南昌大学食品科学国家重点实验室项目资助
南昌大学生物质转化教育部工程研究中心资助
高等教育研究生教材

DATA
ANALYSIS IN FOOD
SCIENC RESEARCH

数据分析
在食品科学研究中的应用

张锦胜　彭　红　编著

中国轻工业出版社　全国百佳图书出版单位

南昌大学食品科学国家重点实验室项目资助
南昌大学生物质转化教育部工程研究中心资助
高等教育研究生教材

数据分析
在食品科学研究中的应用

张锦胜　彭红　编著

中国轻工业出版社

图书在版编目（CIP）数据

数据分析在食品科学研究中的应用/张锦胜，彭红编著．—北京：中国轻工业出版社，2013．5

ISBN 978－7－5019－9177－8

Ⅰ．①数… Ⅱ．①张… ②彭… Ⅲ．①统计数据—统计分析（数学）—应用—食品科学—研究 Ⅳ．①TS201

中国版本图书馆 CIP 数据核字（2013）第 040740 号

简　介

本书结合食品科学研究的特点，针对食品科学研究与开发中常见的数据分析与处理问题，介绍了数据分析的理论、技巧和方法，同时，介绍了如何使用 SPSS16.0 软件进行数据统计分析处理的方法。为了帮助读者掌握解决实际问题的方法，书中列出大量的应用实例，结合科学研究与生产实践中的实际问题，介绍了传统计算、查表分析的原理和方法，以及如何用软件进行同样的数据处理与分析。读者通过阅读教材，可以理解数据统计分析的原理，掌握利用软件分析的方法。

本书为研究生上课讲稿编写而成，参考了很多相关书籍和网上资料。本书可作为大专院校食品科学与工程、农副产品加工与保藏、水产养殖、农业工程、生物工程、化学工程等专业本科高年级学生和研究生的教材，也可作为公司企业的工程技术人员进行质量控制、新产品开发的参考资料。

责任编辑：王　淳　　责任终审：简延荣　　封面设计：锋尚设计
版式设计：宋振全　　责任校对：晋　洁　　责任监印：胡　兵　张　可

出版发行：中国轻工业出版社（北京东长安街 6 号，邮编：100740）
印　　刷：河北省高碑店市德裕顺印刷有限责任公司
经　　销：各地新华书店
版　　次：2013 年 5 月第 1 版第 1 次印刷
开　　本：787×1092　1/16　印张：21.5
字　　数：490 千字
书　　号：ISBN 978－7－5019－9177－8　　定价：42.00 元
邮购电话：010－65241695　传真：65128352
发行电话：010－85119835　85119793　传真：85113293
网　　址：http://www.chlip.com.cn
Email：club@chlip.com.cn
如发现图书残缺请直接与我社邮购联系调换
121099K1X101ZBW

前　言

数据分析在各个行业都非常重要，在食品科学研究与开发的过程中也不例外。在过去的教学中，教学与实际应用还存在一些距离。很多同学在学习概率论与数理统计过程中，对于复杂的分析计算感到头痛、厌倦；很多这方面的教材也与实际应用脱节，还停留在介绍手工查表计算阶段。如今的数据分析已经离不开计算机应用，但软件只能按人的要求进行分析，并不能判断分析是否正确；现在也有很多软件分析方面的教材，但更多的是介绍操作与实例，这使得很多同学，只知道跟着实例的样子用软件分析，而不知道为什么这样分析，甚至对软件分析过程中的报错信息也不理解。只有掌握了数据统计分析的原理，才能理解发挥软件的强大功能。有感于目前将这两者结合起来的书太少，因此，尝试编写一本将理论介绍与实际软件分析相结合的书，重点针对食品科学研究与开发中常见的数据分析与处理问题，介绍数据分析理论的同时，介绍如何使用 SPSS16.0 软件进行数据统计分析。为了帮助读者学习和掌握解决实际问题的方法，书中列出大量的实例，结合科学研究中的实际问题，介绍了用传统计算查表分析的方法，同时，介绍如何用软件解决同样的数据处理与分析问题，使读者知其然更知其所以然，这也是本书的一个特色。否则，读者很容易迷失在软件花样繁多的功能中不知所措。读者通过阅读教材，将能够很好地理解掌握数据统计分析的基本原理与计算方法，同时掌握如何应用软件来进行分析。

