普通高等院校“十二五”规划教材

概率统计基础

（第2版）

主编 颜素容 崔红新

国防工业出版社
National Defense Industry Press

普通高等院校“十二五”规划教材

概率统计基础

（第2版）

主　编　颜素容　崔红新

副主编　刘　芳　缪素芬　贾爱娟　洪全兴

王丽娜

主　审　刘仁权

编　委　陶　欧　毕文斌　赵文峰　杨　洁

陈瑞祥　邬瑞光　安　红　尤海燕

国防工业出版社

·北京·

内 容 简 介

本书在简要叙述了统计学的含义、数据的收集整理及数据特征的基础上，着重介绍了与统计方法原理密切相关的概率论基础知识、统计学重要概念及统计推断的基本思想和方法、实际中常用的基本统计方法以及 Excel 中的常用统计功能等，共 11 部分内容。其中目录中带 * 号的节及书中“附”的内容，是供学有余力的学生进一步理解教学内容及课外阅读的拓展内容。

本书可供高等院校工科类、管理类、医药类等各专业、各层次的学生使用，也可作为参考书或对统计方法及其应用感兴趣的相关人员的自学教材。

图书在版编目(CIP)数据

概率统计基础/颜素容，崔红新主编. —2 版. —北京：国防工业出版社，2013.1

普通高等院校“十二五”规划教材

ISBN 978-7-118-08509-9

Ⅰ.①概… Ⅱ.①颜… ②崔… 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(2012)第 301996 号

国防工业出版社出版发行

(北京市海淀区紫竹院南路 23 号 邮政编码 100048)

北京奥鑫印刷厂印刷

新华书店经售

*

开本 787×1092 1/16 **印张** 13½ **字数** 311 千字

2013 年 1 月第 1 版第 1 次印刷 **印数** 1—4000 册 **定价** 30.00 元

(本书如有印装错误，我社负责调换)

国防书店：(010)88540777 发行邮购：(010)88540776

发行传真：(010)88540755 发行业务：(010)88540717

前　言

目前，高等教育正逐步由纯理论教学向应用型教学转变。衡量教学质量的标准不仅仅是传授理论知识的多少，更体现在对学生能力的培养。教学的目标，从只注重知识、忽略应用背景转变为既注重知识的积累和掌握，又重视对知识的应用和实际问题的分析。特别是统计学，在科学发展的不同领域发挥着巨大作用，统计学的教育更应顺应时代的发展，在注重统计思想教育的基础上，加强培养应用统计学的能力，同时培养学生的自主学习、沟通合作和解决现实问题的能力。由于统计学涉及的理论和方法很多，作者依据教学目标，结合多年教学经验和科研实践，对课程内容的选择做了一些有益的探索和尝试，本教材的主要特色有以下几点。

(1) 统计类基础课程重点在于统计学基础入门知识，由于授课对象在统计方面的知识储备很少，直接将科研工作中用到的很多复杂的统计学理论和方法介绍给学生是不现实的。所以，由实际问题引入，强化学生的统计思想；理解统计原理、明确统计对象、掌握统计方法。在选择内容时，既注重统计的基本概念、基本方法，强调基本的统计思想及原理，又注重统计思想及原理的延伸，为今后学习和工作开拓必要的空间。

(2) 增加了统计描述的相关内容，统计描述是将统计学应用于实际问题的方法之一。根据实例说明如何收集数据，如何将收集到的数据进行整理、展示，使人们从少数的特征数或简单的图表中了解大批数据所蕴藏的信息。使学生能够在刚刚接触统计时，就对统计的应用性有所了解，有利于学生更全面地了解统计学的思想与方法，从而增强实际应用能力。

(3) 对于理论性较强和难度较大的内容，如概率论和统计中的部分内容，重点编排与统计方法原理密切相关的必备知识，减弱其理论性和难度。而与统计学中密切相关的内容和结论，以通俗易懂的方式加以介绍，并将具体内容及相关证明，以“附”的形式写出作为参考。对基本内容的通俗讲解与附加学习材料相结合的方式，有利于突出课程的关键内容，满足不同层次学生的需求。

(4) 统计推断方法、原理，以及常用的基本统计方法是统计中的重要内容。在内容编排上重点强调统计方法的适用条件、应注意的问题、结果分析和解释，同时淡化了某些定理公式的证明及推导过程，增强学生应用统计学方法解决问题的能力。

(5) 目录中带 * 号的内容，虽然不是最基本的教学内容，但对启迪学生的统计思维使其更深入地理解统计学是很有帮助的。

(6) 每章后面增加一些阅读材料，使学生更好地了解概率统计的客观背景，扩大

学生的知识面，提高学生的学习兴趣。

本书的主要教学对象是高等院校工科类、管理类、医药类等各专业、各层次的学生，及对统计方法及其应用感兴趣的相关人员。本书的出版能够帮助他们理解统计学、熟悉统计语言，欣赏到数据是如何被转化为比数据本身更为复杂的知识，并知道如何评估统计结果。如果想研究统计学，本书是这条乐趣无穷道路上的一个起点。

