BIOSTATISTICS FOR ANIMAL SCIENCE

An Introductory Text

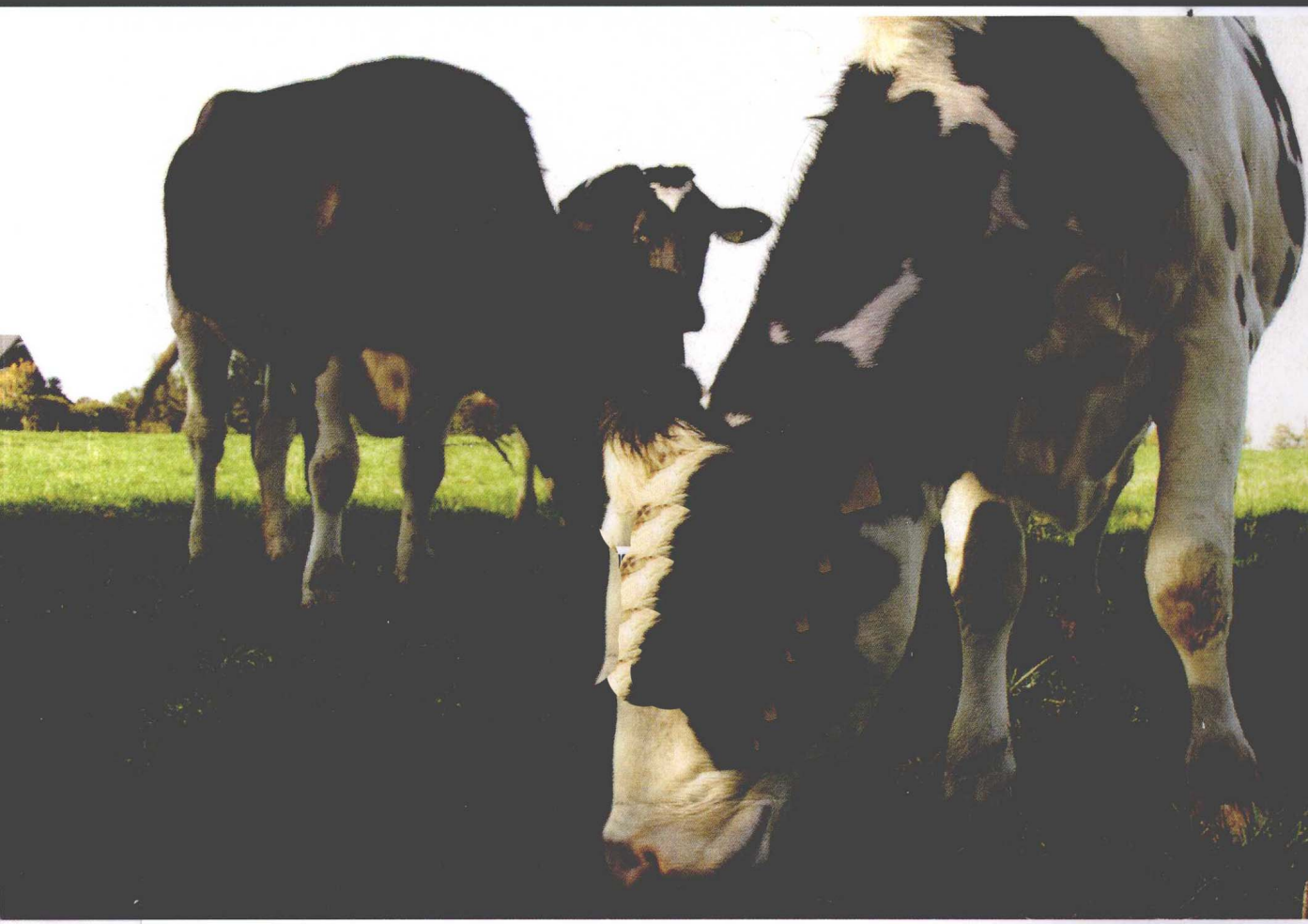
2nd Edition

（第2版）

动物科学生物统计导论

Miroslav Kaps, William Lamberson 编著

于向春 张 豪 张文广 主译



中国农业大学出版社

Biostatistics for Animal Science

an introductory text

2nd Edition

动物科学生物统计导论

第 2 版

Miroslav Kaps, William R. Lamberson 编著

于向春　张　豪　张文广　主译

中国农业大学出版社

· 北京 ·

图书在版编目(CIP)数据

动物科学生物统计导论:第2版/[美]卡普斯(Miroslav Kaps)等编著;于向春,张豪,张文广主译. —北京:中国农业大学出版社,2011.12

英文原著作者及书名:Miroslav Kaps, William R. Lamberson,

Biostatistics for Animal Science, *2nd Edition*

ISBN 978-7-5655-0416-7

Ⅰ.①动… Ⅱ.①于…②张…③张… Ⅲ.①动物学-生物统计-教材 Ⅳ.①Q95-332

中国版本图书馆CIP数据核字(2011)第193182号

书　名　动物科学生物统计导论:第2版

作　者　[美]Miroslav Kaps, William R. Lamberson　于向春　张　豪　张文广　主译

策划编辑	宋俊果	**责任编辑**	洪重光　冯雪梅　田树君　李丽君
封面设计	郑　川	**责任校对**	王晓凤　陈　莹
出版发行	中国农业大学出版社		
社　址	北京市海淀区圆明园西路2号	**邮政编码**	100193
电　话	发行部 010-62818525,8625		读者服务部 010-62732336
	编辑部 010-62732617,2618		出　版　部 010-62733440
网　址	http://www.cau.edu.cn/caup	**e-mail**	cbsszs@cau.edu.cn
