

Testing the significance of the RV coefficient

J. Josse , F. Husson and J. Pagès

¹ Laboratoire de Mathématiques appliquées, UCS 84215, 65 rue de St-Brieuc 35042 Rennes cedex,
(e-mail: julie.josse@agrocampus-rennes.fr)

Keywords: RV coefficient, Permutation tests, Pearson type III distribution, Edgeworth expansion, complexity, napping.

Abstract

The relationship between two data tables X and Y can be evaluated by the RV coefficient. In this presentation, we will compare different strategy to test the significance of the RV coefficient. Since the distribution of the RV statistic is unknown, we use permutation tests to evaluate it. Exact permutation tests can be obtain when the number of individuals is not too important. But, in practice, it is not easy to perform all the simulation (computational cost). Kazi-Aoual (1993) define the first three moments of the RV permutation distribution without doing any permutation at all. We use these moments to approximate the exact distribution of the RV statistic with several methods: the normal approximation, the Pearson type III approximation and the Edgeworth expansion. These approximations are first compared from simulations and then from a real example providing from sensory analysis.

1 Introduction

It is usual to study the links between two sets of variables X and Y defined on the same n individuals. Many authors have already worked on this old question; among them, Escoufier (1970) has proposed the RV coefficient.

Let X a $n \times p$ matrix (of p variables) and Y a $n \times q$ matrix (of q variables); if X and Y are centered by columns, the RV coefficient is defined by:

$$RV = \frac{\langle XX', YY' \rangle}{\|XX'\| \|YY'\|} = \frac{tr(XX'YY')}{\sqrt{tr(XX'^2)tr(YY'^2)}}.$$

The RV coefficient measures the relationship between the variables in X and in Y . It takes values between 0 (each variable of X is uncorrelated to each variable of Y) and 1 (the configurations of the individuals induced by X and Y are homothetic). Between theses two extreme bounds, the value of an RV coefficient brings nothing about its significance. If it is possible (when the number of individuals is not too important), exact permutation distribution can be obtained and an exact test can be computed. However, the main drawback is the computational cost. The standardized RV is used to test the significance of the relationship between the two data matrices. To do that, the standardized RV permutation distribution is approximated by a standard Normal distribution (in concrete terms, two data matrices are significantly linked when the standardized RV is greater than 1.65). Nevertheless, in the framework of permutations test, many authors have shown that permutation distributions are often skewed. Kazi-Aoual (1993) gave the expression to calculate the skewness of the RV permutation distribution. It is then possible to use this coefficient to improve the Normal approximation of the RV distribution with other approximations such as Pearson type III approximation or Edgeworth expansions (both take into account the three first moments).

In this paper, we propose:

- an evaluation, from an user point of view, of the quality of the Normal approximation compared with the Pearson approximation; this will be done by a systematic exploration of various sizes and structures of tables;
- an evaluation of the Edgeworth approximation.
- an implementation of the approximation methods in the free software FactoMineR.

We illustrate our work with simulations and with a real data set providing from sensory analysis.

2 Materials and Methods

2.1 Permutations test

To construct the test of the RV coefficient, we have to specify the two alternative hypotheses. The null hypothesis claims that there are no link between the two matrices. Since the distributions of the RV coefficient is unknown, a reasonable strategy is to use nonparametric approach for testing the hypothesis and thus to use permutations to obtain the RV distribution under H_0 . The permutation test consists in the permutation of the rows of one matrix (for example Y) and then computing the RV coefficient. In this strategy, the null hypothesis becomes: the structures of the two data matrices (X and Y) are given and the individuals of one matrix are labelled at random.

If the number of permutations is not too important, one can perform the exact permutation test. The proportion of the values (including the observed one) more extreme than or equal to the observed one is the p-value. In practice, it is not always possible to perform the $n!$ permutations. Thus, there are two approaches to approximate the permutation distribution:

- random sampling from all possible permutations;
- approximate the distribution by a continuous distribution by using the analytical moments of the exact permutation distribution under the null hypothesis.