本书主要由北京中医药大学和河南中医学院组织编写。本书的编写得到了参编单位各级领导的关心与支持。在本书编写过程中，借鉴和吸收了国内外相关的文献和科研资料。全体参编者对本书提出了宝贵意见，付出了辛勤劳动。在此，对各级领导及参编人员等的大力支持与帮助表示最衷心的感谢。

由于编者能力所限，时间较紧，教材中难免会存在不足之处，恳请广大师生及读者提出修改意见。

编 者

目 录

绪论 …… 1
0.1 统计学的含义 …… 1
0.1.1 什么是统计学 …… 1
0.1.2 统计学的主要思想 …… 2
0.1.3 统计的应用 …… 4
0.2 数据的收集与整理 …… 6
0.2.1 数据的收集 …… 6
0.2.2 数据的整理 …… 7
0.3 数据特征的描述 …… 10
0.3.1 描述集中趋势的数据特征 …… 10
0.3.2 描述离散趋势的数据特征 …… 12
0.3.3 相对数 …… 14
阅读材料 …… 15
习题 …… 18
第1章 概率论的基本概念 …… 20
1.1 随机事件 …… 20
1.1.1 随机试验 …… 20
1.1.2 样本空间 …… 21
1.1.3 随机事件 …… 21
1.1.4 事件间的关系与运算 …… 21
1.2 频率与概率 …… 23
1.2.1 频率 …… 23
1.2.2 概率的统计定义 …… 24
1.3 等可能概型 …… 26
1.4 条件概率 …… 28
1.4.1 条件概率概述 …… 28
1.4.2 乘法定理 …… 30
1.4.3 全概率公式与贝叶斯公式 …… 30
1.5 独立性 …… 32
1.5.1 事件的独立 …… 32
1.5.2 相互独立事件至少发生其一的概率 …… 34
阅读材料 …… 34
习题一 …… 35
第2章 随机变量的概率分布和数字特征 …… 37
2.1 随机变量 …… 37
2.2 离散型随机变量 …… 38
2.2.1 离散型随机变量的概念 …… 38
2.2.2 三种重要的离散型随机变量 …… 39
2.3 随机变量的分布函数 …… 43
2.4 连续型随机变量 …… 45
2.4.1 连续型随机变量的概念 …… 45
2.4.2 正态分布 …… 46
*随机变量的函数的分布 …… 50
*多维随机变量及其分布 …… 51
2.5 随机变量的数字特征 …… 52
2.5.1 数学期望 …… 53
2.5.2 方差 …… 57
阅读材料 …… 61
习题二 …… 61
第3章 随机样本及抽样分布 …… 63
3.1 随机样本 …… 63

3.1.1 总体与样本 …… 63
3.1.2 总体分布 …… 64
3.1.3 有限总体与无限总体 …… 65
3.1.4 样本的二重性 …… 65
3.1.5 简单随机样本 …… 65
3.1.6 统计量 …… 66
3.2 抽样分布 …… 67
3.2.1 χ^2 分布 …… 67
3.2.2 t 分布 …… 68
3.2.3 F 分布 …… 68
3.2.4 正态总体的样本均值与样本方差的分布 …… 69
阅读材料 …… 74
习题三 …… 75
第4章 参数估计 …… 76
4.1 点估计 …… 76
4.2 估计量的评价标准 …… 77
4.2.1 无偏性 …… 77
4.2.2 有效性 …… 79
4.2.3 一致性 …… 79
4.3 区间估计 …… 80
4.4 正态总体均值和方差的区间估计 …… 82
* 离散总体参数的区间估计 …… 87
阅读材料 …… 90
习题四 …… 91
第5章 假设检验 …… 92
5.1 假设检验的基本思想 …… 92
5.2 正态总体均值的假设检验 …… 97
5.2.1 单个正态总体均值的假设检验 …… 97
5.2.2 两个正态总体均值差的假设检验 …… 100
5.2.3 基于成对数据的假设检验 …… 102
5.3 正态总体方差的假设检验 … 104
5.3.1 单个正态总体方差的假设检验 …… 104
5.3.2 两个正态总体方差的假设检验 …… 106
* 置信区间与假设检验之间的关系 … 107
* 离散总体参数的假设检验 …… 108
阅读材料 …… 110
习题五 …… 112
第6章 χ^2 检验 …… 114
6.1 χ^2 统计量 …… 114
6.2 独立性检验 …… 115
6.2.1 2×2 列联表独立性检验 …… 115
6.2.2 $r\times c$ 列联表的独立性检验 …… 118
阅读材料 …… 120
习题六 …… 122
第7章 方差分析 …… 124
7.1 相关术语 …… 124
7.1.1 试验指标 …… 124
7.1.2 因素 …… 124
7.1.3 水平 …… 124
7.1.4 试验处理 …… 125
7.2 单因素方差分析 …… 126
7.2.1 单因素方差分析的数学模型 …… 127
7.2.2 平方和分解 …… 128
7.2.3 SS_E, SS_A 的统计特征 …… 129
7.2.4 假设检验问题的拒绝域 …… 130
7.3 双因素方差分析 …… 131
7.3.1 双因素等重复试验的方差分析 …… 131
7.3.2 双因素无重复试验的方差分析 …… 134
* 两两均值多重比较 …… 136

