普通高等教育“十三五”规划教材

概率论与数理统计

（第二版）

西南交通大学数学学院统计系　编

科学出版社

普通高等教育“十三五”规划教材

概率论与数理统计

(第二版)

西南交通大学数学学院统计系　编

科学出版社

北京

内 容 简 介

本书针对工科专业的特点，以统计建模为侧重点，突出统计方法的基本思想和实用性，并兼顾对理论基础的理解和掌握. 绪论部分详细介绍了教材编写的基本结构和特点，以及每一章的主要内容和教学安排的建议. 全书分为 9 章：第 1 章主要介绍了常用的描述性统计方法，第 2 章至第 4 章介绍了相关的概率论知识，第 5 章介绍简单随机样本及抽样分布，第 6 章至第 9 章介绍了常用统计方法的思想及具体分析过程. 全书主要内容包括对数据的描述性统计分析、随机事件及其概率、随机变量及其分布、多维随机变量及其联合概率分布、简单随机样本及抽样分布、点估计、单个总体的区间估计与假设检验、多个正态总体的统计推断以及回归分析等.

本书可用于高等院校工科类各专业概率论与数理统计课程的教材，也可作为自学和相关科研工作者参考书.

图书在版编目(CIP)数据

概率论与数理统计/西南交通大学数学学院统计系编. —2 版. —北京：科学出版社，2017.8

普通高等教育“十三五”规划教材

ISBN 978-7-03-053868-0

Ⅰ. ①概… Ⅱ. ①西… Ⅲ. ①概率论-高等学校-教材 ②数理统计-高等学校-教材 Ⅳ. ①O21

中国版本图书馆 CIP 数据核字(2017) 第 146514 号

责任编辑：王胡权 / 责任校对：彭 涛
责任印制：白 洋 / 封面设计：迷底书装

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

保定市中画美凯印刷有限公司 印刷
科学出版社发行 各地新华书店经销
*
2011 年 6 月第 一 版 开本：720×1000 1/16
2017 年 8 月第 二 版 印张：19 1/2
2017 年 8 月第一次印刷 字数：391 000

定价：39.00 元

(如有印装质量问题，我社负责调换)

第二版前言

在当今大数据的时代背景下, 对数据的统计分析处理提出了更多的新问题, 从而也形成了统计学科进一步发展的重要推动力. 作为随机数学的基础课程, 如何在概率统计课程教学中让学生领悟统计学的精髓, 是我们一直在努力探索的问题.

与第一版相比, 本书增加了绪论部分, 其目的就是更清楚地说明教材编写的基本想法和结构组成, 以便教师或学生在使用过程中可以更好地把握主线. 此外, 第一版是将多维随机变量与数理统计基本概念放在一章中, 实际使用中感觉内容太多, 综合考虑了教师与学生的反馈意见之后, 我们将这一部分内容分为了第 4 章和第 5 章. 在第二版的修订编写中, 我们还修改并增加了部分例题, 以方便学生在自学过程中的使用.

与第一版的出发点一致, 我们依然是重点突出以统计实践为主导的结构框架, 整体思路可以概括为"统计初步→统计推断", 强调统计学对数据的分析和解读, 让学生自己体会到中学时代所学的统计知识在专业统计学中的基本作用, 引导学生自己发现运用更系统的统计方法从数据中获取更多更深层的信息.

本书的编写与修订主要由西南交通大学数学学院统计系的几位教师完成, 具体分工为: 袁代林负责第 1 章、第 2 章和第 3 章, 程世娟负责第 4 章和第 5 章, 刘赫负责第 6 章、第 7 章、第 8 章以及绪论部分, 赵联文负责第 9 章以及全书的修订工作, 全书统稿工作由刘赫完成.

特别要感谢何平教授、李裕奇教授, 在框架体系的搭建以及许多知识点的具体处理技巧等方面为我们提出了宝贵的意见和建议.

