Applied Mathematics, 2014, 5, 3411-3420
Published Online December 2014 in SciRes. http://www.scirp.org/journal/am
http://dx.doi.org/10.4236/am.2014.521318

Scientific Research

The Influence Function of the Correlation Indexes in a Two-by-Two Table*

Giovanni Girone, Fabio Manca, Claudia Marin

University of Bari "Aldo Moro", Bari, Italy
Email: giovanni.girone@uniba.it, fabio.manca@uniba.it, claudia.marin@uniba.it

Received 5 October 2014; revised 28 October 2014; accepted 11 November 2014

Abstract

In this paper we examine 5 indexes (the two Yule's indexes, the chi square, the odds ratio and an elementary index) of a two-by-two table, which estimate the correlation coefficient ρ in a bivariate Bernoulli distribution. We will find the compact expression of the influence functions, which allow the quantification of the effect of an infinitesimal contamination of the probability of any pair of attributes of the bivariate random variable distributed according to the above-mentioned model. We prove that the only unbiased index is the chi square. In order to determine the indexes, which are less sensitive to contamination, we obtain the expressions of three synthetic measures of the influence function, which are the maximum contamination (gross sensitivity error), the mean square deviation and the variance. These results, even if don't allow a definitive assessment of the overall optimum properties of the five indexes, as not all of them are unbiased, nevertheless they allow to appreciating the synthetic entity of the effect of the contaminations in the estimation of the parameter ρ of the bivariate Bernoulli distribution.

Keywords

Two-by-Two Table, Influence Function, Correlation Indexes, Gross Sensitivity Error, Mean Square Deviation, Asymptotic Variance

1. Introduction

In this paper we analyze the influence of a minimal contamination of the bivariate Bernoulli distribution on the values of the index measuring the association in a two-by-two table, having as a scenario the estimation of the correlation parameter of that distribution.

*The paragraphs 1 and 2 are attributed to G. Girone, the paragraphs 3, 4 and 8 to F. Manca and the paragraphs 5, 6 and 7 to C. Marin.

How to cite this paper: Girone, G., Manca, F. and Marin, C. (2014) The Influence Function of the Correlation Indexes in a Two-by-Two Table. *Applied Mathematics*, **5**, 3411-3420. http://dx.doi.org/10.4236/am.2014.521318

2. Bivariate Bernoulli Model

Selecting a Template

Let us suppose that two dichotomous variables, denoted by X and Y, are relevant within a population. These variables take the values 1 and 0, depending on whether one of the dichotomous attributes is present or absent. The corresponding theoretical model is the bivariate Bernoulli distribution [1], reported in **Table 1**.

The mean values of the two variables are

$$E[X] = \gamma + \delta, \tag{1}$$

$$E[Y] = \beta + \delta. \tag{2}$$

The variances of the two variables are

$$Var[X] = (\alpha + \beta)(\gamma + \delta), \tag{3}$$

$$Var[Y] = (\alpha + \gamma)(\beta + \delta). \tag{4}$$

The covariance between the two variables is

$$Cov[X, Y] = \alpha\delta - \beta\gamma. \tag{5}$$

The correlation coefficient is

$$\rho = \frac{\alpha\delta - \beta\gamma}{\sqrt{(\alpha + \beta)(\gamma + \delta)(\alpha + \gamma)(\beta + \delta)}}. \tag{6}$$

3. Properties of the Correlation Parameter Estimation

Several indexes, suggested by various authors (Yule, Quetelet and others), are available for the sample estimation of the above-mentioned correlation coefficient. We refer to such indexes as R_1, R_2 etc. For given indexes, R_h, all variable between −1 and +1, we must take into account *unbiasedness, i.e.*

$$E[R_h] = \rho, \tag{7}$$

efficiency, i.e.

$$Var[R_h] = \text{minimum}, \tag{8}$$

and the limited influence of limited modification of the model.

With regard to this last fundamental property Hampel [2] in 1974 suggested the *influence function* as a tool for evaluating the effect caused on the value of an indexby a minimal contamination of the model. In our case the model is the bivariate Bernoulli distribution, the parameter is the correlation coefficient ρ and the indexes are those proposed by various authors over time.

