度量生物性状广义相关系数的抽样分布及假设检验

董晓萌[1],袁志发[2]

(1 渭南师范学院 数学与信息科学系,陕西 渭南 714000;2 西北农林科技大学 理学院,陕西 杨凌 712100)

[摘　要]　【目的】完善生物性状非线性相关性的广义相关系数理论体系,为实际育种工作提供一套完整的理论体系。【方法】采用构造综合指标,借助于一元线性回归模型,对度量生物性状线性、非线性相关性的广义相关系数的抽样分布及假设检验问题进行研究。【结果】该广义相关系数服从 F 分布或 t 分布,可用 F 检验或者 t 检验对该相关系数与 0 之间的差异显著性进行检验。【结论】建立的广义相关系数检验方法计算简单、无信息损失,线性与非线性广义相关系数的定义、性质、抽样分布及假设检验问题都予以解决,丰富了多元统计学内容。

[关键词]　广义相关系数;非线性相关;抽样分布;假设检验

[中图分类号]　O212.4;S11+4　　[文献标识码]　A　　[文章编号]　1671-9387(2010)04-0215-04

Sampling distribution and hypothesis testing with generalized correlation coefficient of biology trait

DONG Xiao-meng[1], YUAN Zhi-fa[2]

(1 *Department of Math and Information Science, Weinan Teachers University, Weinan, Shaanxi* 714000, *China*;
2 *College of Science, Northwest A&F University, Yangling, Shaanxi* 712100, *China*)

Abstract: 【Objective】The recearch was done in order to complete generalized correlation coefficient theory system of biology trait and provide the whole theory system for breeding operation.【Method】The sampling distribution and hypothesis testing of this non-linear generalized correlation coefficient to measure biology trait was studied with linear regression model.【Result】The testing of this non-linear generalized correlation coefficient is F-testing or t-testing with correlation coefficient 0.【Conclusion】The definition, nature, sampling distribution and hypothesis testing have been solved, the whole theory system for Mutivariate statistics analysis has been enriched and the testing method is simple.

Key words: generalized correlation coefficient; non-linear correlation; sampling distribution; hypothesis testing

性状的相关性度量是进行系统结构与功能分析的基础。但以往的研究,如简单相关、典范相关[1]和广义相关[2]等,均是建立在线性相关[3]的基础上,或者是造成了原有变异信息的损失[4],或者是计算复杂[5],它们能反映 2 个(组)变量之间的线性相关程度,但不能反映 2 个(组)变量之间的非线性相关程度[6]。本研究对笔者定义的计算简单、无信息损失的非线性广义相关系数的抽样分布及假设检验问题进行分析,旨在完善非线性广义相关的理论知识,丰富多元统计学的内容,为实际育种工作者提供了一套完整的多性状间的非线性相关分析理论体系。

1　一种新广义相关系数的定义

假设 2 个多维变量分别为[7]:$X=(x_1,x_2,\cdots,x_m)^T\sim N_m(\mu_x,\Sigma_X)$,$Y=(y_1,y_2,\cdots,y_p)^T\sim N_p(\mu_y,\Sigma_Y)$,其中 $\mu_x=(\mu_{x_1},\mu_{x_2},\cdots,\mu_{x_m})^T$ 为 X 的均值向

* [收稿日期]　2009-09-26
[基金项目]　国家自然科学基金项目(30571072);渭南师范学院研究生项目(09YKZ007)
[作者简介]　董晓萌(1982－),女,陕西渭南人,助教,硕士,主要从事多元统计分析研究。E-mail:mathmandy@sohu.com

量,$\sum_X$ 为 X 的离差阵;$\mu_y=(\mu_{y_1},\mu_{y_2},\cdots,\mu_{y_p})^T$ 为 Y 的均值向量,$\sum_Y$ 为 Y 的离差阵,$\sum_X$ 和 $\sum_Y$ 均是对称正定矩阵。设 $\sum_{XY}$ 是 X、Y 的协差阵。X、Y 的相关阵分别记为:

$$\boldsymbol{R_X}=(\rho_{Xij})_{m\times m},\boldsymbol{R_Y}=(\rho_{Yij})_{p\times p}。$$

