Correlation paradox of binary classification system

Tamar Gadrich[1], and Emil Bashkansky[2]

[1] ORT Braude college, Karmiel, Israel 21982

[2] ORT Braude college, Karmiel, Israel 21982

[1] Corresponding author: Tamar Gadrich, e-mail: tamarg@braude.ac.il

Cost and speed considerations often conduct to measure/classify items according to two comprehensive and exclusive classes/categories (binary scale). Assume items are submitted for sorting according to binary scale (e.g. Type 1 and Type 2) using some sorting machine (SM). Repeatability or consistency testing of the SM is examined by the help of the correlation coefficient between the random variables denoting the number of items that were classified to the first category (for example) in two sequential sorting procedures. Paradoxical results lead to the conclusion that this measure is not suitable for measuring repeatability and further research is needed in order to find methods for checking consistency.

Keywords: Binary scale, Correlation coefficient, Sorting machine (SM), Repeatability

1.   Introduction

Stevens (1946) proposed conventional way to measure objects according to four measurement scales. The first two scales are called categorical (nominal and ordinal scales) while the two others are called numerical (interval and ratio scales). Both in nominal and ordinal scales, the objects are coded with labels/ names that are corresponding to exhaustive and disjoint categories. While in nominal scale there is no ordering between the categories, in ordinal scale the situation is different and the codes that are given to the different categories are imposing some ordering of the objects. So, the only legitimate operations between nominal objects are comparison (equal, unequal) of their codes while between ordinal objects in addition (to equal and unequal) we can compare in form of greater than or smaller than.

This paper deals with classification of items according to binary scale, such as: Type 1/Type 2, defective/non-defective, etc., which can consider as a nominal case but also as an ordinal case. The usage of binary scale is usually the method of choice when rapid results are needed. Here we assume that items are classified by a sorting machine (SM). The repeatability of the SM is examined by two sequential sorting processes. Assume that the distribution of *n* items according to two binary scale types (Type 1/Type 2) is known. These items are submitted for SM examination and only the items that have been classified as Type 1 are re-examined by the same SM (Gertsbakh (1962), Gertsbakh and Friedman (1985)). The repeatability of the SM is examined through a fit (e.g. correlation coefficient) between the results of these two sorting processes (e.g., the amount of items that were classified as Type 1 after the

first and second sorting). As will be shown in Section 3, the results are unexpected and somewhat quite paradoxical.

2.   Preliminaries assumptions

When dealing with binary scale, the SM's performance can be described by the following square stochastic classification matrix $\hat{P}$ (Bashkansky et al., 2007),

$$\hat{P} = \frac{1}{2}\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix},$$

where α is a Type I error rate (false positive or false alarm) and β is a Type II error rate (false negative). Assume n items are given: n_1 items of Type 1 and n_2 items of Type 2, so a total of $n = (n_1 + n_2)$ items are submitted to the SM for classification into two types: Type 1 and Type 2. We emphasize that the classification of the n items is known a-priori.

3.   Results

Due to the SM classification errors, we can find two kinds of items that will be classified as Type 1: (a) an item that is actually Type 1 and was classified as Type 1; and (b) an item that is actually Type 2 but was classified as Type 1 due to the SM's classification error.

Define the following two random variables. Let X_1 be the total number of items that were classified as Type 1 after the first classification. Let Z_1 be the number of items among X_1 that were classified as Type 1 in the first classification and that are also actually Type 1. Obviously, $Z_1 \leq X_1$. One can see, from Fig. 1, that the random variable X_1 is composed from a sum of two random variables.
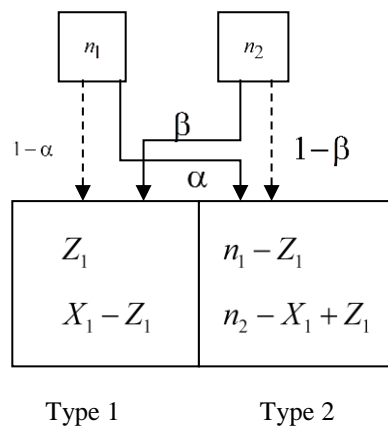