Basically the influence function $If[R]$ referred to the index R is given by

$$If[R] = \lim_{\varepsilon \to 0} \frac{R[H_\varepsilon] - R[H]}{\varepsilon}, \tag{9}$$

where $R[H_\varepsilon]$ is the index computed for the contaminated bivariate Bernoulli distribution H_ε, $R[H]$ is the

Table 1. The bivariate Bernoulli distribution.

Attributes of variable X	Attributes of variable Y		Total
	0	1	
0	α	β	$\alpha + \beta$
1	γ	δ	$\gamma + \delta$
Total	$\alpha + \gamma$	$\beta + \delta$	1

index computed for the non-contaminated bivariate Bernoulli distribution and ε is the weight of the contamination.

It is easily understood that such a function measures the effect of an infinitesimal contamination of the model on the value of the correlation index [3]. From now on we will denote by *a*, *b*, *c* and *d* the empirical frequencies of the four cells of the two-by-two table obtained for a sample of *n* units.

4. Influence Function of the Correlation Indexes

4.1. *C* Index

Let us first consider the elementary index given by

$$C = \frac{a+d-b-c}{a+d+b+c}. \tag{10}$$

A contamination in the cell (0,0) leads to the influence function value

$$If\left[(0,0),C\right] = \frac{2(b+c)}{(a+b+c+d)^2}, \tag{11}$$

while a contamination in the cell (1,1) leads to the influence function value

$$If\left[(1,1),C\right] = \frac{2(b+c)}{(a+b+c+d)^2}. \tag{12}$$

That is, a contamination in one of the two cells indicating concordance increases the value of the *C* index of a quantity, which is proportional to the sum of the frequencies of the two discordance cells.

A contamination in the cell (0,1) leads to the influence function value

$$If\left[(0,1),C\right] = -\frac{2(a+d)}{(a+b+c+d)^2}, \tag{13}$$

while a contamination in the cell (1,0) leads to the influence function value

$$If\left[(1,0),C\right] = -\frac{2(a+d)}{(a+b+c+d)^2}. \tag{14}$$

That is, a contamination in one of the two cells indicating discordance decreases the value of the *C* index of a quantity, which is proportional to the sum of the frequencies of the two concordance cells.

In short, the influence function can be displayed as

$$If\left[(x,y),C\right] = \frac{2v}{(a+b+c+d)^2}, \tag{15}$$

in which *v*, in the case of a concordance cell, is equal to the sum of the discordance frequencies, while, in the case of a discordance cell, is equal to the sum of the frequencies of discordance cells, changed of sign. In other words, the influence of the contamination for each concordance cell is directly proportional to the sum of the discordant frequencies and vice versa, for each discordance cell, provided that it is positive for the concordance cells and negative for the discordance cells [4].

4.2. Yule's *Q* Index

Let us first consider the 1900 Yule's index [5] given by

$$Q = \frac{ad-bc}{ad+bc}. \tag{16}$$

A contamination in the cell (0,0) leads to the following value of the influence function

$$If\left[(0,0),Q\right]=\frac{2bcd}{(ad+bc)^2}, \tag{17}$$

while a contamination in the cell (1,1) leads to the value given by

$$If\left[(1,1),Q\right]=\frac{2abc}{(ad+bc)^2}. \tag{18}$$

That is, a contamination in one of the two cells indicating concordance increases the value of the index Q by a quantity, which is proportional to the product of the frequencies of the three non-contaminated cells.

On the other hand a contamination in the cell (0,1) leads to the value of the influence function given by

$$If\left[(0,1),Q\right]=-\frac{2acd}{(ad+bc)^2}, \tag{19}$$

while a contamination in the cell (1,0) leads to the following value of the influence function

$$If\left[(1,0),Q\right]=-\frac{2abd}{(ad+bc)^2}. \tag{20}$$

That is, a contamination in one of the two cells indicating discordance decreases the value of the index Q by a quantity, which is proportional to the product of the frequencies of the three non-contaminated cells.

