观测数据的分析与处理

胡上序　陈德钊　编著

浙江大学出版社

观测数据的分析与处理

胡上序　　陈德钊　编著

浙江大学出版社

内容简介

本书所叙述的是一切需要试验与观测的各学科门类研究工作者必须掌握的一种基本技术。具体内容包括：数理统计的基础知识，连续和离散的随机变量及其分布，统计量的分布，统计推断，数据的预处理，一元和多元线性回归分析，逐步回归，多项式回归分析，样条回归，非线性回归，岭回归分析，一元与多元方差分析，正交试验设计，序贯试验设计等，共分十一章。

本书的撰写兼顾理论和实用两个方面。对理论的叙述尽可能用简洁明了的形式展开，使非数学专业出身的读者易于接受。同时又注意具体方法的详细介绍，使于读者在实际问题中加以运用。

本书的对象是广大的工程技术、科学研究、经济分析、社会调查等方面的工作人员。可以用作理、工、经、管、各科大学生或研究生课程的教材或教学参考书，也可作为实际工作者自学用书。

观测数据的分析与处理

胡上序　陈德钊　编著

责任编辑　宗贤钧

*　　*　　*

浙江大学出版社出版

（杭州玉古路 20 号　邮政编码 310027）

浙江大学出版社电脑排版中心排版

浙江农业大学印刷厂印刷

浙江省新华书店经销

*　　*　　*

787×1092　16 开　22.5 印张　576 千字

1996 年 3 月第 1 版　1996 年 3 月第 1 次印刷

印数：0001—1000

ISBN 7-308-01542-4/TP・151　定价：22.50 元

前　言

所有应用科学至少有一个共同特点，它们都具有实践性，自然科学或社会科学的各门类都是这样。而所有科学家们又至少有一个共同目标，就是寻找所研究对象的变化规律。

在推动自然科学和工程技术的发展中尤其重要的是进行试验，这是一种有计划的实践。试验的结果会产生大量的观测数据。如何通过对观测数据进行分析和处理，以期获得能反映客观世界变化的规律，就是本书要叙述的主要内容。

所谓规律，一般都具有定性和定量的两个方面。本书的目的是为研究工作者提供各种最常用的寻求规律的工具，且把重点放在定量方法上。这些方法的理论基础是数理统计学。数理统计方法的内涵庞大，新的发展很快。本书作为基本教材，涉及的只能是其中很小一部分。所以，对内容的取舍，颇堪斟酌。我们的原则是，理论与应用并重，经典基础与近代发展兼顾，覆盖的面要稍宽些，重点问题则要深入些。

为了便于各不同领域的研究工作者能既快而又准确地掌握基本方法，根据作者对非数学类理工科学生多年来的教学经验，本书的撰写注意到对基本概念作比较细致的叙述，然而又尽可能避免使用过于抽象和泛化的数学语言。在叙述基础理论的部分，加入一定数量的示例，以帮助读者理解。在叙述方法的部分，则通过具体步骤来说明问题，避免使用繁复的数据示例，以节约篇幅。

本书的内容可以分为五个部分。第一部分介绍数理统计学的理论基础，包括基本概念，离散和连续的随机变量及其分布，统计量的分布，统计推断，分别放在第1、2、3、4四章。第二部分介绍数据的预处理，作为第5章。第三部分介绍回归分析方法，其中第6章包括经典的一元与多元线性回归和逐步回归分析，第7、8、9三章涉及多项式回归，样条回归，非线性回归和岭回归等近代方法。第四部分介绍方差分析，包括一元和多元，固定和随机模型，是本书的第10章。第五部分介绍试验设计，包括正交设计和序贯设计，作为第11章。

本书的写作得到浙江大学教材建设领导小组，国家自然科学基金重点项目和国家教委博士点专项基金的资助，在出版过程中又得到浙江大学出版社等各位的全力支持，谨志由衷谢忱。希望本书对广大的研究工作者在探索自然界的规律中能有所帮助，同时也竭诚欢迎读者提出宝贵意见。

