数据非平衡性对试验分析结果的影响*

胡 希 远
(西北农林科技大学 农学院, 陕西 杨凌 712100)

[摘 要] 利用蒙特卡罗(Monte Carlo)模拟考察了随机区组设计、裂区设计和条区设计因数据缺失形成的数据非平衡性对试验分析结果的影响。结果表明, 数据非平衡性可导致方差组分估计的精准度降低、固定效应误差的偏低估计以及固定效应测验一类统计错误率的增大。缺失数据越多, 这一现象表现越明显。数据缺失对随机区组设计分析精准度的影响小, 而对裂区和条区设计分析精准度的影响大。

[关键词] 数据非平衡性; 方差组分估计; 固定效应测验
[中图分类号] O212.1 [文献标识码] A [文章编号] 1671-9387(2004)01-0103-05

试验分析的精准度直接关系到试验的结论及其可靠性。为了提高试验分析的精准度, 需要对每一试验进行精心设计、认真实施, 并采用正确的统计分析方法。对平衡试验资料一般采用传统的方差分析方法(简称ANOVA)进行分析。用ANOVA 方法分析数据有许多优点, 如计算简便, 可构建准确的 F 测验进行统计检验, 对平衡数据能获得最佳无偏估计等。但是, 通常在试验的实施过程中, 往往由于不可预测的原因出现一个或多个观测值缺失的情况, 使原本平衡的试验设计失去了均衡性。当发生这样的情况时, 传统的ANOVA 方法不能对得到的非平衡资料进行有效分析。

对于非平衡数据, 采用 Henderson 法虽可取得各变异来源的均方值, 但因其不是独立的, 且在零假设成立时它们的期望值也不相等, 因而无法对处理效应进行统计测验。分析非平衡数据的一个可行途径是运用混合线性模型的分析原理。该分析方法不需计算均方, 通过直接估计各项随机效应的方差组分, 并进一步估算固定效应均值及其标准误, 从而进行固定效应的统计测验。但该方法的缺点是, 对小样本非平衡数据分析的精准度低。笔者利用混合线性模型分析原理, 对 3 个常用农业试验设计的小样本非平衡数据进行分析, 为考察其数据分析的效果采用了蒙特卡罗(Monte Carlo)模拟研究。

1 混合线性模型和统计分析方法

1.1 混合线性模型

混合线性模型一般具有下面的矩阵形式[1]:

$$y = X\beta + Zu + e \qquad (1)$$

式中, y 为试验观测值向量; β 为固定效应向量; u 为随机效应向量; e 为剩余效应向量; X 为固定效应试验设计矩阵; Z 为随机效应试验设计矩阵。

对随机效应和剩余效应的分布分别有: $u \sim N(0, G)$, $e \sim N(0, R)$;

对试验观测值的期望值和协方差分别有: $E(y) = X\beta$, $\mathrm{Var}(y) = ZGZ' + R = V$。

G 和 R 分别是随机效应方差组分和剩余方差组分的函数, 如果方差组分已知, 那么 G 和 R 以及试验观测值的协方差阵 V 已知, 这时固定效应 β 的最佳线性无偏估计(BLUE)可由下式给出:

$$\hat{\beta} = (X'V^{-1}X)^{-}X'V^{-1}y \qquad (2)$$

其估计的误差可由下式求得:

$$\mathrm{Var}(\hat{\beta}) = (X'V^{-1}X)^{-} \qquad (3)$$

在实际农业科学研究中, 方差组分的值未知, 需要依据样本数据估计方差组分, 试验观测值的协方差 V 也成为估计值 $\hat{V}$。此时, 由式(2)得到的固定效应估计一般不再是最佳线性无偏估计, 而是经验性最佳线性无偏估计(eBLUE), 由式(3)得到的固定效应估计的误差仅可作为误差的近似计算[2]。在这种情况下, 固定效应及其误差估计的精准度受方差组分估计特性的影响。对非平衡数据, 若仅以估计方差组分为目的, 可采用最小范数二次无偏估计(MINQUE)、最小方差二次无偏估计(MIVQUE)、极大似然估计(ML)或限制性极大似然估计(REML)等方法估计方差组分[3~6]。但在许多试验