经　销	新华书店		
印　刷	涿州市星河印刷有限公司		
版　次	2011年12月第1版　2011年12月第1次印刷		
规　格	787×1092　16开本　29.25印张　756千字		
定　价	68.00元		

图书如有质量问题本社发行部负责调换

本书简体中文版本翻译自 Miroslav Kaps 和 William R. Lamberson 编著的“Biostatistics for Animal Science: an introductory text, 2nd Edition”。

著作权合同登记图字：01 - 2011 - 1698

作　者　**Miroslav Kaps**
University of Zagreb, Croatia
（萨格勒布大学，克罗地亚）

William R. Lamberson
University of Missouri, USA
（密苏里哥伦比亚大学，美国）

主　译　于向春　张　豪　张文广
副主译　黄武仁　柳贤德　肖书奇
吴丽丽　闫晚姝

译者的话

由克罗地亚萨格勒布大学 Miroslav Kaps 和美国哥伦比亚大学 William R. Lamberson 合著的《动物科学生物统计导论》，是专门针对从事动物科学的大学生、研究生及科研人员的教材，也可供动物医学、农学及其他从事生命科学的教学科研人员参考。

该部教材有以下特点：

☑ 内容全面。包括了几乎所有动物科学研究和生产中能够用到的所有问题统计分析方法。

☑ 深奥难懂的生物统计概念以浅显易理解的语言介绍，而且杜绝了一些生物统计教材中前面应用、后面介绍的情况。

☑ 教材编排顺序安排合理，一般首先提出问题，再介绍分析原理，然后给出典型例题，最后介绍 SAS 系统应用。

☑ 统计学原理讲解深浅适当。生物统计学课是学生难学、老师难教的课程。目前农科大学生概率论知识较缺乏，多数没有线性代数基础，给生物统计课程教学带来一定的难度。本书包括了生物统计学用到的概率论和线性代数基础知识，这方面基础较差的学生不用再参考其他书籍。另外，本教材以实用为目的，因此没有过多的统计学原理介绍，一般是通过实例讲解生物统计学基本知识；对于较复杂的内容，如关于离散依变量分析，只是简单介绍了应用统计学软件必须有的统计学知识。

