

Practical
Biostatistics

实用
生物统计学

顾志峰 叶乃好 石耀华 主编



科学出版社

实用生物统计学

顾志峰 叶乃好 石耀华 主 编

科学出版社

北 京

内 容 简 介

本书主要针对生命科学研究领域中的常见问题,以生物统计学的基础理论知识、试验资料的收集整理、资料的统计分析和试验设计为主线来安排主要内容。每章内容先以具体的科学问题引出相应的生物统计学问题,然后对其基本理论进行介绍,结合相应的统计软件(包括 SPSS、DPS、Minitab、Excel 等)对实际问题进行具体分析,同时采用截图法对具体使用的统计方法进行直观形象的介绍,最后结合专业知识对分析结果进行科学阐释。本书的设计和编排使每位读者能够很清晰地理解每个问题的解题要领,科学合理地选用相应的统计软件来分析和解决实际问题,同时能够很直观、形象地掌握每种软件的使用方法和技巧。

本书可供高校的本科、专科生作为生物统计的教材使用,也可供科研工作者、教师和研究生作为工具书使用。

图书在版编目(CIP)数据

实用生物统计学/顾志峰,叶乃好,石耀华主编. —北京:科学出版社, 2012

ISBN 978-7-03-034762-6

I. ①实… II. ①顾…②叶…③石… III. ①生物统计—统计分析—软件包 IV. ①Q-332

中国版本图书馆 CIP 数据核字 (2012) 第 123277 号

责任编辑:马俊 孙青/责任校对:林青梅
责任印制:钱玉芬/封面设计:耕者设计工作室

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>

骏杰印刷厂印刷

科学出版社发行 各地新华书店经销

*

2012年8月第一版 开本:787×1092 1/16
2012年8月第一次印刷 印张:19 3/4

字数:462 000

定价:58.00元

(如有印装质量问题,我社负责调换)

《实用生物统计学》编委会成员

(按姓名汉语拼音排序)

葛长宇	山东大学
顾志峰	海南大学
石耀华	海南大学
王爱民	海南大学
王鸿霞	中国科学院海洋研究所
徐 东	中国水产科学研究院黄海水产研究所
叶乃好	中国水产科学研究院黄海水产研究所
张晓雯	中国水产科学研究院黄海水产研究所
周海龙	海南大学

前 言

生物统计学是应用统计学的原理和方法来分析、理解、推导和探究蕴含在生命科学领域中的各种纷繁复杂现象背后的生命科学规律和本质的一门科学，是生物学和统计学的交叉学科，在生态学、生物化学、农学等传统的生物学分支学科和生物信息学等新兴生物学科中广泛应用。虽然生物统计学是高校生物类专业一门极其重要的基础课程，出版的相关教材数量很多，但是，这些教材主要都注重基本理论和公式的推导，容易使学生将其当做一门数学类课程来学习，在公式推导中迷失了应用性这一学习生物统计学的根本目的，以至于许多学生完成教材的学习后，面临具体的实验设计和数据处理时仍然茫然不知所措。为此，我们编著了本书。

尽管本书仍然是由实验资料的整理、试验设计、显著性检验、相关与回归和协方差分析等章节组成，也对其中的基本概念、基本公式进行了阐述，但是，本书一以贯之的一个根本思想是实用性，即学生通过丰富的实例直观地了解如何选择合适的分析手段、如何采用常规的软件一步一步地完成实验数据的分析，这是本书与已有的生物统计教材的显著不同之处。为了强化和检验教学效果，本书在每一章的最后附上了适当数量的复习思考题。

本书还可以作为生物学科技工作者开展试验设计、数据处理和分析等的参考书，特别是书中介绍了多种常用的生物统计软件的使用方法，在计算机普及的今天，生物统计学的应用更为实用和便捷，把科技工作者从繁杂的数据处理和分析中解放出来。

本书的编著、出版得到了国家自然科学基金项目（41076112，41176153）、国家高技术研究发展计划（“863”计划）项目（2012AA052103、2012AA100814）、教育部科学技术研究重点项目（211143）、海南大学2011年度自编教材资助项目（Hdzbjc1103）、海南大学“211”工程重点学科建设项目、“水产养殖学”国家级特色专业（TS10477）、热带生物资源教育部重点实验室等资助。科学出版社对本书的编著提供了大力支持和大量有益的建议。在此我们一并致以诚挚的感谢。