阅读材料…… 137
习题七…… 139
第 8 章　相关与回归…… 141
8.1　直线相关 …… 141
8.1.1　相关系数 …… 141
8.1.2　相关系数的假设检验 …… 142
8.2　一元线性回归 …… 143
8.2.1　一元线性回归方程的建立 …… 144
8.2.2　一元线性回归方程的检验 …… 145
* 多元相关与回归 …… 147
* 非线性回归方程 …… 148
阅读材料…… 150
习题八…… 152
第 9 章　正交设计…… 154
9.1　正交表 …… 154
9.2　用正交表安排试验 …… 155
9.3　正交试验结果的极差分析 … 157
9.3.1　等水平正交表的极差分析 …… 157
9.3.2　混合水平正交表的极差分析 …… 158
9.4　正交试验结果的方差分析 … 159
9.4.1　2 水平正交表的方差分析 …… 159
9.4.2　3 水平正交表的方差分析 …… 162
阅读材料…… 164
习题九…… 165
第 10 章　Excel 在统计中的应用 … 167
10.1　概述…… 167
10.2　假设检验…… 167
10.2.1　成组 t 检验…… 167
10.2.2　配对 t 检验…… 170
10.3　方差分析…… 172
10.3.1　单因素方差分析…… 172
10.3.2　双因素无重复试验的方差分析 …… 174
10.3.3　双因素可重复试验的方差分析 …… 177
10.4　相关与回归…… 179
10.4.1　散点图 …… 179
10.4.2　相关系数 …… 180
10.4.3　回归方程 …… 182
阅读材料…… 183
附表 …… 186
附表 1　随机数表 …… 186
附表 2　二项分布表 …… 187
附表 3　泊松分布表 …… 188
附表 4　标准正态分布表 …… 189
附表 5　χ^2 分布表…… 190
附表 6　t 分布表 …… 191
附表 7　F 分布表 …… 192
附表 8　Dunnett—t 界值表…… 198
附表 9　q 界值表 …… 199
附表 10　相关系数 r 界值表 …… 200
附表 11　常用正交表 …… 201
附表 12　常用均匀表 …… 204
习题参考答案…… 205
参考文献 …… 208

绪　论

概率论与统计学旨在研究随机现象的统计规律性，是两个密切联系的学科。统计学主要研究怎样有效地收集、整理和分析带有随机性的数据，对所考察的问题作出推断或预测，为采取一定的决策和行动提供依据和建议。鉴于统计学所考察的数据的随机性（偶然性）造成的不确定性，借助概率论的概念和方法成为必要。

0.1　统计学的含义

0.1.1　什么是统计学

统计学随着科学技术中众多问题的出现应运而生，并在人类关注的许多问题上起着重要作用。“*Statistics*”这个词，最早被应用于政府部门对人们出生和死亡信息的记录，它至今在世界上各个层次的政府机构中不可或缺。统计学主要利用概率论建立数学模型，收集所观察系统的数据，进行量化的分析、总结，并进而进行推断和预测，为相关决策提供依据和参考。

统计学可以分成统计描述和统计推断两大类。

1. 统计描述

信息的收集、提炼和展示通常被认为是统计描述。从本质上说，统计描述是通过有目的的、有意义的数据的收集和整理，使人们能够洞察到事物的本质特征。一般情况下，统计描述包括以下内容。

（1）将收集到的数据绘制成图像。

（2）把大量数据提炼成更容易理解的形式（如表格）。

（3）归纳一套简单的度量方法来描述复杂的信息。例如，平均值可以用在一系列数据中提炼出的一个典型数值。

2. 统计推断

统计推断是统计学的核心问题，其理论和方法构成了统计学的主要内容。它提供了分析数据的科学方法，而这些数据是通过统计描述得到的。统计推断涉及的范围很广，主要包含以下内容。

（1）决定某一情况的任一显然特性是否成立。

（2）对未知数进行估计，并决定这些估计值的可靠性。

（3）利用过去发生的事情尝试预测未来。

利用统计推断技术，统计学家可能调查的问题类型的例子如下。

（1）某些疾病或病害与任意特定因素之间是否存在关系？

（2）男人和女人在数学智力上是否存在差距？

(3) 参加课堂学习是否真的影响期末考试的分数?

