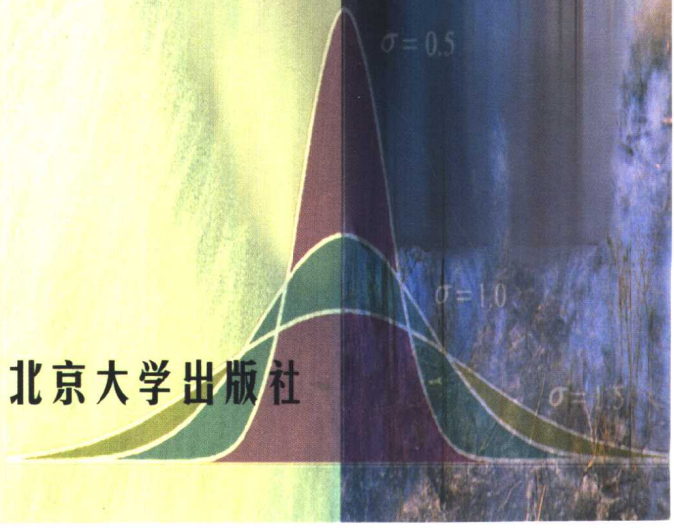
北京大学基础课教材

实用生物统计

李松岗 编著



北京大学出版社

北京大学基础课教材

实用生物统计

李松岗　编著

北京大学出版社
北　京

图书在版编目(CIP)数据

实用生物统计/李松岗编著．－北京：北京大学出版社，2002.3
ISBN 7-301-05472-6

Ⅰ．实… Ⅱ．李… Ⅲ．生物统计 Ⅳ．Q-332

中国版本图书馆 CIP 数据核字(2002)第 005470 号

书　　名：实用生物统计
著作责任者：李松岗
责 任 编 辑：赵学范
标 准 书 号：ISBN 7-301-05472-6/Q·0090
出　版　者：北京大学出版社
地　　址：北京市海淀区中关村北京大学校内　100871
网　　址：http://cbs.pku.edu.cn
电　　话：出版部 62752015　发行部 62754140　编辑部 62752021
电 子 信 箱：zpup@pup.pku.edu.cn
排　版　者：兴盛达打字服务社　62549189
印　刷　者：**北京大学印刷厂**
发　行　者：北京大学出版社
经　销　者：新华书店
850 毫米×1168 毫米　32 开本　14.25 印张　400 千字
2002 年 3 月第 1 版　2002 年 3 月第 1 次印刷
定　　价：22.00 元

内 容 简 介

为适应生命科学研究工作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.

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 …… (6)
1.4　概率的运算 …… (13)
1.5　独立性 …… (16)
1.6　全概公式与逆概公式 …… (20)
习题 …… (24)

第 2 章　随机变量及其数字特征 …… (27)
2.1　随机变量和分布函数 …… (27)
2.2　离散型随机变量 …… (30)
2.3　连续型随机变量 …… (34)
2.4　随机向量 …… (39)
2.5　随机变量的数字特征 …… (46)
2.6　大数定律与中心极限定理 …… (58)
习题 …… (59)

第 3 章　统计推断 …… (61)
3.1　统计学的基本概念 …… (61)
3.2　假设检验的基本方法与两种类型的错误 …… (67)
3.3　正态总体的假设检验 …… (71)
3.4　参量估计 …… (84)
3.5　非参数检验 Ⅰ：χ^2 检验 …… (97)

3.6 非参数检验Ⅱ …… (109)
习题 …… (119)

第4章 方差分析 …… (124)
4.1 单因素方差分析 …… (125)
4.2 多因素方差分析 …… (139)
4.3 方差分析需要满足的条件 …… (169)
习题 …… (180)

第5章 回归分析 …… (183)
5.1 一元线性回归 …… (185)
5.2 相关分析 …… (202)
5.3 多元线性回归 …… (207)
5.4 非线性回归 …… (215)
习题 …… (221)

第6章 协方差分析 …… (223)
6.1 协方差分析的基本原理 …… (225)
6.2 协方差分析的计算过程 …… (229)
习题 …… (234)