* [收稿日期] 2003-07-02
[基金项目] 西北农林科技大学校长基金资助项目
[作者简介] 胡希远(1963-), 男, 陕西蓝田人, 副教授, 博士, 主要从事试验模型与模拟分析研究。

研究中, 试验的最终目的是要对固定效应差异显著性进行测验, 对此需要确定固定效应误差的自由度。在非平衡数据条件下, 混合模型中固定效应误差自由度一般不能由一个均方项的自由度来准确确定, 而是需要通过一定的近似计算才能得到。普遍适用的自由度近似计算方法又依赖于方差组分的估计及其估计的方差和协方差[7~9]。REML 法在提供方差组分估计值的同时, 还能给出方差组分估计的方差和协方差, 为固定效应误差自由度的近似计算提供了条件。鉴于这种情况, 在实际应用中一般采用 REML 法估计方差组分。采用该方法的另一个重要原因是, 在平衡数据条件下, REML 法可获得和 ANOVA 法相一致的估计结果[10]。但是在非平衡条件下, 该方法具有的渐进最佳无偏估计特性对小样本数据方差组分估计效果的改进帮助甚微。

1.2 固定效应显著性测验

线性对比(Linear contrast)是国际上广泛应用的一种效应比较方法, 它可以对两个或多个效应进行比较。在方差组分真值未知而采用样本估计值时, 固定效应的线性对比可根据下面的统计量进行测验[11]。

对两个固定效应的线性对比:

$$t=\frac{L\hat{\beta}}{\sqrt{L'(X'\hat{V}^{-1}X)^{-}L}}\sim t(\nu) \tag{4}$$

对多个固定效应的线性对比:

$$F=\frac{\hat{\beta}'L'(L'(X'\hat{V}^{-1}X)^{-}L)^{-1}L'\hat{\beta}}{\mathrm{rang}(L')}\sim F(\mathrm{rang}(L'),\nu) \tag{5}$$

其中, L 是描述固定效应线性对比的行向量或矩阵, $\mathrm{rang}(L')$ 表示 L 的秩。自由度 ν 由下式计算:

$$\nu=\frac{2\left(\sum_{p=1}^{k}M_p\hat{\sigma}_p^2\right)^2}{\sum_{p=1}^{k}\sum_{q=1}^{k}\mathrm{Cov}\left(\hat{\sigma}_p^2,\hat{\sigma}_q^2\right)M_pM_q} \tag{6}$$

式中, $M_p=L'(X'\hat{V}^{-1}X)^{-}X'\hat{V}^{-1}Z_pZ_p'\hat{V}^{-1}X(X'\hat{V}^{-1}X)^{-}L$; $\mathrm{Cov}\left(\hat{\sigma}_p^2,\hat{\sigma}_q^2\right)$ 为 REML 法估计方差组分的方差协方差; k 为模型中所含随机效应(方差组分)的数目; $\hat{\sigma}_p^2$ 为由 REML 法估计的方差组分值; Z_p 为随机效应试验设计矩阵 Z 中关于第 p 个随机效应的分矩阵。

式(6)中分子表示 2 倍的固定效应线性对比方差的平方, 分母表示固定效应线性对比方差的方差协方差。在这一自由度计算方法中, 利用了方差组分的估计值及其估计的方差协方差。一般情况下, 式(4)和(5)分别近似地服从 t 分布和 F 分布[12]。为区别于方差分析中的 t 测验和 F 测验, 依式(4)和(5)进行的 t 测验和 F 测验也被称作 Wald-t 测验和 Wald-F 测验[13]。

2 蒙特卡罗(Monte carlo)模拟分析

2.1 模拟的试验设计及其模型

随机区组设计:

$$y_{ijk}=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+bl_k+e_{ijk}$$

裂区设计:

$$y_{ijk}=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+bl_k+e_{ik}+e_{ijk}$$

