博思育种
Brains Breeding

主办单位：北京中农博思科技发展有限公司　主编：李奉令　2013 年 2 月 19 日 星期二

《博思育种》，是博思公司推出的一份免费电子版育种技术资料，主要内容专注于数量遗传学、群体遗传学、统计遗传学、分子遗传学、生物信息学、生物统计等实用性育种技术的应用，欢迎浏览，欢迎指教。

本期目录

一 实际育种工作与数量遗传学对接应用思考

1、配合力

数量遗传学中的配合力，给人感觉不伦不类，从遗传学的角度，非要给配合力一个遗传学的解释。而这样的解释又总是可以给出的，一般配合力对应基因型加性效应，特殊配合力对应基因型非加性效应。

配合力从模型上来说，是线性模型，是基于产量平均数的拆分。这一点与统计学中方差分析的线性模型是一脉相承的。

配合力是育种工作实践需要的一种数学方法，在很大程度上，是跟育种试验设计有关联的一种数学分析思路。一个父本与众多母本杂交，看哪一个组合表现更好，也就是说，父本是确定的，来看母本是谁更合适。这时候，我们用配合力的思路来找母本。

一个母本自身表现很好，最终由于配合力不理想淘汰了，其实质是跟其杂交的父本，两个材料杂交，表现不理想，责任是双方的。这个母本真不见得不好。试想，沿着母本可能好，再找别的父本来测试，这样的育种思路很自然，但是这样的话，育种材料选择方向就容易模糊，迷失，这样后面的育种工作怎么做下去呢。坚守一方，执着的选另一方，虽然是无奈之举，亦可算是明智的。

在早代，测配一下配合力，也是保证后面的育种方向性。原因很简单，因为父本是预先就立住了的。

2、平均数和方差

做育种工作，平均数和方差都有用，但怎么用就是个很复杂的问题。为什么说这样说呢？

平均数简单，而且对平均数的理解也是不容易出错的。配合力就是围绕平均数在做文章。

方差是离差的平方和。既去除了离差方向性，又放大了离差。

3.5 , 4 , 5.5

平均数 = 4,

3.5 - 4 = -0.5

5.5 – 4 = 0.5

方差 = ((-0.5)2 + (0.5)2) / (3-1) = 0.25

数量遗传学里有 n 多的关于方差的名词，

表型方差，遗传方差，加性遗传方差，显性遗传方差，上位性遗传方差

数量遗传学将统计学中线性模型的方差剖分思路，应用到遗传学。

统计学中，方差分析的目的，一是看看剖分出来的效应有没有达到统计学上的概率显著标准，如果达到显著了，如，常用的 5%或者 10%的显著水平，就可以基于 95%或者 90%概率说有“效应”，二是进行样本内部的多重比较，看看究竟是哪些个体表现突出，来确定效应的发生原因，或者说将效应对应到表现突出的个体上去。

模型的建立和效应剖分，需要来自于客观实践和科学分析的，统计学的方差剖分模式，

最初就诞生于 fisher 的农业田间试验，有着很好的结合基础，也易于理解和接受。

遗传学的遗传模型，

Y = G + E

表型值=遗传效应+环境效应

G = GA + GD+ GI

遗传效应=加性效应+显性效应+上位性效应

GA = 加性效应 = 育种值

按照统计的思路，建立了遗传模型，进行了模型效应的剖分，从而可以计算出一系列的效应方差出来。如果测验显著，那只能确定出存在某种效应，导致了显著的产生。很显然，这样的结果，对育种实际帮助没有多少。

方差的本意就是变异的幅度。育种者更应该清醒的看到，方差大，说明这种效应的变异幅度大，这是个分离大的群体，不稳定的群体。

考察一个性状表现，要平均数和方差一起看。

3、育种值

GA = 加性效应 = 育种值

数量遗传学中的育种值是一个有用的概念，这样的叫法，更吸引育种者的注意，更生动。

育种值仅有一半传递给后代，因为育种值是基因的效应。

育种值，可以有一半传递给子代，所以育种值可以进行育种预测。育种值由亲代传递给子代，传递了 1/2，子代再传给子子代，就只有 1/4 了。所以确定了群体内部间的亲缘关系，可以计算梳理出来自共同亲本的育种值的大小。基于育种值进行的预测，至少是有可行性的。也是可信的。