In short, the influence function can be displayed as

$$If\left[(x,y),Q\right]=\frac{2abcd}{v(ad+bc)^2}, \tag{21}$$

in which v is equal to one of the frequencies with positive sign if it corresponds to a or to d, and to one of the frequencies with negative sign if it corresponds to b or to c. In other words, the influence of the contamination is inversely proportional to the frequency of the contaminated cell, provided that it is positive for the concordance cells and negative for the discordance cells.

4.3. Yule's *Y* Index

Let us consider now the other index proposed by Yule [6] in 1912,

$$Y=\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}. \tag{22}$$

A contamination in the cell (0,0) leads to the following value of the influence function

$$If\left[(0,0),Y\right]=\frac{\sqrt{bcd}}{\sqrt{a}\left(\sqrt{ad}+\sqrt{bc}\right)^2}, \tag{23}$$

while a contamination in the cell (1,1) leads to the value of the influence function given by

$$If\left[(1,1),Y\right]=\frac{\sqrt{abc}}{\sqrt{d}\left(\sqrt{ad}+\sqrt{bc}\right)^2}. \tag{24}$$

That is, a contamination in one of the two cells indicating concordance increases the value of the index Y by a quantity proportional to the root of the product of the frequencies of the three non-contaminated cells divided by the root of the frequency of the contaminated cell.

A contamination in the cell (0,1) leads to the value of the influence function given by

$$If\left[(1,0),Y\right]=-\frac{\sqrt{acd}}{\sqrt{b}\left(\sqrt{ad}+\sqrt{bc}\right)^2}, \tag{25}$$

while a contamination in the cell (1,0) leads to the following value of the influence function

$$If\left[(0,1),Y\right]=-\frac{\sqrt{abd}}{\sqrt{c}\left(\sqrt{ad}+\sqrt{bc}\right)^2}. \tag{26}$$

That is, a contamination in one of the two cells indicating discordance decreases the value of the index Y by a quantity proportional to the root of the product of the frequencies of the three non-contaminated cells divided by the root of the frequency of the contaminated cell.

In short, the influence function can be displayed as

$$If\left[(x,y),Y\right]=\frac{\sqrt{abcd}}{v\left(\sqrt{ad}+\sqrt{bc}\right)^2}, \tag{27}$$

in which v is equal to one of the frequencies with positive sign if a or d and to one of the frequencies with negative sign if b or c. In other words, the influence of the contamination is inversely proportional to the frequency of the contaminated cell, provided that it is positive for the concordance cells and negative for the discordance cells.

4.4. The Chi Square Index

Let us examine the chi square index

$$\chi^2=\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}. \tag{28}$$

A contamination in the cell (0,0) leads to the influence function value

$$If\left[(0,0),\chi^2\right]=\frac{(b+d)(c+d)\left[(b+c)(ad+bc)+2bc(a+d)\right]}{2\left[(a+b)(a+c)(b+d)(c+d)\right]^{3/2}}, \tag{29}$$

while a contamination in the cell (1,1) leads to the influence function value

$$If\left[(1,1),\chi^2\right]=\frac{(a+b)(a+c)\left[(b+c)(ad+bc)+2bc(a+d)\right]}{2\left[(a+b)(a+c)(b+d)(c+d)\right]^{3/2}}. \tag{30}$$

That is, a contamination in one of the two cells indicating concordance increases the value of the χ^2 by a quantity which is proportional to the product of the sums of the frequency of the other concordance cell with each of the frequencies of the discordance cell.

A contamination in the cell (0,1) leads to the influence function value

$$If\left[(1,0),\chi^2\right]=-\frac{(a+c)(c+d)\left[(a+d)(ad+bc)+2ad(b+c)\right]}{2\left[(a+b)(a+c)(b+d)(c+d)\right]^{3/2}}, \tag{31}$$

while a contamination in the cell (1,0) leads to the influence function value

$$If\left[(0,1),\chi^2\right]=-\frac{(a+b)(b+d)\left[(a+d)(ad+bc)+2ad(b+c)\right]}{2\left[(a+b)(a+c)(b+d)(c+d)\right]^{3/2}}. \tag{32}$$

That is, a contamination in one of the two cells indicating discordance decreases the value of the χ^2 by a quantity which is proportional to the product of the sums of the frequency of the other discordance cell with each of the frequencies of the concordance cells.