(4) 广告花费与销售数量密切程度如何?

(5) 气温和海拔高度是否影响运动成绩?

(6) 第一胎的小孩是否比第二胎的小孩聪明?

(7) 制造商关于产品的主张是否是正确的?

(8) 民意测验和调查是否是民意的准确反映?

(9) 做科学度量的某种技术是否比另外一种好?

(10) 抽烟和肺癌是否存在某种关联?

(11) 某种药物是否能真正治疗某种精神病?

(12) 使用无铅汽油真的可以极大地有助于减少空气污染吗?

这些问题能够用适当的统计检验技术来解答,当然统计学家们有很多这样的检验技术来选择。但在一些情况,很难正确地决定哪一种统计学检验是最恰当的。第 5 章会介绍一些统计检验方法。

0.1.2 统计学的主要思想

1. 统计是一项对随机性中的规律性的研究

当不能预测一件事情的结果时,这件事就有了随机性。例如,当掷硬币时,并不能确定硬币将是正面朝上,还是背面朝上;外出旅游时,也不能确定是否会发生意外。但当把这些随机事件放在一起时,它们会表现出令人惊奇的规律性。如果将同样的硬币掷 100 次,它大概有 50 次正面朝上,50 次背面朝上。与之类似,尽管某一车祸发生的可能性很小,但发生可能性的比率仍有一个稳定的值与之对应。

统计思想的基础知识有助于归纳随机事件可能的规律性,它帮助人们理解随机性和规律性的重要性。因此统计可以看作是一项对随机性中的规律性的研究。

2. 统计是对数据中的偏差问题的研究

然而,规律也表现出某种随机性,如果再将同样的硬币掷 100 次,正面朝上的次数几乎不会和前 100 次一样,在第一个 100 次中,也许有 45 次硬币正面朝上,而在第二个 100 次中,也许有 55 次正面朝上。这表明了统计的一个重要的本质特征,无论重复多少次试验,一般来说,每次得到的结果不尽相同。

这种偏差不仅仅发生于掷硬币案例中,调查、实验和其他任何一组方式的数据收集中都有可能发生。如在某调查中,人们被问到对某一问题的看法,某一比例的人会有某一特定观点;如果对不同的人再做同样调查,支持这一观点的比例则有所不同。这两个比例的差异主要是由数据本身的随机性引起的,后面称其为随机误差。从这种意义上来说,统计就成了对数据中的偏差问题的研究。

根据作为统计基础的概率理论,可以确定一项调查中的某个比例发生的可能性有多大,在下一次的重复调查中,这个比例可能有多大偏差,甚至可以指出这两个比例之间的差异,是否大到随机性本身所不能解决的地步。以后章节将会引申和讨论这些内容。

下面是研究随机性和规律性的两个例子。

实例一 一个说明两个数字之间的差异是否不能仅归因于随机性的例子是,20 世纪

50年代小儿麻痹症疫苗的投入使用。小儿麻痹症是一种可怕的疾病，通常能使患者（大部分是儿童）瘫痪或死亡。这种病持续多年后，一种疫苗最终被研制出来。科学家们希望该疫苗能够预防这种可怕的疾病，但是没有人清楚这种疫苗是否产生预期的效果。尽管实验室和动物实验的结果使人兴奋，然而人体实验才是唯一检验这种疫苗是否能起作用的方法。因为小儿麻痹症是一种较罕见的疾病，疫苗必须试用于相当一大批孩子们的身上，所以研究者们决定在200000个孩子身上做实验。此外，研究者们还决定用另外相同数目的孩子作为对照组。为观察疫苗是否真的起作用，对照组的孩子仅仅得到安慰剂——一种看起来像疫苗的替代品。

当孩子们被注射了疫苗或安慰剂以后，研究者们开始在下一个小儿麻痹症发病时，观察实验结果。在对照组中，有138个孩子感染了此病。这个数字当然有一定的随机性，研究者们并不能确定它意味着什么。如果另外一组的200000个孩子也被注射安慰剂，那么不一定会有同样多的孩子感染此疾病。根据随机性的大小，可能有130或140或其他数目的孩子们染上小儿麻痹症。

在被注射了疫苗的那一组中，有56个孩子患了小儿麻痹症，这个数字当然也有随机性。一个重要的问题是，56和138的差别是否超过了随机性所能解释的程度。如果是，那么研究者们就能够有把握地说，疫苗起作用了。利用第5章介绍的方法可以看到，138和56的差别超出了随机性本身所能解释的范围，因此疫苗被认为是成功的。从此以后，这种疫苗在许多国家根除了小儿麻痹症。全世界的健康组织所做的进一步的努力，将使不发达国家的孩子们在不远的将来就有可能免遭小儿麻痹症的痛苦。从某种重要的意义来说，统计推理为发展和检验疫苗的研究者们提供了有力的支持。