令 $\boldsymbol{Z}\begin{bmatrix}X\\Y\end{bmatrix}\sim N_{m+p}(\boldsymbol{\mu}_z,\sum_Z)$,其中 $\boldsymbol{\mu}_z=\begin{bmatrix}\mu_x\\\mu_y\end{bmatrix}$,

$$\sum_Z=\begin{bmatrix}\sum_X & \sum_{XY}\\ \sum_{YX} & \sum_Y\end{bmatrix}_{(m+p)\times(m+p)}=(\sigma_{Zij})_{(m+p)\times(m+p)}。$$

将 X 与 Y 的线性组合分别记为:$X^{(1)}=x_1+x_2+\cdots+x_m$,$Y^{(1)}=y_1+y_2+\cdots,y_p$,则 $X^{(1)}$ 的方差 $V(X^{(1)})=\sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_{Zij}$,$Y^{(1)}$ 的方差 $V(Y^{(1)})=\sum_{i=m+1}^{m+p}\sum_{j=m+1}^{m+p}\sigma_{Zij}$,$X^{(1)}$ 与 $Y^{(1)}$ 的协方差为 $COV(X^{(1)},Y^{(1)})=\sum_{i=1}^{m}\sum_{j=m+1}^{m+p}\sigma_{Zij}$。

由上述分析可定义如下。

定义1:X 与 Y 的广义相关系数 $r_{(1)XY}$ 为:

$$r_{(1)XY}=\frac{COV(X^{(1)},Y^{(1)})}{\sqrt{V(X^{(1)})V(Y^{(1)})}}=\frac{\sum_{i=1}^{m}\sum_{j=m+1}^{m+p}\sigma_{Zij}}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_{Zij}}\sqrt{\sum_{i=m+1}^{m+p}\sum_{j=m+1}^{m+p}\sigma_{Zij}}}。\quad(1)$$

显然,广义相关系数 $r_{(1)XY}$ 满足相关系数的所有性质,其本质上也是一种线性相关,即将原来的变量赋以权重为1,然后线性组合为2个综合变量,这2个综合变量间的简单相关系数即为2个多维变量间的广义相关系数。

定义2:假设变量 X 的四阶中心矩为 $(\boldsymbol{\mu}_4)_i$[8],四阶混合矩为 $(\boldsymbol{\mu}_4)_{ij}$ $(i,j=1,2\cdots,m$;且 $i\neq j)$;变量 Y 的四阶中心矩为 $(\boldsymbol{v}_4)_i$,四阶混合矩为 $(\boldsymbol{v}_4)_{ij}$ $(i,j=1,2,\cdots,p;i\neq j)$;变量 X、Y 的四阶混合矩为 $(\boldsymbol{\mu v})_{4ij}$ $(i=1,2,\cdots,m;j=1,2,\cdots,p)$,并假设 $0\leqslant\sum_{i=1}^{m}\sum_{j=1}^{p}\leqslant\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}(\mu_4)_{ij}}\sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p}(v_4)_{ij}}$,则称

$$r_{(2)XY}{}^2=\frac{\frac{1}{3}\sum_{i=1}^{m}\sum_{j=1}^{p}(\mu v)_{4ij}}{\sqrt{\frac{1}{3}\sum_{i=1}^{m}\sum_{j=1}^{m}(\mu_4)_{ij}}\sqrt{\frac{1}{3}\sum_{i=1}^{p}\sum_{j=1}^{p}(v_4)_{ij}}}=\frac{\sum_{i=1}^{m}\sum_{j=1}^{p}(\mu v)_{4ij}}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}(\mu_4)_{ij}}\sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p}(v_4)_{ij}}}\quad(2)$$

