21世纪高等学校数学系列教材

（第二版）

应用数理统计与SPSS操作

主编　赵喜林　李德宜　龚谊承

WUHAN UNIVERSITY PRESS
武汉大学出版社

21世纪高等学校数学系列教材

（第二版）

应用数理统计与SPSS操作

■ 主　编　赵喜林　李德宜　龚谊承

■ 副主编　尹水仿　熊　丹　李春丽

■ 编　委　赵喜林　李德宜　龚谊承　尹水仿　熊　丹

李春丽　丁咏梅　何晓霞　张　强

WUHAN UNIVERSITY PRESS

武汉大学出版社

图书在版编目(CIP)数据

应用数理统计与SPSS操作/赵喜林,李德宜,龚谊承主编. —2版. —武汉:武汉大学出版社,2014.8
21世纪高等学校数学系列教材
ISBN 978-7-307-13710-3

Ⅰ.应… Ⅱ.①赵… ②李… ③龚… Ⅲ.①数理统计—高等学校—教材 ②统计分析—软件包—高等学校—教材 Ⅳ.①O212 ②C819

中国版本图书馆CIP数据核字(2014)第150151号

责任编辑:胡 艳 责任校对:鄢春梅 版式设计:马 佳

出版发行:武汉大学出版社 (430072 武昌 珞珈山)
(电子邮件:cbs22@whu.edu.cn 网址:www.wdp.com.cn)
印刷:黄石市华光彩色印务有限公司
开本:787×1092 1/16 印张:19.5 字数:472千字 插页:1
版次:2009年8月第1版 2014年8月第2版
2014年8月第2版第1次印刷
ISBN 978-7-307-13710-3 定价:39.00元

前　言

“数理统计”是工科院校硕士研究生的一门公共基础课，编者多年来一直从事这门课的教学.“数理统计”作为一门应用性很强，并借助计算机技术广泛渗透到各个应用领域的课程，越来越受到重视.本教材是根据工科学生的特点和注重课程的应用性而编写的.全书将统计软件 SPSS 和数理统计的内容有机地结合在一起，书中的例题除了给出手工计算结果外，同时也给出了在 SPSS 中的操作步骤，书中显示的 SPSS 操作界面是 17.0 版本.编写这部分内容的目的是：一方面，使学生会用统计软件解决实际问题，不至于理论与实际脱节；另一方面，使学生明白统计软件背后的理论基础是什么.针对工科硕士生的数学基础，本书省略一些繁琐而复杂的理论推导，取而代之以简单直观的描述.统计模型的引入、统计方法的介绍，尽量做到从实例出发，循序渐进，简明易懂.考虑到多元统计方法的重要性，最后一章介绍了多元正态分布的性质，以及聚类分析、判别分析、主成分分析和因子分析的基本内容.为了提高学习的兴趣，每章前面给出了一个有趣的统计应用实例.

全书共分 6 章，分别介绍了数理统计的基本概念、参数估计、假设检验、方差分析与正交试验设计、回归分析、多元统计分析初步.各章配有适量习题，书末附有答案.

本书可作为非统计专业硕士研究生、高年级本科生的数理统计课程教材，也可作为教师、统计工作者和工程技术人员的参考书.

本书由赵喜林、李德宜、龚谊承任主编，尹水仿、熊丹、李春丽任副主编.由赵喜林、李德宜和尹水仿提出编写思路.第 1 章由龚谊承编写，第 2 章由赵喜林编写，第 3 章由李春丽编写，第 4 章由丁咏梅编写，第 5 章由何晓霞编写，第 6 章由熊丹和张强编写，全书的 SPSS 操作部分由熊丹和赵喜林编写，英文关键词由龚谊承整理，赵喜林对全书做修改和统稿工作，熊丹和李春丽做全书的校对工作.

本次修订主要对上一版中的不妥之处进行了纠正，部分内容作了调整和改写，修订工作由赵喜林完成，李德宜对全书进行了审订.

由于编者水平有限，书中的不妥之处在所难免，恳请读者提出批评和建议.