实例二 另外一个著名的随机性的例子——或者说缺乏随机性，正如这个特定的例子中的情况——发生于军事中。在美国对越南的战争中，为使前线有足够多的士兵，美国政府制订了一个“抓阄”的征兵计划。该计划计算把1到366的号码随机地分配给一年中的每一天，然后由军事部门按分配的号码顺序把生日与之对应的年轻人分批应征入伍。这种方法的目的是为了让大家以相等的概率卷入这场不受欢迎的战争中。

在第一年的征兵计划中，号码1被分配给了9月14日，分配方法是随机抽取一个大容器中的366个写上了日子的乒乓球。结果所有年满18岁且生于9月14日的合格青年将作为第一批被征召入伍。生日被分配为号码2的青年则在第二批被征召入伍，以此类推。可以知道并不是所有人都被征召入伍，因此，生日被分配的号码较大的人也许永远都不用到军队服役。

这种抓阄看起来对决定是否应该被征召入伍是一个相当不错的方法。然而，在抓阄的第二天，当所有的日子和它们对应的号码公布以后，统计学家们开始研究这些数据。经过观察和计算，统计学家们发现了一些规律。例如，本应预期应当有差不多一半的较小的号码(1到183)被分配给前半年的日子，即从1月份到6月份；另外一半较小的号码被分配给后半年的日子，从7月份到12月份。由于抓阄的随机性，前半年中可能不会正好分到一半较小的号码，但是应当接近一半。然而结果是，有73个较小的号码被分配给了前半年的日子，同时有110个较小的号码被分配给了后半年的日子。换句话说，如果生于后半年的某一天，那么，去服兵役的机会，要大于出生于前半年的人。

在这种情况下，两个数字之间只应该有随机误差，而73和110之间的差别超出了随

机性所能解释的范围。这种非随机性是由于乒乓球在被抽取之前没有被充分搅拌均匀而造成的。

3. 概率

在讨论随机性时已经看到，统计学的大部分内容基于一个很重要的概念——概率(probability)，概率为如何从数据中得出结论奠定了基石。统计学可能永远不能确定两个数字的差别是否超出了随机误差，但是可以肯定这种差别发生的概率的大小。有关概率的基础知识及具体如何判断差别的大小将在以后各章详细阐述。

4. 变量

变量(variable)，这个概念是统计研究中的另一基石。人类的性别特征是取两个值的变量：男和女。宗教信仰是一个变量，在西方国家中可能取值为天主教、犹太教、伊斯兰教及其他，不同国家，宗教的取值也不相同。还有其他变量的例子，例如，汽车加每升汽油所能行驶的千米数取值可能在 10km～80km 或者一剂药的药量等。

通常，研究开始时，就要确定他们感兴趣的变量及其取值范围。例如，性别变量是取值为男、女的类别变量；对某一行为的态度变量是取值为有序类别的变量，值为非常赞同、赞同、中立、反对、非常反对等。

5. 数据类型

变量的测量值或观察值称为变量值。变量值的全体构成数据或资料(data)。按照变量值的来源，可将数据分成以下几类。

1) 计量数据

计量数据(measurements)也称为定量数据，是指用仪器、工具或其他定量方法得到的数值，一般带有单位，如身高、体重、血压等。

2) 计数数据

计数数据(counts)为定性数据或无序数据，例如每天失业的雇员数，或者每年发生的交通事故数等。它可分为二分类或多分类资料，例如，性别分为男和女，为二分类的计数数据；人的血型分 A、B、AB、O 四种，为四分类计数数据。

3) 等级数据

等级数据(rank data)为半定性半定量的数据或有序数据。这种数据用级别表示某种现象在表现程度上的大小差别。比如，患者治疗后，疗效可分为治愈、显效、好转、无效或死亡共 5 个等级；消费者按照申请贷款的风险来归类，运动员按身体的适应性来归类等。

数据的类型可以根据需要进行转换。例如，成年男子的血清胆固醇按是否小于某个标准划分为血脂正常和异常两类等。

统计方法的选用是与数据类型密切联系的。在统计方法的学习中特别要注意。

0.1.3 统计的应用

基于统计的特点，以及快速、高效的计算机的出现，使得统计运算变得快速而且有效，统计与数据收集和分析在许多领域都有应用，如政府机关、自然科学、工业、农业、医药、经济、心理学、社会学等。

