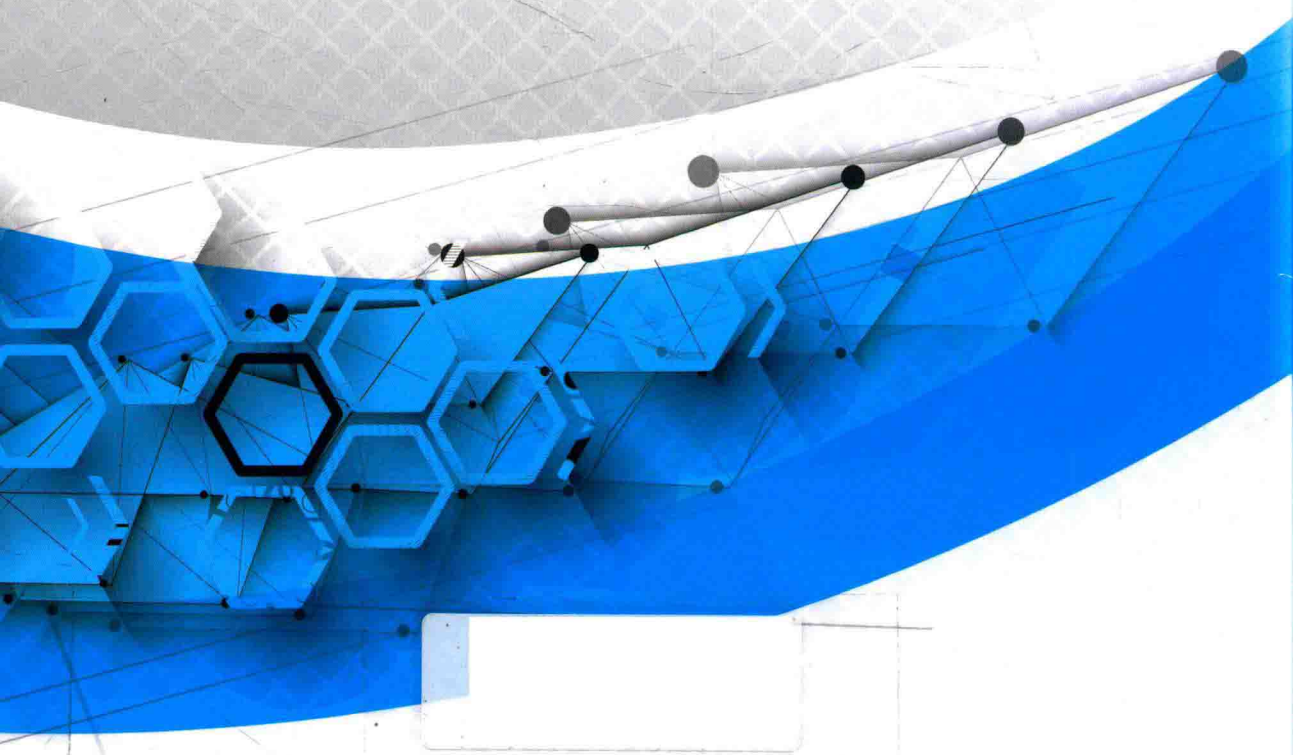
动物育种中的统计计算
——Julia语言应用

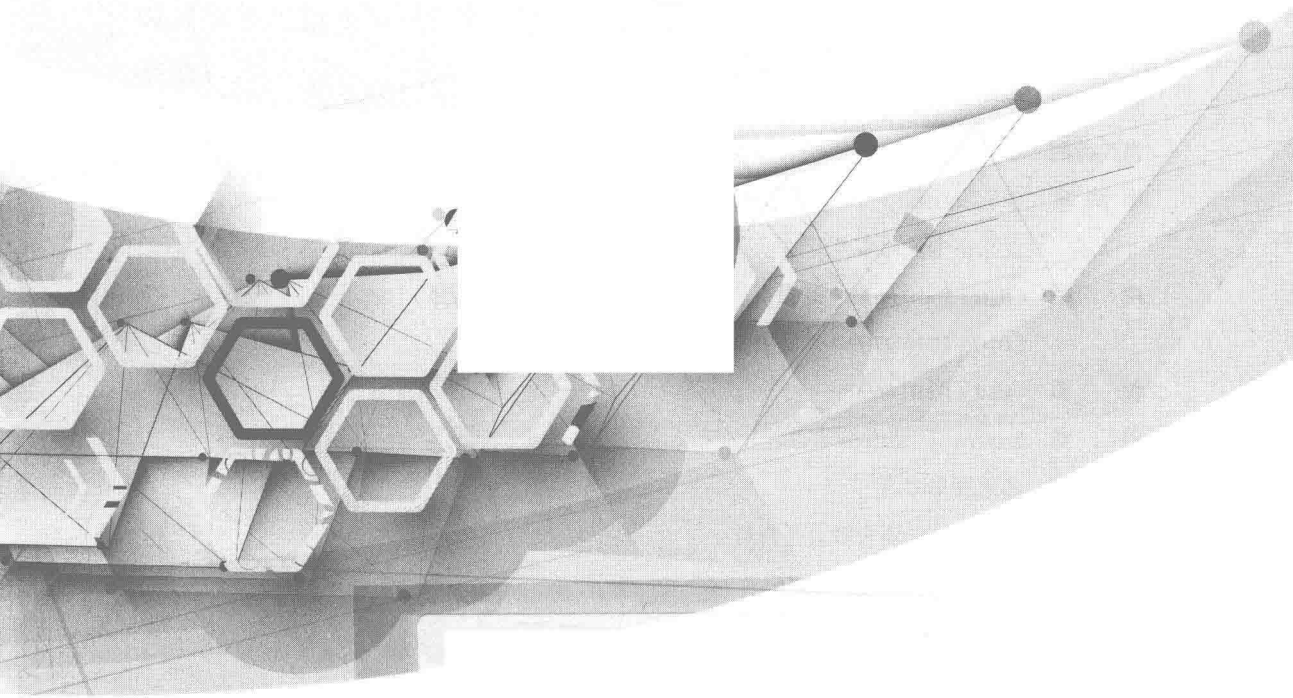
梅步俊 著



中国农业科学技术出版社

动物育种中的统计计算
——Julia语言应用

梅步俊 著



中国农业科学技术出版社

图书在版编目（CIP）数据

动物育种中的统计计算: Julia语言应用 / 梅步俊著. —北京：中国农业科学技术出版社，2016. 8
ISBN 978-7-5116-2700-1

Ⅰ. ①动… Ⅱ. ①梅… Ⅲ. ①程序语言—应用—动物—育种—生物统计—计算方法 Ⅳ. ①Q953-39

中国版本图书馆 CIP 数据核字（2016）第184263 号

责任编辑　张国锋
责任校对　李向荣
出 版 者　中国农业科学技术出版社
　　北京市中关村南大街12号　邮编：100081
电　　话　（010）8210 6636（编辑室）　（010）8210 9702（发行部）
　　（010）8210 9709（读者服务部）
传　　真　（010）8210 6650
网　　址　http://www.castp.cn
经 销 者　各地新华书店
印 刷 者　北京富泰印刷有限责任公司
开　　本　787mm × 1 092mm　1/16
印　　张　21
字　　数　490千字
版　　次　2016年8月第1版　2016年8月第1次印刷
定　　价　78.00元

版权所有·翻印必究

本书的出版发行得到国家自然科学基金“畜禽全基因组关联分析中基因间交互作用检测方法研究”（项目批准号：31460594），国家留学基金委项目“畜禽全基因组关联分析中统计问题研究”（项目批准号：201308155140），河套学院教学研究项目“试验统计学在线智能化考试系统的设计与开发”（项目批准号：HTXYJZ14005）资助。

PREFACE

前 言