编　者

2014 年 4 月于武汉

目　录

第 1 章　数理统计的基本概念

看不见的价值

1941 年，第二次世界大战期间，有一天，美国哥伦比亚大学著名的统计学家沃尔德(Abraham Wald，1902—1950)教授家来了一个意外的访客——英国皇家空军的作战指挥官．他说："沃尔德教授，每次飞行员出发去执行轰炸任务，我们最怕听到的回报是：'呼叫总部，我中弹了'．请协助我们改善这个关系着飞行员生死的难题吧!"沃尔德接下这个紧急研究案，分析德国地面炮火击中联军轰炸机的资料，提出机体装甲应该如何加强才能降低被炮火击落机会的建议．但依照当时的航空技术，机体装甲只能局部加强，否则机体过重，会导致起飞困难及操控迟钝．沃尔德研究发现，机翼是最容易被击中的部位，而飞行员的座舱与机尾则是最少被击中的部位．

沃尔德详尽的资料分析，令英国皇家空军十分满意．但在研究成果报告的会议上，却发生一场激辩．负责该项目的作战指挥官说："沃尔德教授的研究清楚地显示，联军轰炸机的机翼，弹孔密密麻麻，最容易中弹．因此，我们应该加强机翼的装甲．"但沃尔德却坚定而客气地说："将军，我尊敬你在飞行上的专业，但我有完全不同的看法，我建议加强飞行员座舱与机尾发动机部位的装甲，因为那儿最少发现弹孔．"

在全场错愕怀疑的眼光中，沃尔德解释说："我所分析的样本中，只包含顺利返回基地的轰炸机．从统计的观点来看，我认为被多次击中机翼的轰炸机，似乎还是能够安全返航．而飞机很少发现弹着点的部位，并不是真的不会中弹，而是一旦中弹，根本就无法返航．"指挥官反驳说："我很佩服沃尔德教授没有任何飞行经验，就敢做这么大胆的推论．以我个人而言，过去在执行任务时，也曾多次机翼中弹严重受创．要不是我飞行技术老到，运气也不错，早就机毁人亡了．所以，我依然强烈主张应该加强机翼的装甲．"

这两种意见僵持不下，皇家空军部部长陷入苦思，到底要相信这个作战经验丰富的飞行将军，还是要相信一个独排众议的统计学家?

由于战况紧急，无法做更进一步的研究，部长决定接受沃尔德的建议，立刻加强驾驶舱与机尾发动机的防御装甲．不久之后，联军轰炸机被击落的比例果然显著降低．为了确认这个决策的正确性，一段时间后，英国军方动用了敌后工作人员，收集了部分坠毁在德国境内的联军飞机残骸．这些飞机中弹的部位，果真如沃尔德所预料，主要集中在驾驶舱与机尾的位置．

作战指挥官加强机翼装甲的决定看似十分合理，但他忽略了这个事实：弹着点的分布是一种严重偏误的资料，因为最关键的数据是在被击落的飞机身上，但这些飞机却无法被观察到．因此，布满了弹痕的机翼反而是飞机最强韧的部位．空军作战指挥官差点因为太

重视“看得见”的弹痕，反而做出错误的决策. 这个案例有两个特别值得我们注意的地方：

第一，收集更多数据，并不会改善决策质量. 由于弹痕资料的来源本身就有严重的偏误，努力收集更多的资料，只会更加深原有的误解.

第二，召集更多作战经验丰富的飞行员来提供专业意见，也不能改善决策质量. 因为这些飞行员正是产生偏误数据过程中的一环. 他们都是安全回航的飞行员，虽然可能有机翼中弹的经验，但都不是驾驶舱或发动机中弹的“烈士”. 简单地说，当他们越认真凝视那些“看得到”的弹痕，离真相就越远.

样本是统计分析的基础，样本要满足什么条件，读完本章就会有答案.

1.1 导 论

假想你是一位企业工作人员，需要为所在的企业搜集一些相关资料，此时，你会发现，统计信息是一个企业在市场上立足的根本；历届政府工作报告中占据相当篇幅也最有说服力的就是统计数据；历届人民代表大会后的记者招待会上，记者们提问最多的也是相关的统计数据. 这些数据是怎么得来的？经过了哪些处理和分析？该如何解读这些数据呢？

这样的思路是想“透过数据看世界”，此时，实际上已经站在了“数理统计”这座宫殿的大门口了.

1.1.1 数理统计研究对象

在我们的身边，有形和无形的数据形成了一个茫茫数海，这些数据蕴含着什么信息？可以给我们什么启示呢？我们如何从中归纳出精辟的结论，为决策提供依据？这就构成了数理统计学的研究对象.

数理统计学研究怎样有效地收集、处理、分析、解释带有随机性的数据，以对所考察的问题作出推断或预测，直至为采取一定的决策和行动提供依据和建议. 若在这句文字中去掉“ 随机性”这几个字，那就是统计学的研究范围. 统计学也就是数据科学. 由于实践中人们收集到的数据大多受到随机性的影响，因而数理统计学是统计学的主要组成部分. 在许多西方国家(如美国)，这两个名词的含义基本相同. 许多大学里，统计系就是研究数理统计学的，使用“ 数理统计”一词时只是强调统计学中用到很多现代数学知识. 在我国，由于历史的原因，这两个名词在实际使用时有较大差别. 笔者相信，若干年后，这两个名词所指范围的差异将慢慢减小.