Random permutations lead to satisfied estimation of the p-value if the number of randomization is sufficient. Its main shortcoming is its cost in term of computer time.

2.2 Approximating the permutation distribution with distributions based on the first moments

First, we bring to mind how to obtain the first moments of the permutation distributions and how to then approximate the permutation distribution.

Calculating the first moments

Kazi-Aoual (1993, 1995) gives analytic expressions for the first three moments of the trace statistic: $T = tr(X'XY'Y)$ under the null hypothesis (the $n!$ values of T obtained by permutation of the rows of Y are supposed to have the same probability). As $RV = tr(X'XY'Y)/\sqrt{tr(X'X^2)tr(Y'Y^2)}$ and $tr((X'X)^2)tr((Y'Y)^2)$ is invariant under any permutation of the rows of Y , Kazi-Aoual (1993, 1995) obtains the first three moments of the RV coefficient.

The standardized RV

The RV standardized is calculated with only the first two moments:

$$RV_{stand} = \frac{RV - \mathbb{E}(RV)}{\sigma_{RV}}. \quad (1)$$

Cornillon (1993) shows that this previous quantity is asymptotically Normal which can justify a normal approximation.

Assuming a Normal distribution of the permuted RV coefficients, equation 1 can be used to test the null hypothesis. If the value of the observed standardized RV coefficient is greater than 1.65, then the null hypothesis is rejected (Schlich 1996, Abdi 2007).

Approximation of the standardized RV distribution: fitting via moments matching

Many authors (as Mielke, 1984) have noticed that the permutation distribution under the null hypothesis may substantially differ from a Normal distribution and have suggested to use the third moment to take into account that the permutation distribution is often skewed.

Now we are going to approximate the permutation distribution of the RV coefficient with the help of the third moment. Approximations such as Pearson distribution or Edgeworth expansion can be used.

Pearson type III approximation

The Pearson type III distribution, which is a gamma distribution, can approximate a skewed distribution. Moreover, it includes several frequently encountered distributions (such as exponential, the chi squared, etc.) and the permutation distribution often looks like a gamma distribution (Mielke, 1984).

The gamma distribution, with the Johnson and Kotz (1994) notation, can be written:

$$f(x) = \frac{(x - \theta)^{\alpha-1} \exp(-(x - \theta)/\beta)}{\beta^\alpha \Gamma(\alpha)}.$$

With α , β , θ equal to $4/\gamma^2$, $\gamma/2$ and $-2/\gamma$, the Pearson type III distribution have mean 0, variance 1 and skewness γ and can be written:

$$f(x) = \frac{(2/\gamma)^{4/\gamma^2}}{\Gamma(4/\gamma^2)} ((2 + \gamma x)/\gamma)^{(4-\gamma^2)/\gamma^2} \exp(-2(2 + x\gamma)/\gamma^2).$$

Edgeworth Approximation

The asymptotic results concerning the asymptotic normality of the RV coefficient allow to use the Edgeworth expansion. Given the first cumulants of a distribution (which depend on the moments), the Edgeworth approximation is build on the expansion of the Gaussian density function in terms of Hermite polynomials. This approximation can be written:

$$f(x) \approx \Phi(x) \left(1 + n^{-1/2} k_1 H_1(x) + \frac{1}{2} n^{-1} (k_2 + k_1^2) H_2(x) + \frac{1}{6} n^{-1/2} k_3 H_3(x) \right).$$

with $\Phi(x)$ the Normal density, k_1 , k_2 , and k_3 the first three cumulants and H_1, H_2 and H_3 the Hermite polynomials.

2.3 Parameters of the simulation

The different approximations will be compared to the random permutation in different situations (different number of rows and different complexity of the data tables). First, we recall the complexity index.