作者　于浙江大学 1995 年 1 月

目　录

1 基本知识 …… 1
1.1 基本概念和常用术语 …… 1
1.2 样本的定量表示 …… 2
1.2.1 重复观测值的代表 …… 2
1.2.2 重复观测值的变异程度 …… 2
1.3 误差 …… 3
1.3.1 误差的来源和性质 …… 3
1.3.2 误差的表示方法 …… 3
1.3.3 误差的传播 …… 4
1.4 数学期望和中心矩 …… 5
1.4.1 数学期望 …… 5
1.4.2 中心矩 …… 6
1.5 方差 …… 7
2 随机变量及其分布 …… 8
2.1 概率密度和分布函数 …… 8
2.1.1 概率密度分布的表达 …… 8
2.1.2 概率分布的数字特征 …… 10
2.2 离散型随机变量的概率分布 …… 12
2.2.1 离散均匀分布 …… 13
2.2.2 二项分布 …… 13
2.2.3 多项分布 …… 14
2.2.4 负二项分布 …… 15
2.2.5 几何分布 …… 15
2.2.6 超几何分布 …… 17
2.2.7 扩充几何分布 …… 18
2.2.8 泊桑分布 …… 19
2.2.9 几种离散分布模型之间的关系 …… 20
2.3 连续型随机变量的概率分布 …… 21
2.3.1 连续均匀分布 …… 21
2.3.2 指数分布 …… 22
2.3.3 Gamma 分布 …… 24
2.3.4 Beta 分布 …… 27
2.3.5 Weibull 分布 …… 27
2.3.6 Chi 平方分布 …… 28

2.3.7 几种分布之间的关系…… 29
2.4 正态分布 …… 30
2.4.1 正态分布…… 30
2.4.2 标准正态分布…… 31
2.4.3 正态分布和其他分布的关系…… 32
2.4.4 对数正态分布…… 33
2.5 随机变量的函数 …… 34
2.5.1 联合概率分布…… 34
2.5.2 随机变量的函数…… 35
3 统计量的分布…… 37
3.1 样本和统计量 …… 37
3.2 极限定理和几种重要的分布 …… 37
3.2.1 大数定律和中心极限定理…… 37
3.2.2 Chi 平方分布,t 分布和 F 分布…… 39
3.3 几种重要的统计量分布 …… 42
3.3.1 样本均值的分布…… 42
3.3.2 样本均值差的分布…… 43
3.3.3 样本方差的分布…… 44
3.3.4 样本方差比的分布…… 44
3.3.5 统计量分布定理…… 44
4 统计推断…… 48
4.1 参数估计 …… 48
4.1.1 点估计与区间估计…… 48
4.1.2 总体均值的区间估计…… 50
4.1.3 总体均值差的区间估计…… 53
4.1.4 成对观察值差的均值估计…… 54
4.1.5 总体方差的区间估计…… 56
4.1.6 总体方差比的区间估计…… 58
4.1.7 容许区间与容许限…… 59
4.2 假设检验 …… 61
4.2.1 检验的原则…… 61
4.2.2 假设检验的两类错误…… 63
4.2.3 工作特性曲线…… 66
4.2.4 检验能力…… 71
4.2.5 假设检验与区间估计的比较…… 72
4.3 非参数检验 …… 79
4.3.1 两总体的秩和检验…… 79
4.3.2 成对观测值的符号检验…… 83
5 数据的预处理…… 85
5.1 数据预处理的目的 …… 85

5.2 定常观测数据的粗差剔除 …… 85
5.2.1 三倍标准差判别法 …… 86
5.2.2 小概率事件判别法 …… 88
5.2.3 端值判别法 …… 90
5.2.4 t 检验准则剔除异常数据法 …… 90
5.2.5 非参数方法 …… 92
5.3 序列观测数据的噪音平滑 …… 93
5.3.1 线性滑动平滑法 …… 94
5.3.2 二维线性滑动平滑法 …… 97
5.3.3 非线性滑动平滑法 …… 99
5.3.4 二维非线性滑动平滑法 …… 101
5.4 数字滤波方法 …… 103
5.4.1 差分滤波方法 …… 104
5.4.2 剩余值滤波方法 …… 104
6 线性回归分析 …… 106
6.1 回归分析和主要解决的问题 …… 106
6.2 一元线性回归分析 …… 106
6.2.1 一元线性回归方程 …… 106
6.2.2 相关系数和显著性检验 …… 108
6.2.3 线性回归方程的误差 …… 110
6.2.4 线性回归的失拟 …… 112
6.2.5 一元非线性问题的线性化处理 …… 114
6.2.6 回归系数的统计性质 …… 115
6.3 多元线性回归分析 …… 118
6.3.1 多元线性回归方程 …… 118
6.3.2 回归方程的有效性 …… 123
6.3.3 各个自变量的作用 …… 124
6.3.4 偏回归平方和 …… 125
6.3.5 非线性问题的多元化处理 …… 127
6.3.6 回归系数的统计性质 …… 127
6.4 逐步回归分析 …… 128
6.4.1 多元线性回归方程的优选 …… 128
6.4.2 逐步回归分析的数学模型 …… 129
6.4.3 求逆的紧凑格式 …… 131
6.4.4 逐步回归分析的具体步骤 …… 133
7 多项式回归分析 …… 137
7.1 多项式与多元线性回归 …… 137
7.2 正交多项式回归分析 …… 138
7.2.1 基于正交多项式的回归分析 …… 138
7.2.2 一种常用的正交多项式 …… 139