条区设计:

$$y_{ijk}=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+bl_k+e_{ik}+e_{jk}+e_{ijk}$$

$$(i=1,\cdots,a;\ j=1,\cdots,b;\ k=1,\cdots,r)$$

式中, μ 为总体平均值; α_i 为因素 A 第 i 水平的效应(固定); β_j 为因素 B 第 j 水平的效应(固定); $(\alpha\beta)_{ij}$ 为因素 A 第 i 水平与因素 B 第 j 水平交互效应(固定); bl_k 为第 k 个区组的随机效应; e_{ik} 和 e_{jk} 为随机误差效应; e_{ijk} 为剩余误差效应。

$bl_k\sim N(0,\sigma_{bl}^2)$; $e_{ik}\sim N(0,\sigma_{RA}^2)$; $e_{jk}\sim N(0,\sigma_{RB}^2)$; $e_{jik}\sim N(0,\sigma_R^2)$。

根据试验的目的和研究结论适用范围的不同, 区组效应可看作固定效应或随机效应。在上述模型中, 区组被看作随机效应。区组随机的优点是试验结论的适用范围宽[14]。在这 3 个试验设计中, 根据试验因素水平组合 $a\times b\times r$ 的所有处理具有相同观测数目(1 个)可知, 该试验数据在试验设计时是平衡的[14]。这种平衡类型称为分类平衡(Classification balance)。在分类平衡试验设计中, 同一试验因素所有效应的方差是一固定值, 在这种意义上分类平衡试验设计也是方差平衡的试验设计。本研究考虑由于部分数据缺失形成的非平衡数据。这种非平衡数据既是分类非平衡, 又是方差非平衡。

2.2 模拟的数据结构

每一试验设计所考虑因素水平组合为 $a\times b\times r=2\times3\times3$, 所有方差组分 σ_{bl}^2, σ_{RA}^2, σ_{RB}^2 和 σ_R^2 分别取值为 1。关于数据非平衡性考虑一个和两个观测数据缺失这两种情况。

为了准确地考察固定效应测验一类统计错误率的控制情况, 固定效应测验一类统计错误率的估计须达到足够的精度。本文以固定效应测验时经验性一类统计错误率估计值 95% 的置信区间不超过测验显著水平 $\alpha=0.05$ 的 5% 为精度要求, 对所需模拟产生的数据集数进行计划, 由此得出对各试验设

计的每一数据结构应模拟产生 100 000 个数据集用于统计分析。所有数据模拟和统计分析以软件包 SAS 为环境进行编程和运算。

3　模拟分析的结果

3.1　方差组分估计

通过 100 000 次蒙特卡罗模拟计算方差组分估计的相对偏差(Bias/% = $(\hat{\sigma}^2 - \sigma^2)/(\sigma^2 \times 100)$)、均方误差(MSE = Bias + Var($\hat{\sigma}^2$))和效益系数(C.E. = $\sqrt{MSE}/(\sigma^2 + |Bias|)$)见表 1。应用这些统计量可检验方差组分估计的无偏性和有效性。这些统计量越小, 估计效果越好。

由表 1 可见, 在无缺失数据的平衡条件下, 每一试验设计中所有方差组分的估计均是无偏的, 但在非平衡条件下, 方差组分的估计存在着一定的偏差。偏差的出现和增大反映了估计准确性变差。此外, 数据的非平衡性还导致了方差组分估计均方误差和效益系数的增大, 反映了方差组分估计的精确性和估计效益变低。随着缺失数据的增加, 方差组分的相对偏差、均方误差和效益系数都相应地增大。不同方差组分以及同一方差组分在不同试验设计中, 其估计也不是等效的。剩余效应方差组分 σ^2_R 的偏差、均方误差和效益系数最小, 而区组效应方差组分 σ^2_{bl} 的偏差、均方误差和效益系数最大, 另外两个随机效应方差组分的居中。在随机区组设计中, 各方差组分的偏差、均方误差和效益系数小于裂区设计和条区设计中对应的方差组分。