现代动物育种中涉及大量统计问题。由于该领域研究对统计基础依赖性强，系统回顾并梳理动物育种中的统计方法有助于研究者把握这些方法的发展脉络，汲取前人的经验、智慧和教训。统计方法使家畜育种完成了从艺术到科学的变革，在这一过程中，许多科学家做出了杰出贡献。大多数家畜育种问题涉及一系列的定量分析方法和纷繁的数学、统计学计算。例如，选种选配过程可以看作是一个决策问题，可以用线性规划求解；在海量的基因表达数据中挖掘出有生物学意义的基因表达模式实际上是模式识别问题，可以使用聚类分析；预测家畜未来的生产性能或育种值是典型的统计推断问题，育种学家通常使用Henderson的理论解决此类问题。目前，家畜育种中的统计方法依然是许多学术会议的重要议题之一。

一、发展初期成果

将统计方法应用到家畜遗传育种的历史最早可以追溯到Galton（1822–1911）和Pearson（1857–1936）的研究，这些工作实际上早于孟德尔定律被重新发现。1889年，Galton在研究祖先与后代身高之间的关系时发现，身材较高的父母，他们的孩子也较高，但这些孩子的平均身高并没有他们父母的平均身高高；身材较矮的父母，他们的孩子也较矮，但这些孩子的平均身高却比他们父母平均身高高。Galton把这种后代的身高向中间值靠近的趋势称为“回归现象”。Galton的这项研究为“遗传力”和“预期选择反应”等概念奠定了基础。两个极端亲本群体性状平均值的差异类似于选择差，其子代群体平均值的差异等于选择反应。后代和亲本之间的统计回归是遗传力，Falconer（1913–2004）将选择响应（即遗传获得量GS）对选择差的比值称为现实遗传力。同时，Galton的工作也促进了线性模型在动物育种中的应用，即便到21世纪，动物育种中使用的主要还是线性模型。但是使用非线性模型重新分析Galton的数据发现：父-女、父-子、母-女和母-子身高的回归在170.18~172.72cm处有弯曲。这也说明在不知道明确原因的情况下，依然可以使用统计遗传模型准确地估计遗传参数。