为变量 X、Y 的广义相关系数。

其中:

$$(\mu_4)_{ij}=E(X_i-\mu_{x_i})^2(X_j-\mu_{x_j})^2,i,j=1,2,\cdots,m;$$

$$(v_4)_{ij}=E(Y_i-\mu_{y_i})^2(Y_j-\mu_{y_j})^2,i,j=1,2,\cdots,p;$$

$$(\mu v)_{4ij}=E(X_i-\mu_{x_i})^2(Y_j-\mu_{y_j})^2,$$
$$i=1,2,\cdots,m;j=1,2,\cdots,p。$$

因为变量的四阶中心矩与其四阶混合矩组成矩阵中的每个元素,与相关阵的每个元素平方后所得的矩阵中每个元素相差3倍。

故定义2中的式(2)与下式是等价的,即:当 $0\leqslant S_{R_{XY}}\leqslant\sqrt{S_{R_X}}\sqrt{S_{R_Y}}$,

$$r_{(2)XY}{}^2=\frac{\frac{1}{3}\sum_{i=1}^{m}\sum_{j=1}^{p}(\mu v)_{4ij}}{\sqrt{\frac{1}{3}\sum_{i=1}^{m}\sum_{j=1}^{m}(\mu_4)_{ij}}\sqrt{\frac{1}{3}\sum_{i=1}^{p}\sum_{j=1}^{p}(v_4)_{ij}}}\Leftrightarrow$$

$$\frac{S_{R_Z}-S_{R_X}-S_{R_Y}}{2\sqrt{S_{R_X}}\sqrt{S_{R_Y}}}=\frac{S_{R_{XY}}}{\sqrt{S_{R_X}}\sqrt{S_{R_Y}}}。\quad(3)$$

显然,当 $m=p=1$ 时,X、Y 的相关系数是其简单相关系数,即:$r_{(2)XY}{}^2=\rho^2$。

2 广义相关系数的假设检验

2.1 定义1中线性相关性广义相关系数的抽样分布及假设检验问题

将 X 与 Y 的线性组合分别记为:$X^{(1)}=x_1+x_2+\cdots+x_m$,$Y^{(1)}=y_1+y_2+\cdots+y_p$,定义1中2个多维变量 X 与 Y 的广义相关系数实质上是变量 $X^{(1)}$ 与 $Y^{(1)}$ 的简单相关系数。故可以借助于一元线性回归方程的检验方法,对其进行抽样分布及假设检验。

因为 X 与 Y 服从正态分布,由正态分布的性质可得:

$$X^{(1)}=x_1+x_2+\cdots+x_m\sim N(\sum_{i=1}^{m}\mu_{Xi}\sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_{Xij}),$$

$$Y^{(1)}=y_1+y_2+\cdots+y_p\sim N(\sum_{i=1}^{p}\mu_{Yi}\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma_{Yij})。$$

建立 $Y^{(1)}$ 关于 $X^{(1)}$ 的一元线性回归模型,记为:$Y_i^{(1)}=a+bX_i^{(1)}+\varepsilon_i$,其中 ε_i 相互独立,且 $\varepsilon_i\sim N(0,\sigma^2)$。应用最小二乘估计可得,$a=\overline{Y}^{(1)}-b\overline{X}^{(1)}$,式中:$\overline{X}^{(1)}$、$\overline{Y}^{(1)}$ 分别是 $X^{(1)}$、$Y^{(1)}$ 的均值;$b=\frac{l_{X^{(1)}Y^{(1)}}}{l_{X^{(1)}X^{(1)}}}$,其中 $l_{X^{(1)}Y^{(1)}}=\sum(X_i^{(1)}-\overline{X}^{(1)})(Y_i^{(1)}-\overline{Y}^{(1)})$,$l_{X^{(1)}X^{(1)}}=\sum(X_i^{(1)}-\overline{X}^{(1)})^2$,$l_{Y^{(1)}Y^{(1)}}=\sum(Y_i^{(1)}-\overline{Y}^{(1)})^2$。