表 1　数据平衡与非平衡时方差组分估计的相对偏差(%)、均方误差和效益系数

Table 1　Bias(%), MSE and C.E. of variance component estimate for balance and unbalanced data

设计类型 Type	方差组分 Variance component	相对偏差/% Bias			均方误差 MSE			效益系数 C.E.		
		0	1	2	0	1	2	0	1	2
随机区组设计 Block design	σ^2_R	0.0	-1.5	-2.3	0.45	0.46	0.48	0.66	0.67	0.69
	σ^2_{bl}	0.0	2.0	3.0	1.18	1.38	1.42	1.06	1.14	1.18
裂区设计 Split-plot design	σ^2_R	0.0	-3.0	-3.8	0.50	0.50	0.53	0.68	0.69	0.72
	σ^2_{RA}	0.0	17.2	18.3	1.35	1.37	1.41	0.97	0.99	1.17
	σ^2_{bl}	0.0	18.3	19.2	1.96	2.68	2.45	1.19	1.38	1.54
条区设计 Strip-plot design	σ^2_R	0.0	-3.2	-4.2	0.61	0.62	0.64	0.76	0.77	0.80
	σ^2_{RA}	0.0	-18.0	-19.0	1.40	1.47	1.50	1.00	1.02	1.21
	σ^2_{RB}	0.0	5.5	8.0	1.17	1.33	1.36	1.08	1.13	1.15
	σ^2_{bl}	0.0	26.5	28.3	2.36	3.19	3.23	1.21	1.39	1.76

注: 表项中的 0, 1, 2 表示缺失数据数。下表同。

Note: 0, 1 and 2 in the table imply No. of loss values. The same is in other tables.

3.2　固定效应的误差

为了探讨数据非平衡性对固定效应准确性和精确性的影响, 根据模拟得到的固定效应估计值对其偏差和标准误进行了计算。与平衡条件下相同, 非平衡条件下固定效应的估计是无偏的, 但估计的误差不同。表 2 给出了模拟计算的固定效应比较的误差。$a_1 - a_2$, $b_1 - b_2$, $a_1b_1 - a_1b_2$, $a_1b_1 - a_2b_1$ 和 $a_1b_1 - a_2b_2$ 分别代表因素 A 主效应、因素 B 主效应、因素 A 同一水平上 A × B 交互效应、因素 B 同一水平上 A × B 交互效应以及 A 和 B 两因素水平都不同时 A × B 交互效应的比较(以下相同)。从表 2 结果可知, 数据平衡时固定效应估计的误差最小, 随着数据缺失的出现和增加, 所有固定效应比较的误差增大。在无缺失数据的平衡条件下, 固定效应比较的误差在不同效应比较和不同试验设计间存在着差异。根据方差分析可知, 这种差异由固定效应误差组成的不同所致。表 2 表明, 这种差异的趋势在非平衡数据条件下依然存在, 平衡条件下固定效应误差较大的试验设计或效应比较, 其误差在非平衡条件下也较大。

表 2　数据平衡与非平衡时固定效应估计的误差

Table 2　Standard errors of fixed effect contrast for balance and unbalanced data

效应比较 Effect contrast	随机区组设计 Block design			裂区设计 Split-plot design			条区设计 Strip-plot design		
	0	1	2	0	1	2	0	1	2
$a_1 - a_2$	0.22	0.24	0.27	0.89	0.92	0.95	0.88	0.93	1.00
$b_1 - b_2$	0.33	0.37	0.41	0.33	0.37	0.42	1.00	1.07	1.17
$a_1b_1 - a_1b_2$	0.67	0.73	0.81	0.67	0.75	0.85	1.33	1.47	1.66
$a_1b_1 - a_2b_1$	0.67	0.74	0.82	1.33	1.42	1.52	1.33	1.47	1.66
$a_1b_1 - a_2b_2$	0.67	0.74	0.82	1.34	1.41	1.52	1.99	2.13	2.33