动物育种，非常重视育种值的计算。植物育种选择淘汰力度大，与动物育种存在极大不同，但是育种方法理论背景上是一致的。方法是相通的。

BLUP 可以基于亲缘关系计算育种值，可以预测子代育种值。

4、遗传力

这是一个错误的概念。

不少教材给出的遗传力的定义或描述：**亲代某一性状传递给子代的能力**。这是一个错误的表述。

Vp（表现型方差）=Vg（基因型方差）+Ve（环境方差）

某一性状的遗传力 = Vg/Vp

遗传力的定义公式是基于方差的。方差是用于度量群体离散程度的参数。离散程度越大，方差越大。遗传方差（基因型方差）大，遗传力就大，可遗传方差大，不能等同“亲代某一性状传递给子代的能力”就大。这样的表述，会极大的模糊育种者的判断。事实反而是，遗传方差大，亲代的某一性状更不容易传递给子代，因为这一性状很离散，很不稳定。

遗传变异率，这样的名词更适合，

一个考察的性状，所测量观测值的遗传变异率大，其本质就是，观测样本中，这一性状

的遗传变异丰富，从而使得该性状观测值离散明显。

如果举个实例，就是两个纯系杂交，p1xp2，F1 基因型一致，F2 基因型分离，F1 遗传方差（基因型方差）=0，但是，两个亲本的基因 100%传递到了子代，F2 基因型方差<>0。

一个群体某性状的遗传变异率大，只是说明在该研究群体中，该性状有不同的观测表现，比如，株高高低不同，花色不同，这样从该群体中发现创新材料的可能性高，选择到变异材料的可能性高。

无论遗传学还是实际育种实践中，大家都认同，两个纯系材料，杂交后代，F1 基因型一致，生产上表现也一致，F2 代出现分离，这个分离，遗传学上就是同源染色体的分离自由组合导致的基因型变化，而数量遗传学上就对应的是遗传变异率的计算。至于基于遗传变异率的各种育种预测和材料选择，我想从理论研究上也许有意义，实际育种中，很难说多么可信。如果一种选择育种材料的方法用到了“遗传力”这个参数，那你要多考虑一下，这种方法的可靠性了。

配合力在实际育种实践中的大行其道，本身就恰好佐证了这一点。配合力是着眼育种值，或者说是参照育种材料的育种值建立的一种研究育种材料的方法。配合力反应的是一种实实在在的能力。配合力高低，仅能用于衡量配对材料自身。

二 应对育种试验数据缺失的策略

几种常见的缺失数据插补方法

1、个案剔除法(Listwise Deletion)

最常见、最简单的处理缺失数据的方法是用个案剔除法(listwise deletion)，也是很多统计软件（如 SPSS 和 SAS）默认的缺失值处理方法。在这种方法中如果任何一个变量含有缺失数据的话，就把相对应的个案从分析中剔除。如果缺失值所占比例比较小的话，这一方法十分有效。至于具体多大的缺失比例算是“小”比例，专家们意见也存在较大的差距。有学者认为应在 5%以下，也有学者认为 20%以下即可。然而，这种方法却有很大的局限性。它是以减少样本量来换取信息的完备，会造成资源的大量浪费，丢弃了大量隐藏在这些对象中的信息。在样本量较小的情况下，删除少量对象就足以严重影响到数据的客观性和结果的正确性。因此，当缺失数据所占比例较大，特别是当缺数据非随机分布时，这种方法可能导致数据发生偏离，从而得出错误的结论。

2、均值替换法(Mean Imputation)