全书共分十章，第 1 章数据分析基础；第 2 章 SPSS 基本操作；第 3 章描述统计分析；第 4 章统计推断；第 5 章方差分析；第 6 章协方差分析；第 7 章相关分析；第 8 章回归分析；第 9 章非参数检验；第 10 章统计图表的绘制。通过认真学习，读者将能很好地掌握数据统计分析的基本原理和方法，同时能应用计算机进行相关的数据分析。本书借鉴了一些经典统计分析的教材与实例，注意由浅入深、难点分散和严谨论述，力求能通俗易懂地介绍相关的理论和方法。相对于传统的概率与数理统计，本书更注重实际应用。从实际分析应用的需要出发，对于必要的概念和理论予以介绍，同时，书中配有大量图片，手把手教读者应用软件进行相关的分析，因而具有很好的直观性及可读性。本书可作为大专院校食品科学与工程、食品营养与安全、农副产品加工与保藏、水产养殖、生物工程、化学工程、农业工程等专业本科高年级学生和研究生的教材，也可作为公司、企业的工程技术人员进行质量控制、新产品开发的参考资料。

本书获得南昌大学教材出版资金的资助。本书的编写也得到了张继鉴、莫春兰、张华、赵霞、莫彪、卢久洵、王彦勋、李映彤、林建龙、刘靖岩、马彪等的帮助和支持，在此一并感谢。由于时间紧迫和学识有限，书中难免出现不妥之处，敬请读者批评指正。

编者

目　录

第 1 章　数据分析基础

【学习重点】

1. 回顾数据的分类及统计分析的一些基本概念
2. 了解区分集中趋势、离散程度、图形形状的测量值和综合测量值
3. 理解均值、中位数、极差、方差等概念的含义和计算方法
4. 区分样本方差和总体方差、样本标准差和总体标准差含义及计算方法
5. 理解将标准差应用于经验法则和切比雪夫定理的意义
6. 理解正态分布的理论及意义
7. 理解抽样研究与抽样误差的理论及意义

1.1　数据分析的定义和内容

1.1.1　什么是数据分析（data analysis）

西方管理学中流行的一句名言——“我们相信上帝，除此之外我们只相信数据”。其含义是作为一个管理者，任何的决策都应该基于数据分析的结果做出，而不是凭空想象。事实上，不光是管理科学领域，任何一个以实验科学为基础的研究领域，其结论的做出，都应该是基于实验数据分析的结果。

数据也称观测值，是以数值的形式给出的关于自然、社会现象和科学试验的定量或定性的观察记录的结果，是科学研究最重要的基础。研究数据就是对数据进行采集、分类、录入、储存、统计分析、统计检验等一系列活动的统称。所谓数据分析就是指从数据中寻找研究对象的内在规律，提取有意义的信息来帮助评估、预测与决策，从而得出合理的结论。数据分析的过程就是有目的地收集数据、分析数据，使之成为有用信息的过程。在科学研究和生产实践中我们都应该遵循这一点，就是所谓“言之有据”，我们所做出的结论都应该是基于数据分析的结果。

数据分析也是质量管理体系的支持过程。在产品的整个寿命周期，包括从市场调研到售后服务和最终处置的各个过程都需要适当运用数据分析过程，以提升有效性。例如一个企业的领导人要通过市场调查，分析所得数据以判定市场动向，从而制定合适的生产及销售计划；而企业的研发人员也需要数据分析来判断消费群体的口感偏好来设计生产新产品等。因此，数据分析有极广泛的应用范围。

食品科学研究中也经常要对实验结果进行评估。食品企业在新产品的研究开发过程中要对产品的功效、市场等做出评估预测，这些都离不开数据分析。而对于不同类型的数据，我们需要采用不同的分析方法，而且不同的分析方法一般都有其适用范围，因此

需要多加注意。例如，为了评估比较某两种降血脂功能性食品的效果，将这两种降血脂功能性食品 A 和 B 做动物实验，实验方案为四组小白鼠，每组 12 只，1 组为对照组，2 组为用 A，3 组为用 B，4 组为 A、B 同时使用，一个月后测定血脂含量，分析食用降血脂功能性食品的使用结果，这里就要应用数据统计分析的知识和方法——抽样、方差分析、单因素方差分析、因素的主效应和因素的交互效应分析、协方差分析等。又如：某公司想了解顾客对某食品产品的喜爱程度，做了问卷调查，请他对五种不同品牌的某食品按喜爱程度排序，分析顾客对品牌的认同度，这就需要应用非参数估计、多个相关样本检验分析等。实验的数据结果，最后都要通过正确的分析方法才能得出正确的结论。因此，数据分析也可以理解为对数据进行收集、归类、分析和解释，从而发现其中的规律而获取可靠结果的过程。我们这里强调“过程”，是认为数据分析贯穿在整个过程中。作为研究开发人员，我们知道实验设计的好坏关系到数据结果的可靠性。广义的数据分析应该是包括实验设计，而不是仅仅对最终实验结果的数据分析，应充分认识到这一点。由此我们知道，要进行数据分析，要有一定的统计学知识。