在本书第二版的修订与编写过程中, 得到了西南交通大学数学学院相关领导的热情帮助与支持, 特别是统计系的领导和所有教师的大力支持, 在此我们表示衷心的感谢! 同时感谢科学出版社为本书的顺利出版付出的辛苦劳动.

本书在修订过程中, 参考了大量的相关教材和资料, 选用了其中的有关内容和习题, 在此谨向有关编者和作者一并表示感谢.

书中不足之处, 诚恳地希望读者批评指正.

编 者

于西南交通大学

2017 年 3 月 28 日

第一版前言

随着科学技术的进步和各学科的不断发展, 作为数据处理和分析技术的统计方法得到了越来越广泛的应用. 可以说, 只要涉及到数据分析就必然会用到统计分析的方法, 而概率论则为所有统计思想和方法提供了理论支撑. 因此, 概率论与数理统计已经成为高等学校工学、经济学、管理学、社会学等专业本科阶段普遍开设的随机类数学课程. 目前, 结合财经类专业的特点, 国内部分高等财经院校的概率统计教材在内容和体系上都在进行不断地改进. 如何针对工科专业的实际需求编写相应的统计学教材, 一直是我们在教学过程中认真思考和探索的问题.

本书的目的就是针对高等院校工科类专业的实际需求, 强调将实际问题提炼为统计问题的思想和实现过程, 提高学生运用统计方法解决问题的实际能力. 我们在教材的体系安排、内容取舍、教学方法等方面按照上述指导思想作了一些尝试, 主要体现在以下几点:

(1) 在内容安排上, 从常用统计方法的理论根据出发, 同时也兼顾了研究生入学考试中对概率统计部分的要求, 对概率论的部分理论内容做了相应的弱化处理, 突出了概率理论与统计方法的关联, 以利于学生的接受和理解.

(2) 贯彻统计建模的思想: 实际问题→统计模型→求解模型→阐述结果. 具体来讲, 就是从实际问题出发, 建立模型将其转化为统计问题, 然后再提出解决问题的思想, 并利用数学手段实现, 最后再回到实际问题, 对得到的结果进行解释, 引导学生运用所学知识解决实际问题.

(3) 借鉴了国外优秀概率统计教材的经验, 将数理统计部分的结构分为点估计、基于单个总体的统计推断、基于多个总体的统计推断以及回归分析四个部分, 并特别介绍了关于非正态总体的统计推断、Logistic 回归等相关内容.

在编写本书的过程中, 得到了西南交通大学数学学院以及统计系所有领导和同事的热情帮助与支持, 在此我们表示衷心的感谢! 此外, 我们特别感谢西南交通大学教务处教材科为本书的编写工作给予了许多支持和帮助; 真心感谢科学出版社为本书的顺利出版给予的鼎力协助.

本书在编写过程中, 参考了大量的相关教材和资料, 选用了其中的有关内容和

习题, 在此谨向有关编者或作者一并表示感谢.

书中如有不足之处, 诚恳期望读者提出并反馈宝贵意见.

编　者

于西南交通大学

2011 年 3 月 30 日

目　录

绪　论

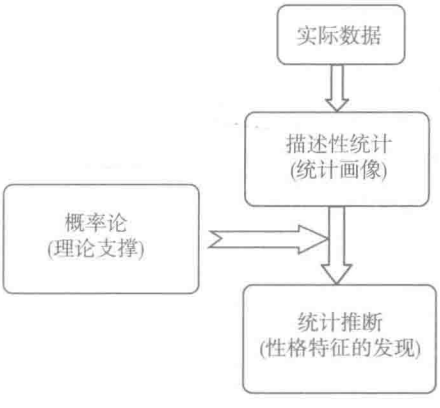
不知不觉间我们就进入了一个大数据的时代，各行各业都在讲着大数据、用着大数据. 大数据的出现，对数据的统计分析处理提出了更多的新问题，从而也形成了统计学科进一步发展的重要推动力. 在“互联网 +”时代，数据迅速进入人们的视野，而统计学的目的就是探寻隐藏于数据中的信息，为数据插上价值的翅膀.