7.2.3 正交多项式回归的特点 …… 141
7.3 各种正交多项式 …… 142
7.3.1 另一种正交多项式 …… 142
7.3.2 不等间距的正交多项式 …… 143
7.3.3 多元正交多项式 …… 143
7.4 分段多项式回归分析 …… 144
7.4.1 分段回归和多项式样条 …… 144
7.4.2 用于回归分析的幂样条 …… 146
7.4.3 样条回归分析 …… 147
8 非线性回归方法 …… 150
8.1 非线性回归问题概述 …… 150
8.1.1 非线性回归方程的系数求解 …… 150
8.1.2 求解非线性代数方程组的方法 …… 152
8.1.3 寻优方法的分类 …… 155
8.1.4 平行法寻优 …… 155
8.1.5 组合法寻优 …… 157
8.2 不用求导的方法 …… 157
8.2.1 不用求导的一维寻优 …… 157
8.2.2 随机走动法 …… 162
8.2.3 网格搜索法 …… 163
8.2.4 单形和复形方法 …… 163
8.2.5 模式搜索法 …… 167
8.2.6 坐标轮换法 …… 168
8.2.7 旋转坐标法 …… 170
8.2.8 共轭方向法 …… 170
8.3 需要一阶导数的方法 …… 175
8.3.1 利用一阶导数值的一维搜索法 …… 175
8.3.2 一阶梯度法 …… 178
8.3.3 线性化迭代校正法 …… 179
8.3.4 阻尼最小二乘法 …… 181
8.3.5 共轭梯度法 …… 183
8.4 需要二阶导数的方法 …… 186
8.4.1 利用二阶导数的一维搜索法 …… 186
8.4.2 二阶梯度法 …… 187
8.4.3 变度量法 …… 188
8.4.4 准二阶梯度法 …… 190
8.5 小结 …… 191
8.5.1 几种非线性方法的比较 …… 191
8.5.2 非线性回归方程的有效性 …… 192
8.5.3 回归问题的分类处理 …… 192

9 岭回归分析 …… 193
9.1 最小二乘估计的性能分析 …… 193
9.2 岭回归估计方法 …… 197
9.3 岭回归估计的性质 …… 199
9.4 岭迹的计算与分析 …… 204
9.5 K 值的选择 …… 207
9.6 广义岭回归 …… 210
10 方差分析 …… 215
10.1 一元方差分析 …… 215
10.1.1 问题的提出 …… 215
10.1.2 数学模型和统计分析方法 …… 216
10.1.3 不等重复数的试验 …… 222
10.1.4 不同水平下试验方差均一性的 Bartlett 检验 …… 223
10.1.5 平均值是否相等的分别检验 …… 226
10.1.6 检验平均值相等的多重比较方法 …… 229
10.1.7 处理平均值与控制标准的比较 …… 232
10.1.8 非均一性试验的实施条件 …… 234
10.1.9 随机效应模型 …… 235
10.2 多元方差分析 …… 238
10.2.1 多因素试验问题的分类 …… 238
10.2.2 二因素多水平交叉组合全面试验的通用情况 …… 240
10.2.3 无交互作用的二因素试验 …… 248
10.2.4 随机模型及多层分组的二因素试验 …… 253
10.2.5 三因素多水平交叉组合全面试验 …… 258
11 试验设计 …… 269
11.1 正交试验设计 …… 269
11.1.1 正交设计方法及其特点 …… 269
11.1.2 正交试验的方差分析 …… 272
11.1.3 正交试验的极差分析 …… 274
11.1.4 交互作用与表头设计 …… 281
11.1.5 有重复的水平数不同的正交试验设计 …… 285
11.2 序贯试验设计 …… 289
11.2.1 几种分布参数的序贯检验 …… 290
11.2.2 模型选择的序贯试验设计 …… 294
附表 …… 298
参考文献 …… 348