数据分析最有用的工具之一就是统计学——一门对数据进行收集、整理、描述、分析、解释的学科。在此，我们所说的数据分析就是指应用统计学的原理与方法，在食品科学研究领域中进行数据收集、整理和分析。统计学的基本原理和方法，包括统计描述（定量资料和分类资料的描述性指标以及常用统计图表）、常见的理论分布及其应用（正态分布、二项分布与 Poisson 分布）、总体参数的估计（分总体均数、总体率和总体平均数）、假设检验（t 检验、z 检验、卡方检验、方差分析、秩和检验等）、回归与相关、多元线性回归与 Logistic 回归等。本书的重点不是数理统计理论的推导，而是更注重统计方法的应用，稍后我们将简单回顾一下有关的统计学基础知识。

1.1.2　什么是 SPSS

SPSS 是软件英文名称缩写，原意为 Statistical Package for the Social Sciences，即“社会科学统计软件包”。随着 SPSS 产品升级，SPSS 公司将英文全称更改为 Statistical Product and Service Solutions，即“统计产品与服务解决方案”，英文缩写仍然不变，但服务领域扩大了。SPSS 是目前世界上公认的三大数据分析软件之一（SAS、SPSS 和 SYSTAT）。

SAS 及 SYSTAT 功能非常强大，一般为专业的统计人员所使用，需要学习专门的统计命令、语言，相比之下，SPSS 的操作最为简单，所有参数设置均采用人机对话的窗口式操作，虽然某些功能不如 SAS、SYSTAT，但能满足绝大多数一般用户的要求。

SPSS 是世界上最早的统计分析软件之一，由美国斯坦福大学的三位研究生于 20 世纪 60 年代末研制。发展至今已有十几个版本，是一种集成化的计算机数据处理应用软件。全球有几十万用户，广泛分布于通讯、医疗、证券、市场研究、科研等多个领域和行业，是世界上应用最广泛的专业统计软件之一，在国际学术界享有很高的声誉。在国际学术交流中，凡是用 SPSS 软件完成的计算和统计分析，可以不必说明算法，由此可见其影响力。

SPSS 也是世界上最早采用图形菜单界面的统计软件，与其他统计软件相比，它最大的特点是对话框式的图形操作界面。它将几乎所有的功能都以大家非常熟悉的统一、

规范的 Windows 的对话窗口方式展现出来，输出结果也美观漂亮。对于大多数计算机用户来说，Windows 操作技能都很熟悉，因此，只要粗通统计分析原理，就很容易上手，可以方便地使用该软件完成特定的科研分析工作。

SPSS 还采用类似 EXCEL 表格的方式输入与管理数据，数据接口较为通用，能方便的从 EXCEL、ORIGIN 等其他数据库拷贝粘贴数据。其统计过程包括了常用的、较为成熟的统计过程，完全可以满足非统计专业人士的工作需要。输出结果可以转存为 HTML 格式和文本格式，可以直接拷贝粘贴到 Word 文档。对于熟悉老版本编程运行方式的用户，SPSS 还特别设计了语法生成窗口，用户只需在菜单中选好各个选项，然后按“粘贴”按钮就可以自动生成标准的 SPSS 程序，极大地方便了中、高级用户。

SPSS 是一个组合式软件包，它集数据整理、分析功能于一身。SPSS 的基本功能包括数据管理、统计分析、图表分析、输出管理等。具体内容包括描述性统计、均值比较、一般线性模型、相关分析、回归分析、对数线性模型、聚类分析、主成分分析、时间序列分析、非参数检验等多个大类，每个类中还有多个专项统计方法，比如回归分析中又分线性回归分析、曲线估计、Logistic 回归、Probit 回归、加权估计、两阶段最小二乘法、非线性回归等多个统计过程，而且每个过程中又允许用户选择不同的方法及参数。SPSS 也有专门的绘图系统，可以根据数据绘制各种图形。本书的重点之一就是介绍如何用 SPSS16.0 进行食品科学研究中的数据分析。

