DOI:10.19789/j.1004-9398.2017.05.001

生物性状不同组合间的相关性度量*

董晓萌
(渭南师范学院数理学院,陕西 渭南 714099)

摘 要

借助于一元线性回归模型,采用矩法,提出描述生物性状间不同组合的非线性广义相关系数. 定义了各性状的总相关信息量及其相关贡献率,利用该广义相关系数对春播菜用大豆生物性状团的相关性进行度量,可使不同性状组合之间的相关程度介于 -1 与 1 之间.

关键词: 广义相关系数, 性状组合, 一元线性回归模型.
中图分类号: O212.4, S11+4

微效多基因控制着生物的数量性状,易受环境的影响. 性状之间的相关性度量是品种改良中对生物体进行育种选择的基本条件. 但作为一个有机整体的生物体,从生理、发育和功能上都可能形成一些由若干相关性状形成的性状团,这些性状团之间亦存在着不同程度的相关[1]. 张尧庭等[2-4]所研究的变量间的广义相关系数均是线性的,对变量间的非线性相关关系[5-7]无法度量. 因此,该相关性度量是有限的. 笔者在前人研究基础之上提出了性状团之间的广义相关系数[8],该广义相关系数只对性状团中不同性状在"等权"之下的线性组合及平方和组合的性状进行相关性度量,但对各性状不同方次组合的相关性度量并未给出. 本文定义了各性状间不同方次组合的广义相关系数,并对其性质进行了研究. 定义了各性状的相关信息及相关贡献,为育种工作提供理论参考.

1 两个随机向量各分量不同组合间的相关系数

假设多维随机向量[9]为: $X=(x_1,x_2,\cdots,x_l,x_{l+1},x_{l+2},\cdots,x_{l+n})^T\sim N_{l+n}(\mu,\sum)$,其中 $\mu=(\mu_1,\mu_2,\cdots,\mu_l,\mu_{l+1},\mu_{l+2},\cdots,\mu_{l+n})^T$ 是随机向量 X 的均值向量,$\sum$ 是随机向量 X 的离差阵,且 $\sum=(\sigma_{ij})_{(l+n)\times(l+n)}$,$R=(r_{ij})_{(l+n)\times(l+n)}$ 是随机向量 X 的相关阵.

定义1 令 $Y=\sum_{i=1}^{l}X_i$,$Z=X_{l+1}+X_{l+2}+\cdots+X_{l+n}+X_{l+1}^2+X_{l+2}^2+\cdots+X_{l+n}^2$,则 Y 与 Z 的相关系数为:

$$r_{YZ}^{(1)}=\frac{\sum_{i=1}^{l}\sum_{j=l+1}^{l+n}\sigma_{ij}}{\sqrt{\sum_{i=1}^{l}\sum_{j=1}^{l}\sigma_{ij}\sum_{i=l+1j=l+1}^{l+n}\sum^{l+n}(\sigma_{ij}+3\sigma_{ij}^2)}}. \quad (1)$$

由于服从正态分布的随机变量的三阶中心距为零,且 Y 的方差 $V(Y)=\sum_{i=1}^{l}\sum_{j=1}^{l}\sigma_{ij}$,$Z$ 的方差 $V(Z)=\sum_{i=l+1j=l+1}^{l+n}\sum^{l+n}(\sigma_{ij}+3\sigma_{ij}^2)$,$Y$ 与 Z 的协方差为 $COV(Y,Z)=\sum_{i=1}^{l}\sum_{j=l+1}^{l+n}\sigma_{ij}$.

定义2 令 $Y=\sum_{i=1}^{l}(X_i+X_i^2)$,$Z=\sum_{j=l+1}^{l+n}(X_j+X_j^2)$,则 Y 与 Z 的相关系数为:

$$r_{YZ}^{(2)}=\frac{\sum_{i=1}^{l}\sum_{j=l+1}^{l+n}(\sigma_{ij}+3\sigma_{ij}^2)}{\sqrt{\sum_{i=1}^{l}\sum_{j=1}^{l}(\sigma_{ij}+3\sigma_{ij}^2)\cdot\sum_{i=l+1j=l+1}^{l+n}\sum^{l+n}(\sigma_{ij}+3\sigma_{ij}^2)}}. \quad (2)$$

收稿日期: 2016-11-30
* 国家自然科学基金项目(30571072);陕西省科技厅项目(2014JQ1006);国家统计局项目(2011LY092).