1.1.2 统计工作诸环节

用数理统计方法解决一个实际问题时，一般有如下几个步骤：建立数学模型；收集整理数据；进行统计推断、作出预测和决策. 需要注意，这些环节不能截然分开，也不一定按上述次序，有时是互相交错的. 具体内容如下：

(1)模型的选择和建立. 模型是指关于所研究总体的某种假定，一般是给总体分布规定某个类型. 另外，建立模型要依据概率的知识、所研究问题的专业知识、以往的经验以

及从总体中抽取的样本(数据).

(2)数据的收集. 一般有三种收集数据的方式，包括全面观测、抽样观测和安排特定的实验. 全面观测又称普查，即对总体中每个个体都加以观测，测定所需要的指标. 抽样观测又称抽查，是指从总体中抽取一部分，测定其有关的指标值. 这方面的研究内容构成数理统计的一个分支学科，叫做抽样调查. 安排特定实验以收集数据，这些特定的实验要有代表性，并使所得数据便于进行分析. 这里面所包含的数学问题，构成数理统计学的又一分支学科，即实验设计的内容.

(3)数据整理. 其目的是把包含在数据中的有用信息提取出来. 整理数据通常有两种形式，一种形式是制定适当的图表(如散点图)，以反映隐含在数据中的粗略的规律性或一般趋势；另一种形式是计算若干数字特征，以刻画样本某些方面的性质，如样本均值、样本方差等简单描述性统计量.

(4)统计推断. 它指根据总体模型以及由总体中抽出的样本，作出有关总体分布的某种论断. 数据的收集和整理是进行统计推断的必要准备，统计推断是数理统计学的主要任务.

(5)统计预测. 统计预测的对象，是随机变量在未来某个时刻所取的值，或设想在某种条件下对该变量进行观测时将取的值. 例如，预测一种产品在未来 3 年内的市场销售量，某个 10 岁男孩在 3 年后的身高、体重，等等.

(6)统计决策. 依据所做的统计推断或预测，并考虑到行动的后果(以经济损失的形式表示)而制定的一种行动方案. 其目的是使损失尽可能小，或反过来说，使收益尽可能大. 例如，一个商店要决定今年内某种产品的进货数量，商店的统计学家根据抽样调查，预测该产品本店今年销售量为 1000 件. 假定每积压一件产品损失 20 元，而少销售一件产品则损失 10 元，要据此作出关于进货数量的决策.

1.1.3 数理统计的一些渊源

数理统计学是伴随着概率论的发展而发展起来的. 19 世纪中叶以前已出现了若干重要的工作，如 C. F. 高斯和 A. M. 勒让德关于观测数据误差分析和最小二乘法的研究. 到 19 世纪末期，经过包括 K. 皮尔森在内的一些学者的努力，这门学科已开始形成. 但数理统计学发展成一门成熟的学科，则是 20 世纪上半叶的事，它在很大程度上要归功于 K. 皮尔森、R. A. 费歇等学者的工作. 特别是费歇的贡献，对这门学科的建立起了决定性的作用. 1946 年，H. 克拉默发表的《统计学数学方法》是第一部严谨且比较系统的数理统计著作，可以把它作为数理统计学进入成熟阶段的标志.

数理统计学的发展大致可分 3 个时期：

(1)20 世纪以前. 这个时期又可分成两段，大致上可以把高斯和勒让德关于最小二乘法用于观测数据的误差分析的工作作为分界线，前一阶段属萌芽时期，基本上没有超出描述性统计的范围；后一阶段可算作是数理统计学的幼年阶段. 首先，强调了推断的地位，而摆脱了单纯描述的性质. 由于高斯等的工作揭示了正态分布的重要性，学者们普遍认为，在实际问题中遇见的几乎所有的连续变量，都可以满意地用正态分布来刻画. 这种观点使关于正态分布的统计得到了深入的发展，但延缓了非参数统计的发展. 19 世纪末，K. 皮尔森给出了以他的名字命名的分布，并给出了估计参数的一种方法——矩法估计.

德国的 F. 赫尔梅特发现了统计上十分重要的χ^2 分布.

(2)20 世纪初到第二次世界大战结束. 这是数理统计学蓬勃发展达到成熟的时期. 许多重要的基本观点和方法,以及数理统计学的主要分支学科,都是在这个时期建立和发展起来的. 这个时期的成就,包含了至今仍在广泛使用的大多数统计方法. 在其发展中,以英国统计学家、生物学家费歇为代表的英国学派起了主导作用.

