普通高等教育“十二五”规划教材

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{u^2}{2}}du$$

概率论与数理统计

刘 赪 程世娟 赵联文 何 平 编

科学出版社

普通高等教育“十二五”规划教材

概率论与数理统计

刘 赪 程世娟 赵联文 何 平 编

科学出版社

北京

内 容 简 介

本书针对工科类专业的特点，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.

全书分为8章，第1章主要介绍常用的描述性统计方法，第2~4章包括了相关的概率论知识和数理统计的基本概念，第5~8章则分别介绍了常用统计方法的思想以及具体分析过程. 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、联合概率分布及简单随机样本、点估计、基于单个总体的区间估计与假设检验、关于多个正态总体的统计推断以及回归分析.

本书可作为普通高等院校工科类各专业本专科学生学习概率论与数理统计课程的教材，也可供自学者和相关科研工作者参考使用.

图书在版编目(CIP)数据

概率论与数理统计/刘赫等编. —北京：科学出版社，2011
普通高等教育“十二五”规划教材
ISBN 978-7-03-031604-2

Ⅰ.①概… Ⅱ.①刘… ②程… ③赵… ④何… Ⅲ.①概率论-高等学校-教材 ②数理统计-高等学校-教材 Ⅳ.O21

中国版本图书馆CIP数据核字(2011)第115119号

责任编辑：胡云志 唐保军/责任校对：冯 琳
责任印制：张克忠/封面设计：华路天然工作室

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
2011年6月第一次印刷 印张：14 3/4
印数：1—4 000 字数：290 000

定价：29.00元

(如有印装质量问题，我社负责调换)

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 $\longrightarrow$ 统计模型 $\longrightarrow$ 求解模型 $\longrightarrow$ 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.

本书的编写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! 尤其感谢西南交通大学数学学院李裕奇教授一直以来的关心和帮助, 并为本书的编写提出了许多宝贵意见和建议. 此外, 我们特别感谢西南交通大学教务处教材科的同事, 他们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的任俊红为本书的顺利出版给予了鼎力协助.

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, 在此谨向有关编者和作者一并表示感谢.

书中难免有不足之处, 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

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.

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. 而如果要研究的是某城市大学在校学生的月生活费支出情况, 那么总体就包含了该城市所有大学的在校学生. 可见, 总体是根据研究范围来确定的.

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. 一般情况下就将所关注的特征量视为总体, 而每个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, 不可能将每个个体都逐一考察. 例如, 为了了解 2008 年某市居民用于食品的平均消费情况, 应该如何做呢? 显然, 我们不可能去调查该市的每位居民, 然后得到所需要的数据. 在实际研究中只需要随机选取该市一部分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, 通过分析这些数据发掘出所需要的信息, 并得到一定的结论. 因此, 在实际调查和研究中, 能够得到的就是从总体中随机抽取的一部分个体, 称其为**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. 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, 每个人成功逃生所耗费的时间为 (单位: 秒)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. 统计描述语言包括图形、表格、各种特征量, 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, 描述性统计可以分为两类: 一类是利用图形的直观性对数据特征进行展示, 如直方图、散点图、折线图等; 另一类则是通过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、标准差、中位数等. 以下将分别介绍一些比较常用的描述性统计方法.

1.2　描述性统计中的图形显示

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.

一般而言, 描述性统计可以分为两类: 一类是图表法, 即利用可视化的工具描述数据; 另一类是数值法, 即利用代表性的数值精确地描述出所给数据的基本特征. 本节将主要介绍描述性统计中与概率和统计推断联系密切的常用图表工具, 而在随后的两节中分别介绍关于数据集中趋势和分散程度的常用描述方法.

① Oxygen consumption and ventilation during escape from an offshore platform. Ergonomics, 1997: 281-292

数据的图形表示是一种简便而又突出主要信息的实用方法, 它能直观地显示数据蕴涵的一些重要信息, 不需要更多的专业背景知识也可以理解. 一个好的统计图形, 能够在最短的时间里传递出最多的信息, 用最少的笔墨给出更多的思维空间. 因此, 如何利用统计图表直观展示数据中所包含的信息, 也是撰写统计分析报告时必须考虑的重要问题. 以下将介绍比较常用的茎叶图、直方图和散点图.

1.2.1 茎叶图

对于未整理的原始数据, 可以利用茎叶图来直观展示数据的分布特征. 关于茎叶图的构造, 下面结合实例进行解释说明.

例 1.2.1 某班 30 名同学的某课考试成绩为

85, 80, 95, 85, 49, 62, 71, 60, 73, 81, 66, 41, 94, 74, 63, 67, 72, 90, 62, 62, 64, 50, 47, 54, 76, 72, 64, 57, 73, 81.

我们根据这组数据给出一个茎叶图 (stem-and-leaf display). 把每个数值分为两部分, 前面一部分 (十位) 称为**茎** (stem), 后面一部分 (个位) 称为**叶** (leaf), 如 85 可以分为茎 (8) 和叶 (5) 两部分, 中间用竖线分开, 即 8|5. 这 30 名同学考试成绩的茎叶图如图 1.1 所示.