政府机关利用统计帮助制定解决各种问题的政策。例如，为决定税收政策，必须了解

现行的税法如何影响各种收入水平的人们，并需要预测税法变化对人们的影响；推行一个农业补贴计划，也必须知道当前农业产量的情况，并预期此计划执行后对将来产量的影响。

各个学术领域的人们在他们的科研中都使用统计，并形成且发展了自己的一套统计方法，如生物统计学、医学统计学、计量经济学、心理统计学等。在学术领域之外，统计也被大量使用。几乎所有的报纸、杂志刊登以统计为基础的文章。在文科方面，一大批历史学家、地理学家、语言学家等利用统计知识得出各种结论，例如，中世纪大鼠疫导致的死亡数；法语在英语国家中的普及程度等。在法律方面，由于统计在社会生活中日益发挥出重要作用，律师们除了要面对法律问题外，还要面对统计问题。

实例三 DNA检验，在这个双螺旋结构上悬着一个故事。在1995年结束的著名的辛普森(O. J. Simpson)谋杀案的审理中，许多证词都涉及DNA样本及它们的收集、分析和确认。从证人在各种层面上收集到的血液样本证据的统计数据中，公众了解到了很多统计知识。问题的关键在于收集的DNA样本有多大的可能性与受害者或被告人的血液相吻合。原告声称，血液样本不是辛普森的概率至少是非常小的，但被告律师反对该结论。

通常，DNA检验的过程是检查DNA链各种指标的模式并计算两个人都有这同一种模式的可能性。一旦这种方法成为可行，公众很快开始在各类案件中引用它。在另一个审判中，一名男子因犯强奸罪已经坐了7年牢，当他的律师证明他的DNA和真正的强奸者的DNA根本不匹配时，这名囚犯终于被释放了。

医药公司为了将一种新药推向市场，必须证明这种药是安全的。公司投入大量资金在动物和人身上做实验，以检验新药的功效。这些公司还雇佣了大批统计学家，他们负责正确安排实验、分析实验结果等。

统计方法在工业上被用来控制质量。从生产线出来的产品并不都是一样的，究其原因，一部分可由随机误差引起，一部分可由在生产过程中某些地方出错引起。统计方法可以研究这种差异，并帮助人们指出错误和错误的原因。

由于统计已被应用于如此多的学科和行业中，统计分析结果无处不在。作为研究者，必须具备统计的知识，而作为消费者，对统计的了解，可以帮助理解和评价对周围的现象或结果的准确性。

实例四 下面的故事有关一次著名的失败的统计调查，它一直是一个统计传奇。在1936年美国总统选举前，一份名为 *Literary Digest* 的颇受人尊重的杂志进行了一次民意调查。调查的焦点是谁将成为下一届总统，是堪萨斯州州长Alf Landon还是现任总统Franklin Delano Roosevelt。为了了解选民意向，民意调查专家们根据电话簿和车辆登记簿上的名单给一大批人发了简单的调查表(电话和汽车在1936年并不像现在这样普遍，但是这些名单比较容易得到)。尽管发出的调查表大约有1000万张，但收回的比例并不高。在收回的调查表中，Alf Landon非常受欢迎。于是，该杂志预测Landon将赢得选举。

如果读者有一些统计知识，他们会对这个声称Alf Landon将赢得选举的预测结果有疑问。因为在经济大萧条时期调查拥有电话和汽车的人们，并不能够很好地反映全体选民的观点。此外，只有少数的调查表被收回，这一点也是值得怀疑的。事实表明，最终是

Franklin Roosevelt 而非 Alf Landon 赢得了这次选举。由此可见，那次的调查结果有多么错误了。当前大多数应用统计不会像上一例子错得那样离谱，但即便在今天，人们也很容易发现统计被误用的情况，尤其在需要考虑选择正确的样本时更易被误用。

0.2 数据的收集与整理

0.2.1 数据的收集

在任何情况下，统计描述和统计推断的价值都由现有数据的价值而定。可靠数据的收集是统计工作的基础要求。

数据收集主要有两种方法，一是调查；二是实验。无论是调查还是实验，统计学对原始数据都要求完整和准确。

1. 调查数据

通过调查或观察而得到的数据称为调查数据。例如，调查某一地区被确诊为艾滋病病毒携带者的人数；调查某市中小学学生近视眼的情况等。调查研究的内容是多种多样，统计在如何收集数据和分析数据两个方面扮演了重要角色。

2. 总体、个体、样本

收集数据的目的是为了从收集研究对象的个体中得出结论。社会学家们收集相关数据以了解人类行为；植物学家收集有关植物的数据以了解它们如何生长；医学家收集有关流行病的数据以了解它们的特点。所有感兴趣的个体就构成了总体。把所研究对象的全体组成的集合称为总体(population)。而组成总体的每个元素称为个体。