1 基础知识

1.1 基本概念和常用术语

为了定量地研究感兴趣的对象，需要采集能反映对象性质的各种观察和测量所得数据，并对数据进行分析处理，取得某种规律性的结论. 数据采集的原则步骤见图 1.1.

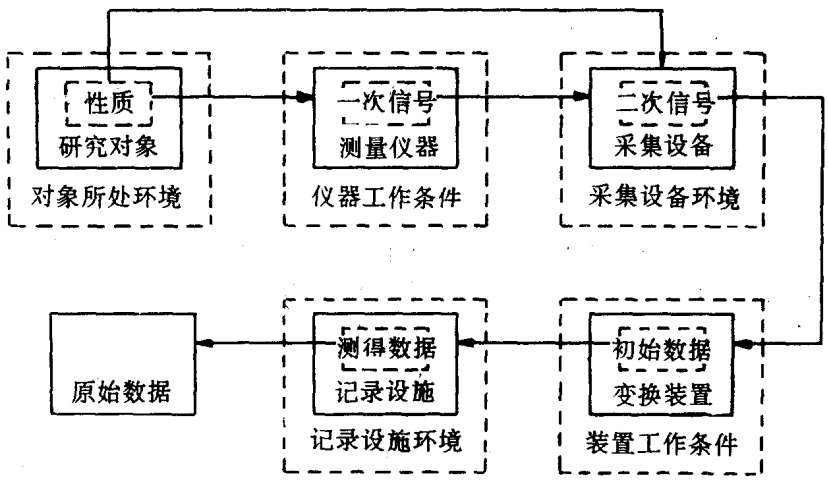

图 1.1 数据采集的原则步骤

一切对象的各种性质可分为两种类型：定常型和变化型. 定常型性质是指在一定的时间和空间范围内原则上应该有某个固定的量值，变化型性质则随着时间和空间位置的变动而有不同的量值. 其中随时间而变化的称为时变性质，随空间位置而变化的称为分布变量. 本书将主要讨论定常型性质观测数据的分析与处理.

实际上对象所处环境、测量仪器的工作条件、以及信号采集、变换、直到数据记录等各个环节的环境条件，经常都会有种种随机的变动. 因此，对同一对象、同一定常型性质的多次观测，将得到一组数值有一定变动的结果，它们是该性质的真值和种种随机因素所造成影响的叠加值. 数据分析的基本任务之一，就是要科学地表达和使用含有随机因素影响的数据，也就是对数据的统计分析.

性质的量值可以是某种分级指标值，也可以是数值. 通过仪器得到的量值称为测量值，不通过仪器得到的称为观察值，统称为观测值，取得观测值的过程称为观测. 由若干次观测得到的、以原始记载形式收集的一组观测值称为原始数据，简称数据. 一组有目标的观测称为试验，对观测结果有影响的因素全属随机性的试验则称为随机试验. 无数次观测值的集合称为总体，有限次观测值的集合称为样本，样本中所包括观测值个数称为该样本的容量，由样本中各个体量计算出来的整体量称为统计量. 随机试验时所有可能出现的结果所存在的位置称为样本空间. 在总体或样本中每一个观测值称为个体或元素，样本空间的一个特定子集称为事件，只包

含一个元素的是简单事件，由多个简单事件联合起来表示的为复合事件. 变量的取值受随机因素影响的称为随机变量.