因此，本教材是非统计学专业的动物学类专业的一部优秀教材。

在翻译过程中，我们力求术语准确无误，一般根据科学出版社的《新英汉数学词汇》翻译术语。个别词汇在《新英汉数学词汇》中没有收录，我们根据英语意义和动物学专业习惯进行翻译。由于书中的一些术语和我们通用教材中存在差别，为了尊重原作者又便于读者阅读和学习，我们采用直译但加脚注的办法。考虑到一些读者可能在学习本书前从来没有接触过 SAS 软件系统，我们对可能会使读者感到困难的 SAS 程序也编写了脚注。

本书初稿完成后，主译分工协作，对初稿进行了通读和校对。这样反复，经过近一年时间，共校对了 7 次。尽管做了最大努力，但由于我们水平所限，译稿中可能还存在一些错误和不妥之处，欢迎读者多提宝贵意见。

第 2 版序言

我们非常感谢使用和阅读过本书第 1 版的读者，感谢评论过或为第 1 版提出修改意见的人士。根据这些建议，在第 2 版中我们对一些内容进行了扩展，增加了新的章节，尽量使得内容和例子更加清晰。

同第 1 版相比，编写本书第 2 版的目的和架构没有发生改变。本书主要是为动物科学专业的学生和科研人员而编写，旨在帮助他们了解并应用合适的试验设计和统计方法。书中的许多章节也适用于农学、生命科学和动物医学的学生和科研人员学习参考。

统计学方法应用于生物科学所形成的一门学科即生物统计学。生物学度量值存在变异，不仅度量值存在误差，而且由于遗传和环境也存在变异。在对生物学资料进行推断时必须考虑变异的来源，同时使得试验设计理论得到发展，如区组、协变量和重复度量值等。

第 2 版的前面几章介绍统计学的基本原理，便于读者熟悉和理解相关统计学方法的应用，不需要再去阅读其他统计学入门书籍。后面章节介绍在动物科学领域中应用最广泛的分析连续性变量和离散性变量的统计学方法。每一章在讲述时都从一个实际问题开始，然后简要介绍问题所涉及的理论背景和简短证明。书中的正文穿插实例，多数例子来自于动物科学及相关领域，以使读者熟悉统计学方法的具体应用。其中的一些例子非常简单，是让读者了解统计学基本原理以及计算原理。这些例子用计算器就可以完成计算。有一些例子比较复杂，尤其是后面章节里的一些例子。多数例子也利用 SAS 软件来计算。SAS 程序和 SAS 输出都有简明扼要的解释。而且通常情况下问题的答案有足够的小数位数，虽然超出了实际需要，但更方便读者通过比较来证实计算结果。

本书的前五章分别是：1)数据的描述和总结；2)概率；3)随机变量及其分布；4)总体和样本；5)参数估计。这五章介绍了生物统计学的基本理论，内容包括统计术语的定义、描述性统计数、数据的图形表示方法、概率的基本规则、参数估计的方法，以及伯努利分布、二项分布、超几何分布、泊松分布、多项分布、均匀分布、正态分布、卡方分布、t 分布和 F 分布等常见分布的描述。第 6 章为假设检验，包括零假设和备择假设的解释，密度函数、临界值、临界区间和 P 值的应用。介绍了多种假设检验方法，如总体均值和总体比例、期望频率和实验频率间差异以及方差同质性检验。对统计显著和实际显著的差异，得出结论时的两类错误，检验功效，以及样本容量进行了讨论。增加了目的是希望处理间无差异的等价检验。检验功效的 SAS 程序中增加了应用POWER过程的程序。

第 7 到第 10 章介绍相关和回归。第 7 章介绍简单线性回归，描述了模型、模型参数及假设。给出了参数的最小二乘法和最大似然法估计。本章还介绍了把总方差剖分成可解释方差和未解释方差的概念。第 8 章介绍了相关系数的定义和一般含义、样本相关系数的估计和假设检验。解释了偏相关。第 9 章描述多重线性回归，第 10 章曲线回归。利用矩阵介绍了这两章的重点知识，顺序与简单线性回归一章相同。介绍了模型建立的知识，包括模型的偏平方和与顺序平方和定义，使用似然函数检验模型是否合适，概念预测和艾特肯(Akaike)标准。描述了回归分析中常见的一些问题，包括异常值和多重共线性，解释了其检测方法及补救措施，包括岭回归和稳健