对于许多大学生而言，统计学可以算作是“熟悉的陌生人”. 说熟悉，是因为翻开任何一本概率统计的相关教材，仅看目录会感觉到大部分概念和内容都是中小学阶段就已经学习过的；至于陌生，则是在课程学习的过程中，学生又会感觉到理论性越来越强，统计学绝不仅仅是计算几个数字而已.

统计学如同一件兵器，使用者的功力不同，其发挥的作用也会有着巨大差异. 我们的目的就是希望从问题出发，从学生所熟悉的描述性统计入手，由浅层的表象分析逐步到达更深层次的信息发掘，在逐步深入的过程中展现统计学的独特魅力!

本书将统计建模作为主线，整体思路可以概括为“统计初步→统计推断”，强调统计学对数据的分析和解读，让学生自己体会到中学时代所学的统计知识在专业统计学中的基本作用，引导学生自己发现运用更系统的统计方法从数据中获取更多更深层的信息.

从全书的结构安排上来看，基本想法可以表示如下图：



首先是结合实际问题，面对具体的数据，基于学生现有的知识储备，系统整理了常见的描述性统计方法，相当于对数据做了一个初步的画像，用直接、感性的可视化方式展示数据中所隐藏的一些基本特征.

进一步, 为了发现数据更深层次的性格和特征, 就必须要用到专业的统计推断方法. 而这些统计方法的学习, 是需要概率论作为理论支撑. 因此, 书中关于概率论部分的知识点都是紧紧围绕着统计学需要的, 特别借鉴了国内外优秀教材的处理方式, 对于概率论中的部分知识点进行了弱化, 以统计学为教材编写的主线.

本书分为 9 章, 各章主要内容如下:

第 1 章　描述性统计

告诉读者在面对样本数据的时候, 如何快速有效地提炼出其包含的信息, 完成数据的可视化描述分析. 同时结合实例, 阐述了概率与统计的关系, 为后续章节的安排留下线索.

第 2 章　随机事件及其概率

借助于集合论的思想和方法对随机事件展开研究. 在分析巩固中学已经学习过的基本概率知识的基础之上, 建立系统的概率论知识体系, 同时也提出研究中“仅仅运用集合论的手段”是很受限制的, 能不能将我们所熟悉的函数用于概率问题的处理?

第 3 章　随机变量及其分布

是将函数引入到概率论研究的重要转折, 导数、积分、级数等数学工具都会出现在概率问题的处理过程中. 在相关知识点的处理上, 强调概率问题是如何转化为积分、求和等数学问题的.

解决上述问题的关键在于: 随机变量, 分布函数.

第 4 章　多维随机变量及其联合概率分布

在数的学习过程中, 我们是先从一维的数轴到平面直角坐标系, 再到三维的立体空间, 然后再推广到更一般的 n 维空间. 这里也是如此, 一维随机变量到多维的推广, 有基本概念的直接引入, 也有多维才会遇到的问题.

第 5 章　简单随机样本及抽样分布

简单随机样本就是一类特殊的多维随机变量, 由此搭建起概率论与数理统计之间的联系. 在完成统计学理论基础搭建的同时, 也将视角重新回到教材的主线——统计学. 从随机变量的角度进一步解读统计学的核心思想——用样本信息对总体作出推断描述.

第 6 章　点估计

重点介绍点估计方法中的矩估计和极大似然估计, 以及对于点估计量优劣性的评价准则. 利用样本对总体的特征指标进行估计, 是一种有科学依据的合理猜测. 这种猜测永远不可能百分百准确, 但对于人们认知事物有着不可估量的作用.