表 2 中固定效应的误差是由所有模拟得到的固定效应估计值计算得出的, 通常称其为固定效应的观测误差值, 它反映了固定效应估计误差的真实大小。在实际试验数据分析时, 固定效应的误差是根据方差组分的估计值由式(3)估算的。为了验证非平衡条件下式(3)估算固定效应误差的准确性, 在这里将式(3)估算的固定效应误差平均值与模拟观测误差值予以比较。表 3 给出了固定效应误差平均值相对于观测误差值的相对偏差, 在数据平衡时所有固定效应误差的偏差为零, 在缺失数据时均为负值。这一结果表明, 在非平衡条件下, 利用方差组分估计值估算的固定效应误差低于实际误差。同时可以看出, 与表 2 中误差大小的变化趋势相似, 缺失数据增加, 误差估计的偏差相应增大。随机区组设计中, 固定效应比较误差的偏差较小, 所有效应比较误差的偏差都未超过 5%; 在裂区和条区设计中, 固定效应比较误差的偏差较大, 不同效应比较间误差偏差的差异也较大, 其中一些效应比较误差的偏差在缺少一个数据时已超过了 5%。相对于裂区设计, 条区设计中效应比较误差的偏差有所增加。

表 3 数据平衡与非平衡时固定效应估计误差的相对偏差

Table 3 Bias % of estimated standard errors of fixed effect contrast for balance and unbalanced data

效应比较 Effect contrast	随机区组设计 Block design			裂区设计 Split-plot design			条区设计 Strip-plot design		
	0	1	2	0	1	2	0	1	2
a_1-a_2	0.0	-1.8	-2.2	0.0	-5.3	-6.5	0.0	-6.0	-7.2
b_1-b_2	0.0	-1.8	-2.2	0.0	-2.8	-4.3	0.0	-2.8	-4.5
$a_1b_1-a_1b_2$	0.0	-1.8	-2.2	0.0	-2.8	-4.3	0.0	-4.3	-7.4
$a_1b_1-a_2b_1$	0.0	-1.9	-2.1	0.0	-3.8	-5.6	0.0	-6.5	-9.8
$a_1b_1-a_2b_2$	0.0	-1.9	-2.2	0.0	-3.9	-5.6	0.0	-4.3	-6.9

3.3 固定效应检验的统计错误率

通常情况下, 固定效应差异显著性统计测验是进行试验的最终目的。在测验准确时, 测验的一类统计错误率等于给定的差异显著性水平 α, 一类统计错误率能得到控制; 但当测验不准确时, 实际的一类错误率可能不等于给定的显著性水平 α, 对一类错误率不能进行有效控制, 给定的显著性水平 α 只是统计测验一类错误率的名义值。通常取显著水平 $\alpha=0.05$ 进行差异显著性测验。图 1~3 是在不同条件下, 给定显著水平 $\alpha=0.05$ 时通过模拟得到的固定效应 t 测验一类统计错误率。