1.2 样本的定量表示

1.2.1 重复观测值的代表

试验所得的任何一个观测值 x_i 包含了所观测性质的真值 μ 和各种随机因素造成的变差 v_i：

$$x_i = \mu + v_i \tag{1.1}$$

真值可由无数次重复观测所得的总体 $x_i, i = 1, 2, \cdots, \infty$ 计算得到的平均值表示，即 x_i 的数学期望 $E(x)$：

$$\mu = \lim_{n\to\infty}\left[\frac{\sum_{i=1}^{n} x_i}{n}\right] = E(x) \tag{1.2}$$

实际条件下只能取有限的 n 次观测值，构成样本 $x_i, i = 1, \cdots, n$. 这样得到样本算术平均值（简称平均值）：

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.3}$$

平均值是最常见的一种统计量，用以整体地代表该组观测值，也是对真值的一种估计.

描述样本代表值的另两个常见统计量是：中值和众值.

将 $x_i, i = 1\cdots, n$ 按数值大小顺序排列，则中值为：

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{，当 } n \text{ 为奇数；} \\ \dfrac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & \text{，当 } n \text{ 为偶数} \end{cases} \tag{1.4}$$

如果并非所有 x_i 都不相等，则 $x_i, i = 1, \cdots, n$ 中出现次数最多的值 $\hat{x}$ 称为众值. 一个样本的众值可能不存在，也可能不止一个.

1.2.2 重复观测值的变异程度

除了用一个数值作为代表定量地表示样本整体性质的一个方面外，为了说明随机因素影响的大小，还需有表示观测值变差程度的度量. 常用的统计量是样本均方差，或简称样本方差：

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)} \tag{1.5}$$

它的平方根取正值称为样本均方根，或称样本标准差.

另外两个表示变异程度的常见统计量是：极差

$$R = x_n - x_1 \tag{1.6}$$

其中 $x_i, i = 1, \cdots, n$ 按数值从小到大顺序排列，以及平均偏差

$$md = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \tag{1.7}$$

1.3 误　差

1.3.1 误差的来源和分类

(1) 误差的来源

如图 1.1 所示从对象性质到数据记录所经各步骤，由于工作环境和条件中各种因素的影响，使记录数据中观测值 x_i，与对象性质的真值 μ 存在某种偏差 ε_i，称之为误差：

$$\varepsilon_i = x_i - \mu \tag{1.8}$$

产生误差的来源至少有四类：

① 环境. 由于工作环境或条件中不可控制的变化造成观测数据的波动.

② 设备. 由于仪器、设备、或装置性能的局限性造成数据的偏差.

③ 处理. 由于信号采集和变换等手段的不同而造成的偏差.

④ 方法. 由于数据记录的方法和过程造成的偏差.

(2) 误差的分类

按性质区分，说差可分为三类：

① 系统误差. 由于观测过程中某种固定因素而造成同一方向的误差，在同一组试验中偏差的大小和符号相同，也可称为恒定误差.

② 随机误差. 由于数据采集过程中各种随机因素造成的偏差，在同一组试验中偏差值忽大忽小，忽正忽负. 但随着观测值个数的增多，误差的算术平均值趋于零.

③ 过失误差. 由于试验进行时数据的观测和采集中不正确的操作引起的错误.

用统计方法分析处理观测数据，主要是针对存在随机误差，即前节所谓变差 v_i 的情况. 同时也可据以判断是否存在个别的过失误差. 至于系统误差的克服，一般不能依靠统计分析方法.

若把讨论限于随机误差，则 $\varepsilon_i = v_i$，因此对总体有：

$$v_i = x_i - \mu$$

相应地对样本可以写为：

$$e_i = x_i - \bar{x} \tag{1.9}$$

而且当样本容量 $n \to \infty$ 时，$\bar{x} \to \mu$，所以 $e_i \to v_i = \varepsilon_i$.

随机误差和系统误差对观测结果造成影响的不同性质见图 1.2 所示，并说明于表 1.1 中.

1.3.2 误差的表示方法

根据不同场合的需要，误差可采用不同的表示方法，常见的有以下几种：

① 绝对误差　$e_i = x_i - \bar{x}$　(1.10)

② 相对误差　$c_i = e_i/\bar{x}$　(1.11)

③ 分贝误差　$b_i = 20 \cdot \lg(1 + c_i)$　(1.12)

④ 引用误差　$a_i = e_i/\text{scale}$　(1.13)