4	1	7	9						
5	0	4	7						
6	0	2	2	2	3	4	4	6	7
7	1	2	2	3	3	4	6		
8	0	1	1	5	5				
9	0	4	5						

茎：十位
叶：个位

图 1.1 茎叶图

容易看到, 不及格的有 6 人, 最低分是 41 分, 最高分是 95 分, 分数主要集中在 60~69 和 70~79 两个分数段. 茎叶图不仅可以直观地读出数据的具体取值, 还能够保留数据的原始信息. 通过茎叶图, 可以看出数据的分布形状, 如分布是否对称、数据是否集中、是否存在极端值等.

综上所述, 可以按照以下步骤构造茎叶图.

(1) 根据数据的具体情况选取茎和叶. 如数据 256, 选择“百位和十位”作为“茎”, “个位”作为“叶”, 即 25|6;

(2) 将所有茎值按大小顺序排成一列;

(3) 对茎值相同的数据, 将其叶值由小到大依次排列在对应茎值的竖线右侧.

如果有两组数据需要对比分析时, 还可以做出它们背靠背的茎叶图, 这是一种简单直观而且有效的对比方法.

例 1.2.2 记上例中的班为甲班, 现有另外一个班 (乙班)30 名同学的考试成

绩为

78, 86, 59, 87, 67, 76, 74, 43, 71, 81, 66, 45, 94, 74, 63, 47, 72, 93, 62, 82, 64, 50, 78, 54, 86, 73, 54, 57, 73, 90.

为了对甲、乙两个班的成绩进行比较, 将这两组数据的茎叶图按图 1.2 方式构造.

乙班									甲班								
					7	5	3	4	1	7	9						
			9	7	4	4	0	5	0	4	7						
		7	7	6	4	3	2	6	0	2	2	2	3	4	4	6	7
8	8	6	4	4	3	3	2	1	7	1	2	2	3	3	4	6	
				7	6	6	2	1	8	0	1	1	5	5			
						4	3	0	9	0	4	5					

图 1.2　甲乙两班成绩的背靠背茎叶图

上图中, 茎在中间, 右边表示甲班的成绩, 左边表示乙班的成绩. 可以看出, 最高分 95 和最低分 41 都在甲班, 乙班不及格的人数多于甲班, 而甲班 70~79 分的人数则少于乙班. 同时, 甲、乙两班的成绩都主要集中在 60~79.

1.2.2　直方图

直方图 (histogram) 是用矩形的宽度和高度来表示频数分布的图形, 可以用于观察数据的分布情况. 具体来讲, 就是在平面直角坐标中, 横轴表示数据分组, 纵轴表示频率, 这样形成的矩形条就称为**直方图**, 其面积恰好就等于数据落在该区间间隔的频率, 因此也称为**频率直方图**.

以例 1.2.1 的数据为例, 做直方图的具体步骤如下.

(1) **对数据进行分组**. 显然, 组的划分对直方图极为重要. 那么在数据的最大值与最小值之间, 如何分组更为合适呢? 通常来讲, 结合问题背景, 组数控制在 5~20 个. H. A. Sturges 建议使用以下经验公式来确定组数:

$$\text{组数 } k = 1 + 3.31 \times \lg n$$

其中 n 是数据总量. 本例中有 30 个数据, 按照上述公式, 可以分为 6 组.

(2) **确定每组组距**. 实际使用中为了便于比较, 通常令各组区间长度相同, 也称其为**组距**, 用所有数据中的最大值与最小值之差除以组数即可得到. 此处, 为方便起见组距选择 10, 与通常对学生成绩考查时的习惯也非常吻合.

(3) **确定分组区间**. 选择略小于最小观测值的数 a, 略大于最大观测值的数 b, 根据所确定的组距将 (a, b) 区间分为 k 个分组区间. 本例中可选择 $a = 40$, $b = 99$, 分组区间为

$$(40, 49], [50, 59], [60, 69], [70, 79], [80, 89], [90, 99]$$

(4) **计算频率**. 统计所有数据中落在每个区间的频数, 并计算相应的频率. 本例中的频率列在表 1.1 中.

表 1.1 频数频率汇总表

组数	区间	频数	频率	累积频率/%
1	(40, 49]	3	0.1	10
2	[50, 59]	3	0.1	20
3	[60, 69]	9	0.3	50
4	[70, 79]	7	0.23	73
5	[80, 89]	5	0.17	90
6	[90, 99]	3	0.1	100

(5) **绘制直方图**. 横坐标表示分组变量, 纵坐标表示频率, 在横轴上以分组区间为底, 以频率/组距为高依次画出长方形, 这样就可以得到单位频率直方图, 简称频率直方图, 本例的直方图如图 1.3 所示.