The complexity index

Kazi-Aoual (1995) writes the expectation of the RV as:

$$\mathbb{E}(RV) = \frac{\sqrt{\beta_x \times \beta_y}}{n - 1}, \quad (2)$$

with,

$$\begin{aligned} \beta_x &= \frac{(\text{tr}(X'X))^2}{\text{tr}((X'X)^2)}, \\ &= \frac{(\sum \lambda_i)^2}{\sum \lambda_i^2}. \end{aligned}$$

β_x can be seen as a measure of complexity (or dimensionality) of the table X . Indeed, it takes values between $1 \leq \beta_x \leq p$. $\beta_x = 1$ if $\text{rank}(X) = 1$ (the only eigenvalue is not null) and $\beta_x = p$ if $\text{rank}(X) = p$ (all the eigenvalues are equals). When there is only one dimension in the data tables (all the correlations between the columns are equal to 1 or -1), the index is one, and when there is p dimensions (all the columns are independent), it is equal to p .

Thus, equation 2 shows that even if the tables X and Y are not linked, a high value of RV is expected if the complexity is high and the number of individuals is small. Here is the interest to use the standardized RV which takes into account the size and the structure of the data set.

The complexity index is also used in other field: for example in analysis of variance with repeated mesures, it is known as a measure of departure from sphericity (epsilon) used to correct the number of degrees of freedom of the F statistic (Box, 1954).

Conducting the simulation

Several data tables X and Y are simulated with different number of individuals using different distributions (Normal, Uniform). Then, a PCA is done for of each table and the standardized eigenvectors are used as new variables. Thus this variables are orthogonal and the complexity of the data table is the number of variables.

We compare the different approximations through the 95 per cent quantile of each distribution: Normal approximation (1.65), Pearson approximation and the approximate one with 5000 permutations. We assume that the distribution obtained with 5000 permutations is very closed to the exact density function.

All the simulations and the computations were run using the free R software (R Development Core Team, 2006).

2.4 The data

An example, providing from the field of sensory data illustrates our purpose. The set of data analyzed concerns eight wines from Jura (France) evaluated twice (two sessions in which the same products are evaluated) by twelve tasters; these wines were difficult to taste. The subset of data was obtained by a data mode of collection presented by Pagès (2005): the napping. In practice, it consists in positioning the products (or objects) on a tablecloth in such a way that two products very similar are closed and two products very different are far from each other. Each taster provides an Euclidean representation of the products which can be represented by a matrix with the products coordinates on the tablecloth. Each matrix has dimensions 8×2 (eight products and two coordinates), and there are as many matrices as couple taster-session ($24 = 12 \times 2$).

The analysis of this type of data showed that it brings interesting elements in the comprehension of the perception of the products. As any method of data collection, the issue of the repeatability arises. Is the products configuration given by a taster stable in time? For that, the 12 tasters repeat the evaluation of the 8 products under the same conditions at a few days of interval. This provides 12 pairs of configurations. In term of repeatability, it is expected that two configurations provided from the same taster are not independent. The likeness between two configurations is measured by the RV coefficient. We build the following test with the two alternative hypotheses: the null hypothesis H_0 "the two configurations are independent" and H_1 "the two configurations are not independent".

3 Results

Simulation results

For each number of individuals (in row) and each complexity (in column), Table 1 gives the 95 per cent quantile of the random permutation and of the Pearson approximation. For the Normal approximation, the quantile is always 1.65.