此为试读，需要完整PDF请访问：www.ertongbook.com

Pearson一生写了大量关于性状进化的论文，Henderson在此基础上发表了预测选择偏差的著名论文。Pearson关于选择如何影响群体方差—协方差结构的论文深刻的影响了Henderson，Henderson发展了在正态分布假设和特定选择强度条件下，如何计算方差减小的公式。选择对遗传方差的影响被称为“Bulmer”效应。但是Pearson的公式只是近似值，只适合于候选个体没有亲缘关系和理想分布的情况。但家畜育种中，候选家畜间往往有亲缘关系，信息量也不相等。如参加后裔测定的公畜可能有几千个有记录的后代，而青年公畜往往没有任何后裔生产记录。因此Pearson只提供了比较理想选择方案时的近似公式。

历史上，遗传学面临的一个重要问题是如何统一连续变异的性状和孟德尔性状。Toyama Kametaro（1867—1918）在研究家蚕时发现了第一个动物中的孟德尔性状；Yule（1871–1951）第一次统一了连续变异和孟德尔性状，虽然他的观点Pearson并不认同。Fisher和Wright无疑是现代家畜数量遗传学的重要奠基人，他们也是数量遗传学历史上著名的Fisher–Wright学术论战的当事人。Fisher（1930）提出了数量遗传学中广泛使用的无穷小模型（infinitesimal model）和方差分析。Wright（1921）使用通径分析和相关分析，提出了近交系数（F）；他还推导出孟德尔群体的特性，还包括存在突变的情况下，有限群体随机交配时等位基因频率的分布。Wright还将物理学中描述扩散现象的Fokker-Planck方程（也称为Kolmogorov向前方程）引入群体遗传学。

Fisher的无穷小模型在动物育种中居于重要地位。假设有K个位点，个体的位点k(k=1,2,⋯,K)贡献A等位基因效应a_k（固定值）到基因型值u（加性值）：

$$u = W_1 a_1 + W_2 a_2 + \cdots + W_K a_K$$

此处，W是随机指示变量，0、1、2对应该位点的aa、Aa和AA。如果群体处于哈代温伯格（HW）平衡，三种基因型的频率分别为$(1-p_k)^2$、$2(1-p_k)p_k$和${p^2}_k$，这里p_k为位点k随机抽取A等位基因的概率。u的边缘分布依赖于K个位点的联合基因型概率分布。由于u是随机变量的线性组合，如果W是相互独立的（基因型间连锁平衡），随着K的增加，u的分布收敛于正态分布，但是连锁不平衡（LD）会降低收敛率。因为u的均值和方差是有限的，$K \to \infty$时单个位点的效应和频率一定变得无限小，取极限时$u \sim N\left(m, \sigma_u^2\right)$，此处$m$的典型值为0，${\sigma^2}_u$是加性遗传方差（多基因）。Wright使用相关分析，Malécot使用概率计算分别建立了“配子相似”概念。在此基础上，20世纪60年代Henderson提出奶牛的动物模型，这个模型实际是Fisher模型的向量扩展形式，加性效应u变为育种值向量$\boldsymbol{u}$，加性遗传方差${\sigma^2}_u$变为$A{\sigma^2}_u$，此处A是个体间没有近交情况下的加性关系矩阵。A矩阵也可以反映亲缘关系，其元素是两个个体随机抽取一个位点，其等位基因是血缘同源（Identity By Descent，IBD）概率的2倍。

育种值概念的提出也得益于Fisher的另一项贡献，即位点平均基因替代效应。Lush在其家畜育种学课上讲授了这一概念，后来Falconer也在其《数量遗传学导论》一书中介绍了

它。和上面相同，假设K个位点处在哈代温伯格（HW）平衡状态，显性效应d_k，$1-p_k=q_k$，u的平均值为$E(u)=\sum_{k=1}^{K}\left[a_k(p_k-q_k)+2d_kp_kq_k\right]$。$k$位点平均基因替代效应为$\alpha_k=a_k+d_k(q_k-p_k)$，其$AA$、$Aa$和$aa$的育种值分别为$2q_k\alpha_k$、$(q_k-p_k)\alpha_k$、$-2q_k\alpha_k$。个体育种值$u$为所有位点育种值之和。育种值依赖于HW假设，其计算公式中的频率和显性偏差是不独立的。因此一般情况下，只有加性效应可以遗传给后代，育种学家最感兴趣的也是a_k，狭义的u是只包含加性效应的随机变量（无穷小育种值），可以被定义为所有a_k之和。在基因组学出现以前，由于观察不到基因和等位基因效应，推断育种值是传统育种学的核心问题。直到今天，将数量遗传学应用到家畜育种实践时也很少考虑基因，统计方法在家畜育种学中依然起着重要作用，在广泛应用的Henderson方法中，也只有A矩阵考虑遗传（基因）因素。即使在基因组时代，由于使用标记检测QTL需要投入大量经费，企业没有利润可言，因此目前对单个基因对复杂性状的影响依然知之甚少。