1.2　实验数据分析的工作步骤

食品科学的研究工作全过程可以分为以下四个步骤：

（1）设计（design）　设计就是针对研究、开发的目的，在着手进行研究开发工作之前做一个周密的设计。是在广泛查阅相关文献资料和研究进展，全面了解现状的基础上，对将要进行的研究开发工作所做的全面设想。其主要内容包括：明确本研究或开发的目的，确定实验对象、观察单位、样本含量和抽样方法，拟定研究方案、预期分析指标、误差控制措施等。设计是整个研究工作中最关键的一环，也是指导以后工作的依据，如果在设计阶段考虑不周，将直接影响到最后结果的准确性和可靠性。自然科学研究的过程就是大胆假设，小心求证！食品科学也不例外。

（2）收集数据（collection）　遵循统计学原理，采取科学必要的措施，及时、准确、完整地获取可靠的原始试验数据。典型的收集数据就是记录实验参数，如反应的时间、温度、压力等。

（3）整理数据（sorting data）　收集来的数据在没有进行整理之前都称为原始数据，原始数据很可能杂乱无章。整理数据就是通过科学的分组、归纳，使之系统化、条理化，为数据分析做准备，以便于进一步计算和分析。

（4）分析数据（analysis of data）　计算各个相关指标，反映数据的综合特征，探索数据的内在联系或规律。数据的统计分析包括统计描述（descriptive statistics）和统计推断（inferential statistics）。前者是用统计指标与统计图（表）等方法对样本数据（实验数据）的数量特征及其分布规律进行描述；后者是指如何抽样，以及如何用样本信息

推断估计总体特征。进行数据分析时，需根据研究目的、设计类型和数据资料的类型，选择恰当的描述性指标和统计推断方法。

数据分析的四个步骤是一个完整的紧密相连、不可分割的有机整体，任何一步的缺陷，都将影响整个研究结果。由于篇幅的原因，本书在介绍基本原理的基础上重点介绍如何用 SPSS16.0 软件进行食品科学研究与开发过程中的相关数据分析，侧重于第四步，而第一步的试验设计由于篇幅的原因，不是本书的重点。

得到数据后，我们该从哪儿开始分析呢？一般来说，典型的数据分析可能包含以下三个步骤：

（1）探索性数据分析　刚取得数据时，可能杂乱无章，尤其在数据较多时，很难看出其中的规律。通过作图、制表，用各种形式的方程拟合，计算某些特征统计量，如均数、方差、标准差、峰度和偏度等手段探索数据分步的特征，探索规律的可能形式，即往什么方向和用何种方式去寻找和揭示隐含在数据中的规律性。

（2）模型选定分析　在探索性分析的基础上提出一类或几类可能的模型，然后通过进一步的分析从中挑选一定的模型。

（3）推断分析　通常使用数理统计方法对所定模型或估计的可靠程度和精确程度做出推断。

1.3　数据统计分析的基本概念

统计学中的基本概念在很多经典的教材中都有介绍，对于概念公式的推导计算不是本书的重点，我们从实用的角度，对于必要的最基本的概念和理论进行回顾之后就进入到实际的应用。只要跟着本书的思路，就能很快地掌握相关的知识和技巧。

1.3.1　总体（population）与样本（sample）

任何研究都必须首先确定观察单位，亦称个体（individual）。观察单位是统计研究中最基本的单位，可以是一个人、一个小白鼠、一个实验样品、一个采样点、一次化学反应的结果等。

总体是根据研究目的确定的同质观察单位的全体，或者说，把要研究对象的全体叫做总体，而把构成总体的每个单元称为个体。从总体中抽取的一部分个体称为总体的一个样本，样本包含的个体的数目称为样本的容量。一个总体可以用一个随机变量来代表它。例如：某奶粉厂欲考察某天某台自动包装机工作是否正常，则这一天该包装机所包装的全部 1kg 装的奶粉（比如 10000 袋）就是一个总体，每一袋奶粉就是一个个体。该包装机生产的奶粉的质量可视为一个随机变量，每袋奶粉的质量就是该随机变量的一个取值，应该在 1000g 左右波动，可能多几克，也可能少几克，数值是随机的。总体又分为有限总体和无限总体。有限总体是指在某特定的时间与空间范围内，同质研究对象的所有观察单位的某变量值的个数为有限个，如上例所示，因为某天某台机器包装的奶粉数量是有限的。无限总体是抽象的，无时间和空间的限制，观察单位数是无限的，如研究某减肥食品对肥胖患者的防治效果，其总体是肥胖患者，该总体应包括已使用和设想

使用该食品的所有肥胖患者的防治效果，没有时间和空间范围的限制，因而观察单位数无限，该总体为无限总体。而如果加上一个时间段、年龄段，以及所处地区限制等，则无限总体就变为有限总体。