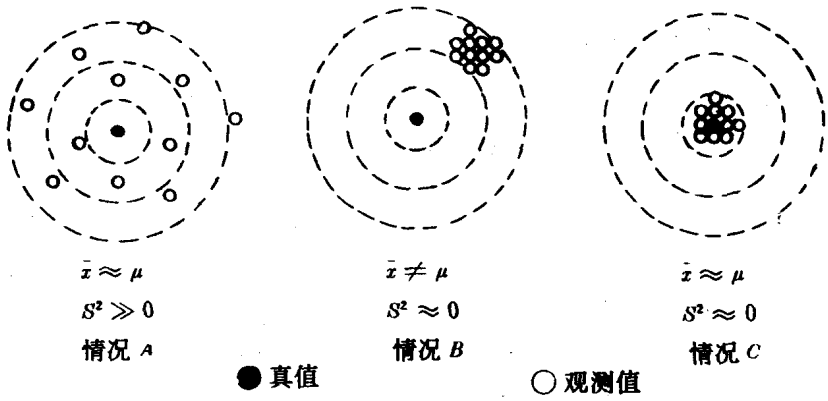

图 1.2　三类不同的观测结果

表 1.1　三类不同观测结果所说明的问题

	情况 A $\bar{x} \approx \mu$ $S^2 \gg 0$	情况 B $\bar{x} \neq \mu$ $S^2 \approx 0$	情况 C $\bar{x} \approx \mu$ $S^2 \approx 0$
系统误差	小	大	小
正确度	高	低	高
随机误差	大	小	小
精确度	低	高	高
准确度	差	差	好

1.3.3　误差的传播

用有误差的观测数据进行运算后就会产生误差传播的现象. 设由 $u_1, u_2, \cdots, u_m$ 各观测值按函数关系 f 计算 F 值:

$$F = f(u_1, u_2, \cdots, u_m) \tag{1.14}$$

令 $\Delta u_1, \Delta u_2, \cdots, \Delta u_m$ 为各观测值的误差,并引起 F 产生误差 ΔF,

$$F + \Delta F = f(u_1 + \Delta u_1, u_2 + \Delta u_2, \cdots, u_m + \Delta u_m)$$

对上式右端作 Taylor 级数展开,有:

$$f(u_1 + \Delta u_1, u_2 + \Delta u_2, \cdots, u_m + \Delta u_m)$$
$$= f(u_1, u_2, \cdots, u_m) + \Delta u_1 \frac{\partial f}{\partial u_1} + \Delta u_2 \frac{\partial f}{\partial u_2} + \cdots + \Delta u_m \frac{\partial f}{\partial u_m} + \frac{1}{2}(\Delta u_1)^2 \frac{\partial^2 f}{\partial u_1^2} + \cdots$$

若只取一阶导数作近似计算,则有:

$$\Delta F \approx \frac{\partial f}{\partial u_1}\Delta u_1 + \frac{\partial f}{\partial u_2}\Delta u_2 + \cdots + \frac{\partial f}{\partial u_m}\Delta u_m \tag{1.15}$$

这就是有误差的数据经运算后误差传播的基本估计算式.

1.4 数学期望和中心矩

1.4.1 数学期望

若某个变量的值是实数，且其取值又因样本空间中每个不同元素而异，则称之为随机变量. 包含有限个元素，或无限个整数(亦称可列无穷个)元素的样本空间称为离散样本空间，在该空间定义的随机变量称为离散随机变量. 包含某个实数区间中的任一元素，因而系由无限个元素构成的样本空间称为连续样本空间，在该空间定义的随机变量称为连续随机变量.

若随机变量 x 取值为 x_i 的概率 P 是 x_i 的函数，并表示为：

$$P(x = x_i) = f(x_i) \tag{1.16}$$

则将函数 $f(x)$ 称为随机变量 x 的概率函数，或概率分布.

其中，对离散随机变量有：

$$\begin{aligned} &f(x_i) \geqslant 0 \\ &\sum_i f(x_i) = 1.0 \end{aligned} \tag{1.17}$$

对连续随机变量则有：

$$\begin{aligned} &f(x_i) \geqslant 0 \\ &\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1.0 \end{aligned} \tag{1.18}$$

随机变量 x 的数学期望，也称 x 的期望值，为：

$$\begin{aligned} &E(x) = \sum_i x_i f(x_i) \\ &E(x) = \int_{-\infty}^{+\infty} f(x)\mathrm{d}x \end{aligned} \tag{1.19}$$