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

由于服从正态分布的随机变量的三阶中心矩等于 0,四阶混合中心矩等于 $3\sigma_{ij}^2$,且 Y 的方差 $V(Y) = \sum_{i=1}^{l}\sum_{j=1}^{l}(\sigma_{ij} + 3\sigma_{ij}^2)$,$Z$ 的方差 $V(Z) = \sum_{i=l+1}^{l+n}\sum_{j=l+1}^{l+n}(\sigma_{ij} + 3\sigma_{ij}^2)$,$Y$ 与 Z 的协方差为 $COV(Y,Z) = \sum_{i=1}^{l}\sum_{j=l+1}^{l+n}(\sigma_{ij} + 3\sigma_{ij}^2)$.

用上述方法可以把 Y 和 Z 构造成各 X 的不同方次的组合,定义不同方次组合间的相关系数 ρ_{YZ}. 在计算中应注意: 在正态分布下,奇次中心矩皆等于 0,偶次中心矩等于 $(k-1)!!\sigma_{ij}^{k/2}$,$k = 2,4,6,\cdots\cdots$,符号 $(k-1)!!$ 表示 1 到 $k-1$ 的一切奇数的乘积.

同样,若对 X 的相关阵 $R = (r_{ij})_{(l+n)\times(l+n)}$ 分析所得结果是一致的,则可以得到与定义 1、定义 2 中相关系数相互等价的式子如下:

$$r_{YZ}^{(1)} = \frac{\sum_{i=1}^{l}\sum_{j=l+1}^{l+n}\sigma_{ij}}{\sqrt{\sum_{i=1}^{l}\sum_{j=1}^{l}\sigma_{ij}\sum_{i=l+1}^{l+n}\sum_{j=l+1}^{l+n}(\sigma_{ij} + 3\sigma_{ij}^2)}} \Leftrightarrow \frac{\sum_{i=1}^{l}\sum_{j=l+1}^{l+n}r_{ij}}{\sqrt{\sum_{i=1}^{l}\sum_{j=1}^{l}r_{ij}\sum_{i=l+1}^{l+n}\sum_{j=l+1}^{l+n}(r_{ij} + 3r_{ij}^2)}},$$

$$r_{YZ}^{(2)} = \frac{\sum_{i=1}^{l}\sum_{j=l+1}^{l+n}(\sigma_{ij} + 3\sigma_{ij}^2)}{\sqrt{\sum_{i=1}^{l}\sum_{j=1}^{l}(\sigma_{ij} + 3\sigma_{ij}^2)\cdot\sum_{i=l+1}^{l+n}\sum_{j=l+1}^{l+n}(\sigma_{ij} + 3\sigma_{ij}^2)}} \Leftrightarrow \frac{\sum_{i=1}^{l}\sum_{j=l+1}^{l+n}(r_{ij} + 3r_{ij}^2)}{\sqrt{\sum_{i=1}^{l}\sum_{j=1}^{l}(r_{ij} + 3r_{ij}^2)\cdot\sum_{i=l+1}^{l+n}\sum_{j=l+1}^{l+n}(r_{ij} + 3r_{ij}^2)}}.$$

2 广义相关系数的性质

任意两个多维随机向量的广义相关系数 $r_{YZ}^{(1)}$,$r_{YZ}^{(2)}$ 的性质如下:

性质 ① 对称性,即 $r_{YZ}^{(1)} = r_{ZY}^{(1)}$,$r_{YZ}^{(2)} = r_{ZY}^{(2)}$;

② 相同的变量之间的相关系数是 1,即 $r_{YY}^{(1)} = r_{ZZ}^{(1)} = 1$,$r_{YY}^{(2)} = r_{ZZ}^{(2)} = 1$;

③ 广义相关系数值在 -1 与 1 之间,即 $-1 \leqslant r_{YZ}^{(1)} \leqslant 1$,$-1 \leqslant r_{YZ}^{(2)} \leqslant 1$. 当 X 与 Y 相互独立时,$r_{YZ}^{(1)} = 0$,$r_{YZ}^{(2)} = 0$.