Normal	c = 1	c = 2	c = 3	c = 4	c = 5	c = 8	c = 10
ind = 6	2.18 - 1.91	1.73 - 1.69	1.66 - 1.68	1.70 - 1.68	1.66 - 1.68		
ind = 8	2.20 - 1.95	1.81 - 1.77	1.75 - 1.71	1.68 - 1.69	1.67 - 1.69		
ind = 10	2.11 - 1.98	1.90 - 1.83	1.79 - 1.76	1.78 - 1.74	1.75 - 1.73	1.77 - 1.74	
ind = 14	2.12 - 1.99	1.90 - 1.84	1.73 - 1.75	1.73 - 1.71	1.71 - 1.69	1.72 - 1.68	1.67 - 1.69
ind = 20	2.08 - 2.01	1.95 - 1.87	1.79 - 1.78	1.73 - 1.74	1.73 - 1.71	1.64 - 1.67	1.69 - 1.67
ind = 25	2.17 - 2.01	1.92 - 1.89	1.79 - 1.81	1.72 - 1.76	1.72 - 1.73	1.68 - 1.68	1.68 - 1.67
ind = 30	2.00 - 2.01	1.91 - 1.90	1.83 - 1.82	1.79 - 1.77	1.75 - 1.74	1.68 - 1.69	1.71 - 1.68

Uniform	c = 1	c = 2	c = 3	c = 4	c = 5	c = 8	c = 10
ind = 6	2.05 - 1.92	1.73 - 1.71	1.64 - 1.68	1.61 - 1.68	1.57 - 1.68		
ind = 8	2.31 - 1.98	1.77 - 1.78	1.72 - 1.72	1.70 - 1.70	1.72 - 1.71		
ind = 10	2.17 - 1.98	1.87 - 1.80	1.74 - 1.71	1.66 - 1.67	1.67 - 1.66	1.67 - 1.68	
ind = 14	2.05 - 2.00	1.90 - 1.85	1.82 - 1.76	1.74 - 1.72	1.70 - 1.70	1.70 - 1.69	1.72 - 1.69
ind = 20	2.10 - 2.01	1.94 - 1.87	1.78 - 1.79	1.72 - 1.74	1.73 - 1.72	1.68 - 1.68	1.65 - 1.68
ind = 25	2.11 - 2.01	1.93 - 1.89	1.77 - 1.80	1.78 - 1.75	1.80 - 1.72	1.68 - 1.68	1.66 - 1.67
ind = 30	2.05 - 2.01	1.87 - 1.89	1.85 - 1.81	1.81 - 1.76	1.70 - 1.73	1.71 - 1.69	1.65 - 1.67

TABLE 1. Comparison of the 95 percentile between the random approximate method based on 5000 simulations (left) and the Pearson distribution (right); for different number of individuals (in row) and different complexity (in column). Data are simulated with a Normal distribution and then with a Uniform distribution.

First, one can remark that the quantiles of the random approximate method are always greater than 1.65 (the quantile of the Normal distribution).

Thus, the decision taken from the Normal approximation leads in these cases to reject the null hypothesis too many often (the associated first kind error is not 5 % but rather 7 or 8 %). The results also suggest that the Pearson approximation is better than the Normal distribution since the difference between the quantile of the Pearson approximation and the approximate method are quite closed. This is the case for all the number of individuals and all the complexities.

One can also notice that the quantile of the approximate and of the Pearson distribution are greater for the smallest complexities (the skewness coefficient increases when complexity decreases). They decrease down to 1.65 when the complexity increase (the Normal approximation is not too bad when the complexity is high). Besides, one can observe that the quantiles increase with the number of individuals.

Results on real data

The objective for this real data example is to evaluate the repeatability of the judgments of each taster. Two 8×2 data matrices (composed by the products coordinates on the tablecloth) are given for each taster. Let's note X (resp. Y) the matrix corresponding to the results of the first (resp. second) session.

For each taster, the relationship between X and Y is evaluated and tested: the RV coefficient and the standardized RV coefficient are calculated and the p-value associated to the test of the RV coefficient significance are estimated from the different methods previously described. We compare the Pearson approximation, the Edgeworth approximation, the normal approximation and the random approximation (5000 permutations) to the exact permutation test since the number of permutations is computationally feasible ($8! = 40320$ permutations).

Table 2 summarizes the results by taster. In addition to the different p-values, the RV and the standardized RV, the first three moments of the RV permutation distribution are given.