由于编著者水平有限，书中疏漏与不妥之处在所难免，敬请读者提出批评指正，以便修改完善。

编 者

2011年3月于海口

目 录

前言

第 1 章 绪论	1
1.1 生物统计学的定义	1
1.2 生物统计学的重要理论基础	2
1.3 生物统计学的作用	4
1.4 生物统计学的特点及其学习方法	5
复习思考题.....	5
第 2 章 数据的整理与分析	7
2.1 常用统计学术语	7
2.2 试验资料的整理.....	10
2.3 试验数据的描述性统计.....	22
2.4 试验数据中异常值的分析.....	38
复习思考题	46
第 3 章 概率分布与抽样分布	48
3.1 概率基础知识.....	48
3.2 概率分布.....	49
3.3 二项分布.....	51
3.4 泊松分布.....	61
3.5 正态分布.....	64
3.6 正态分布的检验.....	76
3.7 统计量的分布.....	82
复习思考题	92
第 4 章 统计推断	93
4.1 假设检验的原理与方法.....	93
4.2 单样本平均数的 u 检验	94
4.3 单样本平均数的 t 检验	97
4.4 成组数据平均数比较的 t 检验	100
4.5 成对数据平均数比较的 t 检验	105
4.6 方差的假设检验	109
4.7 样本频率的假设检验	112
4.8 参数的区间估计与点估计	118
复习思考题.....	125

第 5 章 χ^2 检验	128
5.1 适合性检验	128
5.2 独立性检验	129
复习思考题	139
第 6 章 方差分析	141
6.1 方差分析的相关术语	142
6.2 方差分析的原理	142
6.3 均值间的两两比较	143
6.4 单因素方差分析	147
6.5 二因素方差分析	153
6.6 统分组(嵌套)资料的方差分析	165
6.7 方差分析的基本假定与数据转换	168
复习思考题	172
第 7 章 非参数检验	176
7.1 符号检验	176
7.2 符号秩检验	178
7.3 多个样本比较的非参数检验	182
7.4 Jonkheere-Terpstra 检验	184
7.5 Friedman 检验	185
7.6 Kendall 协同系数检验	187
7.7 二元响应的 Cochran 检验	189
7.8 秩相关	190
复习思考题	193
第 8 章 一元回归与相关分析	195
8.1 回归的概念	195
8.2 一元直线回归与相关	196
8.3 一元曲线回归与相关	201
8.4 一元多项式回归曲线	221
复习思考题	223
第 9 章 多元统计分析	226
9.1 多元方差分析	226
9.2 多元线性回归与多元相关分析	230
9.3 逐步回归	239
9.4 通径分析	242
9.5 聚类分析	243
9.6 判别分析	251
9.7 主成分分析	259
9.8 因子分析	261

复习思考题·····	265
第 10 章 协方差分析 ·····	268
10.1 单因素协方差分析的步骤·····	268
复习思考题·····	272
第 11 章 试验设计 ·····	274
11.1 实验设计的内容与作用·····	275
11.2 试验设计的基本原理·····	275
11.3 试验设计的主要方法·····	276
复习思考题·····	303
参考文献 ·····	305

第 1 章 绪 论

内容提要

生物统计学是指应用统计学的原理和方法分析、理解、推导和探究蕴含在生命科学领域中的各种纷繁复杂现象背后的生命科学规律、本质的一门科学。生物统计学的研究内容主要包括试验设计、统计分析方法及其基本原理。常用的试验设计方法主要有：对比设计、随机区组设计、拉丁方设计、正交设计等；统计分析方法则主要包括：描述性统计、显著性检验、相关与回归、协方差分析、多元统计分析。生物统计学的主要目的是对实验资料进行科学的整理、分析数据、判断试验结果的可靠性、确定事物之间的相互关系、提供试验设计的基本原则，为学习相关学科和科学研究奠定基础。生物统计学是一门较年轻的学科，随着计算机统计软件（Minitab、DPS、SPSS、SAS 等）的开发而得到快速的发展，已广泛应用于各领域，并和其他学科交叉形成了诸如生物信息学等新的学科。