回归。介绍了多项式回归、非线性回归和分段回归等分析方法。文中给出了利用包括稳定水平值生长曲线确定营养需要的一些实例。

第 11 章和第 12 章为单因素方差分析，用固定效应模型定义了假设、用于 F 检验的平方和的剖分、组均值的估计及其检验。给出了 F 检验后平均数的比较方法，包括最小显著差数法，Tukey 检验和对比分析。第 12 章描述了单向分类资料随机效应模型的特点，第 13 章为混合效应模型。

第 14 到第 23 章介绍特殊的试验设计及其分析方法。讨论的专题包括：试验设计的一般概念，区组化，交变设计，析因设计，系统分组设计，双区组化设计，裂区设计，协方差分析，重复度量数据分析，数量水平处理分析。每个专题都给出了例子及其 SAS 分析程序。

最后一章讨论的是不连续依变量数据的分析。包括依变量为二元数据的 logit 模型，依变量为二项分布数据的 probit 模型，以及依变量为计数资料的对数-线性模型。另外，还专门讨论的诊断分析及 ROC 曲线分析。

第 2 版增加了当数据分布未知时一元和二元方差分析的非参数分析方法，讨论了数据不完整和有缺失时的分析方法，SAS 程序中增加了新的有用选项和绘图。还新增了合适的例题。

我们对为本书出版提供过帮助的所有人表示感谢。特别感谢 Matt Lucy、Duane Keisler、Henry Mesa、Kristi Cammack、Marijan Posavi 和 Vesna Luzar-Stiffler 对本书第 1 版进行了审阅以及 Cyndi Jennings、Cinda Hudlow 和 Dragan Tupajic 对第 1 版进行了编辑；感谢 Steven Lukefahr、Denise McNamara、Catherine Selby、Jackie Atkins 和 Laura School 为本书第 2 版提供的帮助。

Miroslav Kaps　　萨格勒布大学，克罗地亚
William R. Lamberson　　密苏里哥伦比亚大学，美国
2009 年 3 月

目　录

第 1 章

数据的描述和总结

1.1 数据和变量

数据(data)是统计学家开展工作的基础,是经度量、计数或观察得到的记录,如肉牛的体重、奶牛的产奶量、动物的性别(雄性或雌性)以及动物眼睛的颜色(蓝色或绿色)等。描述上述数据的变量分别是体重、产奶量、性别和眼睛的颜色等。数据是变量的值,例如体重为 200 kg、日产奶量为 20 kg、性别为雄性或眼睛的颜色为蓝色。描述度量或观察结果的变量值可以不同,即表现出变异性。变量(variable)可分为数量变量(数值变量)和质量变量(属性变量、类型变量、分类变量)。

数量变量(quantitative variable)的值用数值表示,数量变量间的差异表现为数值的差异。动物的体重、窝产仔数、温度和时间等变量都是数量变量。两个数量变量的比值也是数量变量。数量变量分为连续性(continuous)数量变量和离散性(discrete)数量变量。连续性数量变量的值有无穷多个,这些值都是实数;离散性数量变量的值是可计数的(countable),这些值可能是有限的也可能是无限多个,而且取值都是自然数或整数。连续性数量变量如产奶量或体重,离散性数量变量如窝产仔数或月产蛋数。