第 7 章　单个总体的区间估计与假设检验

以正态总体为例, 详细分析区间估计的枢轴量与假设检验中检验统计量之间的关系, 并特别介绍了关于非正态总体的统计推断. 估计与检验是统计推断的两个基

本问题, 在讲述基本结论的同时, 更注重区间估计和假设检验的基本思想, 从问题的提出到统计模型的形式, 从解决问题的初始想法到最终的数学实现, 从模型结果的数学描述到回归实际问题背景的结果阐述, 期望让学生更深刻地了解统计学本质.

第 8 章　多个正态总体的统计推断

在前一章的基础之上, 将讨论范围推广至两个正态总体, 并进一步关注多个正态总体的均值比较问题, 即方差分析.

第 9 章　回归分析

从相关关系的分析出发, 结合实例讲解了线性回归模型的统计思想和方法, 强调了回归分析每一个关键环节的统计问题再现, 并特别讨论了用于分类数据的 Logistic 回归模型.

使用本书有两个建议方案, 具体如下:

如果学时较短, 可以选择部分内容组织教学, 譬如: 描述性统计、随机事件及其概率、随机变量及其分布的主要内容, 多维随机变量的基本概念、统计量的概念, 三大抽样分布、矩估计与最大似然估计, 单正态总体的区间估计与假设检验.

如果学时足够, 可以按照本书的完整结构进行教学, 其中部分知识点可以作为学生课后自学 (如: 大样本置信区间, 非正态总体的统计推断, 成对数据的统计分析, 非线性回归以及 Logistic 回归分析).

第 1 章　描述性统计

在我们了解和认识客观世界的过程中, 统计学的思想和方法经常起着不可替代的作用. 在许多工程及自然科学的专业领域中, 包括可靠性分析、质量控制、生物信息、脑科学、心理分析、经济分析、金融风险管理、社会科学推断、行为科学等诸多领域, 统计分析方法已经成为基本的数据分析与信息分析工具.

在科学研究和实际问题的处理过程中, 往往需要面对数据的分析和处理. 这些数据虽然包含了大量的信息, 但对我们所关心的问题而言, 还需要对数据进行一定的处理才能从中提炼出有用的信息. 那么如何从这些收集到的数据中获取所需要的信息呢? 统计学就提供了相应的思想和方法, 通过对数据的加工和整理, 可以从中提取更有价值的信息. 一般而言, 统计学研究的就是如何有效地收集和整理数据资料, 并通过分析对所研究的对象的性质、特征作出推断. 通常来讲, 利用图、表以及简单计算以实现对数据资料的统计分析统称为描述性统计.

本章将介绍统计中的基本概念, 以及一些常用的描述性统计分析方法.

1.1　总体与样本

对于一个统计问题, 将研究对象的全体称为**总体**(population), 构成总体的每一个元素称为**个体**.

例如, 要考察某大学在校学生的月生活费支出情况, 则该所大学的全体在校学生就构成相应的总体, 而每一个在校学生就是一个个体. 而如果要研究的是成都市大学在校学生的月生活费支出情况, 那么总体就包含了成都市所有大学的在校学生. 可见, 总体是根据研究范围所确定的.

对于不同的研究问题, 通常只对总体中个体的某些特征感兴趣, 如确定一批节能灯泡的使用寿命, 调查某地区 3 岁儿童的身高与体重等等. 一般情况下就将所关注的特征量视为总体, 而所有个体的取值就是总体的所有可能取值. 因此对于要考察的一个或多个特征量就可以定义为一个变量、或一组变量, 后者也可以视为一个多元变量.

由于通常情况下总体中所包含的元素都非常多, 而且有些调查数据是经过破坏性试验获得的, 不可能将每一个个体都逐一考察. 例如, 为了了解 2008 年成都市居民用于食品的平均消费情况, 应该如何做呢? 显然, 我们不可能去调查每一个成都市的居民, 然后得到所需要的数据. 在实际研究中只需要随机选取一部分成都市居

民进行调查获取信息. 统计学的主要目的就是对收集到的数据进行加工和整理, 通过分析从这些数据中发掘所需要的信息, 并得到一定的结论. 因此, 在实际调查和研究中, 所能够得到的就是从总体中随机抽取的一部分个体, 称之为**样本**(sample). 通过对样本的调查或观测所得到的数据, 就是做统计推理时所能利用的信息.

