普通高等教育“十一五”国家级规划教材

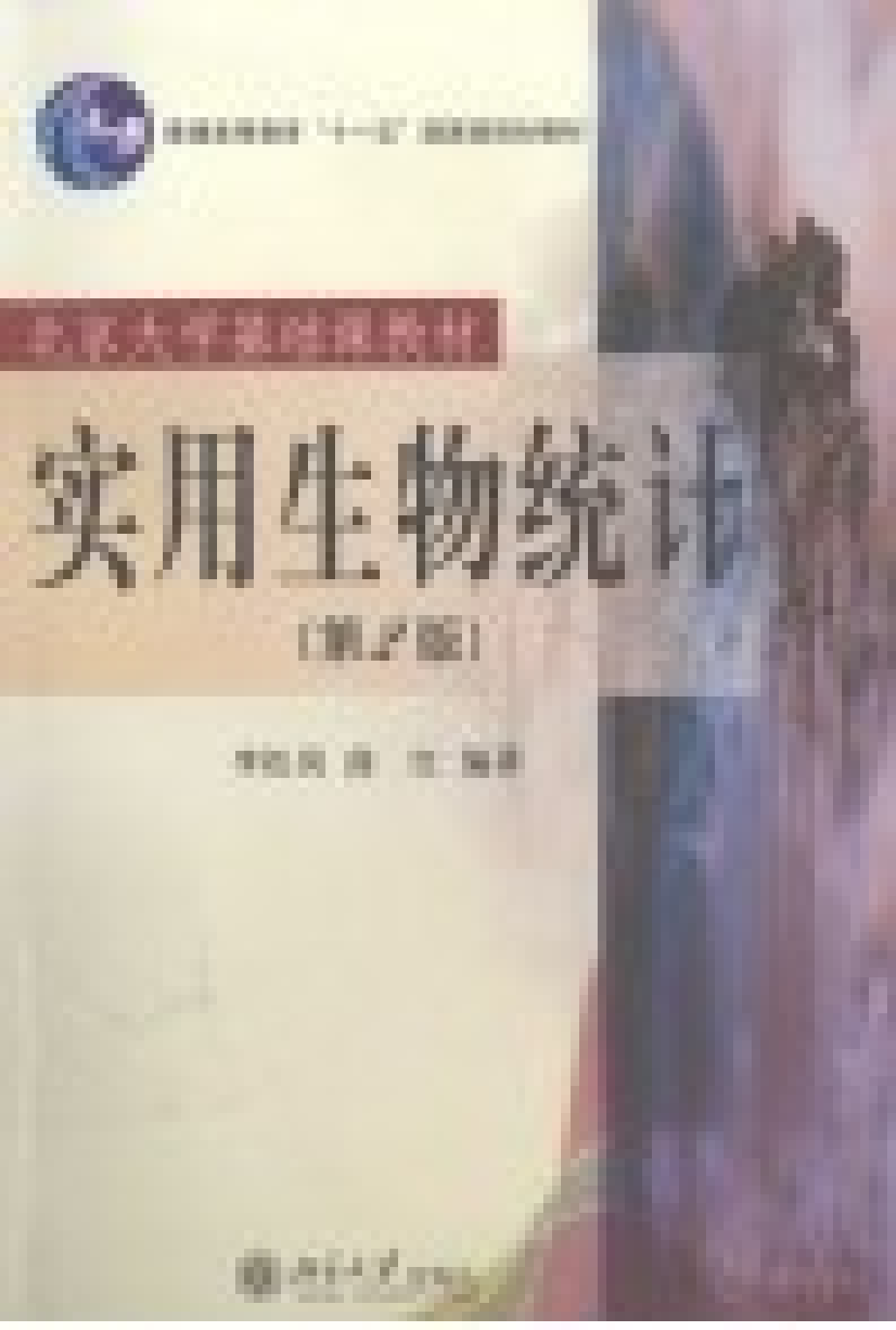
北京大学基础课教材

实用生物统计

（第2版）

李松岗 曲 红 编著

北京大学出版社
PEKING UNIVERSITY PRESS


实用生物统计
(第2版)

普通高等教育“十一五”国家级规划教材

北京大学基础课教材

实用生物统计

（第2版）

李松岗 曲 红 编著

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

实用生物统计/李松岗,曲红编著.—2版.—北京:北京大学出版社,2007.9

(北京大学基础课教材)