1.1 生物统计学的定义

1.1.1 统计学

统计学（statistics）是一门通过搜集、整理、分析、解释统计资料，揭示其内在客观数量规律性的科学。由于它具有准确、客观等特点，已广泛应用于自然科学、社会科学、人文科学以及政府情报决策中。

Statistics 最早起源于拉丁语 *statisticum collegium*；到了 16 世纪，意大利语 *stato* 表示“国家”和“情况”的含义，随后传播到德国、法国、荷兰等国，德国西尔姆斯特大学教授康令（H. Conring）在大学开设了一门 *staatenkunele* 课程，原意是对各国状况的比较，引起了许多学者的关注。随后，德国哥廷根大学教授阿亨瓦尔（G. Achenwall, 1719~1772 年）在康令思想的基础上，把关于人口、财政、军队等事项的学问称为“国势学”，并在 1749 年出版的《近代欧洲各国国势学概论》中首创了“*statistik*”这个词汇，即“统计学”。1787 年，英国学者齐默尔曼（E. A. W. Zimmerman）据语音把“*statistik*”译成英语“*statistic*”。19 世纪，统计学传到日本，日本的学者将其译成“统计学”。

1903 年，我国学者钮永建等翻译出版了日本学者横山雅男所著的《统计讲义录》，统计一词传入我国。1907 年，彭祖植编写的《统计学》是我国最早的一本统计学书籍；1913 年，顾澄翻译了英国学者尤尔的《统计学之理论》，是英美数理统计学传入中国之始；之后又有一些英美统计著作被翻译成中文，Fisher 的理论和方法也很快传入中国。20 世纪 30 年代，《生物统计与田间试验》作为农学系的必修课，1935 年王绶编著出版的《实用生物统计法》是我国出版最早生物统计专著之一。1942 年范福仁出版了

《田间试验技术》等。这些翻译和编著的统计学书籍对推动我国农业生物统计和田间试验方法的应用产生了很大影响。现今，统计学已被延伸到生物学、医学、心理学等领域，相应地又形成了一系列新的学科，如生物统计学、医学统计学、心理统计学等。

1.1.2 生物统计学

在生产实践活动中，人们往往会遇到类似下面的一些问题。例如，转基因动植物的生长速度是否比非转基因动植物快？如何进行判断？吸烟会不会导致得肺癌的概率增大？如何判断一种新疫苗是否有效？如何抽检一部分人来估计某种疾病的流行程度？某种细胞培养方法、抗癌药物疗效，或饲料配方、育苗效果等是否有明显改进等。

这类问题的共同特点，就是人们只能得到他所关心的事情的不完全信息，或者是单个实验的结果有某种不确定性。如何透过纷繁复杂的现象抓住这类生命现象的本质？这需要通过设计相应的试验，开展试验研究，借助于生物统计学理论和方法，透过外界环境条件或其他偶然因素所掩盖的表面现象，从而揭示其生命现象的内在规律。这就是生物统计学研究的内容，由此可知生物统计学（biostatistics）就是运用统计学的原理和方法来分析、理解、推导和探究蕴含在生命科学领域中的各种纷繁复杂现象背后的生命科学规律与本质的学科。

16~17世纪中叶随着数理统计学的发展，18~19世纪正态曲线、最小二乘法等重要理论广泛应用于生物学。1889年，高尔顿（Galton）发表第一篇生物统计论文《自然界的遗传》；1901年，高尔顿和他的学生皮尔逊（Pearson）创办了 *Biometrika*（《生物统计学报》）杂志，首次明确“biometry”（生物统计学）一词。因此，后来大家推崇高尔顿为生物统计学的创始人。近年来，随着相关学科的发展，生物统计学已广泛应用于农学、医学、分子生物学、细胞生物学、生物信息学、生物制药技术、资源保护与利用以及生态学等领域，取得了长足的进步。

1.2 生物统计学的重要理论基础

1.2.1 总体的理论分布

生物统计学中常见的理论分布有正态分布、 t 分布、二项分布、泊松（Poisson）分布、 χ^2 分布和 F 分布，其中前面两种理论分布主要应用于连续型变量概率分布的资料，而另外三种主要应用于离散型变量的概率分布资料。此外， t 分布、二项分布、Poisson分布的极限为正态分布，在一定条件下，可以转化为正态分布进行处理。