而 x 的函数 $g(x)$ 的期望值为：

$$\begin{aligned} &E[g(x)] = \sum_i g(x_i) f(x_i) \\ &E[g(x)] = \int_{-\infty}^{+\infty} g(x) f(x)\mathrm{d}x \end{aligned} \tag{1.20}$$

以上系样本空间为一维的情况. 若研究对象需要不只一个随机变量来说明，则需要用多维样本空间来表达. 例如采用 x 和 y 两个随机变量，且 x 取 x_i，y 取 y_i 值的概率为：

$$P(x = x_i, y = y_j) = f(x_i, y_j) \tag{1.21}$$

习惯上将函数 $f(x,y)$ 称为随机变量 x 和 y 的联合概率分布.

其中，对 x 和 y 为离散随机变量，有：

$$\begin{aligned} &f(x_i, y_i) \geqslant 0 \\ &\sum_i \sum_j f(x_i, y_j) = 1 \end{aligned} \tag{1.22}$$

对 x 和 y 为连续随机变量，则有：

$$f(x_i,y_j)\geqslant 0$$
$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\mathrm{d}x\mathrm{d}y = 1 \tag{1.23}$$

如果在样本空间内对离散和连续的情况分别定义：

$$f_1(x) = \sum_y f(x,y)$$
$$f_2(y) = \sum_x f(x,y) \tag{1.24}$$

和

$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y$$
$$f_2(y) = \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}x \tag{1.25}$$

并称之为 x 和 y 的边缘分布 $f_1(x)$ 和 $f_2(y)$，则当且仅当对所有(x,y) 有：

$$f(x,y) = f_1(x)\cdot f_2(y)$$

时，x 和 y 为统计独立的随机变量，反之则为统计相关.

具有联合概率分布为 $f(x,y)$ 的随机变量 x 和 y 的函数 $h(x,y)$ 的期望值 $E[h(x,y)]$ 在离散和连续的情况下分别为：

$$\sum_i\sum_j h(x_i,y_j)f(x_i,y_j)$$

和

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} h(x,y)f(x,y)\mathrm{d}x\mathrm{d}y \tag{1.26}$$

对于期望值的运算规则，可简述如下：

若 a 和 b 为常数，则：

$$E(ax\pm b) = a\cdot E(x)\pm b \tag{1.27}$$

若 $g(x)$ 和 $h(x)$ 为 x 的两个函数，则：

$$E[g(x)\pm h(x)] = E[g(x)]\pm E[h(x)] \tag{1.28}$$

对于两个随机变量 x 和 y 的情况，有：

$$E[g(x,y)\pm h(x,y)] = E[g(x,y)]\pm E[h(x,y)] \tag{1.29}$$

若 x 和 y 互相独立，则有：

$$E(x\cdot y) = E(x)\cdot E(y) \tag{1.30}$$

1.4.2 中心矩

当 $g(x) = x^k$ 时，所得到的期望值 $E(x^k)$ 称为随机变量 x 的 k 阶原点矩，并记为 μ_k^0，对离散和连续两种情况，分别是：

$$\mu_k^0 = E(x^k) = \sum_i x_i^k\cdot f(x_i)$$
$$\mu_k^0 = E(x^k) = \int_{-\infty}^{+\infty} x^k f(x)\mathrm{d}x \tag{1.31}$$

请注意，当 $k=0$ 时，

$$\mu_0^0 = E(x^0) = E(1) = 1$$

而当 $k = 1$ 时，

$$\mu_1^0 = E(x^1) = \mu \tag{1.32}$$

当 $g(x) = (x - \mu)^k$ 时，所得到的期望值 $E[(x - \mu)^k]$ 称为随机变量 x 的 k 阶中心矩，记为 μ_k，对离散和连续两种情况，分别是：

$$\begin{aligned} \mu_k &= E[(x - \mu)^k] = \sum_i (x_i - \mu)^k f(x_i) \\ \mu_k &= E[(x - \mu)^k] = \int_{-\infty}^{+\infty} (x - \mu)^k f(x)\mathrm{d}x \end{aligned} \tag{1.33}$$

1.5 方　差

在对观测数据用统计方法进行分析和处理时，二阶中心矩 μ_2 具有特别重要的意义，因为它是观测值在平均值附近变异程度的一种度量. 它也被称为随机变量 x 的方差 σ^2，通常可用以下几种符号表示：

$$D(x) = \mathrm{Var}(x) = \sigma_x^2 = \sigma^2 = \mu_2$$

它的计算公式是：

$$\sigma^2 = E[(x - \mu)^2] = E(x^2) - \mu^2 \tag{1.34}$$

方差的平方根 σ 即为标准差.

