^2 \times 1.12^5 \times 1.11^3 \times 1.09^4 \times 1.08} = \sqrt[15]{4.4460474}$$
$$= 1.1046$$

平均年本利率为 110.46%，将其再减去 1，得到平均年利率为 10.46%。

从上面的计算可以看到，银行存款 15 年，该先生所得本金与利息总和是原来的 4.4460474倍，具体计算如下：

$$\begin{aligned} y_{15} &= 15000 \times 1.10^2 \times 1.12^5 \times 1.11^3 \times 1.09^4 \times 1.08 \\ &= 15000 \times 4.4460474 \\ &= 66690.71(\text{元}) \end{aligned}$$

1.2.1.4　平方平均数

平方平均数是变量平方的算术平均数的平方根。

设有 5 个正方形的房间，边长分别为 3.0m，3.2m，3.5m，3.8m 和 4.0m。如果用算术平均法计算平均边长，这个平均数只反映不同长度的平均水平，没有与房间面积大小联系起来。

若要与房间面积大小相联系，就应该首先计算各房间的面积，然后计算各房间的平均面积，最后将面积平均数开平方得到平均边长即平方平均数。

对于未分组的资料，平方平均数的计算式为

$$\bar{x}_Q = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} \tag{1.2.7}$$

如根据上面所举的例子，求得 5 个房间的平均边长为

$$\begin{aligned} \bar{x}_Q &= \sqrt{\frac{3.0^2 + 3.2^2 + 3.5^2 + 3.8^2 + 4.0^2}{5}} \\ &= \sqrt{12.386} \\ &= 3.52(\text{m}) \end{aligned}$$

如果资料是分组的，则用加权平方平均数：

$$\bar{x}_Q = \sqrt{\frac{\sum_{i=1}^{m} x_i^2 n_i}{\sum_{i=1}^{m} n_i}} = \sqrt{\frac{\sum_{i=1}^{m} x_i^2 n_i}{n}} \tag{1.2.8}$$

例 1.6：设有正方形房间 25 个，4 个边长 3.0m，5 个边长 3.2m，10 个边长 3.5m，4 个边长 3.8m，2 个边长 4.0m，求平均边长。

解：根据(1.2.8)式，求得平均边长为

$$\begin{aligned}\bar{x}_Q &= \sqrt{\frac{3.0^2 \times 4 + 3.2^2 \times 5 + 3.5^2 \times 10 + 3.8^2 \times 4 + 4.0^2 \times 2}{25}} \\ &= \sqrt{\frac{299.46}{25}} \\ &= 3.46(\mathrm{m})\end{aligned}$$

1.2.1.5 4 种平均数的关系

上面介绍的算术平均数、调和平均数、几何平均数和平方平均数都是按照严格定义的公式计算的，我们可以用一个统一的幂平均数公式把它们表达出来，即：

$$\bar{x}_K = \sqrt[k]{\frac{\sum_{i=1}^{m} x_i^k n_i}{\sum_{i=1}^{m} n_i}} \qquad (x > 0, -\infty < k < \infty) \tag{1.2.9}$$

当 $k = -1$ 时，得到调和平均数

$$\bar{x}_K = \sqrt[k]{\frac{\sum_{i=1}^{m} x_i^k n_i}{\sum_{i=1}^{m} n_i}} = \left(\frac{\sum_{i=1}^{m} x_i^{-1} n_i}{\sum_{i=1}^{m} n_i}\right)^{-1} = \frac{\sum_{i=1}^{m} n_i}{\sum_{i=1}^{m} \frac{n_i}{x_i}}$$

$$= \bar{x}_H$$

当 $k \to 0$ 时，可以证明(1.2.9)式的极限式即为几何平均数。

当 $k = 1$ 时，得到算术平均数

$$\bar{x}_K = \sqrt[k]{\frac{\sum_{i=1}^{m} x_i^k n_i}{\sum_{i=1}^{m} n_i}} = \frac{\sum_{i=1}^{m} x_i n_i}{\sum_{i=1}^{m} n_i}$$

$$= \bar{x}_A$$