质量变量(qualitative variable)的值可以以类型表示,如动物眼睛的颜色分为蓝色或绿色等、健康状态为患病或健康。质量变量分为次序变量(ordinal variable)和名义变量(nominal variable)。次序变量的类型是有等级的,而名义变量没有等级。名义变量如动物的编号、毛色和性别;次序变量如母牛的产犊情况。通常把产犊情况分为 5 类:1)正常产犊;2)产犊几乎没有困难;3)产犊有困难;4)母牛产犊很困难;5)母牛剖腹产。需要说明的是,次序变量的分类也可用数字(得分)表示,但这些数字并没有数值含义。例如,产犊情况,1)和 2)(正常产犊和产犊几乎没有困难)之间的差异以及 4)和 5)(母牛产犊很困难和母牛剖腹产)之间的差异含义不同。因此,得分描述的是有顺序的类型,而不是得分的含义。根据质量变量的定义,质量变量中可能含有数量变量,如某一分类情况下动物的头数或所占比例。

1.2 数据的图形表示

1.2.1 质量数据的图形表示

在描述质量变量资料时,每个观测值都属于特定的类型,每种类型的数据可用观测值的个数或者观测值个数在总观测值中所占的比例来描述,其中,前者称为频数(frequency),后者称为相对频数(relative frequency)。质量数据可用直条图(bar chart)、柱形图(column chart)或饼图(pie chart)进行表示。

例：克罗地亚各品种产奶牛头数见下表：

品 种	奶牛头数	百分比
西门塔尔牛	62 672	76.7%
荷斯坦牛	15 195	18.6%
褐色牛	3 855	4.7%
总 数	81 722	100%

奶牛头数可以用直条图表示(图 1.1)，其中每一个直条代表一个品种。不同品种奶牛的百分比可用饼图表示(图 1.2)。

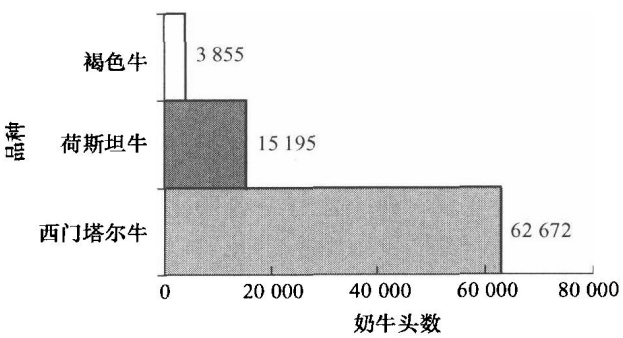


图 1.1 不同品种奶牛头数条形图

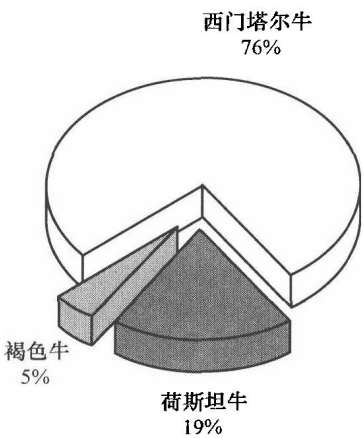


图 1.2 不同品种奶牛百分比饼图

1.2.2 数量数据的图形表示

数量变量资料的分布可以用不同图形进行描述，其中应用最广泛的图形有直方图(histogram)、茎叶图(stem and leaf graph)和箱式图(box plot)。下面介绍直方图和茎叶图，箱式图用来描述连续性数量变量，将在本章 1.3.5 节中介绍。

1.2.2.1 直方图

直方图表示一组数据的频数分布。数量数据先分成不同的组，然后利用直方图表示每一组观测值的个数或相对频数。通常按以下步骤绘制直方图：

1. 计算全距，也称作极差(range)，是全部数据中的最大值与最小值之差；

2. 根据观测值的个数，把数据分成 5～20 组。通过四舍五入使组数(class width)为一整数。通常情况下，最低组的组限要小于所有观测值中的最小值，最高组的组限要大于所有观测值中的最大值；