在变量十分重要而所缺失的数据量又较为庞大的时候，个案剔除法就遇到了困难，因为许多有用的数据也同时被剔除。围绕着这一问题，研究者尝试了各种各样的办法。其中的一个方法是均值替换法(mean imputation)。我们将变量的属性分为数值型和非数值型来分别进行处理。如果缺失值是数值型的，就根据该变量在其他所有对象的取值的平均值来填充该缺失的变量值；如果缺失值是非数值型的，就根据统计学中的众数原理，用该变量在其他所有

对象的取值次数最多的值来补齐该缺失的变量值。但这种方法会产生有偏估计，所以并不被推崇。均值替换法也是一种简便、快速的缺失数据处理方法。使用均值替换法插补缺失数据，对该变量的均值估计不会产生影响。但这种方法是建立在完全随机缺失（MCAR）的假设之上的，而且会造成变量的方差和标准差变小。

3、热卡填充法（Hotdecking）

对于一个包含缺失值的变量，热卡填充法在数据库中找到一个与它最相似的对象，然后用这个相似对象的值来进行填充。不同的问题可能会选用不同的标准来对相似进行判定。最常见的是使用相关系数矩阵来确定哪个变量（如变量 Y）与缺失值所在变量（如变量 X）最相关。然后把所有个案按 Y 的取值大小进行排序。那么变量 X 的缺失值就可以用排在缺失值前的那个个案的数据来代替了。与均值替换法相比，利用热卡填充法插补数据后，其变量的标准差与插补前比较接近。但在回归方程中，使用热卡填充法容易使得回归方程的误差增大，参数估计变得不稳定，而且这种方法使用不便，比较耗时。

4、回归替换法(Regression Imputation)

回归替换法首先需要选择若干个预测缺失值的自变量，然后建立回归方程估计缺失值，即用缺失数据的条件期望值对缺失值进行替换。与前述几种插补方法比较，该方法利用了数据库中尽量多的信息，而且一些统计软件（如 Stata）也已经能够直接执行该功能。但该方法也有诸多弊端，第一，这虽然是一个无偏估计，但是却容易忽视随机误差，低估标准差和其他未知性质的测量值，而且这一问题会随着缺失信息的增多而变得更加严重。第二，研究者必须假设存在缺失值所在的变量与其他变量存在线性关系，很多时候这种关系是不存在的。

5、多重替代法(Multiple Imputation)

多重估算是由 Rubin 等人于 1987 年建立起来的一种数据扩充和统计分析方法，作为简单估算的改进产物。首先，多重估算技术用一系列可能的值来替换每一个缺失值，以反映被替换的缺失数据的不确定性。然后，用标准的统计分析过程对多次替换后产生的若干个数据集进行分析。最后，把来自于各个数据集的统计结果进行综合，得到总体参数的估计值。由于多重估算技术并不是用单一的值来替换缺失值，而是试图产生缺失值的一个随机样本，这种方法反映出了由于数据缺失而导致的不确定性，能够产生更加有效的统计推断。结合这种方法，研究者可以比较容易地，在不舍弃任何数据的情况下对缺失数据的未知性质进行推断。

三《中国玉米品种育种分析报告（1949-2012）》

2011 年，博思公司推出了《中国玉米品种育种分析报告（1949-2011）》，引起了业内的

广泛关注，在众多用户的使用反馈基础上，我们进一步充实完善了报告的数据，2012 年版分析报告现正式推出。

《中国玉米品种育种分析报告（1949-2012）》，收录整理了 1949-2012 年间 5233 个通过审定玉米品种数据，并对其系谱进行了整理研究，梳理提炼出 3170 个有亲本系谱的自交系。报告约 60 万字，近 800 页。

为更好的分析发现我国玉米育种的历史规律和发展趋势，研制了“蓝玉米育种材料分析软件”，从而实现了从品种数据总体上把握中国玉米育种的历史脉络，找出其中的大线条，明显趋势的目标。