第7章 实验设计 …… (235)
7.1 实验设计的基本原理及注意事项 …… (236)
7.2 抽样方法简介 …… (242)
7.3 调查数据的收集与整理 …… (265)
7.4 异常值的判断和处理 …… (272)
7.5 简单实验设计 …… (282)
7.6 随机化完全区组设计 …… (285)
7.7 拉丁方及希腊-拉丁方设计 …… (287)
7.8 平衡不完全区组设计 …… (293)

7.9　裂区设计 …… (299)
7.10　正交设计 …… (304)
习题 …… (315)

附录 A　矩阵基础知识 …… (317)
A.1　矩阵的概念 …… (317)
A.2　矩阵的基本运算 …… (317)

附录 B　采用微软公司的 Excel 软件进行常见的统计计算 …… (321)
B.1　假设检验 …… (321)
B.2　方差分析 …… (336)
B.3　回归分析 …… (352)
B.4　Excel中常用统计函数简介 …… (360)

附录 C　统计用表 …… (367)
C.1　随机数表 …… (367)
C.2a　正态分布密度函数表 …… (370)
C.2b　正态分布函数表 …… (373)
C.2c　正态分布分位数表 …… (376)
C.3　χ^2 分布分位数表 …… (379)
C.4　t 分布分位数表 …… (384)
C.5a　F 分布分位数表($F_{0.95}$) …… (387)
C.5b　F 分布分位数表($F_{0.975}$) …… (389)
C.5c　F 分布分位数表($F_{0.99}$) …… (391)
C.6　Duncan 多重比较 r 值表 …… (393)
C.7a　多重比较 q 临界值表($\alpha=0.05$) …… (396)
C.7b　多重比较 q 临界值表($\alpha=0.01$) …… (398)
C.8a　二项分布 p 的置信区间表($\alpha=0.05$) …… (400)
C.8b　二项分布 p 的置信区间表($\alpha=0.01$) …… (401)

C.9 F_{max}检验临界值表 …… (402)
C.10a 相关系数检验表($\alpha=0.05$) …… (403)
C.10b 相关系数检验表($\alpha=0.01$) …… (404)
C.11 秩和检验表 …… (405)
C.12 符号检验表 …… (407)
C.13a 游程总数检验表($\alpha=0.025$) …… (409)
C.13b 游程总数检验表($\alpha=0.05$) …… (410)
C.13c 游程总数检验表($n_1=n_2$) …… (411)
C.14 奈尔检验法的临界值表 …… (413)
C.15 格拉布斯检验法的临界值表 …… (415)
C.16a 单侧狄克逊检验法的临界值表 …… (417)
C.16b 双侧狄克逊检验法的临界值表 …… (418)
C.17 偏度检验法的临界值表 …… (418)
C.18 峰度检验法的临界值表 …… (418)
C.19a $T_{n(1)}$的临界值表 …… (419)
C.19b $T_{n(n)}$的临界值表 …… (422)
C.20 秩相关系数检验表 …… (425)
C.21 正交拉丁方表 …… (425)
C.22 平衡不完全区组设计表 …… (428)
C.23 常用正交表 …… (434)

参考书目 …… (443)

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，昆虫密度调查，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. 例如早晨太阳从东方升起，水向低处流，万有引力，标准大气压，纯水 100 °C 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，如：

【例 1.1】 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，$\bar{x}=1.436$ kg, $\bar{y}=1.392$ 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作一下比较，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1.2】　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，没有概率论的知识就不能对实验结果作出科学的、有说服力的结论．

（二）频率稳定性

随机事件的结果一般是不可预料的，那又如何研究呢？个别随机事件（结果）在一次实验或观察中可以出现或不出现，但在大量实

验中,它出现的次数与总实验次数之比常常是非常稳定的.这种现象称为频率稳定性,正是随机事件内在规律性的反映.

【例 1.3】 掷币实验的结果列于下表中:

实验者	掷币次数	正面次数	频率
蒲丰	4040	2048	0.5069
皮尔逊	12000	6019	0.5016
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.

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现象的内在规律,因此不可重复的实验对统计学来说是没有多少意义的.

(一) 样本空间的概念

定义 在一组固定的条件下所进行的试验或观察,其可能出现的结果称为样本点,一般用 ω 表示.全体样本点的所构成的集合称为样本空间,一般用 Ω 表示.

