研究生教学用书

全国高等农林院校“十一五”规划教材

高级生物统计

Advanced Biostatistics

明道绪　主编

中国农业出版社

研究生教学用书

全国高等农林院校“十一五”规划教材

高级生物统计

明道绪　主编

中国农业出版社

图书在版编目（CIP）数据

高级生物统计 / 明道绪主编．—北京：中国农业出版社，2006.8
全国高等农林院校“十一五”规划教材
ISBN 7-109-11032-X

Ⅰ．高… Ⅱ．明… 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（2006）第 085124 号

中国农业出版社出版
（北京市朝阳区农展馆北路 2 号）
（邮政编码 100026）
出版人：傅玉祥
责任编辑 彭明喜

中国农业出版社印刷厂印刷 新华书店北京发行所发行
2006 年 9 月第 1 版 2006 年 9 月北京第 1 次印刷

开本：820mm×1080mm 1/16 印张：24.75
字数：592 千字
定价：49.00 元
（凡本版图书出现印刷、装订错误，请向出版社发行部调换）

内 容 简 介

《高级生物统计》是为高等农林院校硕士研究生编写的教材，介绍了近年来试验研究中常用的、先进的、重要的试验设计和统计分析方法。主要内容包括：一元回归分析，多元回归与相关分析，重复数不等资料的最小二乘分析，回归设计与分析，回归方程的优化，特殊试验设计。书后附有统计方法的 SAS 程序、统计数学用表和英汉名词对照表。

本教材可作为高等农林院校种植类、养殖类、园林类、食品科学类、生物技术类等研究生教学用书，对农业科技工作者也有重要的参考价值。

编 审 者

主　编　明道绪（四川农业大学）

副主编　王钦德（山西农业大学）

崔　岘（甘肃农业大学）

参编者　唐章林（西南大学）

刘学洪（云南农业大学）

刘永建（四川农业大学）

田孟良（四川农业大学）

主　审　潘光堂（四川农业大学）

黄玉碧（四川农业大学）

前　言

高级生物统计是我国高等农林院校多数专业硕士研究生开设的一门必修课。该课程针对硕士研究生在大学本科阶段已学习过生物统计的基本内容，有微积分、线性代数等数学基础，介绍近年来试验研究中常用的、先进的、重要的试验设计和统计分析方法。遗憾的是，一直无合适的教材。为了满足我国迅速发展的研究生教育的需要，我们5所农业大学决定在认真总结该门课程讲义编写、教学经验的基础上，根据硕士研究生培养目标的要求，联合编写一本供高等农林院校硕士研究生教学用的教材——《高级生物统计》。该教材经申报获准列为全国高等农林院校“十一五”规划教材（研究生教学用书），由中国农业出版社出版。

《高级生物统计》由四川农业大学明道绪教授、山西农业大学王钦德教授、甘肃农业大学崔岘教授、西南大学唐章林教授、云南农业大学刘学洪教授、四川农业大学刘永建副教授和田孟良博士合作编写，明道绪教授任主编，王钦德教授、崔岘教授任副主编。

本教材在保持本学科的系统性和科学性的前提下，注意与本科生物统计课程教学内容的衔接，编入了本科生物统计教材中由于学时限制而未讲授的重要内容；注意引入本学科发展的新知识、新成果。在编写中力求做到科学性与实用性、先进性与针对性相统一，循序渐进、由浅入深；在正确阐述重要的统计学原理的同时，着重于基本概念、基本方法的介绍；每一种设计和分析方法都安排有步骤完整、过程详细的实例予以说明，各章都配备有习题供读者练习；并结合本教材所介绍的统计方法的实例详细叙述了SAS程序的具体应用。由于多元统计分析已单独作为研究生的一门课程开设，因而本教材不涉及多元统计分析的内容。

本教材包括一元回归分析（刘学洪编写），多元回归与相关分析（崔岘编写），重复数不等资料的最小二乘分析（明道绪编写），回归设计与分析（王钦德、唐章林

编写），回归方程的优化（刘永建编写），特殊试验设计（田孟良、王钦德编写）共六章，附有统计方法的SAS程序（刘永建编写）、统计数学用表和英汉名词对照表。初稿完成后，由主编明道绪教授统稿，对各章进行了仔细审定，作了必要的修改与增删，并请主审——四川农业大学潘光堂教授、黄玉碧教授审阅。