四 蓝玉米育种材料分析软件 2012 上市

蓝玉米 2012 版本在 2011 版本基础上进一步完善了功能，补充了新一年的数据。

如果我们明了了选育出的大品种的种质构成规律，那么对我们育种工作无疑是有重大意义的一件事。我们在热切关注国外育种大公司育种动向的同时，也要坚定自身育种工作的认识和判断。

即便需要借鉴国外先进模式，我们还是需要从脚下的路开始走出第一步。

五 中国玉米自交系直接杂交引用排名 10 大亲本

《中国玉米品种育种分析报告（1949-2012）》 报告内容分享

1、我国玉米审定推广品种在 2006 年出现 1 个高峰，计 548 个，之后下降趋势明显，基本呈逐年下降萎缩态势，2011 年 207 个。

2、玉米审定 10 大省份是，吉林，辽宁，内蒙，河北，四川，山东，山西，北京，贵州，广西。

3、自交系直接杂交引用利用排名 10 大亲本：Mo17（99），吉 853（90），昌 7-2（69），自 330（69），丹 340（68 ），丹 598（58 ），黄早四（55），掖 478（48），k12（45），铁 7922（37）。

4、SS-NSS 理论得到玉米育种数据分析的支持，历史上大品种，好品种多位于 SS-NSS 差值为 0 值附近。

5、自交系分析结果表明，SS-NSS 理论对玉米育种具有指导性作用。

六 大品种 SS-NSS 数据

《中国玉米品种育种分析报告（1949-2012）》部分品种数据摘录，SS-NSS 差值趋近 0 的，不少是大品种。事情胜于雄辩，SS-NSS 理论是指导玉米育种的好理论。

品种	SS-NSS
农大 108	0.0050
豫玉 23	0.0045
安玉 5 号	0.0045
吉单 46	0.0040
吉星 46	0.0040
掖单 12 号	0.0035
吉单 122	0.0035
新玉 8 号	0.0035
郑单 17 号	0.0030
辽单 35 号	0.0020
辽 613	0.0020
龙单 13	0.0020
吉单 27	0.0015
吉引 704	0.0010
丹玉 46 号	0.0010
金海 6 号	0.0010
铁单 17 号(铁 9807)	0.0010
陕单 902	0.0010
丹玉 15 号	0.0005
陕单 5 号	0.0005
陕单 8806	0.0005
掖单 4 号	0.0000
通吉 100	-0.0005
农大 2238	-0.0010
四单 111	-0.0010
掖单 2 号	-0.0015
陕单 911	-0.0020
公引 5 号	-0.0025
丹玉 13 号	-0.0025
鲁单 984	-0.0025
郑单 18 号	-0.0030
农大 0638	-0.0040
吉单 209	-0.0040
九单 209	-0.0040
阿丹 10 号	-0.0060
吉东 2 号	-0.0060

中单 2 号	-0.0075
郑单 19 号	-0.0080
农大 3138	-0.0110
吉单 133	-0.0130
陕单 8813	-0.0160
源玉 3	-0.0295
鲁单 50 号	-0.0305
中单 321	-0.0385
牡丹 9	-0.0405
吉单 303	-0.0435
本玉 9 号	-0.0445
吉新 203	-0.0470
郑单 958	-0.0475

七 向动物育种借鉴之 BLUP 估计育种值

向动物育种借鉴什么？借鉴育种思路，育种方法。

动物育种可以给我们启发和借鉴。搞玉米育种，也要拿出时间了解其他育种领域的进展和取得的成就，以便为我所用，促进自身育种的进展和突破。

自 80 年代以来，随着数理统计学（尤其是线性模型理论）、计算机科学、计算数学等学科领域迅速发展，家畜育种值估计的方法发生了根本的变化，以美国动物育种学家 C.R.Henderson 为代表所发展起来的以线性混合模型为基础的现代育种值估计方法-BLUP 育种值估计法，将畜禽遗传育种的理论与实践带入了一个新的发展阶段。目前在世界各国，尤其在发达国家,这种方法已得到广泛应用,为畜禽重要经济性状的遗传改良做出了重大贡献。