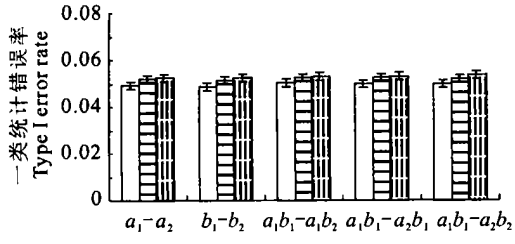

图1 随机区组设计固定效应t测验的一类统计错误率

□.平衡数据; ▣.缺失1数据; ▥.缺失2数据

Fig.1 Type Ⅰ error rate of t-test for block design

□.Balance; ▣.1 missing value; ▥.2 missing values

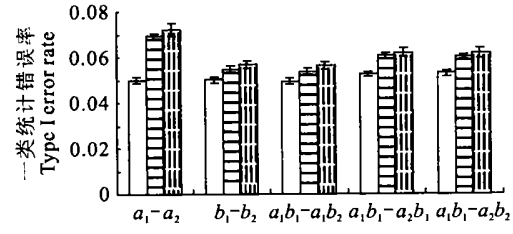

图3 条区组设计固定效应t测验的一类统计错误率

□.平衡数据; ▣.缺失1数据; ▥.缺失2数据

Fig.3 Type Ⅰ error rate of t-test for strip-plot design

□.Balance ▣.1 missing value; ▥.2 missing values

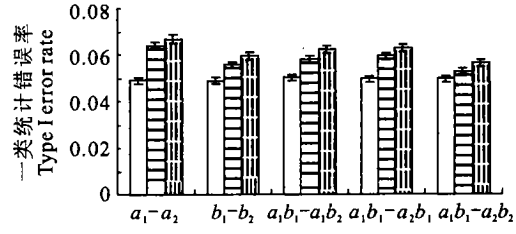

图2 裂区设计固定效应t测验的一类统计错误率

□.平衡数据; ▣.缺失1数据; ▥.缺失2数据

Fig.2 Type Ⅰ error rate of t-test for split-plots design

□.Balance; ▣.1 missing value; ▥.2 missing values

从图 1~3 可以看出, 3 种试验设计在数据平衡时, 固定效应测验的一类统计错误率相当于给定的显著水平 $\alpha=0.05$, 但在缺失数据的非平衡条件下, 固定效应测验的一类统计错误率明显增大。缺失数据越多, 一类统计错误率增大越多。在缺失数据个数相同时, 随机区组设计中一类统计错误率增加较少, 而在裂区设计和条区设计中一类统计错误率增加较多。随机区组设计缺失 2 个数据时, 所有效应比较的一类统计错误率未超过 0.06, 而裂区设计和条区设计缺失 1 个数据时, 一些效应比较的一类统计错误率已超过了 0.06。进一步的模拟研究表明, 随机区

组设计 $a\times b\times r=2\times3\times3$ 在缺失数据超过 4 个时, 固定效应测验的一类统计错误率才会大于 0.06。这一结果表明, 试验数据不平衡时, 统计测验的一类统计错误控制性变差, 实际一类统计错误率比名义一类统计错误率大。数据缺失对随机区组设计的分析结果影响较小, 对裂区设计和条区设计分析结果的影响大。由于固定效应的误差直接影响到固定效应测验统计量的计算结果, 误差值越小, 测验统计量的值越大。前面 3.2 的结果已表明, 固定效应误差的计算值低于实际观测值, 固定效应一类统计错误率增大正是固定效应误差计算值偏低的反映。

4　结论与讨论

试验数据缺失形成数据不平衡时, 采用 REML 法估计方差组分存在着明显的偏差。如果试验的目的只是估计方差组分(例如动植物育种上估算遗传方差组分和环境方差组分), REML 法不是理想的选择。随着缺失数据的增加, 方差组分估计的准确性和精确性都相应降低。

试验数据不平衡导致固定效应误差增大, 而通过方差组分估计值计算固定效应的误差会形成固定效应误差的偏低估计, 这使得固定效应显著性测验的一类统计错误率高于给定的一类错误率(显著水平)。缺失数据越多, 一类统计错误率增加的越多。数据不平衡对随机区组设计分析结果影响小, 而对裂区设计和条区设计分析结果影响大。随机区组设计 $a\times b\times r=2\times3\times3$ 在缺失数据不超过 4 个时, 固定效应测验的一类统计错误率可保持在 0.06 以内, 对应的裂区和条区设计在缺失 1 个数据时, 固定效应测验的一类统计错误率便可超过 0.06。因此, 在试验研究过程中, 要采取严格措施, 尽力避免数据的缺失, 当出现数据缺失时应采用高效的方法估计方差组分, 在进行固定效应测验并作出统计推断时, 应当考虑到实际一类统计错误率与给定统计错误率的差异, 以便作出比较准确的研究结论。

本文只对常用的 3 个试验设计在 $a\times b\times r=2\times3\times3$ 这样的因素水平组合下进行了模拟研究, 模拟得到的固定效应测验统计错误率的大小, 可作为对应试验数据分析时一类统计错误率校正的依据。为了准确了解其他试验设计及其数据结构下, 实际一类统计错误率与给定一类统计错误率差异的大小, 还需要做进一步的模拟研究。

[参考文献]

[1] Henderson C R. Statistical method in animal improvement: historical overview [A]. Advances in statistical methods for genetic improvement of livestock[C]. New York: Springer Verlag, 1990.