在实际工作中，所要研究的总体无论是有限的还是无限的，通常都是采用抽样研究。比如：某奶粉厂欲考察某天某台自动包装机工作是否正常，不可能将这一天该包装机所包装的全部奶粉进行测量，这样工作量太大，尤其如果是破坏性实验，总不能将所有奶粉都打开测量才知道该批次奶粉的各营养要素的含量吧！因此，抽样研究被广泛应用，通过抽样样本来估计总体的特征。样本是按照随机化原则，从总体中抽取的有代表性的部分观察单位的变量值的集合。如从上例的某天某台机器包装的奶粉这个有限总体中，按随机化原则抽取 10 袋称重，它们的质量数据即为一个 10 个个体的样本。从总体中抽取样本的过程称为抽样，抽样研究的目的是用样本信息推断总体特征。抽样方法有多种，由于不是本书重点，有关抽样研究的方法，请参阅相关书籍。

统计学就是总体与样本的桥梁，能帮助人们设计与实施从总体中如何科学地抽取样本，使样本中的观察单位数恰当，信息丰富，代表性好；能帮助人们挖掘样本中的信息，推断总体的规律性。我们要根据样本值对总体的某些特征进行估计和推断。因此需要对如何选取样本值提出一些要求。

1.3.2　同质（homogeneity）与变异（variation）

在上面介绍总体概念时，提到“同质观察单位”一词。所谓“同质”，是指被研究指标的影响因素完全相同。科学实验过程中常把同质理解为对研究对象指标影响较大的、可以控制的主要因素尽可能相同。例如我们在做某功能因子的动物实验的时候，研究含某种功能因子的食品（饲料）对增重的影响时，就尽可能要求实验动物的年龄、性别等对体重影响较大的、易控制的因素要尽量相同，而不易控制的遗传等影响因素可以忽略。之所以如此，是因为实验动物的年龄、性别对实验结果的影响可能会非常大，比如喂食相同重量的食品（饲料），幼年的动物由于处于生长发育期新陈代谢旺盛增重效果要明显高于成年的动物。这种情况下，饲料的增重效果就将明显受到实验动物年龄的影响。因此，在实验之前就应该考虑到将此影响因素消除。这就是实验对象应该尽可能是同质的，即可以控制的主要影响因素应该尽可能相同。

同质基础上的个体差异称为变异。例如相同性别、相同年龄、同一窝的实验小白鼠体重也不会完全相同，会在一个范围内波动；同年龄、同性别的消费者对同一种食品的喜好程度也是不同的。事实上，客观世界变异无处不在，正是这些变异才构成了色彩斑斓的世界。哪里有变异，哪里就需要统计。如果研究的同质群体中所有个体一模一样，比如，所有的实验小白鼠体重完全一样，则只需观察任意一个个体即可，无须进行统计研究，但这实际上是不可能的。因此，单个实验个体之间的差异并不能说明问题，只有这种差异具有统计意义才能说明问题。

1.3.3　数据及其分类

研究对象的总体确定之后，研究者就可以对每个观察单位的某种特征进行测量或观

察。特征也可称为变量，如动物实验的“体重”、“性别”、“血型”、“增重”、“血脂含量”等。变量的测定值或观察值称为变量值或观察值，亦简称为数据。

按变量值是定量还是定性的，其所对应的数据类型可分为不同类型。

1.3.3.1　定类数据

定类数据仅用来对数据进行分类，也称为分类数据。定类数据是指所分类别或属性之间没有程度和顺序的差别分类数据。定类数据又可分为：

（1）二项分类　如实验动物小鼠的性别（雄、雌），减肥效果（有效和无效）等。在数据处理的过程中，可以用数字表示。如用0表示雄性，1表示雌性。

（2）多项分类　如客户类型（男青年、女青年、男中年、女中年）等。

1.3.3.2　定序数据

定序数据比定类数据高一个级别。除了具有定类数据的特征外，定序数据可以将研究对象进行排序，即定序数据有程度的差别。如某种食品的减肥效果按很有效、一般、显效、无效等分类；感官评估时，对某食品的评价按很喜欢、喜欢、一般、不喜欢等分类。由于定类数据和定序数据通常来自不精确的测量，因而定类数据和定序数据统称为非测量型数据，有时也称为定性数据。

当然，变量类型不是一成不变的，根据研究目的的需要，各类变量之间可以进行转化。例如实验对象的血红蛋白量（g/L）原属数值变量，若按血红蛋白正常与偏低分为两类时，即变成二项分类数据；若按重度贫血、中度贫血、轻度贫血、正常、血红蛋白增高分为五个等级时，则成为有等级的分类数据。相反，有时亦可将分类资料数量化，如感官品评时，对口感品评结果：非常喜欢、喜欢、一般、不喜欢、很不喜欢以0、1、2、3、4编码表示，则可按数值变量来进行分析。但是，能够按数值变量来分析并不代表数据间的差值是一定的。比如上述感官品评的五种结果以0、1、2、3、4编码表示喜好程度，并不能认为五个选项间的差异是相等的。