本教材可作为高等农林院校种植类、养殖类、园林类、食品科学类、生物技术类等硕士研究生教学用书；对于农业、畜牧、水产、食品、生物技术等科技工作者也是一本有重要应用价值的工具书。

本教材在编写过程中参考了有关中外文献和专著，编者对这些文献、专著的作者，对大力支持编写工作的中国农业出版社表示衷心感谢！

限于编者水平，错误、疏漏在所难免，敬请生物统计学专家和广大读者批评指正，以便再版时修改。

编　者

2006年5月18日

目　录

第一章　一元回归分析

在农业、畜牧业、水产业等试验研究中常常要研究两个或两个以上相关变量（correlation variable）间的关系。相关变量间的关系分为两种：一种是因果关系，即一个变量的变化受另一个或几个变量的影响，如病虫害发生时期受温度的影响，小麦单位面积产量受单位面积穗数、每穗粒数、千粒重的影响等；另一种是平行关系、它们互为因果或共同受到另外因素的影响，如牛的体长和胸围之间的关系、小麦每穗粒数与千粒重之间的关系等都属于平行关系。

在统计学中采用回归分析（regression analysis）研究呈因果关系的相关变量间的关系。表示原因的变量称为自变量（independent variable），表示结果的变量称为依变量（dependent variable）。研究一个自变量与一个依变量的回归分析称为一元回归分析；研究多个自变量与一个依变量的回归分析称为多元回归分析。一元回归分析又分为直线回归分析与曲线回归分析两种；多元回归分析又分为多元线性回归分析与多元非线性回归分析两种。回归分析的任务是揭示呈因果关系的相关变量间的联系形式，建立它们之间的回归方程（regression equation），利用所建立的回归方程，由自变量（原因）来预测、控制依变量（结果）。

在统计学中采用相关分析（correlation analysis）研究呈平行关系的相关变量之间的关系。对两个变量间的直线关系进行相关分析称为直线相关分析（也称为简单相关分析）。对多个变量进行相关分析时，研究一个变量与多个变量间的线性相关称为复相关分析；研究其余变量保持不变的情况下两个变量间的线性相关称为偏相关分析。在相关分析中，不区分自变量和依变量。相关分析只研究两个变量之间线性相关的程度和性质或一个变量与多个变量之间线性相关的程度，不能用一个或多个变量去预测、控制另一个变量的变化，这是回归分析与相关分析的主要区别。

第一节　直线回归

对于呈因果关系的两个相关变量，自变量用 x 表示，依变量用 y 表示，通过试验（experiment）或调查（survey）获得两个变量的 n 对观测值（observations），记为（x_1，y_1），（x_2，y_2），…，（x_n，y_n）。

如果依变量 y 与自变量 x 的关系是直线关系，根据 n 对观测值，利用最小二乘法（least squares method），可以建立 y 与 x 之间的直线回归方程（linear regression equation）：

$$\hat{y} = a + bx \tag{1-1}$$

其中，

$$b = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{\sum xy - (\sum x)(\sum y)/n}{\sum x^2 - (\sum x)^2/n} = \frac{SP_{xy}}{SS_x} \tag{1-2}$$

$$a = \bar{y} - b\bar{x} \tag{1-3}$$

式（1-2）中的分子 $\sum (x-\bar{x})(y-\bar{y})$ 是自变量 x 的离均差（deviation from mean）与依变量 y 的离均差的乘积和，简称乘积和（sum of products），记作 SP_{xy}；分母 $\sum (x-\bar{x})^2$ 是自变量 x 的离均差平方和，简称平方和（sum of squares），记作 SS_x。

a 叫做样本回归截距（intercept），是回归直线与 y 轴交点的纵坐标，当 $x=0$ 时，$\hat{y}=a$，在有实际意义时，a 表示 y 的起始值；b 叫做样本回归系数（regression coefficient），表示 x 改变一个单位，y 平均改变的单位数量，b 的符号反映了 x 影响 y 的性质，b 的绝对值大小反映了 x 影响 y 的程度；$\hat{y}$ 叫做回归估计值（regression estimate），是当 x 在其研究范围内取某一个值时，y 值平均数（mean）的估计值。