It is impossible to have a unique expression of the influence function as we had for the other indexes, because the expressions for the contaminated concordance cells differ from those related to the discordance ones.

4.5. Odds Ratio, θ

Let us now examine the odds ratio index

$$\theta = \frac{ad}{bc}. \tag{33}$$

A contamination in the cell (0,0) leads to the influence function value

$$If\left[(0,0),\theta\right] = \frac{d}{bc}, \tag{34}$$

while a contamination in the cell (1,1) leads to the influence function value

$$If\left[(1,1),\theta\right] = \frac{a}{bc}. \tag{35}$$

That is, a contamination in one of the two cells indicating concordance increases the value of the index θ by a quantity that is proportional to the frequency of the other concordance cell and inversely proportional to the product of the frequencies of the discordance cells.

A contamination in the cell (0,1) leads to the influence function value

$$If\left[(0,1),\theta\right] = -\frac{ad}{b^2c}, \tag{36}$$

while a contamination in cell (1,0) leads to the influence function value

$$If\left[(1,0),\theta\right] = -\frac{ad}{bc^2}. \tag{37}$$

That is, a contamination in one of the two cells indicating discordance decreases the value of the index θ by a quantity, which is proportional to the product of the frequencies of the concordance cell and inversely proportional to the product of the square of the frequency of the contaminated cell multiplied by the frequency of the other discordance cell.

In short, the influence function can be displayed as

$$If\left[(x,y),\theta\right] = \frac{ad}{bcv} = \frac{\theta}{v}, \tag{38}$$

in which v is equal to the frequency of the contaminated cell, provided that the sign is positive in case of contamination in a concordance cell, and negative in case of contamination in a discordance cell.

5. Unbiasedness of the Indexes

It must be reminded that an index R_h is unbiased if

$$E\left(R_h\right) = \rho, \tag{39}$$

let us examine now the unbiasedness of every index.

5.1. The Chi Square Index

The index

$$\chi^2 = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}, \tag{40}$$

has the mean

$$E\left(\chi^2\right) = E\left(\frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}\right) = \rho, \tag{41}$$

and therefore it is unbiased.

5.2. Other Indexes

Indexes Q, Y, θ e C are biased.

It has to be said that 3 of the considered indexes (Q, Y and θ) are functionally related, as it is shown below:

$$Q=\frac{2Y}{1+Y^2},\quad Y=\frac{1+\sqrt{1-Q^2}}{Q}\quad \text{for}\quad -1<Q<0\quad \text{and}\quad Y=\frac{1-\sqrt{1-Q^2}}{Q}\quad \text{for}\quad 0<Q<1,\quad Q=\frac{\theta-1}{\theta+1},$$

$$\theta=\frac{1+Q}{1-Q},\quad \theta=\frac{(1+G)^2}{(1-G)^2},\quad G=\frac{1-\sqrt{\theta}}{1+\sqrt{\theta}},\ \text{for}\quad 0<\theta<1\quad \text{and}\quad G=\frac{1+\sqrt{\theta}}{1-\sqrt{\theta}},\ \text{for}\quad 1<\theta<\infty.$$

The other 2 indexes (C and χ^2) are not functionally explainable with themselves nor with the above-mentioned ones. The 5 indexes estimate functions of the ρ parameter. More exactly 4 of these indexes (C, Q, Y and χ^2) are estimators of increasing functions of this parameter and, in particular in the points −1, 0 and +1, these functions coincide with the argument. So the index θ can easily lead to the 2 Yule's indexes achieving again its characteristics.

6. Influences of the Indexes

Since that the effects of contaminations in the various cells are balanced, it is necessary to evaluate their overall influence regardless of the sign. This can be done considering the *maximum of the absolute values* of the influence or the *mean absolute deviation* or the *variance* of the said values [7].