1.3.3.3　定距数据

定距数据所定义的连续数据间的距离是有意义的，这一点与定序数据有所不同，读者可以体会一下。定距数据多为数值型数据，表现为数值大小，可经测量取得数值，多有度量衡单位，且连续数据间的差所代表的距离是相等的，也就是说定距数据具有相等的距离。如动物实验中小白鼠的体重（g）、食品添加剂合成实验中的温度、压力数值、反应物浓度等数值数据都是定距数据。

1.3.4　随机事件（random event）与概率（probability）

在自然界中存在着各种各样的事件。在一定条件下必然发生的事件称为必然事件。在一定条件下必然不发生的事件称为不可能事件。除了上述必然事件和不可能事件之外，还有一类事件，它们在一定条件下可能发生，也可能不发生，称为随机现象。科学研究的现象，大多数是随机现象，对随机现象进行实验或观察称为随机试验。随机试验的各种可能结果的集合就是随机事件，亦称偶然事件。例如用相同减肥食品给一批肥胖患者使用，结果可能为好转、无效、有效等多种结果，对于一个随机的肥胖患者，食用这种减肥食品一段时间后发生的减肥效果能达到什么程度是不确定的，可能发生的每一

种结果都是一个随机事件。

对于随机事件来说，在一次随机试验中，某个随机事件可能发生也可能不发生，但在一定数量的重复试验后，该随机事件的发生情况是有规律可循的。概率是描述随机事件发生的可能性大小的数值，常用 P 表示。例如，食品厂检验某批产品的质量时，如果一件一件产品孤立地检验，“产品是次品”这一事件是否发生就带有偶然性，假如把一批产品的检验结果综合在一起，就会发现“产品是次品”这一事件发生的机会和可能性的大小是确定的。

以下以统计学上的经典例子来说明：投掷一枚均匀的硬币，随机事件 A 表示“正面向上”，用 n 表示投掷次数，m 表示随机事件 A 发生的次数，f 表示随机事件 A 发生的频率（$f = m/n$），$0 \leqslant m \leqslant n$，$0 \leqslant f \leqslant 1$。频率 = 频数 / 总次数，频数就是一个变量在各个变量值上取值的个数。频数是事件的实际发生次数，总次数是总事件数。例如：将一枚硬币抛五次，假设正面在上三次，反面在上两次，那么总次数是 5，事件“正面在上”的频数是 3，它的频率 = 3/5；事件“反面在上”发生的频数是 2，它的频率 = 2/5。用不同的投掷次数 n 做随机试验，结果如下：m/n = 8/10 = 0.8、7/20 = 0.35、…、249/500 = 0.498、501/1000 = 0.501、1000/2000 = 0.5，由此看出当投掷次数 n 足够大时，$f = m/n \to 0.5$，称 $P(A) = 0.5$，或简写为：$P = 0.5$。当 n 足够大时，可以用频率 f 估计 P。概率和频率都是一个界于 0 和 1 之间的分数，它们之间的关系，事实上就是古典概率与试验概率之间的关系。当被研究对象是总体的部分单位时，频率只是试验概率。因此可以说，概率是频率的期望值或理论值；频率是概率的估计值或试验值。在试验次数或抽样次数非常大时，频率逼近概率。

随机事件概率的大小在 0 与 1 之间，即 $0 < P < 1$，常用小数或百分数表示。P 越接近 1，表示某事件发生的可能性越大；P 越接近 0，表示某事件发生的可能性越小。$P = 1$ 表示事件必然发生，$P = 0$ 表示事件不可能发生，它们是确定性的，不是随机事件，但可以把它们看成随机事件的特例。

若随机事件 A 的概率 $P(A) \leqslant \alpha$，习惯上，当 $\alpha = 0.05$ 时，就称 A 为小概率事件，“小概率”的标准 α 是人为规定的，其统计学意义是小概率事件在一次随机试验中不可能发生。例如，根据统计，我国乘火车发生交通事故而受伤的发生概率为 1/1000 万，但还是有很多人乘火车，这是因为乘火车“被撞受伤”事件是小概率事件，乘车 1000 万次中只有一次发生事故，所以出行人认为自己乘火车出行这“一次试验”中不会发生“被撞”事件。而如果某天交通管理部门发布信息称乘火车出行发生交通事故的概率为 1/100，那么估计很多人不敢乘火车了。因此，所谓的小概率 α 值在不同情况下是不同的。对于可能引起严重后果的事件，小概率的 α 值一般规定的很小，如药物的不良反应率，一旦发生将产生严重后果，一般规定 $\alpha = 0.001$，甚至更小。