正态分布理论最早由棣莫弗（De Moiver）于1733年发现，后来高斯（K. F. Gauss）在进行天文观察和研究土地测量误差理论时又一次独立发现了正态分布（又称常态分布）的理论方程，提出“误差分布曲线”，后人为了纪念他，将正态分布也称为 Gauss 分布。戈塞特（W. S. Gosset）在生产实践中对样本标准差进行了大量研究，于1908年以

“Student”（学生）为笔名在该年的 *Biometrika* 上发表了《平均数的概率误差》，创立了小样本检验代替大样本检验的理论，即 t 分布，也称为学生式分布。1900年，戈塞特的老师——皮尔逊独立发现了 χ^2 分布，并提出了著名的卡方检验法。1923年，费希尔（R. A. Fisher）提出了 F 分布和 F 检验。1838年，法国数学家泊松（S. D. Poisson）提出了泊松分布。在1713年出版的由瑞士数学家雅各布·伯努利（Jacob Bernoulli）撰写的专著《猜度术》一书中，用组合公式证明了帕斯卡曾经提出的 n 为正数时的二项式定理，即二项分布。此外，棣莫弗在《抽签的计量》一书中首次定义了独立事件的乘法定理，给出了二项分布公式。

1.2.2 显著性检验

生命科学研究中，往往会获得一系列的变异资料，差异产生的原因可能是由于处理间（如不同试剂、不同药物、不同品种、不同浓度间）有本质差异，也可能是一些偶然因素导致的，要找出其中的真实原因，必须进行显著性检验（significance test）。Neyman 和 S. Pearson Fisher 提出了显著性检验理论，为假设检验理论的发展奠定了坚实的基础：根据“小概率事件实际不可能性原理”来接受或否定零假设，从而对最后结果进行推断。常用的显著性检验方法有 t 检验、 χ^2 检验、 F 检验等。

1.2.3 方差分析

方差分析（analysis of variance, ANOVA），又称“变异数分析”或“ F 检验”，用于两个及两个以上样本平均数差别的显著性检验，1923年由英国统计学家费希尔提出。根据分析的因素的数量可以分为单因素方差分析、二因素方差分析和多因素方差分析；如果根据其数学模型则可以分为固定模型（fixed model）、随机模型（random model）和混合模型（mixed model）。方差分析在生命科学工作中极为重要，特别是在多因素试验中，可以帮助大家剖析起主导作用的变异来源，列出方差分析各自的期望均方（expected mean square, EMS），从而估计出各种效应值。

1.2.4 回归与相关

相关（correlation）与回归（regression）是研究变量间相互关系的一种统计分析方法。1886年，高尔顿在“在遗传性身高中向中等身高的回归”一文中，正式提出了回归的概念：两个或多个变量间存在依从关系。根据变量的个数，可分为一元相关回归、二元相关回归及多元相关回归；而根据相关回归的曲线形态则可分为直线相关回归和曲线相关回归。1888年高尔顿又在“相关及其主要来自人体的度量”一文中充分论述了“相关”的统计意义，并提出了相关系数的计算公式。相关是指两个或多个变量间存在平行的关系，主要用于研究两个变量之间相互关系的密切程度，用相关系数表示。

1.2.5 试验设计

试验设计 (experimental design), 广义上是指试验研究课题设计, 也就是整个试验计划的拟定; 狭义上是指试验单位 (如单个细胞、植株、老鼠等) 的选取、重复数目的确定以及试验单位的分组。通过试验设计可避免系统误差, 控制、降低试验误差, 无偏估计处理效应, 从而对样本所在总体作出科学、可靠、正确的推断。在试验设计过程中, 必须遵循试验三原则, 即随机、重复、局部控制。

Fisher 在首次发表《试验研究工作中的统计方法》一书后, 提出了田间试验的基本原则和主要设计方法, 成为试验设计的经典著作。1925 年, Fisher 提出随机区组和正交拉丁方试验设计, 同时, 他还在试验设计中提出“随机化”原则, 并于 1938 年与 Yates 合编了 Fisher-Yates 随机数字表。

1.3 生物统计学的作用

1964 年, 英国著名统计学家 Yates 和 Healy 曾说过: “非常痛心地看着, 因为数据分析的缺陷和错误, 那么多好的生物研究工作面临被葬送的危险”。从这句话, 足可以看出生物统计学对于生命科学研究是何等重要, 其作用主要体现在以下三个方面。