6.1. Maximum of the Absolute Values of the Influence Function (Gross Sensitivity Error)

6.1.1. *C* Index

As, regardless of the sign, the influence function is equal to

$$If\left(v,C\right)=\frac{a+d}{\left(a+b+c+d\right)^2}\ \text{or}\ If\left(v,C\right)=\frac{b+c}{\left(a+b+c+d\right)^2},\tag{42}$$

the maximum of the influence function is therefore

$$GSE\left(C\right)=\frac{2\left(a+d\right)\left(b+c\right)}{\min\left(a+d,b+c\right)\left(a+b+c+d\right)^2}.\tag{43}$$

6.1.2. Yule's *Q* Index

As, regardless of the sign, the influence function is equal to

$$If\left(v,Q\right)=\frac{2abcd}{v\left(ad+bc\right)^2},\tag{44}$$

in which v is one of the four frequencies of the table, the maximum of the influence function is obtained for $\min(v)$; it is therefore

$$GSE\left(Q\right)=\frac{2abcd}{\min\left(a,b,c,d\right)\left(ad+bc\right)^2}.\tag{45}$$

6.1.3. Yule's *Y* Index

As, regardless of the sign, the influence function is equal to

$$If\left(v,Y\right)=\frac{\sqrt{abcd}}{v\left(\sqrt{ad}+\sqrt{bc}\right)^2},\tag{46}$$

in which v is one of the four frequencies of the table, the maximum of the influence function is obtained for $\min(v)$; it is therefore

$$GSE\left(Y\right)=\frac{\sqrt{abcd}}{\min\left(a,b,c,d\right)\left(\sqrt{ad}+\sqrt{bc}\right)^2}.\tag{47}$$

6.1.4. Chi Square Index

An empirical analysis allows to asses that the maximum absolute value of the influence function is obtained in correspondence of the minimum frequency. Thus,

$$GSE\left(\chi^2\right) = If\left[(0,0),\chi^2\right] \text{ if } a < \min(b,c,d), \tag{48}$$

$$GSE\left(\chi^2\right) = If\left[(1,1),\chi^2\right] \text{ if } d < \min(a,b,c), \tag{49}$$

$$GSE\left(\chi^2\right) = If\left[(0,1),\chi^2\right] \text{ if } b < \min(a,c,d), \tag{50}$$

$$GSE\left(\chi^2\right) = If\left[(1,0),\chi^2\right] \text{ if } c < \min(a,b,d). \tag{51}$$

6.1.5. Odds Ratio

As, regardless of the sign, the influence function is equal to

$$If\left(v,\theta\right) = \frac{ad}{vbc}, \tag{52}$$

in which v is one of the four frequencies of the table, the maximum of the influence function is obtained for $\min(v)$; it is therefore

$$GSE\left(\theta\right) = \frac{ad}{\min\left(a,b,c,d\right)bc}. \tag{53}$$

6.2. Variability of Influence Functions: Mean Absolute Deviation

6.2.1. *C* Index

A few algebraic steps allow us to obtain

$$MD(C) = \frac{4(a+d)(b+c)}{(a+b+c+d)^3}. \tag{54}$$

6.2.2. Yule's *Q* Index

$$MD(Q) = \frac{8adbc}{(ad+bc)^2}. \tag{55}$$

6.2.3. Yule's *Y* Index

$$MD(Y) = \frac{4\sqrt{adbc}}{ad+bc+2\sqrt{adbc}}. \tag{56}$$

6.2.4. Chi Square

$$MD\left(\chi^2\right) = \frac{\left[2bc(a+d)+(b+c)(ad+bc)\right]\left[2ad(b+c)+(a+d)(ad+bc)\right]}{\left[(a+b)(a+c)(b+d)(c+d)\right]^{3/2}}. \tag{57}$$

6.2.5. Odds Ratio

$$MD(\theta) = \frac{4ad}{bc}. \tag{58}$$

It can be seen that the mean deviation for all indexes is a symmetric function either of the concordant fre-

quencies or of the discordant frequencies.