1.3.5　统计指标与统计推断

统计指标提供了从有用的数据中获取有意义的信息的一种直接、有效的方法。描述统计学（descriptive statistics）是来描绘（describe）或总结（summarize）观察量的基本情况的统计总称。描述统计学研究如何取得反映客观现象的数据，并通过图表形式对所

收集的数据进行加工处理和显示，进而通过综合概括与分析得出反映客观现象的规律性数量特征。涉及一系列表述数据的定量指标和方法，包括频数分布，直方图，集中趋势指标——平均数（mean）、中位数（median）、众数（mode）的功能等，以及离散趋势指标（极差、方差、标准差）等。这些指标描述了数据的特征，使人们能够通过上述指标对整体数据有个大体的认识。

对数据的收集、整理和描述通常称为描述统计学，统计推断（statistical inference）是在样本数据的基础上，推断总体未知特征的过程。样本的这些指标称为统计量，整体的这些指标称为参数。根据样本的统计量来推断总体的参数，就是参数估计。预测统计学（predictive statistics）是在历史数据的基础上预测未来值，这是统计方法论的第三个重要组成部分。

1.3.6　均值（average）

均值表示的是某变量所有取值的集中趋势或平均水平。它是一种集中趋势指标，说明一组数据的分布集中在什么地方。均值对于每一组数据或一个数据集来说都是唯一的，对定距数据和定比数据来说都是有意义的。但均值常常会受到异常值（outliers）的影响。异常值就是与数据集中的其余观测值有很大差异的少数观测值。一般来说，异常值是一组测定值中与平均值的偏差超过两倍标准差的测定值。与平均值的偏差超过三倍标准差的测定值，称为高度异常的异常值。在处理数据时，应剔除高度异常的异常值。异常值是否剔除，视具体情况而定。在统计检验时，一般指定检出异常值的显著性水平 $\alpha=0.05$，为检出水平；指定检出高度异常的异常值的显著性水平 $\alpha=0.01$ 为舍弃水平，又称剔除水平（reject level）。

均值的计算公式：

（1）总体平均数　若一组数据 $X(X_1,X_2,\cdots,X_n)$ 代表一个大小为 N 的有限总体，总体一般以大写字母表示，则其总体均数 μ 为

$$\mu=\frac{\sum_{i=1}^{N}X_i}{N}$$

（2）样本平均数　若一组数据 $x(x_1,x_2,\cdots,x_n)$ 代表一个大小为 n 的有限样本，样本一般以小写字母表示，则其样本平均数为

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

样本来自总体，样本均值是总体均值的无偏估计。所谓无偏估计是参数的样本估计值的期望值等于参数的真实值。样本的统计描述量可以反映总体数据的特征，但由于抽样等原因，使得样本数据不一定能够准确反映总体的特征值，可能与真实值之间存在一定的差异。但是这个误差是可以通过统计方法估计出来的。无偏估计就是系统误差为零的估计。简单的理解就是，样本均值是总体均值的好的估计。

1.3.7　众数（mode）

众数是指一组数据中，出现次数最多的那个变量值。在统计分布上具有明显集中趋

势点的数值，代表数据的一般水平（众数可以不存在或多于一个）。众数在描述数据集中趋势方面有一定的意义。如果有两个数据发生的次数最多且相同，就有两个众数，叫双众数，也称为双峰的。如果该组数据中虽然没有双众数，但是有两个数据发生的次数明显比其他数据多，也称此两个数为广义的双众数。众数如果不是唯一的，我们就称数据为多峰的（multimodal）。例如：1、2、3、3、4 的众数是 3。如果有两个或两个以上个数出现次数都是最多的，那么这几个数都是这组数据的众数，例如：1、2、2、3、3、4 的众数是 2 和 3。如果所有数据出现的次数都一样，那么这组数据没有众数，例如：1、2、3、4、5 没有众数。在高斯分布中，众数位于峰值。

用众数代表一组数据，可靠性较差。不过，众数不受极端数据的影响，并且求法简便。在一组数据中，如果个别数据有很大的变动，选择中位数表示这组数据的“集中趋势”就比较适合。