(1) 提供科学的试验设计方法。科学的试验设计可用较少的人力、物力和时间取得丰富可靠的试验资料。因此, 在开展任何一项生命科学试验之前, 都必须科学地进行试验设计, 包括样本容量的确定、抽样方法的挑选、处理水平的选择、重复数的设置以及试验的安排等, 都必须严格遵循试验三原则。

(2) 提供科学的试验分析方法。在生命科学试验过程中, 常常可以获取大量的非常复杂的第一手资料, 我们如何透过纷繁复杂的信息得出客观科学的结论、抓住蕴含在其中的生命科学的本质规律呢? 在数据收集、整理、分析过程中, 我们必须根据实际资料, 选取一套科学而严密的生物统计学分析方法。例如, 研究某转基因水稻品种的产量特征, 我们可获得不同植株、不同地区、不同批次的结穗数。从这些杂乱的数据中, 很难直接看出其规律性, 如果采用生物统计学方法对其进行整理、分析, 就可以了解转基因水稻与非转基因水稻产量之间的关系, 以及不同地区该转基因水稻的产量是否存在显著差异, 为进一步的深入研究提供科学依据。

(3) 正确评价试验结果的可靠性。在生命科学研究中, 试验单元间的差异除了试验处理外, 还存在一些无法控制的偶然因素对试验结果的影响。例如, 在比较转基因与非转基因水稻产量时, 可以采用对比设计、分两个区块进行试验, 尽管施肥、管理、浇水等耕作技术都一致, 但由于不同区块土壤肥力必然存在一定的差异、每株水稻的生理生化特性的差异, 以及其他一些不可控的因素对水稻产量均存在一定的影响。因而, 每株水稻的产量必然包括转基因的真实效应以及偶然因素两个效应。因此, 我们在分析结果时, 只有正确区分这两种效应, 判断它们各自效应的大小, 才能对试验结果作出科学可靠的结论。根据不同的资料类型, 生物统计学提供了推断试验结果可靠性的不同方法。

1.4 生物统计学的特点及其学习方法

在生物统计学的教学中,一些学生反映生物统计学比较难懂,这主要是因为生物统计学是生物科学与数理统计的交叉学科,涉及的数学公式、数学概念、数学符号及数学用表等较多,需要付出较多的时间和精力来掌握这些基本概念、基本原理和基本方法,记忆一些常用的公式和符号,理解并推导一些基本的数学表达式。事实上,生物统计学是一门逻辑性和应用性较强的学科,作为一个并非专门从事生物统计学原理与方法研究的人员,其主要目的是应用生物统计学方法解决生产和研究中遇到的实际问题,无需一一推导和记忆复杂的数学公式。因此,为了便于学习,本教材中基本原理和分析方法均有具体软件的操作实例,可操作性强,并在每一章的后面附有复习思考题。在具体学习过程中,应首先掌握基本概念、基本原理,看懂例题,理解分析结果;然后结合常用的生物统计分析软件(Minitab、DPS、SPSS等)练习复习思考题,达到掌握各种统计分析方法的目的。在学习中,应注意以下5个方面。

第一,必须理论联系实际,结合专业知识,不可忽视专业知识的重要性。生物统计对试验数据提供收集、整理、分析的方法,并作出结论。因此,生物统计学在使相关的研究省时、省事、经济、有效的同时,可以对试验结果作出科学的推断,但必须结合专业知识进行全面综合分析才能获得客观、科学的结论。例如,某研究者测得某地区200名正常成年男性转氨酶和身高的数据,用简单相关与回归对这两个变量间的关系进行分析,并开展了正确的显著性检验,得出的结论是:可用身高去预测转氨酶的含量。显然这个结论与实际不符,缺乏专业理论知识的支撑。

第二,要理解生物统计学的基本理论和基本公式。要正确理解每一个公式和生物统计分析方法的实际含义和应用条件。例如,进行 t 检验时,要求原始资料服从 t 分布,否则可能会得出错误结论;又如,进行方差分析时必须满足其三个基本假定。如果不考虑应用条件,生搬硬套,会得出与事实不符,甚至截然相反的结果。

第三,掌握各种符号及其意义,定期整理所学过的公式与符号,不必深究其数学推导过程。因此,平时需多留意国内外杂志、书籍中的图表、数据及其分析方法和对结果的解释,从而熟悉表达方法及其应用。

第四,要及时认真地做好课后习题。结合常用的生物统计分析软件(Minitab、DPS、SPSS等)加深对统计的基本理论和基本方法的理解和掌握,达到能熟练运用生物统计方法分析实际问题的目的,提高生物统计分析的效率;及时认真地复习是保证学习效果的关键。

第五,应注重培养科学的统计思维方法。生物统计意味着一种新的思考方法——从不确定性或概率角度来思考问题和分析科学试验的结果。

复习思考题

1. 什么是统计学?什么是生物统计学?