6.3. Variability of the Influence Function Asymptotic Variance (A.S.V.)

6.3.1. *C* Index

Let us consider the asymptotic variance of the indexes. A few algebraic steps lead us to the following expression

$$ASV(C)=\frac{4(a+d)(b+c)}{(a+b+c+d)^4}. \tag{59}$$

6.3.2. Yule's *Q* Index

$$ASV(Q)=\frac{4adbc\left[ad(b+c)+bc(a+d)\right]}{(ad+bc)^4}. \tag{60}$$

6.3.3. Yule's *Y* Index

$$ASV(Y)=\frac{ad(b+c)+bc(a+d)}{\left(ad+bc+2\sqrt{adbc}\right)^2}. \tag{61}$$

6.3.4. Chi Square Index

$$ASV(\chi^2)=\frac{\left\{(a+d)(b+c)\left(6adbc+b^2c^2+a^2d^2\right)+4adbc\left[(a+d)^2+(b+c)^2\right]\right\}}{\left[2(a+b)(a+c)(b+d)(c+d)\right]^2}. \tag{62}$$

6.3.5. Odds Ratio

$$ASV(\theta)=\frac{ad\left[ad(b+c)+bc(a+d)\right]}{(bc)^3}. \tag{63}$$

It can be seen that the asymptotic variance is a symmetric function of the concordance and discordance frequencies as well.

7. Example

Let us consider a practical example in which 1071 persons are classified on 2 dichotomic characters: "does he/she smoke" and "is he/she suffering from bronchitis?" both with yes or no response (see **Table 2**).

There were 1071 cases of which 135 smoke and have bronchitis and 547 don't smoke and don't have bronchitis.

Elementary indexes	*C*	*Q*	*Y*	χ^2	θ
G.S.E.	0.00119	0.00399	0.00232	0.00179	0.02473
M.D.	0.00086	0.00152	0.00089	0.00076	0.00942
A.S.V.	0.0000008	0.0000035	0.0000012	0.0000009	0.0001338
S.D.	0.00090	0.00186	0.00109	0.00094	0.01157

As it can be noticed, between the 4 indexes whose values go between −1 and +1, the ones which are less sensitive to contamination are *C* and chi square indexes; on the other hand, the more sensitive ones are Yule's indexes, *Q* and *Y*. The greater sensitivity of the odds ratio is due to the fact that such index measures a function of the correlation of the model that goes in the range from 1 to ∞.

Table 2. Smoke versus bronchitis.

Smoke	Bronchitis		Total
	Yes	No	
Yes	135	287	422
No	102	547	649
Total	237	834	1071

Source: Survey at the University Hospital of Bari, Department of Pulmonology.

8. Conclusions

In this paper we analyzed the indexes of a two-by-two table, which allow the estimation of the correlation coefficient ρ in the bivariate Bernoulli model. More precisely, we considered the two Yule's indexes, the chi square, the odds ratio and a further elementary index. We obtained, for these indexes, the compact expressions of the influence functions, which allow the quantification of the effect of an infinitesimal contamination of the probability of any pair of attributes of the bivariate random variable distributed according to the above-mentioned model.

In order to determine the indexes which are less sensitive to contamination, we obtained the expressions of three synthetic measures of the influence function, specifically the maximum contamination (gross sensitivity error), the mean absolute deviation and the variance. These expressions, even if don't allow a definitive assessment of the overall optimum properties of the five indexes considered, as not all of them are unbiased, nevertheless they allow to appreciating the synthetic entity of the effect of the contaminations in the estimation of the parameter ρ of the bivariate Bernoulli model.

References

[1] Barnard, G.A. (1981) Two by Two (2 × 2) Tables. *Encyclopedia of Statistical Sciences*, **9**, 367-372.

[2] Hampel, F.R. (1974) The Influence Curve and Its Role in Robust Estimation. *Journal of the American Statistical Association*, **69**, 383-393. http://dx.doi.org/10.1080/01621459.1974.10482962

[3] Kendall, M.G. and Stuart, A. (1977) The Advanced Theory of Statistics. Vol. 2, C. Griffin, London, 566-571.

[4] Pearson, K. (1904) On the Theory of Contingency and Its Relation to Association and Normal Correlation. Biometric Series, Drapers' Co. Memoirs, London.

[5] Yule, G.U. (1900) On the Association of Attributes in Statistics. *Philosophical Transaction*, **194**, 257. http://dx.doi.org/10.1098/rsta.1900.0019

[6] Yule, G.U. (1912) On the Methods of Measuring Association between Two Attributes. *Journal of the Royal Statistical Society*, **75**, 579. http://dx.doi.org/10.2307/2340126

[7] Yule, G.U. and Kendal, M.G. (1958) An Introduction to the Theory of Statistics. C. Griffin, London, 271-272.