ISBN 978-7-301-05472-7

Ⅰ.实… Ⅱ.①李…②曲… Ⅲ.生物统计—高等学校—教材 Ⅳ.Q-332

中国版本图书馆CIP数据核字(2007)第131183号

书　　名:**实用生物统计(第2版)**

著作责任者:李松岗　曲　红　编著

责任编辑:赵学范

封面设计:张　虹

标准书号:ISBN 978-7-301-05472-7/Q·0090

出版发行:北京大学出版社

地　　址:北京市海淀区成府路205号　100871

网　　址:http://www.pup.cn

电　　话:邮购部62752015　发行部62750672　编辑部62752038

出版部62754962

电子信箱:zpup@pup.pku.edu.cn

印　刷　者:北京大学印刷厂

经　销　者:新华书店

850毫米×1168毫米　32开本　16.25印张　480千字

2002年3月第1版

2007年9月第2版　2007年9月第1次印刷(总第4次印刷)

印　　数:9001～13000册

定　　价:26.00元

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有,侵权必究

举报电话:(010)62752024　电子信箱:fd@pup.pku.edu.cn

北京高等教育精品教材

BEIJING GAODENG JIAOYU JINGPIN JIAOCAI

此为试读,需要完整PDF请访问: www.ertongbook.com

内 容 简 介

为适应生命科学研究工作者进行数据分析的需要，本书较全面地介绍了常用的概率论知识和统计方法.

第1章主要介绍了概率论的基础知识，特别是古典概型的一些计算方法.这些方法比较古老，但在今天生活和工作中都还有许多应用；第2章介绍了随机变量及其数字特征，主要是为学习以后的统计打下基础；第3章～第6章介绍了常用统计方法，包括假设检验、参数估计、非参数检验、方差分析、回归分析、协方差分析等；第7章介绍了实验设计的基本方法，包括抽样方法.书后的附录介绍了矩阵的基本知识，采用Excel进行统计计算的方法，以及常用统计表.全书内容紧紧围绕应用的目的，尽可能做到深入浅出，同时也有适量的理论推导，使读者能在理解的基础上掌握各种方法的适用条件、应用范围、优缺点等.在对各种方法的介绍中均辅以例题，各章后附有习题.

本书适合作为生命科学各领域本科生的教材，也可用于自学.书中的例题和习题除来自作者本人的工作外，也有一些引自书后列出的参考书，在此向原作者致以深深的谢意.

第 2 版前言

本书的读者绝大多数都是从事生命科学和医学的科技人员，他们需要的不是系统地掌握统计理论，而是能根据变化的情况，正确使用统计方法处理工作中得到的数据，并得到可靠的科学结论. 根据读者的这种实际需要，本书突出了实用性的特点. 主要考虑以下几个方面：(1) 本书不仅介绍必要的统计方法，还希望通过这些方法帮助读者树立用统计观点看问题的习惯：工作中接触到的数据都是有误差的，作出判断时必须考虑这一点. (2) 对每种统计方法都不仅介绍使用方法，还重点讨论其适用条件及优缺点，使读者能针对不同问题作出正确选择. (3) 配合本书内容，尽量选择有代表性的生物学问题为例题和习题，培养读者分析和解决实际问题的能力. (4) 本书涉及的内容尽可能全面，包括：从古典概型计算到统计方法；从单个实验的结果分析到大规模的科学调查；从实验设计到数据处理等各种工作中可能用到的知识. 读者可根据自己的情况有选择地阅读，同时本书可作为读者常备的有用工具书. (5) 统计理论的介绍要服从实用性的目标，即主要是帮助读者在理解的基础上掌握各种方法的适用范围和优缺点，而不是死记硬背. (6) 统计计算常是比较繁杂的. 书后附录中介绍了使用最常见的软件 Excel 表进行统计计算的方法，从而能在任何地点借助计算机解决一般的统计问题，大大减少了读者数据处理的繁杂工作. 书中的章节安排和内容较好地体现了上述目标.

本书出版以来，受到读者的好评，并于 2005 年被评为北京市精品教材. 目前本书已经拥有上万名读者，包括许多学校使用本书作为生物统计课程的教材. 不少读者中肯地指出了书中的一些不足，如：

部分章节逻辑性仍有改进余地；有些较重要的概念如统计检验的功效等没有提及；对数据的直观分析方法介绍不够等. 鉴于读者对本书的基本特点与结构还是认同的，在第 2 版的修订中对全书结构没有进行大的调整，而是重新改写了一些内容（如：第 3 章的第二节、第四节；增加了第 7 章的第五节等）；对全书进行了文字上的修改，加强了教学中学生问题较多部分的解释与分析；适当补充了例题和习题，并增加了习题参考答案；等等. 希望这些改进能够得到读者的认同.