(3)第二次世界大战以后时期. 这一时期中,数理统计学在应用和理论两方面继续获得很大的进展.

1.1.4 数理统计的应用领域

数理统计的应用涉及工农业生产、自然科学和技术科学以及社会经济等众多广泛的领域,只要有数据的地方,就会用到数理统计方法. 具体地说,大致可以分为如下几个方面:

(1)在农业中,对田间试验进行适当的设计和统计分析.

(2)在工业生产中,实验设计法、回归设计和回归分析、方差分析、多元分析等统计方法,对于试制新产品和改进老产品、改革工艺流程、寻求适当的配方等方面,起着广泛的作用;对于统计质量管理,在控制工业产品的质量中起着十分重要的作用.

(3)医学是较早使用数理统计方法的领域之一. 在防治一种疾病时,需要找出导致这种疾病的种种因素. 统计方法在发现和验证这些因素上,是一个重要工具. 另一方面的应用是,用统计方法确定一种药物对治疗某种疾病是否有用、用处多大,以及比较几种药物或治疗方法的效力.

(4)在自然科学和技术科学中,统计方法可用于地震、气象和水文方面的预报、地质资源的评测等.

(5)在社会、经济领域方面的应用,如人口调查和预测,心理学中能力方面的分析,等等.

◎ 思考:

1. 数理统计是研究什么的?
2. 数理统计与概率论、统计学有什么样的关系?

1.2 数理统计的基本概念

1.2.1 总体、样本、参数

1. 总体(population)

在数理统计学中,把研究对象的全体称为总体或母体,构成总体的每个元素称为个体. 如要研究某大学的学生身高情况,则该大学的全体学生就是该问题的总体,每个学生是个体;要研究某厂生产的显像管的质量时,总体是该厂生产的所有显像管,每一个显像管是个体.

每个个体有很多特征,比如学生有性别、年龄、身高、体重、籍贯等,我们并不是关

心个体的所有特征，而只是关心某些数量指标值，比如身高．抛开实际背景，总体就是一堆数，这堆数有它的分布规律．若考察的数量指标用 X 表示，则 X 是一个随机变量，X 的可能取值就是总体里的数，研究总体的分布规律实际上就是研究随机变量 X 的分布规律，所以，总体就是一个随机变量．

定义 1.1 一个随机变量 X 或其相应的分布函数 $F(x)$ 称为一个总体．

如果我们对每一研究对象要观测两个或多个数量指标，此时用多维随机向量来表示总体，这是多元分析研究的对象·.

根据总体中所含元素的个数是有限还是无限的，又可以分为有限总体和无限总体．我们讨论的主要是无限总体．

2. 样本（sample）

为了研究总体的分布规律或某些特征，我们从总体中随机地抽取 n 个个体，观测它们的数量指标，分别记为 X_1，…，X_n，这些通过观察或试验得到的数据称为总体的一个样本或子样，n 称为样本容量（sample size）．这些观察或试验过程称为抽样（sampling）．

例如，用同一架天平称某重物 5 次，得到一组 5 个数据 x_1，x_2，x_3，x_4，x_5，称它们是一个样本，样本容量为 5.

每个容量为 n 的样本都可认为是 n 维空间的一个点，样本所有可能的取值构成了 n 维空间的一个子集，称为样本空间（sample space）．注意，“数据”一词在这里是广义的，它可以是实数值，例如 X_i 表示称得某重物的重量；也可以是事物的属性，例如 X_i =“正品”（或“废品”），等等．通常，为了方便研究，也常将这些属性数量化，例如用“1”表示“废品”，用“0”表示“正品”，当然这不是本质的问题．

对于样本，需要强调两点：

（1）二重性：样本并非一堆杂乱无章无规律可循的数据，而是受随机性影响的一组数据，因此，用概率论的话说，就是每个样本既可以视为一组数据，又可视为一组随机变量，这就是所谓样本的二重性．当通过一次具体的试验，得到一组观测值，这时样本表现为一组数据；但这组数据的出现并非是必然的，它只能以一定的概率（或概率密度）出现，这就是说，当考察一个统计方法是否具有某种普遍意义下的效果时，又需要将其样本视为随机变量，而一次具体试验得到的数据，则可视为随机变量的一个观察值．今后为行文方便，我们常交替使用上述两种观点来看待样本，而不去每次声明此处样本是指随机变量还是其观察值．