由上所述, 统计就是要研究客观现象总体的数量特征和数量关系, 统计学主要是根据样本信息对总体进行推断. 当我们用试验或观察的方法研究一个具体问题时, 首先就是从总体中抽取一定的样本, 要通过适当的观察或试验获取必要的数据. 通过对样本的研究, 才能进一步对总体的实际情况做出相应的推断.

例 1.1.1 有 26 名海上石油工人被随机选中参加一项模拟逃生试验, 他们每个人成功逃生所耗费的时间 (单位: 秒) 如下:①

389	356	359	363	375	424	325	394	402
373	373	370	364	366	364	325	339	393
392	369	374	359	356	403	334	397	

作为所获取的初始信息, 对这样一组数据如果不做任何整理和分析, 很难从中直接得到有价值的结论. 所以, 当我们拿到具体数据之后, 首先会希望对数据进行一些基本的汇总、整理, 并对数据的一些基本特征给以简单描述和总结. 在这一过程中所用到的方法就属于描述性统计, 可以说描述性统计分析不仅是进行统计分析的第一步, 同时也是对数据进行更深层次分析的基础.

简单地说, 描述性统计分析就是对所收集的大量数据进行加工整理, 用统计语言去描述这些数据的特征, 提取它们包含的信息, 从而揭示研究对象的内容和本质. 统计描述语言包括反映数据分布特点的各种特征量及图形、表格等, 概括和表现研究对象的统计性质, 包含了全面分析的研究过程. 因此, 描述性统计可以通过样本数据有关特征的计算给出一些具体的数字来描述数据的一些显著特点, 如均值、中位数、标准差等; 也可以利用图形的直观性对数据特征进行展示, 如直方图、散点图等.

我们一般地将描述性统计方法分为两类: 一类是数值法, 即利用代表性的数值精确地描述出所给数据的基本特征; 一类是图表法, 即利用可视化的工具描述数据. 以下先介绍数值法, 包括关于数据集中趋势和分散程度的常用描述方法, 然后介绍描述性统计中与统计推断联系密切的常用图表工具, 包括茎叶图、直方图、散点图和箱线图.

① "Oxygen Consumption and Ventilation During Escape from an Offshore Platform", Ergonomics, 1997:281-292

1.2　中心位置的描述

统计分析的目的是研究总体特征. 但一般情况下, 我们能够得到的只是从总体中随机抽取的一部分观察对象, 这些观察对象就构成了样本. 通过对样本的研究, 才能进一步对总体的实际情况做出相应的推断. 而描述性统计分析是进行统计分析的第一步, 也是许多统计分析方法的前期预处理过程.

如果要分析样本数据蕴含总体特征的信息, 则需要对反映数据分布特征的一些指标进行计算和解释. 也就是说, 面对一个个的数据, 我们希望从中提取出一些指标, 其数值大小可以反映出这个数据集的某些特征. 本节主要关注的是那些能够刻画数据分布位置的特征量, 特别是分布的中心位置.

对于一组具体的数据, 通常会通过计算均值、中位数和四分位数等特征量, 了解它们的取值主要集中在什么位置, 即这些数据分布的集中趋势.

1.2.1　均值 (Mean)

这里讲的均值是指样本均值, 它是全部样本数据的算术平均, 也称为算术平均值. 假设有 n 个样本数据 $x_1,x_2,\cdots,x_n$, 其均值 $\overline{x}$ 定义为

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

显然, 均值只适用于数值型数据.

例 1.2.1　根据调查, 某集团公司的中层管理人员的年终奖金数据如下 (单位: 千元):

40.6	39.6	37.8	36.2	40.8
38.6	39.6	40.0	34.7	41.7
38.9	37.9	37.0	35.1	36.7
37.1	37.7	39.2	36.9	38.3

由于 $\sum_{i=1}^{20}x_i=764.4$, 根据式 (1.2.1) 可以计算得到均值

$$\overline{x}=\frac{764.4}{20}=38.22$$

说明这 20 名中层管理人员的平均年终奖是 38.22 千元.