由一元线性回归方程的统计检验可得[7]:

$$\frac{l_{Y^{(1)}Y^{(1)}}}{\sigma^2}\sim\chi^2(n-1),\frac{U}{\sigma^2}\sim\chi^2(1),\frac{Q}{\sigma^2}\sim\chi^2(n-2)。$$

式中:σ^2 是随机误差 ε_i 的方差,$\hat{\sigma}^2=\frac{Q}{n-2}$,$Q=l_{Y^{(1)}Y^{(1)}}-l^2_{X^{(1)}Y^{(1)}}/l_{X^{(1)}X^{(1)}}=l_{Y^{(1)}Y^{(1)}}(1-r^2)$,其中 r 为

$X^{(1)}$ 与 $Y^{(1)}$ 的简单相关系数，即是定义 1 中的 X 与 Y 的广义相关系数；$U=l_{Y^{(1)}Y^{(1)}}r^2$。

因为 $U/\sigma^2\sim\chi^2(1)$，$Q/\sigma^2\sim\chi^2(n-2)$，则由 F 分布的定义可知：

$$F=\frac{U/1}{Q/(n-2)}=\frac{l_{Y^{(1)}Y^{(1)}}r^2}{l_{Y^{(1)}Y^{(1)}}(1-r^2)/(n-2)}=\frac{r^2}{(1-r^2)/(n-2)}\sim F(1,n-2)。\tag{4}$$

再由 t 分布和 F 分布的关系可得：

$$t=\sqrt{F}=\sqrt{\frac{r^2}{(1-r^2)(n-2)}}=\frac{r}{\sqrt{(1-r^2)/(n-2)}}\sim t(n-2),\tag{5}$$

其中 r 的标准差 $S_r=\sqrt{(1-r^2)/(n-2)}$。

以上分析表明，定义 1 中的广义相关系数即是 $X^{(1)}$ 与 $Y^{(1)}$ 之间的简单相关系数，所以定义 1 中的广义相关系数 $r=0$ 的检验即可应用 t 检验或者 F 检验进行。

2.2　定义 2 中非线性相关性广义相关系数的抽样分布及假设检验问题

记 $X^{(2)}=X_1^2+X_2^2+\cdots+X_m^2$，$Y^{(2)}=Y_1^2+Y_2^2+\cdots+Y_p^2$，建立 $Y^{(2)}$ 关于 $X^{(2)}$ 的一元线性回归模型，记为：$Y_i^{(2)}=\beta_0+\beta X_i^{(2)}+\varepsilon_i$，且 $\varepsilon_i\sim N(0,\sigma^2)$。应用最小二乘估计可得 $\beta_0=\bar{Y}^{(2)}+b\bar{X}^{(2)}$，其中 $\bar{X}^{(2)}$、$\bar{Y}^{(2)}$ 分别是 $X^{(2)}$、$Y^{(2)}$ 的均值；$\beta=\frac{l_{X^{(2)}Y^{(2)}}}{l_{X^{(2)}X^{(2)}}}$，其中 $l_{X^{(2)}Y^{(2)}}=\sum(X_i^{(2)}-\bar{X}^{(2)})(Y_i^{(2)}-\bar{Y}^{(2)})$；$l_{X^{(2)}X^{(2)}}=\sum(X_i^{(2)}-\bar{X}^{(2)})^2$；$l_{Y^{(2)}Y^{(2)}}=\sum(Y_i^{(2)}-\bar{Y}^{(2)})^2$。

由一元线性回归方程的统计检验可得：

$$\frac{l_{Y^{(2)}Y^{(2)}}}{\sigma^2}\sim\chi^2(n-1),\frac{U}{\sigma^2}\sim\chi^2(1),\frac{Q}{\sigma^2}\sim\chi^2(n-2)。$$