二、动物育种中主要问题

在缺乏性状的遗传背景知识时，数量遗传学可以作为获得家畜遗传价值概括性评价的基础。随着人类对生物体代谢途径、基因网络和基因组结构等知识的不断增加，传统数量遗传学方法就略显简单。由于性状之间遗传和环境因素的关联性，我们要使用统计方法合理的分析影响选择的多种效应，就必须使用复杂的多元分析方法。Ronald Fisher（1890—1972）奠定了自然选择的基本理论。动物育种学认为选择进展和加性方差-协方差成正比，在这一观点的启发下，Alan Robertson（1920—1989）进一步发展了自然选择理论，Crow、Kimura和Edwards在文章中给出了该理论较为容易理解的描述。统计方法也是这些自然选择理论的基础，模型的参数估计强依赖于加性遗传假设前提。如果存在非加性遗传变异，为了在模型中考虑未知基因间复杂的交互作用，许多理论的假设都是不切实际的。由于小群体和选择导致的LD使剖分遗传方差组分变得很困难。如果基因网络正好处在LD中，推断特定基因对遗传方差的贡献也会变得很麻烦。变异可能产生于直接的代谢途径，也可能间接来源于由LD引起的基因间的相关性。群体遗传学创始人之一的Sewall Wright（1889—1988）引入通径分析来区分直接效应和间接效应，但是这种方法实际上需要考虑基因之间相互关系的背景知识。

现在，我们使用生产性能记录、系谱记录和分子标记信息研究性状的遗传基础，推断家畜遗传价值，寻找基因组区域和表型之间的关联性（即基因组选择）。动物育种中常见的生产性能数据包括：肉用家畜的生长率、采食量；绵羊和山羊的剪毛量和品质；乳用家畜的产奶量、乳成分、繁殖性能和长寿性；多胎品种（如鸡和猪）的产蛋量和产仔数。家畜患病记录（如奶牛乳房炎）往往很难获得，常使用替代变量进行研究，如牛奶的体细胞

数（SCC）、体表的寄生虫数量。其他性状，如生存或长寿性状可用删失数据统计方法来处理，即只知道家畜在t时刻存活，t时刻以后的状态未知；再比如计数性状（如产仔数）或分类性状（如产犊难易性，疾病发展阶段）。因此，家畜育种中的统计模型除使用正态分布外，也使用其他分布，如使用双指数或t分布可以使分析更具鲁棒性。

三、动物育种中主要方法

现代育种学之父Lush（1896—1982）认为：可能所有的基因都影响复杂性状。即使在基因组学飞速发展的今天，我们依然不太清楚大多数复杂性状的基因数量，基因的作用机制、等位基因频率及效应等。统计方法将基因组对某个表型的全部效应概括为“基因型值”。表型可由一些数学模型来表示，其中最重要的就是模型中的加性遗传值部分，也被称为育种值。但是，遗传值或模型的其他组分不能被直接观察到，必须由家畜个体及其亲属数据来推断。因为线性模型易于使用，较非线性模型计算强度小，结果便于解释、应用，所以家畜育种中的统计推断过程往往使用线性模型。如果使用大量的基因组标记，理论上可以由此计算家畜的分子相似性，而不再需要详细的系谱记录。但是标记的基因组相似性并不能完全代表致因变异的遗传相似性，除非标记和QTL间有强的LD。QTL也是表示基因组区域和表型有统计显著性关系的抽象概念。动物育种中的标记辅助推断可能最早是Neiman-Sorensen 和Robertson在分析牛群体变异时提出的。

虽然许多性状是多基因遗传模式，但是标准的全基因组关联分析（GWAS）却基于表型和单个标记间的回归分析。GWAS结果往往不会出现大量的统计显著性变异，只能解释部分性状变异。不能拒绝GWAS中的零假设往往被认为是多基因模型的佐证，但是从因果论证的角度看是不充分的。动物育种数据集可能非常大，如奶牛泌乳记录，且是多元变量（同一模型同时考虑多个性状），多数变量是正态分布（牛奶中的体细胞数浓度和乳房炎指标对数变换后近似为正态分布），但是少数为非正态分布（如离散性状）。数据结构为横断面或纵向数据（肉鸡生长曲线），而且极度不平衡，存在不随机缺失数据。例如，由于选择、生殖障碍或疾病，有第一泌乳期数据的奶牛不一定有第二泌乳期数据。由于一些优秀公牛有更多的后代，数据不完全是随机的，遗传效应的真值不能从环境效应中完全区分出来。家畜育种中的另一个难题是限性性状。

Lush首先将数学模型用在动物育种中，他使用通径分析处理模型中的隐变量。动物育种中的模型往往包括固定效应和随机效应。随机效应包括无穷小模型的u，或加性遗传模型的显性和上位效应，群效应、重复测量数据的永久环境效应、窝效应。随机效应是表型之间相关和重复测量数据之间相关性的原因。随机效应的分布由遗传和环境因素的分布参数（方