For example, for the first taster, the observed value of RV is 0.552 and the value of the standardized RV is 2.078. The $\mathbb{E}(RV)=0.245$, $\sigma_{RV} = 0.0218$ and $\gamma_{RV} = 0.790$. The tasters 1, 5 and 7 are the three only tasters who gave repeatable configurations of the products from one session to the other. This is surprising that nine tasters gave two configurations with no link but

this sensory evaluation of these wines was very difficult.

The p-values of the Normal approximation are often far from the real p-values. The Pearson approximation, which takes into account the expectation of the RV, the variance of the RV but also the skewness, appears to be an excellent approximation since its p-values are closed to the real p-values. The Edgeworth approximations (which also takes into account the first three moments) is quite better than the normal approximation but in this example, it is not as good as the Pearson approximation.

We also notice that the p-values estimated from a sample of 5000 random permutations are very closed to the exact values.

Two figures compares the exact density functions with its approximations. Figure 1 shows the performance of the Pearson distribution approximation compared with the normal distribution and the Edgeworth expansion for the taster 10.

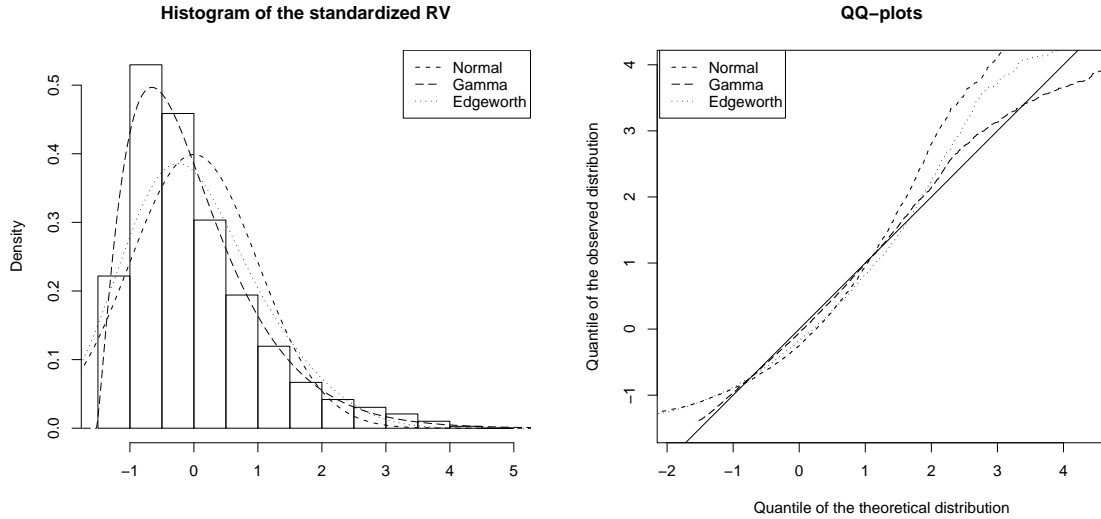


FIGURE 1. Normal, Edgeworth and Pearson approximations of the standardized RV.

On the left, the histogram of the standardized RV statistic shows that approximating the per-

Taster	RV	RVstd	$\mathbb{E}(RV)$	σ_{RV}	γ_{RV}	Norm	Pears	Edge	Rand	Exact
1	0.552	2.078	0.245	0.0218	0.790	0.019	0.035	0.032	0.034	0.039
2	0.222	-0.335	0.263	0.0152	0.467	0.631	0.605	0.614	0.581	0.580
3	0.357	0.944	0.216	0.0225	1.066	0.173	0.160	0.186	0.160	0.162
4	0.127	-0.579	0.217	0.0242	0.944	0.719	0.684	0.694	0.648	0.642
5	0.640	2.727	0.222	0.0234	1.141	0.003	0.017	0.009	0.023	0.020
6	0.144	-0.320	0.190	0.0205	1.127	0.625	0.559	0.596	0.535	0.528
7	0.794	3.841	0.157	0.0275	1.571	6E-05	0.006	4E-04	0.004	0.004
8	0.058	-0.873	0.195	0.0239	1.238	0.810	0.817	0.791	0.828	0.830
9	0.492	2.080	0.144	0.0280	1.609	0.019	0.044	0.039	0.057	0.055
10	0.280	0.373	0.221	0.0252	1.312	0.354	0.287	0.339	0.270	0.276
11	0.224	0.046	0.217	0.0256	1.221	0.482	0.401	0.454	0.380	0.376
12	0.190	-0.262	0.231	0.0240	1.039	0.603	0.539	0.575	0.523	0.520