式中：σ^2 是随机误差 ε_i 的方差，$\hat{\sigma}^2=\frac{Q}{n-2}$，$Q=l_{Y^{(2)}Y^{(2)}}-l^2_{X^{(2)}Y^{(2)}}/l_{X^{(2)}X^{(2)}}=l_{Y^{(2)}Y^{(2)}}(1-r^2)$，其中 r 为 $X^{(2)}$ 与 $Y^{(2)}$ 的简单相关系数，亦是定义 2 中 X 与 Y 的广义相关系数；$U=l_{Y^{(2)}Y^{(2)}}r^2$。

因为 $U/\sigma^2\sim\chi^2(1)$，$Q/\sigma^2\sim\chi^2(n-2)$，则由 F 分布的定义可知：

$$F=\frac{U/1}{Q/(n-2)}=\frac{l_{Y^{(2)}Y^{(2)}}r^2}{l_{Y^{(2)}Y^{(2)}}(1-r^2)/(n-2)}=\frac{r^2}{(1-r^2)/(n-2)}\sim F(1,n-2)。\tag{6}$$

再由 t 分布和 F 分布的关系可得：

$$t=\sqrt{F}=\sqrt{\frac{r^2}{(1-r^2)/(n-2)}}=\frac{r}{\sqrt{(1-r^2)/(n-2)}}\sim t(n-2),\tag{7}$$

其中 r 的标准差 $S_r=\sqrt{(1-r^2)/(n-2)}$。

以上分析表明，定义 2 中的广义相关系数即是 $X^{(2)}$ 与 $Y^{(2)}$ 之间的简单相关系数，所以定义 2 中广义相关系数 $r=0$ 的检验即可应用 t 检验或者 F 检验进行。

2.3　假设检验步骤

为了方便工作者应用，现将检验步骤列举如下。

第 1 步：由定义 1(或定义 2)中的公式计算出广义相关系数值 r；

第 2 步：将 r 代入式(4)或(5)(式(6)或(7))中，算出相应的 F 值和 t 值；

第 3 步：给定置信水平 α，查表得 $F_\alpha(1,n-2)$ 或者 $t_\alpha(n-2)$：若 $|F|<F_\alpha(1,n-2)$ 或者 $|t|<t_\alpha(n-2)$，则接受原假设，即广义相关系数 r 与 0 无显著性差异，否则原假设不成立，即广义相关系数 r 与 0 有显著性差异。

3　应用实例

本例原始数据来自参考文献[9]，对春播菜用大豆 5 个性状团之间线性与非线性相关性的广义相关系数进行计算，结果如表 1 所示[10]。

表 1　春播菜用大豆不同性状团之间的广义相关系数

Table 1　Generalized correlation coefficient between trait groups of spring vegetable soybean

性状团 Ttrait group	X_1(生育期 Growth period)	X_2(植株形态 Plant shape)	X_3(产量性状 Yield trait)	X_4(荚粒形态 Pod shape)	X_5(品质性状 Quality trait)
X_1(生育期 Growth period)	1.00	−0.69	−0.91	−0.01	0.04
X_2(植株形态 Plant shape)	0.53	1.00	0.40	0.71	0.10
X_3(产量性状 Yield trait)	0.82	0.38	1.00	0.61	−0.11
X_4(荚粒形态 Pod shape)	0.78	0.53	0.58	1.00	−0.01
X_5(品质性状 Quality trait)	0.66	0.67	0.58	0.50	1.00

表 1 中对角线上面的数据是由定义 1 得到的广义相关系数，对角线下面的数据是由定义 2 得到的广义相关系数[10]。

现在按照 2.3 中给定的检验步骤对表 1 中的广义相关系数进行假设检验，以 t 检验为例($t_{0.01}(50-2)=2.68$)，计算结果如表 2 所示。表 2

中对角线上面的数据为定义1中由检验公式(5)所得 t 值,对角线下面的数据为定义2中由检验公式(7)所得 t 值。

表2 春播菜用大豆不同性状团之间广义相关系数假设检验的 t 值

Table 2 Hypothesis testing of generalized correlation coefficient between trait groups of spring vegetable soybean