众数在数值或被观察者没有明显次序（常发生于非数值性资料）时特别有用，因为可能无法良好定义算术平均数和中位数。例如：{鸡、鸭、鱼、鱼、鸡、鱼}的众数是鱼。

1.3.8　全距（range）

全距也称为极差，是数据的最大值和最小值之间的绝对差。反映总体标志值之间的范围，它是一个比较粗略的测量值，只是描述数据两端的差值。相同样本容量的情况下的两组数据，全距大的一组数据更为分散。在一定程度上，极差能够反映极端值的信息，因为它是由极端值计算出来的。极差的一个重要用途是进行数量控制，绘制控制图。缺点是受极端值的影响，因为是由极端值计算出来的，在描述数据离散程度方面比较粗略，不能反映总体数据的差异程度，应用受到限制。

例如：一组数据 1、2、2、3、3、4，全距为 4 - 1 = 3。

1.3.9　分位数

1.3.9.1　四分位数（quartiles）

四分位数是将一组数据由小至大排序后，用 3 个点将全部数据分为四等份，与 3 个点上相对应的变量称为四分位数，分别记为：Q_1（第一四分位数）、Q_2、Q_3。其中 Q_3 和 Q_1 之间的距离的一半又称为四分位差，记为 Q。四分位差越小，说明中间数据越集中；四分位数越大，则意味中间部分的数据越分散。

1.3.9.2　十分位数（deciles）

十分位数是将一组数据由小至大排序后，用 9 个点将全部数据分为十等份，与 9 个点上相对应的变量称为十分位数，分别记为：D_1（第一个十分位数）、D_2、$D_3 \cdots D_9$，表示 10% 的数据落在 D_1 下、20% 的数据落在 D_2 下、90% 的数据落在 D_9 下。

1.3.9.3　百分位数（percentiles）

百分位数是将一组数据由小至大排序后，用 99 个点将全部数据分为一百等份，与 99 个点上相对应的变量称为百分位数，分别记为：P_1（第一个百分位数）、P_2、$P_3 \cdots P_{99}$，表示 1% 的数据落在 P_1 下、2% 的数据落在 P_2 下、99% 的数据落在 P_{99} 下。分位数都是用于对数

据的整体观测，能够很快知道数据的大致分布。

1.3.10　方差（variance）和标准差（standard deviation）

均值表示的是某变量所有取值的集中趋势或平均水平。由均值可以知道一组数据分布的平均的集中趋势，但对于数据偏离均值的情况无法了解。比如：两组数据（12.5，12.3，12.4，12.8）和（8.5，16.2，10.4，14.5）均值差不多，但是，两组数据的离散程度显然是有差别的。此时，如何衡量数据的离散程度？方差就是用来度量随机变量和其数学期望（即均值）之间的偏离程度的一个统计量。标准差可以指出数据在平均数的附近什么地方聚集。样本中各数据与样本平均数的差的平方和的平均数叫做样本方差，表示一组数据分布的离散程度的平均值。样本方差的算术平方根叫做样本标准差，表示一组数据关于平均数的平均离散程度。方差和标准差越大，说明变量之间的差异越大，距离平均数这个中心的离散趋势越大。

方差和标准差的计算公式：

（1）总体方差

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

（2）总体标准差

$$\sigma = \sqrt{\sigma^2}$$

（3）样本方差

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

（4）样本标准差

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

其中，μ 为总体平均数；$\bar{x}$ 为样本平均数；N 为总体的个数；n 为样本的个数。

计算方差的时候我们需要知道真实的均值，而计算样本方差时我们不知道这个真实值，而是用样本的平均值代替总体的真实均值，这样的替换本身带来了误差，因此除以（$n-1$）而不是除以 n 来修正这个误差。样本方差是用来估计总体方差的，样本方差除以（$n-1$）的这个式子才是总体方差的无偏估计。具体的证明可参见相关的统计教材。

1.3.10.1　标准差的含义

标准差是方差的算术平方根，是用上述计算公式计算出来的一组数据的某个特征值。由于其在统计学上有非常重要的意义，有必要详细地进一步阐述。

统计学上有一个经验法则：当一组数据是正态分布时，标准差可以用来确定数据在给定的离差范围内分布的大致百分比。如果该组数据是正态分布的，则该组数据中大约有 68% 的数据分布在距离均值正负一个标准差的范围内，95% 的数据分布在距离均值正负两个标准差的范围内，而几乎 100% 的数据都在均值正负三个标准差的范围内。

举例来说，有一种水果的成熟果实质量是平均为 80g，标准差 5g 的正态分布，那么根据经验法则，就可以知道大约 68% 的果实，质量在 75～85g，即（$80 \pm 1\sigma$）；有 95%