证明: ① 因为

$$r_{YZ}^{(1)} = \frac{\sum_{i=1}^{l}\sum_{j=l+1}^{l+n}\sigma_{ij}}{\sqrt{\sum_{i=1}^{l}\sum_{j=1}^{l}\sigma_{ij}\sum_{i=l+1}^{l+n}\sum_{j=l+1}^{l+n}(\sigma_{ij} + 3\sigma_{ij}^2)}},$$

又因为 $r_{ZY}^{(1)} = \frac{\sum_{i=l+1}^{l+n}\sum_{j=1}^{l}\sigma_{ij}}{\sqrt{\sum_{i=1}^{l}\sum_{j=1}^{l}\sigma_{ij}\sum_{i=l+1}^{l+n}\sum_{j=l+1}^{l+n}(\sigma_{ij} + 3\sigma_{ij}^2)}}$. 由于 $\sum = (\sigma_{ij})_{(l+n)\times(l+n)}$ 是对称的,所以 $\sum_{i=1}^{l}\sum_{j=l+1}^{l+n}\sigma_{ij} = \sum_{i=l+1}^{l+n}\sum_{j=1}^{l}\sigma_{ij}$,则 $r_{YZ}^{(1)} = r_{ZY}^{(1)}$.

② 因为

$$r_{XX}^{(1)} = \frac{\sum_{i=1}^{l}\sum_{j=1}^{l}\sigma_{ij}}{\sqrt{\sum_{i=1}^{l}\sum_{j=1}^{l}\sigma_{ij}\sum_{i=1}^{l}\sum_{j=1}^{l}\sigma_{ij}}} = 1,$$

即 $r_{XX}^{(1)} = 1$,同理可证 $r_{YY}^{(1)} = 1$.

③ 假设 Y,Z 相互独立时,$\sum_{i=1}^{l}\sum_{j=l+1}^{l+n}\sigma_{ij} = 0$. 故其相关系数为:

$$r_{YZ}^{(1)} = \frac{\sum_{i=1}^{l}\sum_{j=l+1}^{l+n}\sigma_{ij}}{\sqrt{\sum_{i=1}^{l}\sum_{j=1}^{l}\sigma_{ij}\sum_{i=l+1}^{l+n}\sum_{j=l+1}^{l+n}(\sigma_{ij} + 3\sigma_{ij}^2)}} = 0,$$

即 $r_{YZ}^{(1)} = 0$.

当变量 Y,Z 相互不独立时,则 Y,Z 的广义相关系数为

$$r_{YZ}^{(1)} = \frac{\sum_{i=1}^{l}\sum_{j=l+1}^{l+n}\sigma_{ij}}{\sqrt{\sum_{i=1}^{l}\sum_{j=1}^{l}\sigma_{ij}\sum_{i=l+1}^{l+n}\sum_{j=l+1}^{l+n}(\sigma_{ij} + 3\sigma_{ij}^2)}}.$$

因为 $-\sqrt{V(Y)V(Z)} \leqslant COV(Y,Z) \leqslant \sqrt{V(Y)V(Z)}$,则

$$-1 \leqslant \frac{\sum_{i=1}^{l}\sum_{j=l+1}^{l+n}\sigma_{ij}}{\sqrt{\sum_{i=1}^{l}\sum_{j=1}^{l}\sigma_{ij}\sum_{i=l+1}^{l+n}\sum_{j=l+1}^{l+n}(\sigma_{ij} + 3\sigma_{ij}^2)}} \leqslant 1,$$

从而 $-1 \leqslant r_{YZ}^{(1)} \leqslant 1$. 同理可以证明定义 2 中的相关系数 $r_{YZ}^{(2)}$ 也满足上述的三个性质.