由于作者水平所限，书中难免还有疏漏与不妥之处. 希望使用本书的教师、学生与读者不吝赐教. 借此机会，作者也对广大读者的支持与厚爱表示深深的感谢.

作　者

2007 年 5 月

第 1 版前言

在人们的实践活动中，常常会遇到类似下面的一些问题，如：一种新的疫苗，如何判断它是否有效？吸烟会不会使得肺癌的机会增加？如何抽检几百或几千人来估计某种病的流行程度？某批产品中合格品究竟有多少？该不该报废？某种实验方法或饲料配方，是否有明显的改进？等等. 总之，人们面临的这类问题可以归结为如何消耗最少的资源和人力来得到所需要的某种信息.

这一类问题的共同特点，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. 例如，为了知道产品合格与否或它的使用寿命，我们常常需要对它作破坏性检验，此时我们显然不能把所有的产品都检验一遍，而只能完成对少数几个样品的抽检，这样获得的信息显然是不完全的；再比如，要检验疫苗的有效性，但一般来说，接种过疫苗的动物不一定全不发病，而未接种的也不会全发病. 那么发病与不发病的差别究竟到多大时我们才能认为接种是有效的呢？同时，即使我们采用完全一样的实验条件再次进行实验，发病与不发病的动物数量也会有所变化，这说明类似实验的结果具有某种内在的不确定性. 要想在这种情况下正确判定疫苗的有效性，就涉及到了我们如何评价一些并不确定的实验结果的问题.

要从这样一些问题中得出科学可靠的结论，就必须依靠统计学. 有人干脆给统计学下了这样的定义："统计学就是从不完全的信息里取得准确知识的一系列技巧"，这个定义还是有一定道理的.

另外，当必须根据有限的，不完全的信息作出决策时（例如决定一批产品是出厂还是报废，某种新药是否有效，等等），统计学可以提

供一种方法，使我们不仅能做出合理的决策，而且知道所冒风险的大小，并帮助我们把可能的损失减至最小.

其次，如何花费最小代价取得所关心的信息，也是统计学的一大课题(实验设计). 不注意这一点，可能使辛辛苦苦的工作成为一种浪费.

生物学是一门实验科学. 不管你从事的是生物学的哪一个分支，都不可能完全脱离实验，只进行逻辑推理. 而实验所得到的结果几乎无例外地都带有或多或少的不确定性，即实验误差. 在这种情况下，不用统计学而想要得出正确的结论是不可能的. 可以毫不夸张地说，作为一个实验科学工作者，离开了统计学就寸步难行. 希望大家通过这门课程的学习，能够掌握常用的统计方法，尤其是它们的条件、适用范围、优缺点等，从而能够应用它们去解决实践中遇到的问题.

本书是在给北京大学生命科学学院本科生多年讲授“生物统计”课程的讲义基础上改编而成. 书稿曾经北大数学学院耿直和孙山泽教授认真审阅，并提出了宝贵的修改意见. 北京大学出版社编审赵学范在本书编辑过程中在层次、版式等方面进行了大量工作，付出了艰辛的劳动，使本书增色不少. 同时，本书还荣幸地得到北京大学“九五”教材出版基金的支持和资助. 在此一并致以深深的谢意.

作 者

2001 年 10 月

目　录

第 1 章　概率论基础…… (1)
1.1　随机现象与统计规律性 …… (1)
1.2　样本空间与事件 …… (3)
1.3　概率 …… (7)
1.4　概率的运算…… (16)
1.5　独立性…… (19)
1.6　全概公式与逆概公式…… (24)
习题 …… (29)
第 2 章　随机变量及其数字特征 …… (32)
2.1　随机变量和分布函数…… (32)
2.2　离散型随机变量…… (35)
2.3　连续型随机变量…… (39)
2.4　随机向量…… (45)
2.5　随机变量的数字特征…… (51)
2.6　大数定律与中心极限定理…… (64)
习题 …… (65)
第 3 章　统计推断 …… (68)
3.1　统计学的基本概念…… (68)
3.2　假设检验的基本方法与两种类型的错误…… (75)
3.3　正态总体的假设检验…… (81)
3.4　离散分布的假设检验…… (95)
3.5　参量估计…… (99)