依变量 y 与自变量 x 间是否存在直线关系，可用 F 检验法或 t 检验法进行检验。

1. F 检验法 统计学已证明

$$\sum (y-\bar{y})^2 = \sum (\hat{y}-\bar{y})^2 + \sum (y-\hat{y})^2 \tag{1-4}$$

$\sum (y-\bar{y})^2$ 反映了 y 的总变异程度，称为 y 的总平方和，记为 SS_y；$\sum (\hat{y}-\bar{y})^2$ 反映了由于 y 与 x 间存在直线关系所引起的 y 的变异程度，称为回归平方和（sum of squares of regression），记为 SS_R；$\sum (y-\hat{y})^2$ 反映了除 y 与 x 存在直线关系以外的原因（包括试验误差）所引起的 y 的变异程度，称为离回归平方和（sum of squares due to deviation from regression）或剩余平方和（residual sum of squares），记为 SS_r。于是，式（1-4）又可表示为：

$$SS_y = SS_R + SS_r \tag{1-5}$$

与此相对应，y 的总自由度（degrees of freedom）df_y 也划分为回归自由度 df_R 与离回归自由度 df_r 两部分，即

$$df_y = df_R + df_r \tag{1-6}$$

在直线回归分析中，y 的总自由度 $df_y = n-1$；回归自由度 $df_R = 1$；离回归自由度 $df_r = n-2$。

F 检验的计算公式为：

$$F = \frac{MS_R}{MS_r} = \frac{SS_R/df_R}{SS_r/df_r} = \frac{SS_R}{SS_r/(n-2)}, \quad df_1 = 1, \quad df_2 = n-2 \tag{1-7}$$

回归平方和还可用下面的公式计算得到：

$$SS_R = bSP_{xy} \tag{1-8}$$

$$= \frac{SP_{xy}^2}{SS_x} \tag{1-9}$$

根据式（1-5），可得到离回归平方和计算公式为 $SS_r = SS_y - SS_R$。

2. t 检验法 t 检验的计算公式为：

$$t = \frac{b}{s_b}, \quad df = n-2 \tag{1-10}$$

$$s_b = \frac{s_{yx}}{\sqrt{SS_x}} \tag{1-11}$$

其中，s_b 为回归系数标准误（standard error of regression coefficient）；s_{yx} 为离回归标准误（standard error due to deviation from regression），其计算公式为：

$$s_{yx}=\sqrt{\frac{\sum(y-\hat{y})^2}{n-2}} \tag{1-12}$$

离回归标准误 s_{yx} 的大小表示了回归直线与实测点偏差程度的大小，即表示了回归估计值 $\hat{y}$ 与实际观测值 y 偏差程度的大小。

在直线回归分析中，F 检验与 t 检验等价，可任选一种进行检验。

从等式 $\sum(y-\bar{y})^2=\sum(\hat{y}-\bar{y})^2+\sum(y-\hat{y})^2$ 不难看到：y 与 x 直线回归效果的好坏取决于回归平方和 $\sum(\hat{y}-\bar{y})^2$ 与离回归平方和 $\sum(y-\hat{y})^2$ 的大小，或者说取决于回归平方和在 y 的总平方和 $\sum(y-\bar{y})^2$ 中所占的比例的大小。这个比例越大，y 与 x 的直线回归效果就越好，反之则差。我们把比值 $\sum(\hat{y}-\bar{y})^2/\sum(y-\bar{y})^2$ 称为 x 对 y 的决定系数（coefficient of determination），记为 r^2，即

$$r^2=\frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2}=\frac{SS_R}{SS_y} \tag{1-13}$$