（2）独立同分布性：我们要求每一个个体都有同等机会被选入样本，这就意味着样本中每个 $X_i(i=1,2,\cdots,n)$ 具有与总体 X 相同的分布，使样本具有代表性；同时，要求样本中各数据的出现互不影响，$X_i(i=1,2,\cdots,n)$ 相互独立．或者说，抽取样本时应该是在相同条件下独立重复地进行．

【例 1.1】 设一组抽奖券共 10000 张，其中有 5 张有奖．问：连续抽取 3 张均有奖的概率为多少？

为了讨论这个问题，不妨设

$$X_i=\begin{cases}1,\text{第 } i \text{ 次抽到奖}\\0,\text{第 } i \text{ 次未抽到奖}\end{cases}$$

要求该事件的概率，实际上即是求联合概率分布

$$P\{X_1 = x_1, X_2 = x_2, X_3 = x_3\}\ (x_i = 0 \text{ 或 } 1)$$

在 $x_1 = x_2 = x_3 = 1$ 处的值. 但题中没有说明“连续抽取”是“有放回的”还是“无放回的”,我们不妨都计算一下:

(1)无放回时:

$$P\{X_1 = 1, X_2 = 1, X_3 = 1\} = \frac{5}{10000} \cdot \frac{4}{9999} \cdot \frac{3}{9998}$$

(2)有放回时:

$$P\{X_1 = 1, X_2 = 1, X_3 = 1\} = \frac{5}{10000} \cdot \frac{5}{10000} \cdot \frac{5}{10000} = \left(\frac{5}{10000}\right)^3$$

显然,(1)中的抽样方式不是独立的,每次抽样的结果都将影响下一次抽样的分布,这种抽样不是我们通常研究的抽样. 而(2)中的抽样则是多次独立的抽样,它们是同分布的,即我们通常称为的随机抽样(random sampling). 这样得到的数据,即是我们常研究的简单随机样本(simple random sample),或简称为样本.

定义 1.2 如果随机变量

$$X_1, \cdots, X_n \tag{1.1}$$

相互独立,且 $X_i (i = 1, 2, \cdots, n)$ 都服从与总体 X 相同的分布,则称 $X_1, \cdots, X_n$ 是来自于总体 X 的简单随机样本,或简称为样本.

【例 1.2】 用两台车床车同一批产品,分别车 m 及 n 件,尺寸为 $X_1, \cdots, X_m$ 及 $Y_1, Y_2, \cdots, Y_n$,这时,我们得到的样本是

$$X_1, \cdots, X_m, Y_1, Y_2, \cdots, Y_n \tag{1.2}$$

它们显然通常不会是同分布的,可视 $X_1, \cdots, X_m$ 来自于总体 X, $Y_1, Y_2, \cdots, Y_n$ 来自于总体 Y 的样本.

由此可以看出,如果总体 X 的分布为 $F(x)$,则其联合分布为

$$F(x_1)F(x_2)\cdots F(x_n) \tag{1.3}$$

相应地,若总体 X 有概率密度 $f(x)$,则样本(1.1)的联合概率密度为

$$f(x_1)f(x_2)\cdots f(x_n) \tag{1.4}$$

【例 1.3】 设总体(1) $X \sim N(\mu, \sigma^2)$;(2) $X \sim b(1, p)$;(3) X 服从参数为 λ 的泊松分布, $X_1, \cdots, X_n$ 为来自总体的样本,则

(1)总体的密度

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

样本 $(X_1, \cdots, X_n)$ 的联合概率密度为

$$f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i, \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

(2)总体的分布律为

$$P\{X = x\} = p^x (1-p)^{1-x} \qquad (x = 0, 1)$$

样本 $(X_1, \cdots, X_n)$ 的联合分布律为

$$P\{X_1 = x_1, \cdots, X_n = x_n\} = \prod_{i=1}^{n} P\{X = x_i\} = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i} (1-p)^{n-\sum_{i=1}^{n} x_i}$$

$$(x_i = 1,0, i = 1,2,\cdots,n)$$

(3)总体泊松分布的分布律为

$$P\{X = x\} = \frac{\lambda^x e^{-\lambda}}{x!} \qquad (x = 0,1,2,\cdots)$$

样本 $(X_1,\cdots,X_n)$ 的联合分布律为

$$P\{X_1 = x_1,\cdots,X_n = x_n\} = \prod_{i=1}^{n} P\{X = x_i\} = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{x_1!\cdots x_n!}$$

$$(x_i = 0,1,2,\cdots; i = 1,2,\cdots,n)$$

有了总体、样本的概念后,我们就可以将统计推断的基本任务概括为由样本推断总体.如在例1.2中,我们就可以从样本(1.2)中推断出总体 X 与 Y 是否有显著差别.关于这一基本任务,我们今后可以慢慢体会到.由于推断总体实质上是推断总体的分布,即解决一个实际统计问题,往往归结为总体分布的确定,所以我们也常称总体的分布是该问题的统计模型(statistics model).