【例 1.4】 本例以 ω、Ω 表述投 1 枚硬币和投 2 枚硬币的情况:

	ω	Ω
投 1 枚硬币	{正},{反}	{正,反}
投 2 枚硬币	{正正},{正反},{反正},{反反}	{正正,正反,反正,反反}

样本点和样本空间是严格依赖于我们的实验设计的，不同的实验设计可能有不同的样本点和样本空间．每一个最基本、最简单的结果称为一个样本点，所有可能的样本点构成样本空间，而部分样本点的集合则构成了事件．

定义　样本点的集合称为事件．

显然有：必然事件 Ω；不可能事件 Φ，这里 Φ 表示空集．

注意：上述定义不严格，如果 Ω 中有不可列*个样本点，则不能把 Ω 的一切子集都看成事件，否则无法在其上定义概率．关于这些问题的详细讨论超出了本书的范围．

(二) 事件间的关系

设 A、B 均为事件，则它们可能有以下关系：

包含　若 A 发生，则 B 必然发生，此时称 A 包含于 B，或 B 包含 A．记为：$A\subset B$，或 $B\supset A$．例：{正正}⊂{两币相同}．

相等　若 $A\supset B$，且 $B\supset A$，则称 A 与 B 相等，记为 $A=B$．例：{反反}={正面不出现}．

对立　由所有不包含在 A 中的样本点所组成的事件称为 A 的逆事件，或 A 的对立事件，记为 $\overline{A}$（也可称为“非 A”）．例：$\overline{\{\text{两币相同}\}}=\{\text{正反，反正}\}=\{\text{两币不同}\}$．

显然 A 逆的逆等于 A，即 $\overline{\overline{A}}=A$．

(三) 事件的运算

1. 运算的种类

已知事件 A、B，我们可以通过它们构成一些新的事件：

交　同时属于 A 及 B 的样本点的集合．记为：$A\cap B$ 或 AB，此时 A 与 B 同时发生．

若 $A\cap B=\Phi$，则称 A 与 B 互不相容，也可称为相离．样本点一

---

* 一个无穷集合，若它的元素可与自然数集建立一一对应，则称其为可列集，否则称为不可列集．详细讨论可参见有关测度论的书籍．

定是互不相容的.

并 至少属于 A 或 B 两事件中一个的全体样本点的集合,记为 $A \cup B$.

此时可能 A, B 都发生,也可能只发生一个.

若 $A \cap B = \Phi$,则可把并称为**和**,且记为 $A + B$.

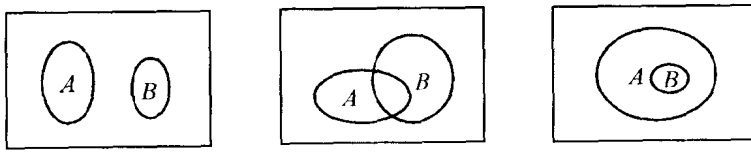
注意:在集合论的运算中,和只是并的特例,要明确它们的不同,原因是在集合论中,同一个元素只能计算一次,所以一个集合中不能有两个相同的元素.

差 包含在 A 事件中且不包含在 B 事件中的样本点的集合.记为 $A - B$.

注意:这是三种运算中惟一不满足交换律的运算.

显然:$A - B = A\overline{B}$, $A \cup \overline{A} = \Omega$, $A \cap \overline{A} = \Phi$, $\overline{A} = \Omega - A$.

可用图解的方法表示集合间的关系,如 Venn 图:



(a) 相离 (b) 相交 (c) 包含

图 1.1 两集合 A、B 的三种关系

显然,两事件 A 与 B 的关系只有上述三种,这种图解的方法对我们搞清事件间的关系是很有好处的.

2. 运算顺序

(1) 逆, (2) 交, (3) 并或差.

与算术运算比较,"逆"的运算优先级相当乘方,"交"相当乘法,"并或差"相当加减法.

3. 运算规律

集合论的运算规律与算术运算类似,但又不完全相同.它们包括

(1) 交换律：$A\cup B=B\cup A, A\cap B=B\cap A$.

(2) 结合律：$(A\cup B)\cup C=A\cup(B\cup C),(AB)C=A(BC)$.