此为试读,需要完整PDF请访问: www.ertongbook.com

3.6 非参数检验Ⅰ：χ^2 检验 …… (112)
3.7 非参数检验Ⅱ …… (123)
习题 …… (133)
第 4 章 方差分析 …… (139)
4.1 单因素方差分析 …… (140)
4.2 多因素方差分析 …… (155)
4.3 方差分析需要满足的条件 …… (188)
习题 …… (199)
第 5 章 回归分析 …… (205)
5.1 一元线性回归 …… (207)
5.2 相关分析 …… (228)
5.3 多元线性回归 …… (232)
5.4 非线性回归 …… (241)
习题 …… (248)
第 6 章 协方差分析 …… (252)
6.1 协方差分析的基本原理 …… (254)
6.2 协方差分析的计算过程 …… (259)
习题 …… (264)
第 7 章 实验设计 …… (266)
7.1 实验设计的基本原理及注意事项 …… (267)
7.2 抽样方法简介 …… (276)
7.3 调查数据的收集与整理 …… (299)
7.4 异常值的判断和处理 …… (306)
7.5 数据的描述性分析 …… (316)
7.6 简单实验设计 …… (324)
7.7 随机化完全区组设计 …… (328)
7.8 拉丁方及希腊-拉丁方设计 …… (330)
7.9 平衡不完全区组设计 …… (335)

7.10　裂区设计 …… (342)
7.11　正交设计 …… (347)
习题 …… (359)
附录A　矩阵基础知识 …… (363)
A.1　矩阵的概念 …… (363)
A.2　矩阵的基本运算 …… (363)
附录B　采用微软公司的 Excel 软件进行常见的统计计算 … (367)
B.1　假设检验 …… (367)
B.2　方差分析 …… (382)
B.3　回归分析 …… (398)
B.4　Excel 中常用统计函数简介 …… (406)
附录C　统计用表 …… (413)
C.1　随机数表 …… (413)
C.2a　正态分布密度函数表 …… (416)
C.2b　正态分布函数表 …… (419)
C.2c　正态分布分位数表 …… (422)
C.3　χ^2 分布分位数表 …… (425)
C.4　t 分布分位数表 …… (430)
C.5a　F 分布分位数表($F_{0.95}$) …… (433)
C.5b　F 分布分位数表($F_{0.975}$) …… (435)
C.5c　F 分布分位数表($F_{0.99}$) …… (437)
C.6　Duncan 多重比较 r 值表 …… (439)
C.7a　多重比较 q 临界值表($\alpha=0.05$) …… (442)
C.7b　多重比较 q 临界值表($\alpha=0.01$) …… (444)
C.8a　二项分布 p 的置信区间表($\alpha=0.05$) …… (446)
C.8b　二项分布 p 的置信区间表($\alpha=0.01$) …… (447)
C.9　F_{max}检验临界值表 …… (448)

C.10a 相关系数检验表($\alpha=0.05$) …… (449)
C.10b 相关系数检验表($\alpha=0.01$) …… (450)
C.11 秩和检验表 …… (451)
C.12 符号检验表 …… (453)
C.13a 游程总数检验表($\alpha=0.025$) …… (455)
C.13b 游程总数检验表($\alpha=0.05$) …… (456)
C.13c 游程总数检验表($n_1=n_2$) …… (457)
C.14 Nair(奈尔)检验法的临界值表 …… (459)
C.15 Grubbs(格拉布斯)检验法的临界值表 …… (461)
C.16a 单侧 Dixon(狄克逊)检验法的临界值表 …… (463)
C.16b 双侧 Dixon(狄克逊)检验法的临界值表 …… (464)
C.17 偏度检验法的临界值表 …… (464)
C.18 峰度检验法的临界值表 …… (464)
C.19a $T_{n(1)}$的临界值表 …… (465)
C.19b $T_{n(n)}$的临界值表 …… (468)
C.20 秩相关系数检验表 …… (471)
C.21 正交拉丁方表 …… (471)
C.22 平衡不完全区组设计表 …… (474)
C.23 常用正交表 …… (480)
附录 D 习题参考答案(部分) …… (489)
附录 E 常用统计术语中英文对照 …… (499)
参考书目 …… (505)

第 1 章　概率论基础

1.1　随机现象与统计规律性

(一) 概率论是研究随机现象的数量规律的数学分支

所谓随机现象，就是在基本条件不变的情况下，各次实验或观察会得到不同的结果的现象，而且这一结果是不能准确预料的. 例如血球计数板上某一格中的血球数；昆虫密度调查时某一个样方中目标昆虫的数量；某一时刻车间中开动的车床数，优秀选手射击弹着分布，抽样时某一样品合格与否，等等.