3. 参数(parameter)

参数是用来描述总体特征的概括性数字度量,参数是研究者想了解的总体的某种特征值.总体的分布一般来说是未知的,所以很多情况下,参数是总体分布中的未知常数,需要通过样本来推断.例如,正态总体的分布密度 $f(x,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ 中,μ,σ^2 是参数,分别表示总体的均值和方差.

对于每个总体,我们称其分布中参数的一切可能取值的集合为参数空间(parameter space),记为 Θ,如在例1.3中:(1)参数 $(\mu,\sigma^2) \in \Theta = \mathbf{R} \times \mathbf{R}^+$,其中 $\mathbf{R} = (-\infty, +\infty)$,$\mathbf{R}^+ = (0, +\infty)$;(2)参数 $p \in \Theta = [0,1]$;(3)参数 $\lambda \in \Theta = \mathbf{R}^+ = (0, +\infty)$.

1.2.2 统计量

1. 统计量(statistic)

在利用样本推断总体时,往往不能直接利用样本,而需要对它进行一定的加工,这样才能有效地利用其中的信息,否则,样本只是呈现为一堆"杂乱无章"的数据.

【例1.4】 从某地区随机抽取50户农民,调查其年收入情况,得到下列数据(每户人均元):

924	800	916	704	870	1040	824	690	574	490
972	988	1266	684	764	940	408	804	610	852
602	754	788	962	704	712	854	888	768	848
882	1192	820	878	614	846	746	828	792	872
696	644	926	808	1010	728	742	850	864	738

试对该地区农民收入的水平和贫富悬殊程度做个大致分析.

显然,如果不进行加工,面对这堆大小参差不齐的数据,你很难得出什么印象.但是,对这些数据做出大致分析:如记各农户的年收入数为 $X_1,\cdots,X_{50}$,则考虑

$$\bar{X} = \frac{1}{50}\sum_{i=1}^{50} X_i = 809.55$$

$$S=\sqrt{\frac{1}{49}\sum_{i=1}^{50}(X_i-\bar{X})^2}=155.849$$

这样,我们就可以从 $\bar{X}$ 得出该地区农民平均人均收入水平大致为809元,从 S 可以得出该地区农民贫富悬殊不大的结论. 由此可见,把分散在样本中我们关心的信息提炼出来,针对不同的研究目的构造不同的样本函数,是统计推断的基础;同时,为了使提炼的信息是已知的,这个函数不能含有未知参数,这样的函数在统计学中称为统计量.

定义 1.3 设 $X_1,\cdots,X_n$ 是总体 X 的一个样本,如果 $T(X_1,\cdots,X_n)$ 是样本 $X_1,\cdots,X_n$ 的一个不含未知参数的函数,则称

$$T=T(X_1,\cdots,X_n)$$

为统计量.

【例 1.5】 设总体 $X\sim N(\mu,2^2)$,其中 μ 是未知参数,$X_1,\cdots,X_n$ 是总体 X 的一个样本,判断下面哪些是统计量:

(1) $T_1=\frac{1}{n}\sum_{i=1}^{n}X^2{}_i$;

(2) $T_2=X_1$;

(3) $T_3=\max\{X_1,\cdots,X_n\}$;

(4) $T_4=\sum_{i=1}^{n}\left(\frac{X_i-\mu}{2}\right)^2$.

解: T_1,T_2,T_3 都是样本的函数,且不含未知参数,所以是统计量. T_4 虽然是样本的函数,但含有未知参数 μ,所以不是统计量.

因为样本具有二重性,而统计量又是样本的函数,所以统计量也具有二重性:一方面可以看成随机变量,另一方面它的值又是可以观测的. 统计量究竟表示的是随机变量还是它的观察值,可根据具体环境确定.

2. 常用统计量

设 $X_1,\cdots,X_n$ 是来自总体 X 的样本,常用统计量如下:

(1)样本均值(sample mean):

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

样本均值 $\bar{X}$ 反映了总体 X 的期望的信息.

(2)样本方差(sample variance):

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)$$

样本方差 S^2 或记为 S_n^2. 为了消除样本方差与总体量纲的差别,通常取 $S=\sqrt{S^2}$,称 S 为样本标准差(sample standard deviation). 样本标准差与总体的量纲一致. 样本方差描述了样本的离散程度,反映了总体 X 的方差的信息.