有时，能够收集到总体中所有个体的数据，此时对总体做普查，如人口普查、健康普查等。然而，在现实生活中，由于总体个数庞大，或资金时间有限，以及不断变化的环境条件，做普查通常是不可能的。这时，就需要从总体中随机抽取一部分个体进行研究，此时对总体做了抽样调查。从总体中抽取的若干个体所构成的集合称为样本(sample)，样本中个体的个数称为样本容量(sample size)。

如何选择样本是统计研究者面临的一个关键问题，他们希望由研究样本而得出的结论能适应该样本所属的总体。如果没有一个“好”的样本，这是不能实现的。在前面提到的对越战争的例子中，选择士兵的征兵计划就是一个“不好”的样本。由于选择样本对于结果的可信度有重要作用，所以根据正确的统计原理选择样本是非常重要的。研究从总体中抽取样本的方法很多，第 3 章将介绍其中一种抽样方法，其他方法可参考有关抽样理论的书籍。

3. 参数、统计量

根据总体的统计学定义，统计学关心的是全部研究单位某个观测值(随机变量)的统计学特性。如某地 7 岁男童身高的平均值、某地全部高血压患者血清总胆固醇的平均值等。根据全部研究单位某个观测值计算的平均值也称总体均数。反映总体特征的统计指标称为参数(parameter)。重要的参数除总体均数之外，还有总体方差、总体标准差、总体相关系数等。由于大多数研究得不到总体数据，所以参数通常是未知的。

虽然多数情况下不知道参数，但可以从总体中抽取样本，通过计算样本的特征数，对

相应的未知参数作出估计。如用样本观测值计算的平均值(样本均数)估计总体均数;用样本观测值计算的标准差(样本标准差)作为总体标准差的估计值等。通过样本计算的、反映样本的统计学特征且不依赖未知参数的量称为统计量。

4. 抽样误差

当用统计量估计参数时,参数是固定的常数,而统计量则随着样本的观察值的变化而变化。例如,用抽样的方法估计某地 10 岁儿童身高的总体均数,如样本容量为 100,第一个样本的 100 个儿童身高的样本均数,一般不会等于第二个样本的;另外 100 个儿童身高的样本均数,也不会恰好等于总体均数(参数)。这种由于样本的随机性引起的统计量与参数的差别,或同一总体的相同统计量之间的差别,称为抽样误差(sample error)。抽样误差的大小用标准误度量,详见第 4 章。

5. 实验数据

通过在实验中控制一个或多个变量并测量结果而得到的数据,称为实验数据。在实验中,研究者希望控制某一情形的所有相关方面,对少数感兴趣的变量进行规划设计及实验,并观察实验结果。例如,前面研究小儿麻痹症疫苗是否有效的例子中,研究者们给一组儿童服用此疫苗,称这组为实验组;给另外一组服用安慰剂,称这组为对照组。几乎所有设计好的实验,都有一个对照组和一个或多个实验组。如果没有对照组,就没有比较疫苗是否产生作用的基础。

0.2.2 数据的整理

数据的整理是统计研究的基础。前面介绍了收集数据的方法,一旦数据被收集,就必须在这些数据中寻找所包含的信息。面对如此多的数据,以至于使人们无法把它们全部理解。因此,需要一些方法使人们能够从数据中提取信息,并转化成可用的形式。通常采用具有一定特点的表格(tables)、绘制一定形式的图(charts),如折线图(graphs)、饼图(pie)、柱状图(bar)、直方图(histogram)等,以及计算来整理数据。

最常用的数据整理方法是编制频数分布表,并根据需要作出样本的频数直方图,简称直方图。其他绘制图的方法各有不同特点,需要的时候可以选择相应的统计工具绘制。下面根据例子说明作直方图的步骤。

例 0.2.1 测量 100 个某种机械零件的质量,得到样本观测值如下(单位:g)

247 251 260 254 246 253 237 252 250 251
249 244 249 244 243 246 256 247 252 252
250 247 255 249 247 252 252 242 245 240
260 263 254 240 255 250 256 246 249 253
246 255 244 245 257 252 250 249 255 248
258 242 252 259 249 244 251 250 241 253
250 265 247 249 253 247 248 251 251 249
246 250 252 256 245 254 258 248 255 251
249 252 254 246 250 251 247 253 252 255
254 247 252 257 258 247 252 264 248 244

试编制零件质量的频数分布表并作频数直方图。

编制步骤如下。

1. 求极差

极差(range)也称全距，即样本观测值 $x_1, x_2, \cdots, x_n$ 中的最小值和最大值之差，记为 R。本例 $R=265-237=28$(g)。

2. 确定组段数和组距

组段数通常取 8 组～15 组，分组过多计算烦琐，分组过少难以显现分布特征。

组距可通过极差除以组段数求得，一般取方便阅读和计算的数字。本例中，将数据分成 10 个组，组距 $d=28/10=2.8\approx3$。

3. 根据组距写出组段

适当选取略小于最小值的数 a 与略大于最大值的数 b，将区间(a,b)按照组距分成所需的组段。每个组段的下限为 L、上限为 U，变量 X 值的归组统一定为 $L\leqslant X<U$，除最后组段写出上限以外，其他各组段可不用写上限。起始组段和最后组段应分别包含全部变量值的最小值和最大值，如表 0.2.1 第(1)栏所列。