摘自 2001 年 张沅主编《动物育种学》

说到 BLUP，必须要提到张世煌老师。

2010 年，我第一次跟张老师见面的时候，张老师就跟我说要去考虑在植物育种中应用 BLUP，并形象的用“公牛产奶量的计算”来说明 BLUP 在动物育种中估计育种值的应用，并谈到了国外已经在应用了，我们又要落后了。当时我对 BLUP 的认识尚浅，粗略知道一些，因为考虑在植物育种实践中没有广泛得到采用，所以育种家软件中没有纳入进来，就没有用心钻研它。中间耽搁了一些，一直没有开始研究 BLUP，但张老师的指导意见一直没忘，有机会就关注了解一些 BLUP 的研究文章。在认识到 BLUP 在动物育种中的巨大成效后，自然看到了这一技术在植物育种上的应用潜力。对张老师的指导就完全领悟了。

2 年后，BLUP 作为一项软件功能，在育种家软件中实现了。张老师得知这一消息，还是很高兴的，我们希望用我们的努力，回报先生的无私指导，为中国育种尽点微力。

之后，参加张老师那里组织的一个育种信息化培训班，听到了马东辉博士的精彩报告，

也了解到了国际种业巨头也在向动物育种借鉴学习，招聘动物育种专业人员充实育种队伍。所以，觉得很有必要将 BULP 植物育种应用多做宣传和推广，至少引起育种工作者的注意。

向动物育种学什么？

BLUP 在动物育种中的应用是着眼于育种值的估计。

显性效应是指同一基因座内不同等位基因之间的互作，上位效应是不同基因座间的基因的互作，显性效应和上位效应统称为非加性效应，由于亲代只将其基因而不是基因型传递给后代，而显性效应和上位性效应都与特定的基因型有关，所以只有加性效应才能遗传给后代，因而对于种畜的选择来说，我们主要感兴趣的是 GA，它也常被称为家畜个体的育种值。
摘自 2007 年 张勤《动物遗传育种中的计算方法》

准确估计出育种值，就更容易把握育种的选择方向。育种值高，说明亲本所含的优良基因多，因此后代的表现优良。

数量遗传学告诉我们，育种值是亲本个体一般配合力效应的两倍。说到这，有的读者会想，哦，原来动物育种和植物育种一样没有不同嘛，我们不也是在用一般配合力特殊配合力嘛，那有什么好学的。

别急，往下看。

《植物数量遗传学》（孔繁玲主编）中对育种值给出了理论和实际两种定义：

理论的定义：一个个体的育种值 A（breeding value）就是它所携带的基因的平均效应的总和。在一对等位基因的情况下，A1A1 个体的育种值为 2a1，A1A2 个体的育种值为 a1+a2，A2A2 个体的育种值为 2a2。（这与动物育种学中对育种值的定义是一致的）。
实际的定义：如果一个个体与来自群体内的许多个体随机交配，则该个体的育种值为其子代均值与群体平均离差的两倍。

实际育种中，是使用的后一种方法在进行配合力的计算。

$$gi = \overline{yi.} - \overline{y..}$$（亲本 i 一般配合力效应）

$$Sij = yij - \overline{y..} - gi - gj$$　（个体 S 的特殊配合力效应）

在《DPS 数据处理系统》一书中，有这样的说明：

“对双列杂交设计进行配合力分析之前，首先要进行基因型间的方差分析，检验遗传型（组合）间的差异显著性，如果基因型间差异显著，然后才能进行配合力分析”。

可见，按照实际定义进行的配合力计算，是采用方差分析的思想进行的变异剖分，因此存在先进行 F 测验检验差异是否显著，显著之后才能进行配合力分析的论述，这样就不合理反而又合理了，但实际上，无论差异显著与否，谁又能否定遗传加性效应的存在呢？配合力、育种值本是数量遗传学中内容，通过方差分析的方法进行配合力分析，这是最优的分析方法吗？