在这个定义中,$\sum_{i=1}^{n}(X_i-\bar{X})^2$ 称为偏差平方和,$n-1$ 称为偏差平方和的自由度,其含义是:在 $\bar{X}$ 确定后,n 个偏差 $X_i-\bar{X}(i=1,2,\cdots,n)$ 中,只有 $n-1$ 个可以自由变动,因为这

n 个数之间有一个约束 $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$.

定理 1.1 设总体 X 具有二阶矩，$EX = \mu, DX = \sigma^2 < +\infty$，$X_1, \cdots, X_n$ 为来自总体 X 的样本，则

$$E\bar{X} = \mu, D\bar{X} = \frac{\sigma^2}{n} \tag{1.5}$$

$$ES^2 = \sigma^2 \tag{1.6}$$

证明：

$$E\bar{X} = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{n\mu}{n} = \mu$$

$$D\bar{X} = \frac{1}{n^2}D\left(\sum_{i=1}^{n} X_i\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = E\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = n(\mu^2 + \sigma^2) - n\left(\mu^2 + \frac{\sigma^2}{n}\right) = (n-1)\sigma^2$$

两边除以 $n-1$，即得 $ES^2 = \sigma^2$.

(3)变异系数(coefficient of variation)：

$$V = \frac{S}{\bar{X}}$$

变异系数 V 是相对标准差，也是反映数据离散程度的统计量. 变异系数大，说明数据的离散程度大；变异系数小，说明数据的离散程度小.

(4)样本 k 阶原点矩(sample k -th moment about the origin)：

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

样本一阶原点矩就是样本均值.

(5)样本 k 阶中心矩(sample k -th moment about the mean)：

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$$

有时，也称样本的二阶中心矩 B_2 为样本方差，称 $\sqrt{B_2}$ 为样本均方差，记 $S^* = \sqrt{B_2}$. 当 n 很大时，B_2 与 S^2 差别不大.

(6)样本偏度(sample skewness)：

$$\gamma_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{S^*}\right)^3 = \frac{B_3}{B_2^{\frac{3}{2}}}$$

样本偏度 γ_1 反映了总体分布密度曲线的对称性信息. $\gamma_1 = 0$ 表示样本对称；$\gamma_1 > 0$ 表示样本右尾长，即样本中有几个较大的数，反映总体分布是右偏的；$\gamma_1 < 0$ 表示样本左尾长，样本中有几个特小的数，反映总体分布是左偏的.

(7)样本峰度(sample kurtosis)：

$$\gamma_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{S^*}\right)^4 - 3 = \frac{B_4}{B_2^2} - 3$$

样本峰度 γ_2 反映了总体分布密度曲线在其峰值附近的陡峭程度. 当 $\gamma_2 > 0$ 时，分布密

度曲线在其峰值附近比正态分布陡,称为尖顶型;当 $\gamma_2 < 0$ 时,分布密度曲线在其峰值附近比正态分布平坦,称为平顶型.

【例 1.6】 用 SPSS 计算例 1.4 中数据的均值、标准差、样本偏度及样本峰度.

SPSS 操作步骤:

(1)输入数据,定义变量名为"收入".

(2)在菜单栏中选择 Analyze→ Descriptive Statistics→ Frequencies,进入 Frequencies 对话框(图 1.1).将"收入"选入 Variable(s)框,单击 Statistics 按钮进入 Statistics 对话框(图 1.2),选中需要的统计指标,单击 Continue 按钮返回 Frequencies 对话框.

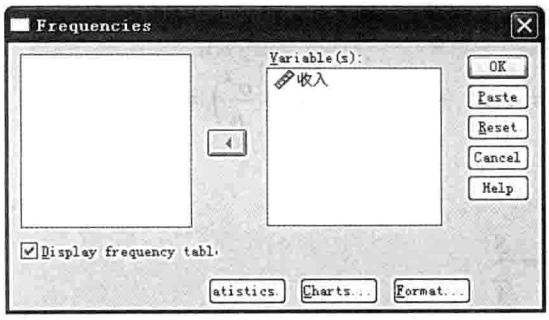

图 1.1 Frequencies 对话框

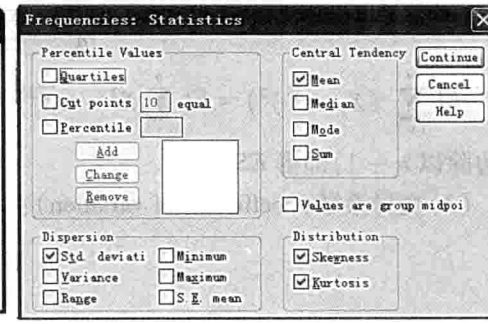

图 1.2 Statistics 对话框

(3)单击 OK,可得到计算结果,见表 1.1.