Fig.1. The first sorting and classification errors

Only the X_1 items are sorted again using the same SM and the same binary scale. Let Y_1 be the number of items classified as Type 1 after the second classification. Clearly, $Y_1 \le X_1$. As above, Y_1 is a sum of two random variables (see Fig. 2): W_1 denotes the number of items that are actually Type 1 and classified as Type 1 during the second sorting and $(Y_1 - W_1)$ is the number of items that are actually Type 2, but were classified as Type 1 in the second sorting.
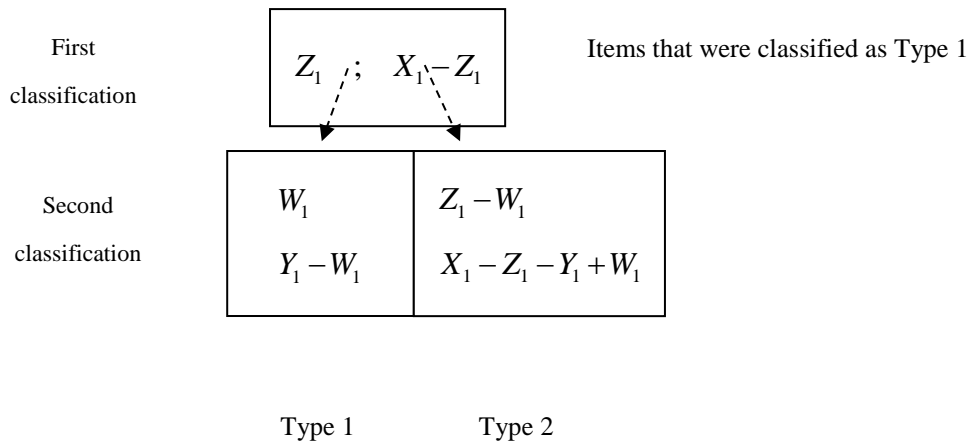
| First classification | Z_1 ; $X_1 - Z_1$ | Items that were classified as Type 1 |

| Second classification | W_1 | $Z_1 - W_1$ |
| | $Y_1 - W_1$ | $X_1 - Z_1 - Y_1 + W_1$ |

Type 1   Type 2

Fig.2. Two sequential sorting processes

3.1 Two sequential sorting processes in the case when only one item type is submitted to the SM

Let's us begin with a private case. Assume that all the n items that are submitted for SM examination are of Type 1 only (i.e., $n_2 = 0$, $n_1 = n$). The random variable $X_1 (= Z_1)$, that denotes the number of items that were classified as Type 1 (success), i.e., reflecting the number of successes in n independent Bernoulli trials. Now, the X_1 items are sorted again using the same SM and the same binary scale. Let $Y_1 (= W_1)$ be the number of items classified as Type 1 after the second classification. Clearly, $Y_1 \le X_1$. Some calculations lead to the following results:

$$X_1 \sim Bin(n, 1-\alpha); \quad Y_1 | X_1 = k \sim Bin(k, 1-\alpha),$$

$$E(X_1) = n(1-\alpha); \quad VAR(X_1) = n\alpha(1-\alpha),$$

$$E(Y_1) = n(1-\alpha)^2; \quad VAR(Y_1) = n(1-\alpha)^2\left(1-(1-\alpha)^2\right),$$

$$COV(X_1, Y_1) = n\alpha(1-\alpha)^2; \quad \rho = \sqrt{\frac{1-\alpha}{2-\alpha}}.$$

Fig. 3 presents the values of ρ as a function of α. The naïve hypothesis is that as

$\alpha \to 0$, the value of $\rho \to 1$, but as seen from Fig. 3, its maximal value is only

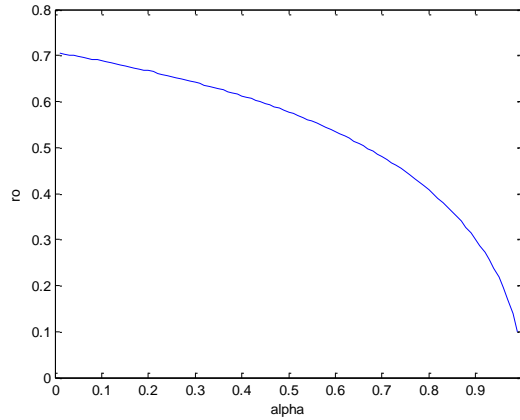$$\frac{1}{\sqrt{2}} \approx 0.707 .$$

Fig. 3. The correlation coefficient as a function of a Type I error ($n_2 = 0$).

Return now to the general case. Assume we are given n items such that n_1 items of

Type 1 and n_2 items of Type 2, a total of $n = (n_1 + n_2)$ items are submitted to the

SM for classification into two types: Type 1 and Type 2

3.2   First application of the sorting machine

As follows from Fig. 1, X_1 is distributed according to the convolution of the two

random variables, $Z_1 \sim Bin(n_1, 1-\alpha)$ and $X_1 - Z_1 | Z_1 = k \sim Bin(n_2, \beta)$ $(0 \le k \le n_1)$.

leads to the following results:

$$E(X_1) = n_1(1-\alpha) + n_2\beta ,$$

$$VAR(X_1) = n_1(1-\alpha)\alpha + n_2\beta(1-\beta) .$$

Now, in order to evaluate the consistency of the SM, only the items that were classified as Type 1 after the first classification undergo a second classification by the same SM. In the following section we deal with the issue of evaluating the repeatability of this SM.