决定系数的大小表示了回归方程估测可靠程度的高低。显然有 $0\leqslant r^2\leqslant 1$。

由于直线回归分析与直线相关分析关系十分密切，两种分析所进行的显著性检验都是回答 y 与 x 间是否存在直线关系的问题，因而两者的检验是等价的。即相关系数（coefficient of correlation）显著，回归系数亦显著；相关系数不显著，回归系数亦不显著。由于利用查表法对相关系数进行显著检验十分简便，因此在实际进行直线回归分析时，可用相关系数显著性检验代替直线回归关系显著性检验，即可先计算出相关系数 r 并对其进行显著性检验，若检验结果 r 不显著，则用不着建立直线回归方程；若 r 显著，再计算回归系数 b、回归截距 a，建立直线回归方程，此时所建立的直线回归方程代表的直线关系是真实的，可利用来进行预测和控制。相关系数的计算公式为：

$$r=\frac{SP_{xy}}{\sqrt{SS_xSS_y}}=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} \tag{1-14}$$

$$=\frac{\sum xy-\dfrac{(\sum x)(\sum y)}{n}}{\sqrt{\left[\sum x^2-\dfrac{(\sum x)^2}{n}\right]\left[\sum y^2-\dfrac{(\sum y)^2}{n}\right]}} \tag{1-15}$$

一、加权回归

如果通过试验或调查获得的两个变量 x 与 y 的 n 对观测值（x_1，y_1），（x_2，y_2），…，（x_n，y_n）分别是由 m_1，m_2，…，m_n 个重复（replication）观测值计算得来的平均数，此时 n 对观测值（x_1，y_1），（x_2，y_2），…，（x_n，y_n）在整个资料中所处的地位是不同的，重复数大的（x_i，

y_i），在整个资料中所占的比重大；重复数小的（x_i，y_i），在整个资料中所占的比重小。这样在用这些观测值进行回归分析时就不能把它们同等看待，应以 m_i 为“权”进行加权回归（weighted regression）分析。此时，先利用加权法计算出 $\bar{x}$，$\bar{y}$，SP_{xy}，SS_x，SS_y：

$$\bar{x}=\frac{1}{N}\sum_{i=1}^{n}m_ix_i,\quad \bar{y}=\frac{1}{N}\sum_{i=1}^{n}m_iy_i$$

$$SP_{xy}=\sum_{i=1}^{n}m_ix_iy_i-\frac{1}{N}(\sum_{i=1}^{n}m_ix_i)(\sum_{i=1}^{n}m_iy_i)$$

$$SS_x=\sum_{i=1}^{n}m_ix_i^2-\frac{1}{N}(\sum_{i=1}^{n}m_ix_i)^2 \qquad (1-16)$$

$$SS_y=\sum_{i=1}^{n}m_iy_i^2-\frac{1}{N}(\sum_{i=1}^{n}m_iy_i)^2$$

其中，$N=\sum_{i=1}^{n}m_i$。

然后利用式（1-2）、（1-3）计算回归系数 b、回归截距 a，即

$$b=\frac{SP_{xy}}{SS_x},a=\bar{y}-b\bar{x}$$

建立直线回归方程后，可用 y 与 x 的相关系数检验 y 与 x 线性关系的显著性（注意，此时查临界 r 值的自由度 $df=N-2=\sum_{i=1}^{n}m_i-2$），用决定系数 $r^2=\frac{SS_R}{SS_y}$ 度量回归方程估测可靠程度的高低。

【例 1.1】 为了研究某品种水稻中蛋白质和赖氨酸含量的关系，把不同地区的水稻进行分组，每组抽测若干个样品的蛋白质和赖氨酸，结果如表 1-1 所示，进行回归分析。

表 1-1 水稻蛋白质和赖氨酸测定结果

组号	1	2	3	4	5	6	7	8	9	10
m_i	3	5	4	8	11	7	4	6	2	9
x_i	8.90	8.41	9.80	8.09	9.00	10.22	8.56	8.78	10.08	9.90
y_i	0.283	0.320	0.276	0.299	0.267	0.255	0.290	0.295	0.263	0.270

其中，m_i 为第 i 组样品数，x_i，y_i 为第 i 组 m_i 个样品的蛋白质和赖氨酸测定值的平均数。此例各 m_i 不完全相同，应以 m_i 为“权”进行加权回归分析。先计算出 m_ix_i，m_iy_i，$m_ix_iy_i$，$m_ix_i^2$，$m_iy_i^2$，列于表 1-2。