2. 生物统计学的主要理论基础有哪些？各部分之间有何联系？
3. 试从基础型生命科学研究的过程分析生物统计学在科学实践中的地位。
4. 试从应用基础型生命科学研究的过程分析生物统计学在科学实践中的地位。
5. 学习生物统计学过程中需要注意哪些问题？

第 2 章 数据的整理与分析

内容提要

生物统计学分析的第一步是资料的收集和整理。收集资料主要有调研和开展生物学试验两种方法，而资料的整理主要通过对原始资料的核查、校对，制作次数分布表和次数分布图来完成。生命科学领域的试验资料一般都具有集中性、离散性及分布形态三个基本特征：集中性主要利用算术平均数、中位数、几何平均数等反映；离散性主要通过标准差、方差、变异系数等特征数进行度量；分布形态则主要借助偏度和峰度体现。本章首先介绍总体与变量等最基本的生物统计学名词术语，继而结合实例，应用软件来阐明实验原始数据整理的具体方法，并对实验数据的特征进行统计分析，全面阐明数据资料的整理分析方法。

在科学试验与调查中，常常会得到大量的原始数据，这些对某种具体事物或现象观察的结果称为资料（data）。这些资料在统计分析前，一般是分散的、零星的和孤立的，是一堆无序的数字。为了揭示这些资料中所蕴含的科学意义，需要对其进行必要的整理分析，揭示其内在的规律。

2.1 常用统计学术语

为了更好地学习和理解后续章节的生物统计学知识，首先必须掌握以下几组生物统计学基本概念。

2.1.1 总体、个体与样本

总体（population）是指研究对象的全体，其中的每一个成员称为个体（individual）。依据构成总体的个体数目的多寡，总体可以分为有限总体（finite population）和无限总体（infinite population）。例如，研究珠母贝的壳高，因为无法估计出珠母贝的具体数量，可以认为珠母贝是无限总体。

总体的数目往往非常庞大，全部测定需要耗费大量的时间、人力和物力，甚至根本无法完全测定每一个个体；另外，有时候数据的获取过程对研究对象具有破坏性，如要测定贝壳硬度，需要压碎贝壳。因此，只能通过研究总体中的一部分个体来反映总体的特征。

从总体中随机获得部分个体的过程，称为抽样（sampling）。为了使抽样的结果具有代表性，需要采取随机抽样（random sampling）的方法，如对一个生物的总体，机会均等地抽取样本，估计其总体的某种生物学特性。简单的随机抽样的方法有抽签、抓

阍、随机数字表法等。

从总体中抽取的一部分个体所组成的集合称为样本 (sample)。样本中个体的数量称为样本容量、样本含量或样本大小 (sample size), 通常记为 n 。如果 $n \leq 30$, 则该样本为小样本; $n > 30$, 该样本则为大样本。

例如, 2009 年 3 月, 某珍珠养殖场为了调查 2007 年繁育的 100 万只马氏珠母贝生长情况, 随机取 10 笼, 共 227 只马氏珠母贝。这里需要研究的 100 万只马氏珠母贝是总体, 其中的每只珠母贝则是个体, 随机抽取的全部 227 只马氏珠母贝是一个样本。该样本的样本容量为 227, 远大于 30, 属于大样本。

2.1.2 变量与常量

变量是研究对象所反映的指标, 如海水中叶绿素 a 的含量, 动物的体重、体长, 鱼的摄食量, 酶活力, 细胞的直径, DNA 分子的大小等。变量通常记作 X 或 Y 等大写的英文字母, 而变量的观测值可以标记为 x , 称为资料或数据。例如, 测量一批鱼的体长 X , 我们可以随机抽取 10 尾鱼作为一个样本, 测量它们的体长 (x , cm), 得到 10 个观测值 14.2、15.4、13.6、15.8、15.5、16.1、14.9、15.3、14.8、15.7, 这里体长是变量 X , 而这 10 个观测值就是样本数据 x 。