上述构建的广义相关系数,对于在“等权”条件下,不同性状组合的相关程度进行了度量. 利用该广义相关系数[10]可以对不同生物的各个性状团之间不同性状组合的相关性进行刻画,用具体的数值来

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表示其相关程度.

3　性状的相关信息及相关贡献

将随机向量 X 的离差阵 $\sum$ 的各元素的平方和 $I_{\sum}=\sum_{i=1}^{l+n}\sum_{j=1}^{l+n}\sigma_{ij}^2$ 称为随机向量 X 的总变异信息; 将随机向量 X 的相关阵 R 的各元素的平方和 $I_R=\sum_{i=1}^{l+n}\sum_{j=1}^{l+n}r_{ij}^2$ 称为随机向量 X 的总相关信息; 对于实对称阵 $\sum$ 和 R 都有一个重要性质,它们经过正交相似变换得 $L^T\sum L=(p_{ij})_{(m+p)(m+p)}$ 和 $U^TRU=(q_{ij})_{(m+p)(m+p)}$ 其总信息不变,即

$$I_{\sum}=\sum_{i=1}^{m+p}\sum_{j=1}^{m+p}p_{ij}^2,\qquad I_R=\sum_{i=1}^{m+p}\sum_{j=1}^{m+p}q_{ij}^2.$$

在随机向量 X 的相关阵 R 中,第 l 行元素的平方和 $I_l=\sum_{j=1}^{n}r_{lj}^2(l=1,2,\cdots,n)$ 表示变量 x_l 与其他各个变量 x 的相关信息之和,称其为变量 x_l 的相关信息,显然各性状的相关信息之和就等于 X 的总相关信息. 相关信息大的性状,其变化可引起其它性状的比较显著的变化. 相关信息小的性状,其变化所引起的其它性状的变化是平稳的. 这对育种很有意义.

将性状 x_l 的相关信息 I_l 与 X 的总相关信息 I_R 的比值 $\eta_l=\frac{I_l}{I_R}$ 定义为性状 x_l 的相关贡献,相关贡献大的性状的变化对其他性状的变化将起到显著性影响的作用.

4　应用实例

本研究采用文献[11]中的数据,根据生物学意义将春播菜用大豆性状[11]分成了 5 个性状团(X_1, X_2, X_3, X_4, X_5),其中: X_1(生育期), X_2(植株形态), X_3(产量性状), X_4(荚粒形态), X_5(品质性状).

根据定义 1 计算春播菜用大豆 5 个性状团的方差及其之间的协方差[12-13]. 结果见表 1.

表 1　各个性状团的方差及其之间的协方差(定义 1)

性状团 Trait groups	X_1	X_2	X_3	X_4	X_5
X_1	129	-29.445	-231.092	-2.673	0.145
X_2	-29.445	14.060	33.758	0.813	0.176
X_3	-231.092	33.758	509.440	4.240	-1.303
X_4	-2.673	0.813	4.240	0.094	-0.023
X_5	0.145	0.178	-1.303	-0.023	0.276

根据定义 2 计算春播菜用大豆 5 个性状团的方差及其之间的协方差[14]. 结果见表 2.

表 2　各个性状团的方差及其之间的协方差(定义 2)

性状团 Trait groups	X_1	X_2	X_3	X_4	X_5
X_1	4 151.384	9.895	14.299	12.529	10.829
X_2	9.895	829.724	7.748	10.986	10.757
X_3	14.299	7.748	82 331.150	15.203	6.738
X_4	12.529	10.986	15.203	0.100	4.612
X_5	10.829	10.757	6.738	4.612	0.418

由春播菜用大豆不同形状团的方差与其协方差,可计算出各个性状团之间的广义相关系数值[15],见表 3.