表 1.1 **收入的描述统计指标 Statistics**

N	Valid	50
	Missing	0
Mean		809.52
Std. Deviation		155.849
Skewness		0.247
Std. Error of Skewness		0.337
Kurtosis		1.446
Std. Error of Kurtosis		0.662

SPSS 结果解释:

(1)样本容量 N=50,数据丢失 0 个;

(2)均值 Mean 为 809.52,标准差为 155.849;

(3)样本偏度为 0.247,标准误差为 0.337;

(4)样本峰度为 1.446,标准误差为 0.662.

3. 次序统计量(order statistic)

定义 1.4 设 $X_1,\cdots,X_n$ 是取自总体 X 的样本,定义 $X_{(i)}(i=1,2,\cdots,n)$ 取值为将 $X_1,\cdots,X_n$ 的观察值由小到大排列后得到的第 i 个观察值,则称 $X_{(i)}$ 为该样本的第 i 个次序统

计量. 其中，$X_{(1)}=\min\limits_{1\leqslant i\leqslant n}\{X_i\}$ 称为最小次序统计量，$X_{(n)}=\max\limits_{1\leqslant i\leqslant n}\{X_i\}$ 称为最大次序统计量.

定理 1.2 设总体 X 的密度函数为 $f(x)$，分布函数为 $F(x)$，$X_1,\cdots,X_n$ 是总体 X 的样本，则第 k 个次序统计量 $X_{(k)}$ 的密度函数为

$$f_k(x)=\frac{n!}{(k-1)!\,(n-k)!}[F(x)]^{k-1}[1-F(x)]^{n-k}f(x) \tag{1.7}$$

证明见参考文献[1].

特别地，令 $k=1$ 和 $k=n$ 得到最小次序统计量 $X_{(1)}$ 和最大次序统计量 $X_{(n)}$ 的密度函数分别为

$$f_1(x)=n[1-F(x)]^{n-1}f(x)$$

$$f_n(x)=n[F(x)]^{n-1}f(x)$$

【例 1.7】 设总体 $X\sim U(0,1)$，$X_1,\cdots,X_n$ 为样本，求第 $k(1\leqslant k\leqslant n)$ 个次序统计量的密度函数 $f_k(x)$.

解：总体 X 的分布函数为

$$F(x)=\begin{cases}0, & x\leqslant 0\\ x, & 0<x\leqslant 1\\ 1, & 1<x\end{cases}$$

密度函数为

$$f(x)=\begin{cases}1, & 0<x<1\\ 0, & \text{其他}\end{cases}$$

代入式(1.7)，得

$$f_k(x)=\begin{cases}\dfrac{n!}{(k-1)!\,(n-k)!}x^{k-1}(1-x)^{n-k}, & 0<x<1\\ 0, & \text{其他}\end{cases}$$

下面介绍两个用次序统计量构造的常用统计量：

(1)样本中位数(sample median)：是次序统计量的函数，定义如下：

$$\tilde{X}=\begin{cases}X_{\left[\frac{n+1}{2}\right]}, & n\text{ 为奇数}\\ \dfrac{1}{2}\left[X_{\left[\frac{n}{2}\right]}+X_{\left[\frac{n}{2}+1\right]}\right], & n\text{ 为偶数}\end{cases}$$

即样本容量为奇数时，样本中位数就是取样本按从小到大次序排列位于最中间的那个数；样本容量为偶数时，取位于最中间的两个数的算术平均值. 样本中位数与样本均值一样，是刻画样本的位置特征的量，但它不受样本中异常值的影响.

(2)样本极差(sample range)：定义为

$$R=X_{(n)}-X_{(1)}=\max_{1\leqslant i\leqslant n}\{X_i\}-\min_{1\leqslant i\leqslant n}\{X_i\}$$

样本极差与样本方差一样，是反映样本的离散程度的量，但样本极差计算方便.

1.2.3 直方图和经验分布函数

对连续型总体的分布的描述工具是分布函数 $F(x)$ 或者概率密度函数 $f(x)$，由于总体分布的未知性，$F(x)$ 或 $f(x)$ 的精确表达式也是未知的. 我们推断总体的信息来源于样本，如何由样本来推断 $F(x)$ 或 $f(x)$ 呢？下面介绍的直方图和经验分布函数是分别用来推断