按照其可能取得的值, 可将变量分为连续型变量 (continuous variable) 和离散型变量 (discrete variable)。

连续型变量是指在某一个区间内可以取任何数值的变量, 其测量值可无限细分, 数值之间是连续不断的。例如, 50~60cm 的水稻株高为连续变量, 因为在该范围内可取出无数个值, 同样, 分子运动速度、鱼的体重、贝类的壳高、酶活力的大小、DNA 分子的大小等都属于连续型随机变量。连续型变量需通过测量才能获得, 其观测值称为连续型数据 (continuous data), 也称为度量数据 (measurement data), 如长度值、时间、重量值等。

如果变量可能取值的数值为自然数或整数, 这种变量称为离散型变量, 其数值一般通过计数获得, 如鱼、贝的怀卵量等。离散型变量的观测值称为离散型数据 (discrete data), 也称为计数数据 (count data)。

如果变量的取值, 在一定的范围内是一个相对稳定的数值, 那么这种变量称为常量 (constant)。例如, 在一个小的时空范围内, 重力加速度是一个常量。常量的取值是一个常数, 具有相对稳定性。

2.1.3 参数与统计量

统计学中把总体的指标统称为参数 (parameter), 把由样本算得的相应的总体指标称为统计量 (statistic)。例如, 2009 年 3 月, 某珍珠养殖场为了调查 2007 年繁育的 100 万只马氏珠母贝生长情况, 随机取 10 笼, 共 227 只马氏珠母贝, 测量壳高。这 227 只马氏珠母贝的壳高平均数就是统计量, 而 100 万只马氏珠母贝的壳高平均数就是参

数。参数常用希腊字母表示，如 μ 表示总体的平均数， σ 表示总体的标准差；统计量常用小写英文字母表示，如 \bar{x} 表示样本平均数， s 表示样本标准方差等。

2.1.4 准确性与精确性

由于测量方法或测量仪器等原因，实验获得的数据和变量的真实值之间会存在一定的差异；此外，数据之间也会有差别。这种差异，通常以准确性和精确性来描述。

准确性 (accuracy) 也称准确度，描述量测值平均偏离目标值的程度，指在试验中某一指标的观测值与真实值的接近程度。假设试验指标的真实值为 μ ，观测值为 x ，则 $|x - \mu|$ 越小， x 值的准确性就越高。

精确性 (precision) 也称精确度、再现性 (repeatability) 或再生性 (reproducibility)，指在相同条件下对同一对象重复测量所获得的数值间的接近程度，用来描述量测值的变异程度大小。

准确性、精确性合称为正确性。由于真值 μ 常常不知道，所以准确性不易度量，但利用统计方法可度量观测的精确性。“准确”和“精确”两个概念间的相互关系如图 2-1 所示。

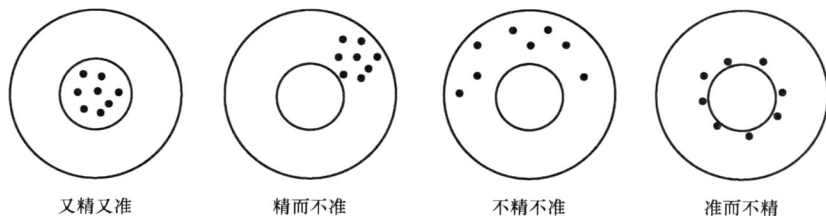


图 2-1 观测实验结果的准确性和精确性

2.1.5 随机误差与系统误差

随机误差 (random error)，是由无法控制的偶然因素导致的误差，其中最重要的是抽样误差 (sampling error)，即由随机抽样导致的样本与总体间的差异。扩大样本容量或增加试验次数是减小随机误差的有效途径，但不能完全消除随机误差。随机误差会影响试验的精确性。

系统误差 (systematic error)，又称片面误差 (lopsided error)，是由测量工具没校准、试验方法不完善、试验人员操作习惯或精神状态不同等因素所导致的差异。系统误差在一定程度上可以通过改进测量工具、完善实验设计、提升试验人员操作水平等手段来降低或消除。