1.2.2　中位数 (Median)

中位数是将一组数据从小到大排序后, 处于中间位置的数据值, 通常用 M_e 表

示. 假设有 n 个样本数据 $x_1,x_2,\cdots,x_n$, 将其按照从小到大的顺序排列, 记为

$$x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$$

若 n 为奇数, 则中位数为 $x_{(\frac{n+1}{2})}$; 若 n 为偶数, 则中位数为 $x_{(\frac{n}{2})}$ 和 $x_{(\frac{n}{2}+1)}$ 的平均值. 即

$$M_e = \begin{cases} x_{(\frac{n+1}{2})}, & n\text{为奇数} \\ \dfrac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right), & n\text{为偶数} \end{cases} \tag{1.2.2}$$

例 1.2.2 (续例 1.2.1) 计算这 20 名中层管理人员年终奖金的中位数.

解 将这 20 名中层管理人员的年终奖金从低到高排列如下

34.7 35.1 36.2 36.7 36.9 37 37.1 37.7 37.8 37.9

38.3 38.6 38.9 39.2 39.6 39.6 40 40.6 40.8 41.7

由于一共有 20 个数据, 所以中位数就等于排序后第 10 个和第 11 个数据的平均值, 即

$$M_e = \frac{37.9 + 38.3}{2} = 38.1$$

需要注意的是, 极大值和极小值对中位数没有影响, 而对均值则会造成一定影响. 如上例中, 若将较大的两个值 40.8 和 41.7 分别替换为 42.5 和 44, 那么中位数没有改变, 仍然是 38.1, 而均值则变为 38.42. 因此相对于均值, 中位数具有一定的稳健性或耐抗性.

1.2.3 四分位数

中位数是从中间点将全部数据等分为两部分. 为了更详细地反映数据的分布位置, 还可以将数据做更多的等分. 简单来讲, 四分位数是将所有的数据等分为四部分, 处在各分点位置的数据就是四分位数.

通常情况下, 称第一个四分位数为**下四分位数**, 记为 Q_L; 第三个四分位数为**上四分位数**, 记为 Q_U; 而第二个四分位数恰好就是中位数, 记为 Q_M. 四分位数的计算方法与中位数的计算类似, 如上例中, $Q_L = 36.95$, $Q_M = M_e = 38.1$, $Q_U = 39.6$.

如果处理的是分组数据, 则先确定 Q_L 和 Q_U 的位置以及它们各自所在的组, 然后再仿照中位数的计算公式确定 Q_L 和 Q_U 的具体数值. 具体计算公式如下:

$$Q_L = L_L + \frac{\frac{n}{4} - S_L}{f_L} \times i_L \tag{1.2.3}$$

$$Q_U = L_U + \frac{\frac{3n}{4} - S_U}{f_U} \times i_U \tag{1.2.4}$$

其中 n 是数据的总个数, L_L 和 L_U 分别是 Q_L 和 Q_U 所在组的下限值; f_L 和 f_U 分别是 Q_L 和 Q_U 所在组的频数; i_L 和 i_U 分别是 Q_L 和 Q_U 所在组的组距; S_L 和 S_U 分别是 Q_L 和 Q_U 所在组以前各组的累积频数.

1.3　离散程度的描述

通常情况下, 对数据资料的基本分析仅仅关注其集中趋势的描述还不够, 还需要对数据的离散趋势作出有效的描述. 中心位置只能反映数据集的部分特征, 不同的数据集即使具有相同的中心位置 (如均值或中位数), 可能仍然存在很大的差异, 如下例.

例 1.3.1　比较下面 A、B 两个小组 (各有 10 人) 的身高 (单位: cm).