为了讨论方差的性质，取 $g(x)$ 为随机变量 x 的函数，令 $g = g(x)$，而它的平均值和方差分别为 μ_g 和 σ_g^2，根据方差定义有：

$$\sigma_g^2 = E[(g - \mu_g)^2] \tag{1.35}$$

由上式可导出方差的基本性质，若 $g = x + c$，c 为常数，则有：

$$\sigma_{x+c}^2 = \sigma_x^2 = \sigma^2 \tag{1.36}$$

若 $g = a \cdot x$，a 为常数，则有：

$$\sigma_{a \cdot x}^2 = a^2 \cdot \sigma_x^2 = a^2 \cdot \sigma^2 \tag{1.37}$$

设 x 和 y 两个随机变量具有联合分布 $f(x,y)$，$g = a \cdot x + b \cdot y$，$a$ 和 b 都是常数，则有：

$$\sigma_{ax+by}^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\sigma_{xy} \tag{1.38}$$

其中 σ_{xy} 称为随机变量 x 和 y 的协方差，它的定义是：

$$\sigma_{xy} = E(x \cdot y) - \mu_x \cdot \mu_y \tag{1.39}$$

协方差还有另一种表示形式是：

$$\mathrm{cov}(x,y) = \mathrm{cov}(y,x) = E[(x - \mu_x)(y - \mu_y)] \tag{1.40}$$

如果 x 和 y 为独立随机变量，则：

$$\sigma_{xy} = E(x \cdot y) - \mu_x \cdot \mu_y = 0 \tag{1.41}$$

所以在此情况下有：

$$\sigma_{a \cdot x + b \cdot y}^2 = a^2\sigma_x^2 + b^2\sigma_y^2 \tag{1.42}$$

2 随机变量及其分布

本章将介绍有关随机变量和概率分布的基本概念，重点讨论各种常见的有实用价值的分布函数，并以实际示例加以具体说明.

2.1 概率密度和分布函数

这一节要介绍的是用统计方法分析和处理观测数据的理论基础.

2.1.1 概率密度分布的表达

从一个具体问题出发. 设某工厂产品中成分 A 的含量受不可控的随机因素影响而有波动，也就是前面所称的变异、或变差. 工厂每两小时测量一次 A 的百分含量，记为 x. 现摘取一个时间段的数据作为示例列于表 2.1.

表 2.1 某产品中成分 A 的百分含量数据，x(%)

日期	成分 A 的百分含量 x_i (%)											
1	1.40	1.28	1.36	1.38	1.44	1.40	1.34	1.54	1.44	1.46	1.80	1.44
2	1.46	1.50	1.58	1.54	1.50	1.48	1.52	1.58	1.52	1.46	1.42	1.58
3	1.70	1.62	1.58	1.62	1.76	1.68	1.68	1.66	1.62	1.72	1.60	1.62
4	1.46	1.38	1.42	1.38	1.60	1.44	1.46	1.28	1.34	1.38	1.24	1.36
5	1.58	1.38	1.34	1.28	1.18	1.08	1.36	1.50	1.46	1.28	1.18	1.28

本例中 x 为连续变量，但在样本中只能包括有限个测量值. 为了研究数据的变异(波动)特性，可根据某种人为确定的分界值，把随机变量 x 的整个取值范围，分有限个区段，称为分段或分级. 每个级段的取值范围，称为级宽或段宽 $\Delta x_j = (x_b - x_a)_j$，右边项表示第 j 个级段内上界值 $(x_b)_j$ 和下界值 $(x_a)_j$ 的差. 每个级段中数据值出现的次数 ΔF_j，称为级频数. 将级频数被样本中数据总个数 $S_F = \sum \Delta F_j$ 相除得到相对频数，或称频率 $\Delta f_j = \Delta F_j / S_F$. 将表 2.1 的数据从 $x = 1.00$ 到 $x = 1.90$ 取级宽为 0.10 分为 9 级后，得到观测数据分布在各级中的频数和频率列于表 2.2.