Rostra
rubrica a cura di
Benito V. Frosini

THE GEOMETRIC INTERPRETATION OF THE CORRELATION RATIO*

Renato Leoni

Dipartimento Statistico dell'Università di Firenze – Firenze.

It is well known that the simple linear correlation coefficient r admits two interesting geometrical interpretations as a cosine of an angle. The purpose of this paper is to give a geometric interpretation of the correlation ratio h along lines parallel to those now mentioned for r.

1.

Given the centered values $\tilde{y}_i, \tilde{x}_i$ (i=1,2,...,n) assumed by the quantitative characteristics (variables) Y, X on n objects (individuals), it is well known that several geometric interpretations of the simple linear correlation coefficient r are possible. One of these consists in regarding $\tilde{y}_i, \tilde{x}_i$ (i=1,2,...,n) as vectors, say $\tilde{\mathbf{y}}, \tilde{\mathbf{x}}$, in E^n and in looking at the cosine of the angle α they form. Alternatively, considering in E^2 the concentration ellipse

$$\mathbf{z}' \mathbf{R}^{-1}\mathbf{z} = 1 \ , \ \mathbf{R} = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} , -1 < r < +1$$

it can be shown that

$$r = \begin{cases} \cos(T_2T_1T_4) & \text{if } r \geq 0 \\ \cos(T_1T_2T_3) & \text{if } r \leq 0 \end{cases}$$

---

* *Based on a research partially funded by CNR.*

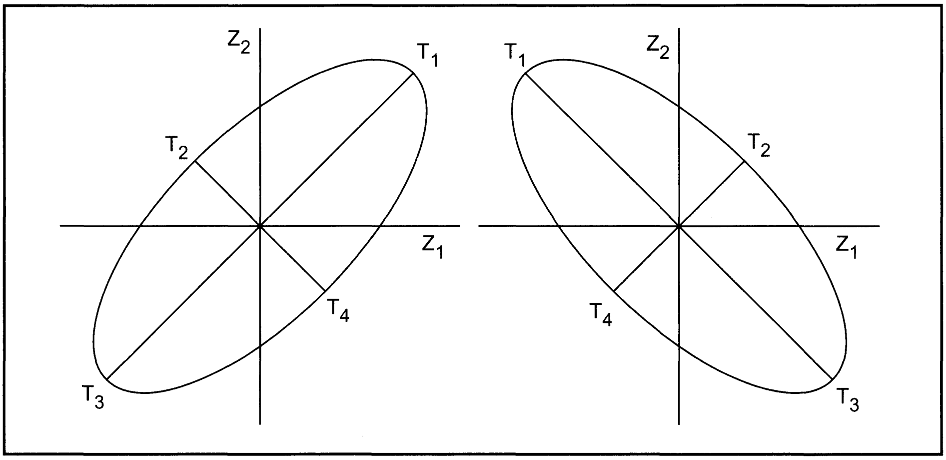

Fig. 1, 2.

where T_1, T_3 and T_2, T_4 are the intersections of the ellipse, respectively, with the major and minor axis (cf. Figures 1,2)[1].

But what about the correlation ratio η? The purpose of this paper is to give a geometric interpretation of η along lines parallel to those mentioned above for r.

2.

Let us suppose that n objects are classified according to q values $y_1,\ldots,y_k,\ldots,y_q$ of a variable Y and to p attributes $a_1,\ldots,a_h,\ldots,a_p$ of a qualitative characteristic A. The results are usually summarized as in Table I, in which n_{hk} (h=1,2,...,p; k=1,2,...,q) denotes the number of objects taking the attribute a_h of A and the value y_k of Y and

$$n_{h\cdot}=\sum_{k=1}^{q} n_{hk}>0\ ,\ n_{\cdot k}=\sum_{h=1}^{p} n_{hk}>0\ ,\ \sum_{h=1}^{p} n_{h\cdot}=\sum_{k=1}^{q} n_{\cdot k}=n.$$

Setting

$$\bar{y}_h=\frac{\sum_{k=1}^{q} y_k n_{hk}}{n_{h\cdot}}\ ,\ \bar{y}=\frac{\sum_{k=1}^{q} y_k n_{\cdot k}}{n}$$

[1] *In these interpretations it is assumed that E^n and E^2 are Euclidean spaces with scalar products represented – with respect to the natural bases of E^n and E^2 – by unit matrices of appropriate order.*

Tab. I.

A	Y					Totals
	y_1	…	y_k	…	y_q	
a_1	n_{11}	…	n_{1k}	…	n_{1q}	$n_{1\cdot}$
.	.	…	.	…	.	.
.	.	…	.	…	.	.
.	.	…	.	…	.	.
a_h	n_{h1}	…	n_{hk}	…	n_{hq}	$n_{h\cdot}$
.	.	…	.	…	.	.
.	.	…	.	…	.	.
.	.	…	.	…	.	.
a_p	n_{p1}	…	n_{pk}	…	n_{pq}	$n_{p\cdot}$
Totals	$n_{\cdot 1}$	…	$n_{\cdot k}$	…	$n_{\cdot q}$	n

the correlation ratio h is defined by the expression

$$\eta = \left(\frac{\sum_{h=1}^{p} (\bar{y}_h - \bar{y})^2 n_{h\cdot}}{\sum_{k=1}^{q} (y_k - \bar{y})^2 n_{\cdot k}} \right)^{1/2}$$

and interpreted as the square root of the ratio between the "interclass" variance and the total variance.