A	171	167	168	173	172	170	169	173	169	168
B	165	174	162	180	159	183	165	172	175	165

经过计算可以得到

$$\overline{x}_{\mathrm{A}} = 170\,\mathrm{cm}, \quad \overline{x}_{\mathrm{B}} = 170\,\mathrm{cm}$$

也就是说, 这两个小组的平均身高是相同的, 都是 170 cm. 但是仔细分析可以发现, 相对于 B 组, A 组中每个成员都比较接近平均身高, 而 B 组中成员的身高差异比较大. 那么如何将这种区别更加精确地描述出来呢?

1.3.1　极差和四分位数间距

对于一个样本的观测数据, **极差**(Range)定义为最大值与最小值之差, 通常用 R_n 表示. 记观测数据的最小值和最大值分别记为 $x_{(1)} = \min\{x_1, x_2, \cdots, x_n\}$、$x_{(n)} = \max\{x_1, x_2, \cdots, x_n\}$, 则

$$R_n = x_{(n)} - x_{(1)} \tag{1.3.1}$$

例 1.3.1 中 A、B 两组数据的极差分别为

$$\begin{aligned} R_{n\mathrm{A}} &= 173 - 167 = 6\ \mathrm{cm} \\ R_{n\mathrm{B}} &= 183 - 159 = 24\ \mathrm{cm} \end{aligned}$$

说明 B 组成员的高矮差别比较大. 可以看到, 极差的取值直接反映出样本数据取值范围的大小, 其最大优点就是实际操作简单. 但是极差的缺点也非常明显, 它只依赖于所有数据中的两个极端取值, 而对这两个极端值之间的数据分布情况一无所知.

四分位数间距f_s 定义为上四分位数 Q_U 与下四分位数 Q_L 之差, 它反映了中间一半数据的分布范围, 即

$$f_s = Q_U - Q_L \tag{1.3.2}$$

例 1.3.1 中 A、B 两组数据的四分位数间距分别为

$$f_{s\mathrm{A}} = 172 - 168 = 4 \text{ cm}$$
$$f_{s\mathrm{B}} = 175 - 165 = 10 \text{ cm}$$

也说明了 B 组成员的高矮差别比 A 组成员大. 四分位数间距只与取值在下四分位数和上四分位数之间的数据有关, 而与取值落在这个范围之外的数据无关. 四分位数间距也称为半极差, 等于中间一半数据的极差, 它作为分散性的度量比极差 R_n 更稳健.

1.3.2 样本方差和样本标准差

为了描述所有数据偏离中心位置的程度, 最直接的想法就是给出所有数据相对于均值的偏差 (deviations from the mean), 即 $x_1 - \overline{x}$, $x_2 - \overline{x}$, $\cdots$, $x_n - \overline{x}$. 如果将这些偏差全部加起来, 由于

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

所以

$$\sum_{i=1}^{n} (x_i - \overline{x}) = \sum_{i=1}^{n} x_i - n\overline{x} = 0$$

这是因为这些偏差中有正有负, 直接求和之后出现了正负抵消的结果. 那么如何消除这种影响呢? 通常可以采用对所有偏差取绝对值或者平方之后再求和. 如果取绝对值, 即 $|x_1 - \overline{x}|$, $|x_2 - \overline{x}|$, $\cdots$, $|x_n - \overline{x}|$, 则会带来计算与分析上的困难, 因为在许多问题的分析过程中, 首先考虑的就是如何去绝对值的问题. 因此, 考虑对所有的偏差取平方, 即 $(x_1 - \overline{x})^2$, $(x_2 - \overline{x})^2$, $\cdots$, $(x_n - \overline{x})^2$, 并称

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{S_{xx}}{n-1} \tag{1.3.3}$$

为**样本方差**, 其中 $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$. 相应的**样本标准差**为

$$s = \sqrt{s^2} \tag{1.3.4}$$

s 的量纲与样本数据 x_i 的量纲一致, 在实际问题中使用更为方便.