为什么动物育种在用 BLUP 进行育种值的计算估计，而植物育种却在用方差分析思想

进行配合力估计，然后估计育种值呢？

《应用数量遗传》（翟虎渠 王建康）一书推荐植物育种的朋友买来读读，书中从动物育种 BLUP 介绍，到给出植物育种配合力分析的 BLUP 方法，让人耳目一新，看到了植物育种配合力分析的新进展。

比较 BLUP 配合力估计和方差分析模型配合力估计，发现二者有一个明显差别，就是 BLUP 配合力估计在线性模型建立中引入了加性遗传相关矩阵（共祖先系数矩阵）。

方差分析也是基于线性模型理论建立的分析模型，但限与从平方和变异来源进行分解，对材料亲属关系缺乏考虑。BLUP 配合力估计，引入了加性遗传相关矩阵与育种材料的亲缘关系建立了联系，可以估计出更加准确的育种值。

向动物育种学习，一个内容就是学习，动物育种中的，采用 BLUP 线性模型理论估计育种值的技术。

这也是动物育种取得成就的重要原因之一。

八 BLUP 估计育种值中植物应用与动物应用的区别

1、区别

（1）、模型不同

BLUP 动物育种值估计采用的是动物模型：

y=Xb+Za+e （a 是育种值）

BLUP 植物育种值估计采用的模型是：

y=Xb+Ug1+Wg2+Zs+e （g1 父本一般配合力，g2 母本一般配合力，s 特殊配合力）

（2）、加性遗传相关矩阵 A 的构建方法有不同

动物育种中对 A 的构建采用依据系谱的方法建立。但植物育种采用此方法会有问题。主要原因是，动物植物系谱记载所反映的内容不同。

动物系谱中，明确记录了每一代的父母关系，交配的真实记录，所以根据系谱可以很好的追溯遗传物质的传递，明晰育种后代间的后裔同样的概率。

但植物育种则不是这样，系谱通常反映的仅是一个育成材料的来源，自交系选育在起初的父母本杂交后，经历了多代的自交，并且加入了人为选择，这样就不能通过系谱去计算加性遗传相关矩阵 A。

那这样通过系谱无法构建出加性遗传相关矩阵 A，BLUP 估计育种值岂不就不能实现了，好在有分子遗传标记。

2、分子标记在 A 构建中的作用

加性遗传相关矩阵 A，反映的是测试材料间的共祖先关系，因此也叫共祖先系数矩阵。既然这样，不能通过系谱还原材料间的亲缘关系，通过分子标记技术来实现，也是可以的。

分子标记可以通过对材料的分析检测，进行材料的亲缘关系分析。张老师的博客中，就提供了基于分子标记分析结果的玉米常见自交系亲缘关系数据。那次见到张老师，张老师还特意让谢传晓博士给我们介绍了此项工作的一些情况，我们感谢张老师对我们育种家软件开发的指导、支持和帮助。

分子标记技术与常规育种总感觉距离很远，在亲缘关系界定方面，可以对常规育种起到极大的帮助，这一点，育种工作者要充分认识到并要利用上。

在张老师提供的数值化系谱中，分子标记仅给出一个自交系与 6 大群体

（来源 Reid PA LRC SPT PB Lan）的距离，

需要进行必要的数据数值加工变换才能反映自交系材料之间的亲缘关系远近。

正是上面的加性遗传相关矩阵 A 构建障碍才限制了 BLUP 在植物育种领域的应用，而动物育种却在 20-30 年的时间内，借助 BLUP 很好地推动了育种发展。

北京中农博思科技发展有限公司
地址：北京海淀区彰化路曙光宾馆写字间 202 室(北京农林科学院南门)
邮编：100097
电话：010-88435130
传真：010-88432569
电邮：nbs777@nbs.net.cn
http://www.nbs.net.cn