表 1-2 水稻蛋白质和赖氨酸测定结果计算表

组号	1	2	3	4	5	6	7	8	9	10	$\sum$
m_ix_i	26.7	42.05	39.2	64.72	99	71.54	34.24	52.68	20.16	89.1	539.39
m_iy_i	0.849	1.6	1.104	2.392	2.937	1.785	1.16	1.77	0.526	2.43	16.553
$m_ix_iy_i$	7.556	13.456	10.819	19.351	26.433	18.243	9.930	15.541	5.302	24.057	150.688
$m_ix_i^2$	237.63	353.64	384.16	523.58	891.00	731.14	293.09	462.53	203.21	882.09	4 962.08
$m_iy_i^2$	0.240 3	0.512 0	0.304 7	0.715 2	0.784 2	0.455 2	0.336 4	0.522 2	0.138 3	0.656 1	4.664 5

由式（1-16），得

$$N=\sum_{i=1}^{n}m_i=59\ ,\ \bar{x}=\frac{1}{N}\sum_{i=1}^{n}m_ix_i=\frac{539.39}{59}=9.14\ ,\ \bar{y}=\frac{1}{N}\sum_{i=1}^{n}m_iy_i=\frac{16.553}{59}=0.281$$

$$SP_{xy}=\sum_{i=1}^{n}m_ix_iy_i-\frac{1}{N}(\sum_{i=1}^{n}m_ix_i)(\sum_{i=1}^{n}m_iy_i)=150.688-\frac{539.39\times16.553}{59}=-0.6429$$

$$SS_x=\sum_{i=1}^{n}m_ix_i^2-\frac{1}{N}(\sum_{i=1}^{n}m_ix_i)^2=4\,962.08-\frac{539.39^2}{59}=30.8669$$

$$SS_y=\sum_{i=1}^{n}m_iy_i^2-\frac{1}{N}(\sum_{i=1}^{n}m_iy_i)^2=4.6645-\frac{16.553^2}{59}=0.0204$$

于是，

$$b=\frac{SP_{xy}}{SS_x}=\frac{-0.6429}{30.8669}=-0.0208$$

$$a=\bar{y}-b\bar{x}=0.281-(-0.0208)\times9.14=0.4711$$

回归方程为：

$$\hat{y}=0.4711-0.0208x$$

因为

$$r=\frac{-0.6429}{\sqrt{30.8669\times0.0204}}=-0.8102**\quad(r_{0.01(57)}=0.3337)$$

表明 x 与 y 之间存在极显著的线性关系。

决定系数

$$r^2=\frac{SS_R}{SS_y}=\frac{bSP_{xy}}{SS_y}=\frac{(-0.0208)\times(-0.6429)}{0.0204}=\frac{0.0134}{0.0204}=0.6569$$

该回归方程的估测可靠程度达到 65.69%。

二、有重复观测值的回归

直线回归关系显著性检验（test of significance）显著，表明相对于其他因素、x 的高次项及试验误差（experimental error）来说，因素 x 的一次项对 y 的影响是显著的，但未回答：影响 y 的除 x 外是否还有其他不可忽略的因素，y 与 x 是否确是线性关系。也就是说，还须检验直线回归方程的失拟性（lack of fit）。这个问题可以通过做一些重复试验从而估计出真正的试验误差来解决。

（一）部分试验有重复的回归

设一个试验有 n 个处理（treatment）：x_1，x_2，…，x_{n-1}，x_n，其中 x_1，x_2，…，x_{n-1} 重复 1 次，x_n 重复 m 次，观测结果如下：

$$\begin{array}{cccccc} x_1 & x_2 & \cdots & x_{n-1} & x_n & x_{n+1}\cdots x_{n+m-1} \\ y_1 & y_2 & \cdots & y_{n-1} & \underbrace{y_n} & \underbrace{y_{n+1}\cdots y_{n+m-1}} \\ & & & & m\text{ 个重复观测值} & \end{array}\qquad(x_n=x_{n+1}=\cdots=x_{n+m-1})$$