差和协方差）决定。可以将公畜作为固定效应也可以作为随机效应，除非公畜完全近交，公畜的育种值是固定值，但形成配子时不同的等位基因是随机抽样的，会导致遗传上不同的后代。将公畜作为随机效应可以估计育种值，估计的均方误差更稳定，减少预测的过拟合，甚至可以估计没有记录个体的育种值。动物模型中需要估计育种值的个体超过样本数，在基因组时代情况依然一样。但基因组分析模型与数量遗传基本假设本质上是有冲突的，基因组分析模型使用固定的基因型数据和随机标记效应。大多数动物育种模型认为数据是正态的，有大量的加性基因和微小的替代效应。但是如果认为有无限多的位点或等位基因，发现显著效应的概率就应该是0，但是这明显与分子生物学结果不符，所以MAS（辅助标记选择）将QTL概念引入到动物育种中。

理论上有两种非加性基因效应，显性和上位效应。Comstock和Robinson提出北卡罗林那设计Ⅰ、Ⅱ、Ⅲ估计基因平均显性效应。实际育种中，显性效应主要应用在交配方案问题。但是当显性效应作为随机效应时，因为难以收集携带两个家系等位基因的亲属数据，如全同胞或堂（表）兄妹数据，所以很难获得精确的方差估计。在非近交情况下，加性方差可由A阵构建的显性关系矩阵估计，在近交情况下计算较为复杂。杂交品种往往使用固定效应模型，也可以使用SNP标记估计显性基因组方差，但是由于标记不等于QTL，标记显性方差和遗传方差是有区别的。假设两个等位基因之间无显性，且处于哈代温伯格平衡和LE状态，表型y和两个位点等位基因数的线性回归模型为：

$$E(y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2$$

此处X_1和X_2表示给定位点A等位基因的数量，$E(.|.)$是条件期望。如果回归系数β_{12}为0，则模型变为加性模型。位点1的等位基因替代效应为：

$$\frac{\partial E(y \mid X_1, X_2)}{\partial X_1} = \beta_1 + \beta_{12} X_2$$

上式表示其决定于位点2的拷贝数。整个群体该性状的平均值为：

$$\left[E(y \mid X_1, X_2)\right] = \beta_0 + 2p_1\beta_1 + 2p_2\beta_2 + 4p_1p_2\beta_{12}$$

因此

$$\frac{\partial E(y)}{\partial p_1} = 2(\beta_1 + 2p_2\beta_{12})$$

和育种值类似，上位效应也依赖于等位基因频率。除非β_{12}非常大，当一个等位基因为稀有基因时，基因频率的改变对平均值的影响主要依赖于加性效应项。即使上位效应对性状有影响，大部分遗传方差也是加性的。因为复杂性状实际上是不同基因编码的酶协同代谢反应的结果，Michaelis-Menten动力学表明底物浓度和反应速率之间是非线性关系，并以非线性

方式影响基因产物。近来的文献报道了使用基因组数据发现数量性状中大量基因上位作用的证据。研究中轻易忽略高阶上位作用是不正确的，Taylor和Ehrenreich报道酵母中五个基因之间的交互作用。但是Hill等指出大量上位作用的上位方差非常小，可能的原因是：如果上位作用具有重要的生物学意义，但是上位效应方差却小于加性效应方差的原因可能是方差组分解释遗传结构的能力是有限的。Lush指出因为基因间的重组，所以针对上位效应的选择是无效的。因此，育种学家也主要关注育种值对遗传进展的影响，而忽略上位作用在育种中的作用。虽然，Fisher早已提出上位作用的概念，但直到Cockerham和Kempthorne才将这种交互作用剖分为上位组分。Cockerham使用正交多项式，Kempthorne使用IBD概率，他们假设在大的随机群体，且不存在连锁的情况下研究上位作用。上位方差依据影响性状的位点数，可以被剖分为若干正交组分。例如两个位点时，上位方差是加性×加性、加性×显性、显性×加性、显性×显性效应之和。Henderson使用以上结论推断显性和上位遗传效应，并且用BLUP预测总的遗传值。

20世纪60年代，许多家畜或家禽的母体遗传效应逐渐引起育种学家的兴趣。20世纪80年代，动物育种学的主要研究内容是不同环境的方差异质性。表观遗传学一直没有引起统计家畜育种学家的注意，但是Neugebauer建立了以系谱为基础的模型，考虑了父系和母系印记加性效应及其协方差，发现基因组印记可以解释高达25%的加性方差。