3. 统计每组的观测值数即频数；

4. 计算相对频数，相对频数$=\dfrac{\text{频数}}{\text{总观测值数}}$；

5. 画直方图，以柱状图或条形图表示，一个轴表示组限，另一个轴表示频数。

例:绘制 100 头小牛的 7 月龄体重(kg)的直方图。

233	208	306	300	271	304	207	254	262	231
279	228	287	223	247	292	209	303	194	268
263	262	234	277	291	277	256	271	255	299
278	290	259	251	265	316	318	252	316	221
249	304	241	249	289	211	273	241	215	264
216	271	296	196	269	231	272	236	219	312
320	245	263	244	239	227	275	255	292	246
245	255	329	240	262	291	275	272	218	317
251	257	327	222	266	227	255	251	298	255
266	255	214	304	272	230	224	250	255	284

最小值=194

最大值=329

全距=329−194=135

把所有数据分为 15 组,组距是$\left(\frac{135}{15}\right)=9$,可以令组数等于 10,得下表:

组和组限	组中值	小牛数(频数)	相对频数(%)	累积频数
185～194	190	1	1	1
195～204	200	1	1	2
205～214	210	5	5	7
215～224	220	8	8	15
225～234	230	8	8	23
235～244	240	6	6	29
245～254	250	12	12	41
255～264	260	16	16	57
265～274	270	12	12	69
275～284	280	7	7	76
285～294	290	7	7	83
295～304	300	8	8	91
305～314	310	2	2	93
315～324	320	5	5	98
325～334	330	2	2	100

图 1.3 是描述小牛体重的直方图,横轴代表小牛分组情况,纵轴代表每组小牛头数即频数。不同组的值以每组的组中值(class midrange)表示,组中值是每一组的组下限与组上限的中间值,但不同组的值也可用组限表示。

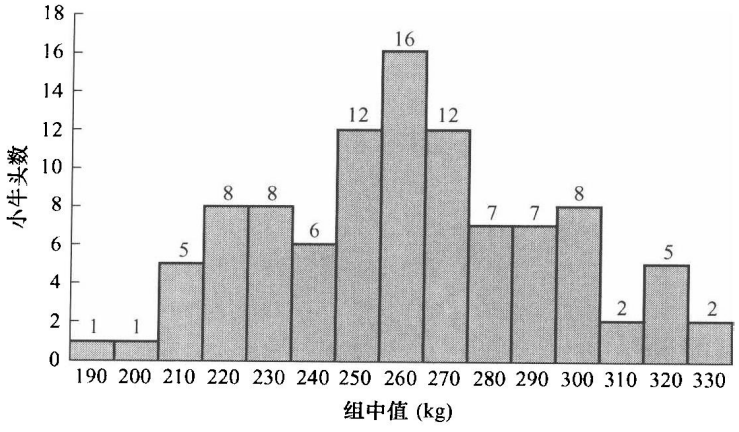


图 1.3 7 月龄小牛体重的直方图(n=100)

1.2.2.2 茎叶图

另外一种常见的数量变量资料的表示方法是茎叶图,具体步骤如下:

1. 将数据资料中的每个值分成"茎"和"叶"两部分,"茎"与数值的较高小数位对应,"叶"与数值的较低小数位对应。如上面小牛体重的例子,每个重量的前两个数字表示茎,第三个数字表示叶;

2. 然后在第一列,将"茎"按升序排列;

3. 在第一列的后面,与"茎"相对应的"叶"按升序依次排列。

由此绘制出小牛体重的茎叶图如下:

茎	叶
19	4 6
20	7 8 9
21	1 4 5 6 8 9
22	1 2 3 4 7 7 8
23	0 1 1 3 4 6 9
24	0 1 1 4 5 5 6 7 9 9
25	0 1 1 1 2 4 5 5 5 5 5 5 6 7 9
26	2 2 2 3 3 4 5 6 6 8 9
27	1 1 1 2 2 2 3 5 5 7 7 8 9
28	4 7 9
29	0 1 1 2 2 6 8 9
30	0 3 4 4 4 6
31	2 6 6 7 8
32	0 7 9

例如,倒数第二行的"茎"是 31,"叶"分别是 2、6、6、7 和 8,表明这一组的度量值分别是 312、316、316、317 和 318。当数量数据适合用茎叶图描述时,说明茎叶图的分布同直方图的分布类似,而且茎叶图还能够直接列出所有的观测值。