TABLE 2. RV, standardized RV, first three moments of the RV and the p-values of the Normal approximation, the Pearson approximation, the random approximation and the exact p-value.

mutation distribution of the standardized RV statistics with the Pearson distribution (using the first three moments) is more precise than with a Normal distribution (using only the first two moments). The use of the third moment is an improvement on the classical Normal approximation because permutation distributions are often skewed. The Edgeworth approximation seems to be between the two others approximations. On the right, the QQ-plots of the three approximations distributions show that the Pearson approximation is the best one but the Edgeworth approximation is roughly similar. For these examples, the exact distribution of the standardized RV is known and we have tested the Edgeworth expansion with the first four moments (the fourth moment is calculated from all the permutations) but this approximation is not better than the one done with the first three moments.

4 Conclusion and perspectives

Approximation with Pearson type III distribution provides an efficient way of implementing permutations tests without doing explicitly any permutation. Edgeworth approximation seems to be less good than the Pearson one. These approximations allow to perform significance testing with a first kind error better controlled than the Normal approximation. This is especially true for low complexities.

The density approximations proposed in this article are entirely specified by the first few moments of a given distribution and are unimodal which can be a limitation. Indeed we sometimes encountered bimodal distribution of the RV. In that case, it is possible to approximate the distribution with random permutations. However, it would be better to approximate the distribution without doing any permutation.

We propose an algorithm implemented in a free software <http://factominer.free.fr>, to obtain the p-value associated with the Pearson approximation in order to know the statistical significance of an RV coefficient.

References

- Abdi, H. (2007). RV Coefficient and Congruence Coefficient. In: Neil Salkin (Ed.), *Encyclopedia of Measurement and Statistics*, Thousand Oaks (CA): Sage. pp. 849–853.
- Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and of correlation between errors in the two-way classification *The Annals of Mathematical Statistics*, 25, 484–498.
- Cornillon, P.A (1998). Prise en compte de proximités en analyse factorielle et comparative. *Thesis*, pp 101–118.
- Escouffier, Y. . Le traitement des variables vectorielles. (1973) *Biometrics* **29** 751–760.
- Johnson N.L., Kotz S. (1994). Continuous univariate distributions. *Wiley, New-York*, **Vol 1**.
- Kazi-Aoual, F., Hitier S., Sabatier R., & Lebreton J-D. (1995). Refined approximations to permutation tests for multivariate inference. *Computational Statistics and Data Analysis*, **20** 643–656.
- Kazi-Aoual, F. (1993). Defining and validating assessor compromises about product distances and attribute correlation. *Handbook of statistics* **Vol 4** 813–830.
- Mielke, P.W (1984). Meteorological Applications of Permutation techniques based on Distances Functions. *Multivariate Analysis of Data in Sensory Science*.
- Meyners, M. (2001). Permutations tests: Are there differences in product liking? *Food Quality and Preference* **Vol 12** 345–351.

- Pagès, J. (2005). Collection and analysis of perceived product inter-distances using multiple factor analysis; application to the study of ten white from the Loire Valley. *Food quality and preference*, **Vol 16** 642–649.
- R Development Core Team (2006). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.*, URL <http://www.R-project.org>.
- Schlich, P. (1996). Approximations to permutation tests for data analysis. *Rapport de recherche*, Unité de Biométrie Montpellier **93-06**.