3.3   Determining the repeatability of the sorting machine

In the second sorting, only the X_1 items are sorted again using the same SM and the same binary scale. The naïve attitude is to assume that precise classification is

expressed by high correlation between X_1 and Y_1, however calculations show the following results.

The joint probability distribution function of X_1 and Y_1 is given by:

$$P(X_1 = x_1, Y_1 = y_1) =$$

$$\sum_{k=\max\{0, x_1-n_2\}}^{\min\{x_1, n_1\}} \sum_{h=\max\{0, y_1-x_1+k\}}^{\min\{k, y_1\}} \binom{n_1}{k}\binom{n_2}{x_1-k}(1-\alpha)^k \alpha^{n_1-k}\beta^{x_1-k}(1-\beta)^{n_2-x_1+k}$$

$$\cdot \binom{k}{h}\binom{x_1-k}{y_1-h}(1-\alpha)^h \alpha^{k-h}\beta^{y_1-h}(1-\beta)^{x_1-k-y_1+h} \qquad (x_1 = 0,1,...,n; \; y_1 = 0,1,...,x_1)$$

The expectation and variance of the random variable Y_1 are given by:

$$E(Y_1) = n_2\beta^2 + n_1(1-\alpha)^2,$$

$$VAR(Y_1) = n_2\beta^2(1-\beta^2) + n_1(1-\alpha)^2(1-(1-\alpha)^2).$$

The correlation coefficient between X_1 and Y_1 is given by:

$$\rho(X_1, Y_1) = \frac{n_1\alpha(1-\alpha)^2 + n_2\beta^2(1-\beta)}{\sqrt{n_1(1-\alpha)\alpha + n_2\beta(1-\beta)}\sqrt{n_2\beta^2(1-\beta^2) + n_1(1-\alpha)^2(1-(1-\alpha)^2)}} \quad .$$

The correlation coefficient never exceeds $\frac{1}{\sqrt{2}} \approx 0.707$. In the following two cases:

$\alpha = \beta \to 0, \quad \alpha = \beta \to 1$, the correlation coefficient tends to 0.5. In the case of $\alpha = 0, \beta = 0$ (or $\alpha = 1, \beta = 1$), however, the correlation coefficient is actually undefined because its value depends on the direction in which we are approaching this particular point. It is clearly from Table 1, where the correlation coefficient is calculated for some tiny values of α and β. If $\alpha = 0$ and $\beta \to 0$, then $\rho(X_1, Y_1)$ tends to zero, but if $\beta = 0$ and $\alpha \to 0$, then $\rho(X_1, Y_1)$ tends to $\frac{1}{\sqrt{2}}$, and finally, if $\alpha = \beta \to 0$ (by the diagonal of Table 1), then $\rho(X_1, Y_1)$ tends to 0.5. So the point $\alpha=0, \beta=0$ is the correlation coefficient's point of discontinuity.

Table 1. The correlation coefficient for some tiny values of α and β

β → α ↓	0	0.0001	0.0002	0.0003	0.0004	0.0005
0	undefined	0.01	0.0141	0.0173	0.02	0.0224
0.0001	0.707089	0.5	0.4084	0.3538	0.3166	0.2893
0.0002	0.707071	0.5773	0.5	0.4474	0.4085	0.3783
0.0003	0.707054	0.6123	0.5477	0.5001	0.4631	0.4333
0.0004	0.707036	0.6324	0.5773	0.5345	0.5001	0.4716
0.0005	0.707018	0.6454	0.5975	0.5590	0.5271	0.5001

4. Conclusions

The paper examines the problem of evaluating repeatability tests of a given SM. For simplicity, we assume that each item that is submitted for examination by the SM is classified according to binary scale. The usual way to check SM accuracy is to conduct sequential sorting processes. For simplicity—again—we used only two sequential processes. As shown in the paper, even for small classification errors, the correlation coefficient between the results of two sequential sorting is much less than 1. These paradoxical results lead to the conclusion that the correlation coefficient is not suitable for evaluating repeatability. Moreover, even for tiny error rates, predicting the results of a second sorting based on those of the first sorting's results is inaccurate. Further research is needed in order to generalize these results to more than two categories and more than two sequential sorting processes. Other methods of checking repeatability must be considered for cases when the sorting is based on an ordinal or nominal scale.

References

Bashkansky, E. Dror, S. Ravid, R. and Grabov, P. (2007) "Effectiveness of a product quality classifier," *Quality Engineering*, 19(3), 235–244.

Gertsbakh, I.B. (1962) "Determining the error of an automatic measurement, testing and sorting machine," *Measurement Techniques*, 5(7), 533–537.

Gertsbakh, I.B. and Friedman, L. (1985) "Estimation of a standard measuring error by repeated measurements and sortings," *Statistical Planning and Inference* , 11(1), 1–14.

Stevens S.S. (1946) "On the theory of scales of measurement," *Science*, 103(2684), 677–680.