3.

In order to obtain a geometric interpretation of η, let us consider Table II where, besides Y, we have the new variables $X_1, \ldots, X_h, \ldots, X_p$. X_h ($h=1,2,\ldots,p$) denotes the variable "number of times with which the attribute a_h of A is present in an object", and hence

x_{ih} = value taken by X_h on the ith ($i=1,2,\ldots,n$) object

$$= \begin{cases} 1 \text{ if the ith object takes the attribute } a_h \text{ of A} \\ 0 \text{ otherwise.} \end{cases}$$

In turn,

y_i = value taken by Y on the ith object.

Tab. II.

Objects	X_1	...	X_h	...	X_p	Y
1	x_{11}	...	x_{1h}	...	x_{1p}	y_1
.	.	...	.	...	.	.
.	.	...	.	...	.	.
.	.	...	.	...	.	.
i	x_{i1}	...	x_{ih}	...	x_{ip}	y_i
.	.	...	.	...	.	.
.	.	...	.	...	.	.
.	.	...	.	...	.	.
n	x_{n1}	...	x_{nh}	...	x_{np}	y_n
Totals	$n_1.$	...	$n_h.$	...	$n_p.$	$n\bar{y}$

Of course, since there is a one–to–one correspondence between Table I and Table II, their informative contents are the same.

Moreover, writing

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1h} & \cdots & x_{1p} \\ . & \cdots & . & \cdots & . \\ x_{i1} & \cdots & x_{ih} & \cdots & x_{ip} \\ . & \cdots & . & \cdots & . \\ x_{n1} & \cdots & x_{nh} & \cdots & x_{np} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ . \\ y_i \\ . \\ y_n \end{bmatrix}$$

we have

$$\mathbf{u}'\mathbf{X} = \left[n_{1.} \ldots n_{h.} \ldots n_{p.}\right], \quad \mathbf{u}'\mathbf{y} = n\bar{y}$$

where $\mathbf{u}' = [1 \ldots 1 \ldots 1]$.

Therefore, the centered matrix and vector, corresponding to $\mathbf{X}$ and $\mathbf{y}$, are

$$\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{u}\frac{1}{n}\mathbf{u}'\mathbf{X} = \mathbf{X} - \mathbf{u}\left[\frac{n_{1\cdot}}{n} \ldots \frac{n_{h\cdot}}{n} \ldots \frac{n_{p\cdot}}{n}\right], \quad \tilde{\mathbf{y}} = \mathbf{y} - \mathbf{u}\bar{y}.$$

Notice that – since we have assumed $n_h. > 0$ $(h=1,2,\ldots,p)$ – $r(\mathbf{X})=p \leq n$ and $r(\tilde{\mathbf{X}})=p-1$; thus, $\tilde{\mathbf{X}}$ is not of full column rank.

4.

With these premises, we are in a position to give a geometric interpretation of η. Firstly, let

$S(\mathbf{u})$: the vector space spanned by the vector $\mathbf{u}$;

$S(\mathbf{X})$: the vector space spanned by the column vectors of $\mathbf{X}$;

$S(\tilde{\mathbf{X}})$: the vector space spanned by the column vectors of $\tilde{\mathbf{X}}$;

and

P(**u**) : the orthogonal projection matrix onto S(**u**);

P(**X**) : the orthogonal projection matrix onto S(**X**);

P($\tilde{\mathbf{X}}$) : the orthogonal projection matrix onto S($\tilde{\mathbf{X}}$).

Then, since S(**u**) and S($\tilde{\mathbf{X}}$) are orthogonal complements in S(**X**)[(2)], $\forall\, \mathbf{x} \in E^n$ we have

$$\mathbf{P}_{(\mathbf{X})}\mathbf{x} = \mathbf{P}_{(\mathbf{u})}\mathbf{x} + \mathbf{P}_{(\tilde{\mathbf{X}})}\mathbf{x}$$

and hence

$$\mathbf{P}_{(\tilde{\mathbf{X}})} = \mathbf{P}_{(\mathbf{X})} - \mathbf{P}_{(\mathbf{u})} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{u}(\mathbf{u}'\mathbf{u})^{-1}\mathbf{u}'.$$

Therefore, the orthogonal projection $\hat{\tilde{\mathbf{y}}}$ of $\tilde{\mathbf{y}}$ onto S($\tilde{\mathbf{X}}$) is given by

$$\hat{\tilde{\mathbf{y}}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{y}} - \mathbf{u}(\mathbf{u}'\mathbf{u})^{-1}\mathbf{u}'\tilde{\mathbf{y}} = \mathbf{X}\begin{bmatrix} \bar{y}_1 - \bar{y} \\ \cdot \\ \bar{y}_h - \bar{y} \\ \cdot \\ \bar{y}_p - \bar{y} \end{bmatrix}.$$

Now, denoting by α the angle between the vectors $\tilde{\mathbf{y}}$ and $\hat{\tilde{\mathbf{y}}}$, it is very easy to show that $\cos^2\alpha$ is just η^2.