设分为 k 个组段，即$[a,t_1)$，$[t_1,t_2)$，$\cdots$，$[t_k,b)$，子区间的长度 $\Delta t_i=t_i-t_{i-1}(i=1,2,\cdots,k)$，即为组距。

4. 分组统计频数

计算样本观测值落在各子区间内的频数 m_i 及频率 $f_i=\dfrac{m_i}{n}(i=1,2,\cdots,k)$。各组段的频数如表 0.2.1 第(2)栏所列，然后求频数合计，完成频数分布表。表 0.2.1 第(3)栏和第(4)栏用于后面的统计计算。其中组中值按下面公式计算，即

$$\text{缺上限的开口组组中值} = \text{下限} + \text{邻组组距}/2$$

表 0.2.1　100 个某种机械零件的质量频数

组段 (1)	频数 m (2)	组中值 X (3)	$m\cdot X$ (4)=(2)×(3)
236.5～	1	238	238
239.5～	5	241	1205
242.5～	9	244	2196
245.5～	19	247	4693
248.5～	24	250	6000
251.5～	22	253	5566
254.5～	11	256	2816
257.5～	6	259	1554
260.5～	1	262	262
263.5～266.5	2	265	530
合计	100	—	25060

根据表 0.2.1，以各组段零件质量为横坐标、频数 m 为纵坐标，可绘制频数分布图(graph of frequency distribution)，如图 0.2.1 所示。它比频数表更要直观和形象。

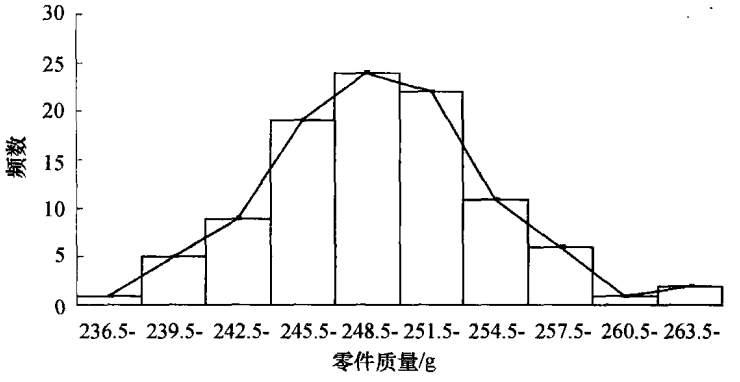

图 0.2.1

从频数分布图可以看出频数分布的类型，是对称分布(即正态分布)还是不对称分布(即偏态分布)。也可以看出频数分布的特征，如变量的变化范围、集中区域等，以便进一步作统计分析和处理。

附 如果在 X 轴上截取各子区间，以 $\frac{f_i}{\Delta t_i}$ 为高作小矩形，各个小矩形的面积 ΔS_i 就等于样本观测值落在该子区间内的频率，即

$$\Delta S_i = \Delta t_i \cdot \frac{f_i}{\Delta t_i} = f_i (i = 1,2,\cdots,k)$$

所有小矩形的面积总和等于 1，这样作出的所有小矩形就构成频率密度直方图(图 0.2.2)。

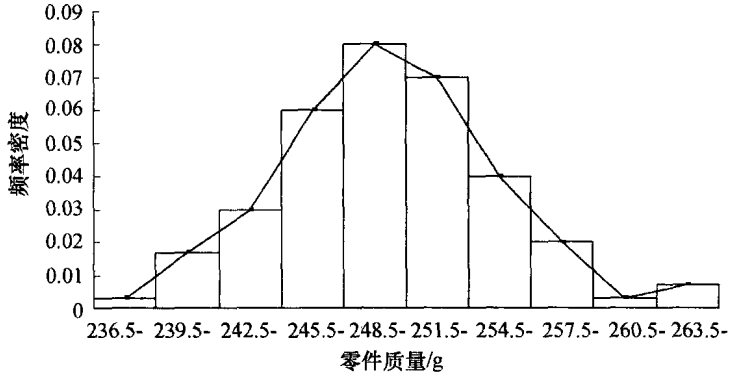

图 0.2.2

因为当样本容量 n 充分大时，随机变量 X 落在各个子区间 (t_{i-1}, t_i) 内的频率近似等于其概率，即 $f_i \approx P\{t_{i-1} \leqslant X < t_i\} (i=1,2,\cdots,k)$，所以直方图大致地描述了总体 X 的概率分布。概率分布在第 2 章可以详细看到。