性状团 Ttrait group	X_1(生育期 Growth period)	X_2(植株形态 Plant shape)	X_3(产量性状 Yield trait)	X_4(荚粒形态 Pod shape)	X_5(品质性状 Quality trait)
X_1(生育期 Growth period)	(**)	−6.53(**)	−15.40(**)	−0.01	0.26
X_2(植株形态 Plant shape)	4.27(**)	(**)	3.01(**)	6.97(**)	0.63
X_3(产量性状 Yield trait)	9.74(**)	2.81(**)	(**)	5.39(**)	−0.77
X_4(荚粒形态 Pod shape)	8.72(**)	4.27(**)	4.87(**)	(**)	−0.01
X_5(品质性状 Quality trait)	6.04(**)	6.25(**)	4.95(**)	4.04(**)	(**)

注:(**)表示某2个性状团之间的相关系数与0之间有显著性差异,其中主对角线上的元素(即表格中无数据只标出(**)的地方)表示性状团自身的相关系数与0之间有显著差异。

Note:(**)show signifies difference with significance of correlation coefficient between trait groups and 0,the null of the table but note (**)show signifies difference with significance of correlation coefficient between trait groups and 0.

表2显示,在由定义1所得出的广义相关系数检验中,生育期与荚粒形态和品质性状的差异不显著,品质性状与生育期、植株形态、产量性状和荚粒形态的差异均不显著;但在由定义2所得出的广义相关系数检验中,5个形状团之间的差异均极显著。因为定义1中的广义相关系数是刻画性状团间的线性相关性,而定义2中的广义相关系数是刻画性状团间的非线性相关性。

4 小 结

本研究采用构造综合指标,借助于一元线性回归模型,对生物性状之间的广义相关系数的抽样分布及假设检验问题进行了分析,该广义相关系数可用 F 检验或者 t 检验进行检验。该广义相关系数的定义、性质、抽样分布及假设检验问题都予以解决,这为实际育种工作提供了一套完整的理论体系。

[参考文献]

[1] 张尧庭.广义相关系数及其应用[J].应用数学学报,1978(4):33-39.
Zhang Y T. Generalized correlation coefficient and its application[J]. Journal of Applied Mathematics,1978(4):33-39.(in Chinese)

[2] 胡永宏.一种广义相关系数[J].统计与信息论坛,1997(1):20-23.
Hu Y H. A kind of generalized correlation coefficient[J]. Statistics and Information Tribune,1997(1):20-23.(in Chinese)

[3] 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.
Zhang Y T. How to measure the correlation among random variables[J]. Journal of Shanghai University of Finance and Economics,1999(2):60-63.(in Chinese)

[4] Nelsen N B. An Introduction to copulas,lectures notes in statistics[M]. New York:Spring Verlag,1998:139 .

[5] Kullback S. Information theory and statistics[M].[s. l.]:John Wiley&Sons Inc,1959.

[6] 刘垂玗.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(增刊):52-57.
Liu C Y. Genetic correlation information and additivity of crop quantity[J]. Journal of Anhui Agricultural Science,1981(S1):52-57.(in Chinese)

[7] 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.
Yuan Z F,Zhou J Y. Mutivariate statistics analysis[M]. Beijing:Science Press,2002:172-180,241-256.(in Chinese)

[8] 陈希孺.概率论与数理统计[M].合肥:科学出版社,2002:126-140.
Chen X R. Probability theory and mathematics statistics[M]. Hefei:Science Press,2002:126-140.(in Chinese)

[9] 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.
Zhou Y F,Zhou D Y. Canonical correlation analysis of growing period,agronomic character,yield and quality character in the spring vegetable soybean[J]. Journal of Fujian Agricultural and Forestry University,2005(3):11-17.(in Chinese)

[10] 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.
Dong X M,Cao B J,Luo F J et al. Non-linear generalized correlation coefficient of biology trait[J]. Journal of Northwest A&F University:Natural Science Edition,2008,36(5):191-195.(in Chinese)