[2] Henderson C R. Application of linear models in animal breeding[M]. Guelph: University of Guelph, 1984.

[3] Rao C R. Estimation of variance and covariance components in linear models[J]. J Amer Statist Ass, 1972, 67: 112-115.

[4] Lamotte L R. Quadratic estimation of variance components[J]. Biometrics, 1973, 29: 311-330.

[5] Hartley H O, Rao C R. Maximum likelihood estimation for the mixed analysis of variance model[J]. Biometrica, 1967, 54: 93-108.

[6] Harville D A. Maximum likelihood approaches to variance component estimation and to related problems[J]. Journal of the American Statistical Association, 1977, 72: 320-338.

[7] Fai A H T, Cornelius P L. Approximate *F*-Tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments[J]. J Statist Comput Simul, 1996, 54: 363-378.

[8] Giesbrecht F G, Burns J C. Two-stage analysis based on a mixed model: large-sample asymptomatic theory and small-sample simulation results[J]. Biometrics, 1985, 41: 477-486.

[9] Kenward M G, Roger J H. Small sample inference for fixed effects from restricted maximum likelihood[J]. Biometrics, 1997, 53: 983-997.

[10] Searle S R, Casella G, McCulloch C E. Variance components[M]. New York: John Wiley and Sons Inc, 1992.

[11] Verbeke G, Molenberghs G. Linear mixed models in practice[M]. New York: Springer-Verlag, 1997.

[12] Mclean R A, Sanders W W, Stroup W W. A unified approach to mixed linear models[J]. The American Statistician, 1991, 145: 54-65.

[13] Piepho H P. Analysis of randomized block design with unequal subclass numbers[J]. Agronomy of Journal, 1997, 89: 718-723.

[14] Searle S R, Casella G, McCulloch C E. Variance Components[M]. New York: John Wiley and Sons Inc, 1992.

(下转第 112 页)

Thinking on the measurement problems of information

WANG Nai-xin, ZHANG Hui-feng

(*College of Life Sciences, Northwest Sci-Tech University of Agriculture and Forestry, Yangling, Shaanxi* 712100, *China*)

Abstract: In order to eliminate the limitation of the Shannon entropy, the axiom system for the Shannon entropy was modified, and an information measure system with the axiomatization method was deduced. That could be extended to all discrete random variable and vector with the infinited countable distribution series. The paper proved the theorem of maximum information measure, the theorem of additivity, the theorem of generalized additivity, the theorem of mutual information measurement and the the theorem of independency. Suggestions to choose the value of the parameters in the information mearurement system were put forward.

Key words: Shannon entropy; measurement of information; information measurement system

(上接第 107 页)

Influence of data imbalance on the analysis results of experiment

HU Xi-yuan

(*College of Agronomy, Northwest Sci-Tech University of Agriculture and Forestry, Yangling, Shaanxi* 712100, *China*)

Abstract: This paper examined the influence of data imbalance on the analysis results for random block design, split-plot design and strip-plot design using the Monte Carlo simulation. The simulation results showed that the imbalance of data due to missing observations led to lower precision and lower exactness of variance component estimation, to underestimation of the error of fixed effects as well as to higher type Ⅰ error rate for fixed effect test. With more missing observations these influences would be heavier. The imbalance of data influenced the analysis results in split-plot design and strip-plot design much heavier than in random block design.

Key words: imbalance of data; variance component estimate; fixed effect test