In fact, since ($\mathbf{X}'\mathbf{X} = \text{diag}\,[n_{1.} \dots n_{h.} \dots n_{p.}]$)

$$\hat{\tilde{\mathbf{y}}}'\hat{\tilde{\mathbf{y}}} = \left[\bar{y}_1 - \bar{y} \dots \bar{y}_h - \bar{y} \dots \bar{y}_p - \bar{y}\right]\mathbf{X}'\mathbf{X}\begin{bmatrix} \bar{y}_1 - \bar{y} \\ \cdot \\ \bar{y}_h - \bar{y} \\ \cdot \\ \bar{y}_p - \bar{y} \end{bmatrix} = \sum_{h=1}^{p}\left(\bar{y}_h - \bar{y}\right)^2 n_h.$$

---

2 *Actually, for $\forall\, a \in R$ and $\forall\, \mathbf{b} \in R^p$, we have*

(i) $X = \tilde{X} + u\frac{1}{n}u'X \Rightarrow Xb = \tilde{X}b + u\left(\left[n_{1.} \dots n_{h.} \dots n_{p.}\right]b\right)$

(ii) $a\mathbf{u}'\tilde{\mathbf{X}}\,\mathbf{b} = 0$;

then, $S(X) = S(u) \oplus S(\tilde{X})$ and $S(u) \perp S(\tilde{X})$.

and

$$\tilde{\mathbf{y}}'\tilde{\mathbf{y}} = (y - u\bar{y})'(y - u\bar{y}) = \mathbf{y}'\mathbf{y} - \mathbf{u}'\mathbf{n}\bar{y}^2 = \sum_{k=1}^{q}(y_k - \bar{y})^2 n_{\cdot k}$$

we have[(3)]

$$\cos^2\alpha = \frac{\hat{\tilde{\mathbf{y}}}'\hat{\tilde{\mathbf{y}}}}{\tilde{\mathbf{y}}'\tilde{\mathbf{y}}} = \eta^2.$$

5.

Another geometric interpretation of η is easily obtained by considering in E^2 the ellipse

$$\mathbf{z}'\mathbf{N}^{-1}\mathbf{z} = 1\ ,\ \ \mathbf{N} = \begin{bmatrix} 1 & \eta \\ \eta & 1 \end{bmatrix}\ ,\ \ 0 \le \eta < +1.$$

Denoting by d and D the lengths of the minor and major axes of the ellipse (cf. Figure 1), since

$$d = 2\sqrt{1-\eta}\ ,\ \ D = 2\sqrt{1+\eta}$$

we have

$$\eta = \frac{D^2 - d^2}{D^2 + d^2} = \frac{1 - \dfrac{d^2}{D^2}}{1 + \dfrac{d^2}{D^2}} = \frac{1-t^2}{1+t^2}.$$

But $t = d/D = \tan(T_2T_1T_4/2)$; thus,

$\eta = \cos(T_2T_1T_4)$.

Of course, if the angle $T_2T_1T_4$ is rectangular then the correlation ratio is 0. In turn, narrow ellipses correspond to angles close to 0 and, therefore, to correlation ratios close to +1.

---

3 *Notice that this is equivalent to the squared multiple linear correlation coefficient between Y and the set of variables* $X_1, \ldots, X_h, \ldots, X_p$.

REFERENCES

Châtillon G.: *The Balloon Rules for a Rough Estimate of the Correlation Coefficient.* The American Statistician, pp. 58–60; 1984.

Châtillon G.: *Reply.* Ther American Statistician, pp. 330–331; 1984.

Gypen L.M.J.: *Comment on Rogers and Nicewander.* The American Statistician, pp. 291, 1988.

Marks E.: *A note on the Geometric Interpretation of the Correlation Coefficient.* Journal of Educational Statistics, pp. 233–237; 1982.

Ozer D.J.: *Correlation and the Coefficient of Determination.* Psychological Bulletin, pp. 307–315; 1985.

Rodgers J.L., Nicewander W.A.: *Thirteen Ways to Look at the Correlation Coefficient.* The American, Statistician, pp. 59–66; 1988.

Rodgers J.L.: *Reply.* The American Statistician, p. 291, 1988.

Schilling M.F.: *Some Remarks on Quick Estimation of the Correlation Coefficient.* The American Statistician, p. 330, 1984.

Thöni H.: *Comment on Rogers and Nicewander.* The American Statistician, p. 290, 1988.

SUMMARY

È ben noto che il coefficiente di correlazione lineare semplice r ammette due interessanti interpretazioni geometriche come coseno di un angolo. Lo scopo di questa nota è quello di mostrare la possibilità di interpretare in termini analoghi il rapporto di correlazione η.