表 3　各个性状团之间的广义相关系数

性状团 Trait groups	X_1	X_2	X_3	X_4	X_5
X_1	1	-0.076	-0.459	-0.383	-0.013
X_2	0.207	1	0.243	0.250	0.047
X_3	0.281	0.183	1	0.271	-0.042
X_4	0.276	0.291	0.378	1	-0.019
X_5	0.340	0.406	0.239	0.183	1

由上述定义 1 计算的广义相关系数值见表 3 中上三角形中的数值. 从表中数据可知: 生育期性状与产量性状的广义相关系数为 -0.459,达到最大负相关,表明生育期性状对产量性状的负作用最大; 生育期性状与品质性状的广义相关系数为 -0.013,达到最小负相关,表明生育期性状对品质性状的负作用最小; 荚粒形态性状团与产量性状的相关性达到最大的正相关,表明荚粒形态性状团对产量性状的正作用最大,荚粒形态与品质性状的相关性为最小的负相关,表明荚粒形态性状团对品质性状起到一定的负作用; 植株形态与品质性状的相关性为最小的正相关. 由上述定义 2 计算的广义相关系数值见表 3 中下三角形中的数值,从表中数据可知: 品质性状与植株形态的相关性最强,与生育期的相关性次之; 产量性状与荚粒形态的相关性最强,与品质性状的相关性次之; 在 5 个性状团中,品质性状与荚粒形态的相关性最小.

5　结论与讨论

本研究借助于一元线性回归模型,采用矩法,提出描述生物性状间不同组合的非线性广义相关系数,并定义了各性状的相关信息及相关贡献. 将该广义相关系数应用于春播菜用大豆生物性状分析,可使不同性状组合之间的相关程度介于 -1 与 1 之间. 相关信息大的性状,其变化可引起其它性状的比较显著的变化,相关信息小的性状,其变化所引起的其它性状的变化是平稳的,这对育种很

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

有意义.并且为育种工作者进行育种时提供参考依据.为了进一步完善该理论,文章下一步将对上述构建的相关系数的抽样分布问题及假设检验的问题进行系统研究.

参 考 文 献

[1] 袁志发,周静芋.多元统计分析[M].北京:科学出版社,2002:172-180,241-256.
[2] 张尧庭.广义相关系数及其应用[J].应用数学学报,1978(4):33-39.
[3] 胡永宏.一种广义相关系数[J].统计与信息论坛,1997(1):20-23.
[4] 张尧庭.关于度量变量之间的相关程度[J].上海财经大学学报,1999(2):60-63.
[5] 何晓群.多元统计分析[M].北京:中国人民大学出版社,2010:150-158.
[6] 朱建平.应用多元统计分析[M].北京:科学出版社,2006:134-142.
[7] 刘垂玕.作物数量性状的遗传相关信息及其可加性[J].安徽农业科学,1981(S1):52-57.
[8] 董晓萌,曹彬婕,罗凤娟,等.一种度量生物性状非线性相关性的广义相关系数[J].西北农林科技大学学报:自然科学版,2008,36(5):191-195.
[9] 陈希孺.概率论与数理统计[M].北京:科学出版社,2002:126-140.
[10] 盛骤,谢式千,潘承毅.概率论与数理统计[M].北京:高等教育出版社,2010:74-82.
[11] 周以飞,周德银.春播菜用大豆生育期,农艺性状和品质性状的典范相关分析[J].福建农林大学学报,2005(3):11-17.
[12] 董晓萌,袁志发.度量生物性状广义相关系数的抽样分布及假设检验[J].西北农林科技大学学报:自然科学版,2010,38(4):215-218.
[13] 高惠璇.应用多元统计分析[M].北京:北京大学出版社,2004:241-250.
[14] 梅长林,周家良.实用统计方法[M].北京:科学出版社,2002:220-229.
[15] 汤银才.R 语言与统计分析[M].北京:高等教育出版社,2008:233-240.

The Correlation Measurement Among Different Biology Trait Groups

Dong Xiaomeng

(College of Math and Physics, Weinan Normal University, Weinan Shaanxi 714099)

Abstract

A kind of new generalized correlation coefficient based on the moment methods was built with linear regression model, which could describe the correlation of different biology trait groups. By defining total correlation information and correlation contribution of biology trait, it was applied in the analysis of trait groups with vegetable soybean of spring seeding, and got the correlation measurement between -1 and 1.

Key words: generalized correlation coefficient, biology trait group, simple linear regression model.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net