必然现象(或不可能事件)则是指在一定条件下必然会发生(或不发生)的事件，也可称为决定性事件. 例如早晨太阳会从东方升起；水向低处流；万有引力定律下的天体运行；纯水在标准大气压(1.01×10^5 Pa)下会在 100℃沸腾，等等.

大部分科学实验的结果都属于随机事件，分析它们就需要概率的知识，因此概率与统计就成为了所有科学工作者都应该掌握的基础知识.

(二) 频率稳定性

随机事件的结果是不可预料的，那又如何研究呢？经过长期观察，人们发现个别随机事件在一次实验或观察中可以出现或不出现，但在大量重复实验中，它出现的次数与总实验次数之比总是非常稳定的. 这种. 现象称为频率稳定性，它正是随机事件内在规律性的反映.

【例 1.1】 掷币实验：

实验者	掷币次数	正面次数	频率
Buffon(蒲丰)	4040	2048	0.5069
Pearson(皮尔逊)	12000	6019	0.5016
Pearson(皮尔逊)	24000	12012	0.5005

从上述实验结果可知，随着投掷次数的增加，正面出现的次数越来越接近一个常数：0.5. 这一实验结果很好地反映了多次重复的随机实验中频率的稳定性.

直观上，我们用一个数 $P(A)$ 来表示随机事件 A 发生可能性的大小，$P(A)$ 就称为 A 的概率. 一般来说，当实验次数 n 越来越大，直至趋于无穷时，频率也会逐渐趋近于概率.

(三) 统计的基本思想

大部分科学实验的结果都属于随机事件，即所得数据都是有误差的. 要正确地分析它们并得出可靠的结论，就必须要依靠统计知识. 下面的例子可以让我们对统计的基本思想有一个直观的了解.

【例 1.2】 试验配方 1(x)和配方 2(y)两种不同饲料配方对鸡增重的影响. 饲养 5 周后，增重如下：

	增重/kg
配方 1 (x)	1.49, 1.36, 1.50, 1.65, 1.27, 1.45, 1.38, 1.52, 1.40
配方 2 (y)	1.25, 1.50, 1.33, 1.45, 1.27, 1.32, 1.60, 1.41, 1.30, 1.52
$\bar{x}=1.436$ kg, $\bar{y}=1.392$ kg	

在例 1.1 中，$\bar{x}=1.436\,\text{kg}$，$\bar{y}=1.392\,\text{kg}$，我们是否可以说配方 1 比配方 2 好呢？也许有人会说："$\bar{x}>\bar{y}$，当然就说明配方 1 好啦."实际问题却不是这样简单. 由于鸡的个体差异等因素都会影响实验的结果，因此上述实验中包含着一些无法排除的随机误差. 在这种情况

下，我们怎么能判断 $\bar{x}$ 与 $\bar{y}$ 之间的差异是随机误差造成的，还是配方 1 真的优于配方 2？或者换句话说，$\bar{x}$ 与 $\bar{y}$ 的差异大到何种程度，我们就可以较有把握地得出配方 1 优于配方 2 的结论？要科学地回答这一类问题，靠我们以前学过的数学知识是解决不了的，必须依靠统计学的知识. 由于吃同一种饲料的一组鸡的生活条件基本上是一致的，它们之间的差异应该是随机误差大小的一种估计，因此我们可以把上述两组鸡之间的差异与组内的差异做一下比较，如果组间差异明显大于组内的差异，则认为配方 1 比配方 2 好；否则，就只能认为这两种配方差不多. 根据这样的统计学理论，我们只能认为这两个配方间没有明显差异，原因是它们组内差异比较大，说明随机因素的影响很大，平均数间的差异可能是随机因素引起的.

【例 1.3】 如果上例中的结果变成下表中的数据：

	增重/kg
配方 1 (x)	1.40，1.42，1.50，1.39，1.46，1.45，1.51，1.44，1.41，1.38
配方 2 (y)	1.38，1.41，1.35，1.50，1.36，1.33，1.42，1.38，1.37，1.41
$\bar{x}=1.4365$ kg，$\bar{y}=1.391$ kg	

此时两组数据的平均值变化不大，直观上结果应与上题相同，但统计结论却完全变了——配方 1 明显优于配方 2. 这是因为组内差距变小了，x 与 y 之间的差别不能仅用随机因素的影响来解释.

从上述例子可看出，没有统计学的知识就不能对实验结果作出科学的、有说服力的结论.

1.2 样本空间与事件

我们假定试验或观察可在相同的条件下重复进行. 这是因为一次随机实验的结果不可预料，我们主要依靠频率稳定性来研究随机

此为试读，需要完整PDF请访问：www.ertongbook.com