四、目前的主要成就

Lush使用通径系数，建立了评估奶牛公畜遗传值的公式，该模型假设遗传和环境方差是已知的。Robertson研究表明Lush的统计量是群体信息和数据的加权平均值，实际上体现了贝叶斯统计思想。假设公畜的传递力（TA）为$s\sim N(m,v_s)$，如果公畜有n个后代其平均生产性能减去群体平均值为$\bar{y}-\mu$，TA的估计为加群平均数：

$$\hat{s}=\left[\frac{1}{\vartheta_s}+\frac{n}{\vartheta_e}\right]^{-1}\left[\frac{1}{\vartheta_s}m+\frac{n}{\vartheta_e}(\bar{y}-\mu)\right]$$

$$=m+n(n+\alpha)^{-1}(\bar{y}-\mu-m)$$

此处，$\alpha=\dfrac{4-h^2}{h^2}$，$h^2$为狭义遗传力。上式是公畜TA条件分布的平均值，在已知后代记录时，回归系数$b=n\left(n+\dfrac{v_e}{v_s}\right)^{-1}$依赖于公畜信息数（$n$）和不确定性测度$v_{cond}=\left[\dfrac{1}{v_s}+\dfrac{n}{v_e}\right]^{-1}=v_e\left[n+\dfrac{v_e}{v_s}\right]^{-1}$，等于$Var(s-\hat{s})$，Henderson将其称为预测误方差，此外Henderson引入了最佳预测（BP）、最佳线性预测（BLP）、最佳线性无偏预测

（BLUP）。估计遗传参数常用的方法有最小范数二次无偏估计、ML和REML等。

20世纪，频率学派和似然函数为基础的方法在动物育种中居于主要地位。随着MCMC方法的出现，贝叶斯方法的灵活性和功效也体现在动物育种中。最著名的MCMC方法是Gibbs，虽然其只适用于某些特定的情况。Sorensen使用Gibbs模拟选择过程中加性遗传方差的变化。随后，贝叶斯方法被用在遗传学的许多领域，如基因定位、QTL检测、群体分化、系统发育分析、序列比对和动植物基因组选择。一些非线性方法被用来分析动物育种中的分类或计数性状、生存数据和纵向数据。鲁棒分布和混合模型（Mixture Model）也出现在动物育种研究中。除了在科学研究中，实践中动物育种极少出现完全随机交配的情况，群体在历史上的选择过程也不完全知道。选择和选配如何影响遗传参数估计和预测育种值依然是育种中的重要问题。

随着基因组测序技术的发展，大量的二等位基因标记数据出现，动物育种学进入基因组选择时代。Meuwissen提出了基因组选择的Bayes A和Bayes B方法。通过将数据分为训练集（模型拟合）和测试集（预测），可以由训练集估计标记效应或遗传值，预测测试集的表型值。Meuwissen的工作为其后的贝叶斯基因组预测奠定了基础，其后又出现了一系列贝叶斯线性回归方法，如Bayesian Lasso、Bayes C和Bayes R，这些回归模型基本相同，只是标记先验分布的假设不同。Meuwissen的另一项贡献是引入交叉验证。为了结合非测序家畜数据和测序家畜数据，提出单步BLUP（SS-BLUP）。但是这些方法并没有考虑到非加性遗传方差，检测交互作用模型需要密集的计算。由于$n < p$问题或缩减模型，上位效应的回归系数接近于0，但是基因组分析比传统的数量遗传学分析存在更多的交互作用。再生核希尔伯特空间回归（RKHS）和神经网络可以利用非加性效应。实际上，BLUP和G-BLUP可以看做RKHS的特例。

五、关于本书

本书是我在美国爱荷华州立大学（ISU）动物科学系访学期间所著，较为系统的收集了一些国际上该领域最新的科研、教学成果。在美国期间，我有幸聆听了多位ISU教授的课程。几位教授深入浅出的讲解加深了我对动物遗传育种学科知识体系的了解，在庆幸自己专业上有所收获的同时，也深深地感觉到我国动物遗传育种教学与科研方面的相对落后。与国内的动物遗传育种专业课程相比，ISU的课程更为详实，也更注重知识间的内在联系和与科研前沿的交互，许多我们司空见惯的基本概念、原理实际上有着更为深刻的内涵；在强调学生基本功的同时，ISU的课程注重学生实际操作能力的培养，学生所掌握的知识、技能几乎和学科前沿没有距离。ISU的研究生在经过1~2年的系统培养后，多数人就已具备从事该领域

此为试读，需要完整PDF请访问：www.ertongbook.com

前沿研究的能力，其人才培养的效率之高也是国内许多高校所不能比拟的。ISU动物遗传育种学科所从事的许多研究项目往往植根于整个知识体系，来源于学科知识内在的不完善性或相互矛盾的部分，可以说其项目具有很强的“内生性”和“原创性”。这种状况与国内许多项目往往机械跟踪他人成果，依赖于生物学试剂、仪器的发展，科研项目较强的路径、资源依赖而缺乏真正创新性的状况截然不同。