(3) 分配律：$(A\cup B)\cap C=(A\cap C)\cup(B\cap C)$,

$$(A\cap B)\cup C=(A\cup C)\cap(B\cup C).$$

(4) 德莫根(De Morgan)定理：

$$\overline{A\cup B}=\overline{A}\cap\overline{B},\ \overline{A\cap B}=\overline{A}\cup\overline{B}$$

对于 n 个事件，甚至对可列个事件，上述定理仍成立，可写为

$$\bigcup_i\overline{A_i}=\overline{\bigcap_i A_i},\ \overline{\bigcup_i A_i}=\bigcap_i\overline{A_i}$$

注意：上述集合论运算规律与算术运算的规律很相似. 若把并比做算术加法，把交比做算术乘法，则交换律与结合律是相同的. 但分配律有差异：集合论运算中除有交对并的分配律外，还有并对交的分配律，而后者在算术运算中是不成立的. 算术运算中没有与德莫根定理相对应的规律.

【例1.5】 A,B,C 是三个事件，请用运算式表示下列事件：

(1) A 发生，B 与 C 不发生：$A\overline{B}\,\overline{C}$，或 $A-B-C$，或 $A-(B\cup C)$.

(2) A 与 B 都发生而 C 不发生：$AB\overline{C}$，或 $AB-C$，或 $AB-ABC$.

(3) 至少发生一个：$A\cup B\cup C$.

(4) 恰好发生一个：$A\overline{B}\,\overline{C}+\overline{A}B\overline{C}+\overline{A}\,\overline{B}C$.

(5) 恰好发生二个：$AB\overline{C}+A\overline{B}C+\overline{A}BC$.

1.3 概　　率

(一) 古典概型

从17世纪中叶，人们就开始研究随机现象. 当时这种兴趣或需要主要是由赌博引起的，因此人们首先注意的是这样一类随机事件：它们只有有限个可能的结果，即只有有限个样本点，同时这些样本点出现的可能性相等. 这样的概率空间称为古典概型. 由于样本点是等可能的，很自然地，人们就把事件 A 的概率定义为 A 所包含的样本点数(常称为 A 的有利场合)与样本点总数的比值，即

$$P(A)=\frac{m}{n}=\frac{A\text{ 包含的样本点数}}{\text{样本点总数}}=\frac{A\text{ 的有利场合数}}{\text{样本点总数}}$$

显然,这样的定义同时也给出了概率的计算方法.这种方法今天还有着广泛的用途,尤其是在产品的抽样检查方面.这样建立起来的概率有如下的性质.

[1] 非负性:对任意事件 A,$P(A)\geqslant 0$.

[2] 规范性:$P(\Omega)=1$.

[3] 可加性:若 $A_1,A_2,\cdots,A_n$ 两两互不相容,则

$$P(A_1+A_2+\cdots+A_n)=P(A_1)+P(A_2)+\cdots+P(A_n)$$

注意:上述可加性称为有限可加性.它主要适用于样本空间只包含有限个样本点的情况.如果样本空间含有无穷多样本点,则上述可加性也应推广为可列可加性(或称完全可加性),即若 $A_1,A_2,\cdots,A_n,\cdots$ 互不相容,则

$$P\left(\sum_{i=1}^{\infty}A_i\right)=\sum_{i=1}^{\infty}P(A_i)$$

【例 1.6】 5个身高不同的人,随机站成一排,问恰好是按身高顺序排列的可能性有多大?

解 5个人随机排列,则排法共有 5! 种.有利场合则为从高到矮,或从矮到高,共两种.因此所求概率为

$$p=2/5!=2/120=1/60$$

【例 1.7】 100 块集成电路中混有 5 块次品.任取 20 块检测,问至多发现 1 块次品的概率为多大?

解 样本空间:C_{100}^{20}

有利场合:20 块样品中没有次品:C_{95}^{20}

20 块样品中有一块次品:$C_{95}^{19}C_5^1$

$$\therefore\quad p=(C_{95}^{20}+C_{95}^{19}C_5^1)\Big/C_{100}^{20}$$

【例 1.8】 10个同样的球,编号为 1～10.现从中任取 3 个,求恰有一个球编号小于 5,一个球等于 5,另一个大于 5 的概率.

解 样本空间:$C_{10}^3=\dfrac{10\times 9\times 8}{2\times 3}=120$