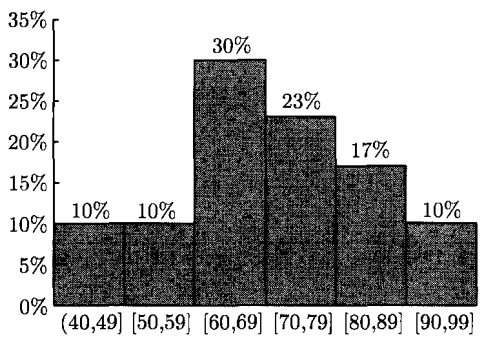

图 1.3 直方图

直方图可以直观的反映出数据的分布情况, 相对于茎叶图更容易被大众所理解和接受. 因此, 直方图也是对数据进行描述性分析时十分常用的处理工具.

1.2.3 散点图

将每个数据在坐标图上用相应的点表示所得到的图, 称为散点图 (scatter plot). 通过散点图, 可以直观地展示数据的分布特征和变化趋势. 多数情况下, 散点图也是对数据进行统计分析的第一步.

例如, 根据某市统计年鉴中相关年份的数据, 可以描绘出 1978~2004 年某市不同产业从业人员的总人数散点图, 如图 1.4 所示. 为了更加清晰直观地表现出不同产业从业人员的变化趋势, 可以用折线将各产业所对应的数据点连接起来, 形成折线图, 如图 1.5 所示.

特别是对于二维数据, 在分析两个变量之间的相互关系时, 通常都是先做散点

图, 为统计模型的构造提供直观思维, 并为选择恰当的分析方法提供参考.

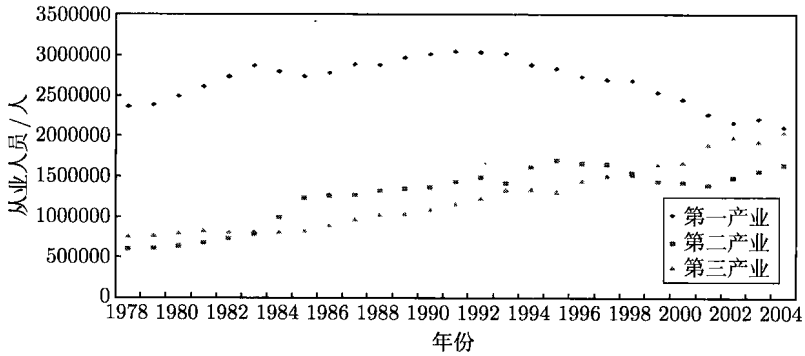

图 1.4　某市 1978~2004 年分产业从业人员的总人数散点图

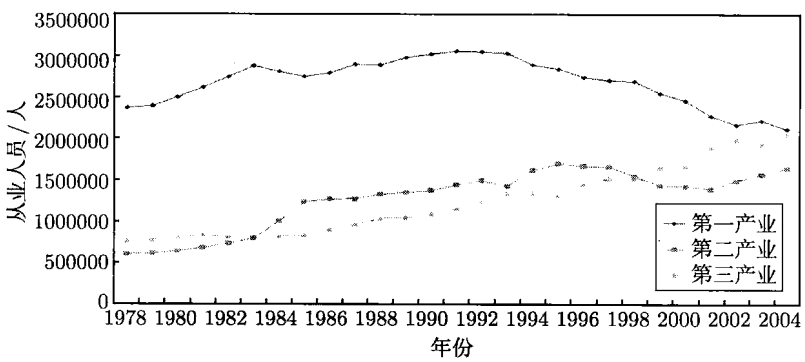

图 1.5　某市 1978~2004 年分产业从业人员的变化趋势折线图

以上介绍的直方图和散点图可以利用 SPSS 统计软件或 Excel 办公软件绘制, 帮助我们对收集到的观测数据进行初步整理和直观展示, 同时也为进一步的分析提供思路.

1.3　中心位置的描述

利用图表等可视化工具可以从数据中获得一些初步的认知和信息. 如果做进一步分析, 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.

1.3.1 均值

均值 (mean) 也称为算术平均值, 是指全部样本数据的算术平均. 假设有 n 个样本数据 $x_1,x_2,\cdots,x_n$, 其均值 $\overline{x}$ 定义为

$$\overline{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{1.3.1}$$

显然, 均值只适用于数值型数据.

例 1.3.1 根据调查, 某集团公司的中层管理人员的年终奖金数据为 (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.

首先可以绘制出茎叶图, 如图 1.6 所示.

34	7	茎: 整数部分
35	1	叶: 小数部分
36	2 7 9	
37	0 1 7 8 9	
38	3 6 9	
39	2 6 6	
40	0 6 8	
41	7	

图 1.6 茎叶图

由于 $\sum x_i=764.4$, 根据式 (1.3.1) 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. 与茎叶图相比, 均值所提供的信息更加精确, 同时也更具有针对性.

1.3.2 中位数

中位数 (median) 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

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

例 1.3.2　续例 1.3.1, 计算这 20 名中层管理人员年终奖金的中位数.

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, 若将两个最大的两个值 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.

1.3.3　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 4 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. 具体计算公式为

$$Q_L = L_L + \frac{\frac{n}{4} - S_L}{f_L} \times i_L \tag{1.3.2}$$

$$Q_U = L_U + \frac{\frac{3n}{4} - S_U}{f_U} \times i_U \tag{1.3.3}$$

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.

1.4　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