我国介绍动物遗传育种统计方法的著作较少，主要是中国农业大学动物科技学院张勤教授的《动物遗传育种中的计算方法》。张勤老师计算方法的课程我断断续续地听了三遍，这门课程难度较大，再加上我天资愚钝，聆听了这门课这么多遍之后，才逐渐觉得自己在这方面的基础扎实起来。相信这本书也是国内动物遗传育种领域许多研究者的必读书籍之一。这本书我读过很多遍，每读一次都会觉得受益匪浅。但是书中的方法在理论上都很抽象，需要较高深的统计学和概率论基础，所以较难理解。相信许多初学者和我当年一样面对书中抽象的公式、算法时也会一头雾水，不知道如何用计算机语言有效地实现这些方法；张勤教授的书成书于2007年，八年来动物育种学已经从主要利用表型数据的“BLUP”世代，发展到基因组选择时代。虽然不能割裂传统育种学与基因组数据分析的关系，但是目前动物遗传育种中的许多研究内容、方法已经和以前不尽相同。基于以上这几点，我觉得有必要将近些年动物育种中新出现的统计方法做一下总结；为了弥合实际应用和抽象公式、算法之间的“鸿沟”，本书大部分内容均配有Julia语言代码。这些代码既有便于读者理解，但运行效率较低的示意性代码，也有经过一定优化的代码，并尽可能为程序增加注释，书中的许多代码可以直接用于科学研究。衷心希望本书能成为一本对广大动物育种工作者有价值的参考书。

本书的出版发行得到国家自然科学基金“畜禽全基因组关联分析中基因间交互作用检测方法研究”（项目批准号：31460594），国家留学基金委项目“畜禽全基因组关联分析中统计问题研究”（项目批准号：201308155140），河套学院教学研究项目“试验统计学在线智能化考试系统的设计与开发”（项目批准号：HTXYJZ14005）支持。由于作者水平有限，书中一定有许多错误和不足，希望广大读者不吝赐教。

梅步俊

2015年12月26日于美国ISU

目录 Contents

第一章　Julia语言使用说明

随着基因组测序技术的发展，基因组测序成本不断降低，基因组测序数据逐渐在动物育种中广泛应用，这些进展增加了我们对分子水平数量性状的遗传机理的认识，为进一步提高育种效率奠定了基础，特别是对那些使用现行的育种方法效率不高或不能获得理想改良效果的性状。然而，新的理论和方法一般都会涉及大量复杂的运算，这一方面有赖于高性能的计算机硬件设备的发展，另一方面也需要有适应动物育种特点的先进的计算方法。同时，伴随着遗传育种理论和方法的不断发展，新的计算方法也不断出现。一方面，动物育种理论和方法的发展产生了新的计算问题；另一方面，不断涌现的新的计算方法又催生了动物育种理论和方法的新发展。因此，计算技术、方法的研究一直是动物育种理论研究和应用研究中不可或缺的关键技术领域。不掌握这些技术方法，就不具备真正理解现代动物育种理论和方法的基础，也就难以开展较为深入的研究。

虽然有许多现成的软件或程序可以解决动物育种中的诸多问题，但是由于实践中会出现林林总总的计算问题，编写程序仍然是育种工作者或育种理论研究者的必备技能。同时，由于新的计算理论、技术和算法层出不穷，所以在很多情况下，没有现成的软件或程序可以实现动物育种学研究者创造新的方法或改善现有计算效果的意图。因此，掌握若干计算机语言，并能用其解决育种学问题，往往是研究动物育种学前沿问题的基础。据统计，目前仍被广泛使用的计算机语言约有91种，依据这些语言的不同特点及不同研究领域的传统，特定领域会使用不同的计算机语言。美国农业部资助的“Animal Genome”数据库项目（http://www.animalgenome.org/）收集了329种遗传学分析软件。在动物育种中，广泛使用的计算机语言有C++（包括C）、Fortran、Java、MATLAB、AWK、Python、Visual Basic、R、Perl等。这些语言可粗略的分为编译型语言和解释型语言。前者程序执行速度快，但对于一般的动物育种学研究者而言学习及编写程序的难度较大，开发周期也相对较长。后者对不同系统平台的兼容性较好，借助特定的函数库，开发特定程序的周期较短，但此类语言在运行程序时需要专门有一个解释器，每个语句都是在执行的时候才翻译，执行一次就要翻译一次，因此效率比较低。但这些区别也不能一概而论，部分解释型语言的解释器，通过在运行时动态优化代码，甚至能够获得超过编译型语言的性能。

第一节 Julia语言简介

1. 关于Julia

Julia语言是高性能、动态编译的高级计算机语言。它具有极强的灵活性，适合于解决数值和科学计算问题，拥有与传统的静态型语言相媲美的执行速度。Julia语言的开发目的是创建一个功能强大、易用性好和高效的单一语言环境。

Julia 语言受NumFOCUS资助，其创始人为若干精通Matlab科学计算的编程人员，创立此项目的初衷据称是由于不满意现有的编程工具。该项目大约于2009年中开始，目前的版本为v0.3.11，其源代码，及各种平台的可执行文件及专业编译器Juno可在http://julialang.org网站下载。Julia语言可以通过基于网页的Jupyter（IJulia）交互环境执行，方便在教学等情景下展示执行结果。Julia是新的高性能、编译型、动态交互式的高级编程语言，Julia 集中了许多计算机语言的优点，“它拥有类似于C语言一样的执行速度，拥有如同Ruby语言的动态性，又有Matlab般熟悉的数学记号和线性代数运算能力，兼具像Python般的通用性，又像R语言一样擅长于统计分析，并有Perl般处理字符串的能力和shell等胶水语言的特点，并易于学习”。目前已有多所国际知名高校的数值计算或统计学课程结合Julia语言进行讲解，如斯坦福大学的“应用矩阵方法（Introduction to Matrix Methods；课程代码：EE103）”和麻省理工学院的“线性代数（Linear Algebra；课程代码：18.06）”。爱荷华州立大学动物科学系2015年5月在其开设的“家畜基因组预测（Genomic Prediction in Livestock）”短期课程中结合Julia语言进行了讲解。使用七种标准检查程序，Julia语言的运行速度接近于C及Fortran语言（图1-1），但其编写数值计算程序的速度却快得多。一般情况下，Julia语言运行数值计算程序时的速度也接近于C++，是R语言速度的100倍，MATLAB语言的1 000倍。

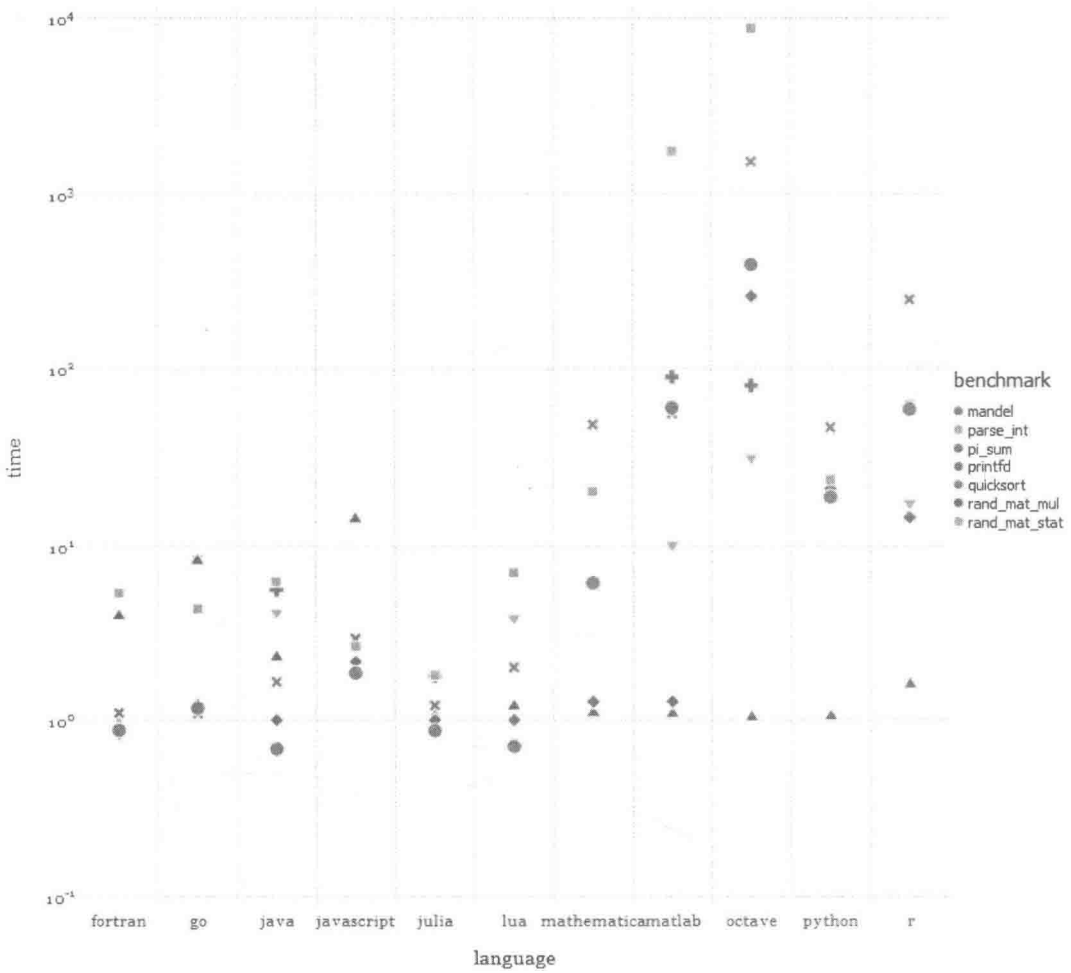


图1-1 11种常见计算机语言相对于C语言运行标准检查程序时间

注：设C语言的运行时间为1.0；C和Fortran语言使用gcc 4.8.2进行编译；C、Fortran和Julia使用OpenBLAS v0.2.13；Python运行rand_mat_stat和rand_mat_mul使用NumPy v1.9.2库函数。

2. Julia语言的下载

Julia语言是属于GNU系统的一个自由、免费、源代码开放的软件，是一个主要用于科学计算的高性能语言。作为一个免费软件，它有Windows、Linux（Ubuntu和Fedora等）、Mac OS X版本，均可免费下载和使用。Julia的官方网站是http://julialang.org/。在官方网站可以下载到Julia的安装程序、外挂程序、文档和课程（图1-2）。Julia的安装程序中只包含基础模块，其他外在模块可以通过http://pkg.julialang.org/获得。

此为试读，需要完整PDF请访问：www.ertongbook.com