Correlations in Compositional Data without Log-Transformations

Yury Monich and Yury Nechipurenko *

Posted Date: 10 May 2023

doi: 10.20944/preprints202305.0716.v1

Keywords: compositional data; mathematical expectation shift; loss of degrees of freedom; hybrid model

Article

Correlations in Compositional Data without Log-Transformations

Yury V. Monich [1] and Yury D. Nechipurenko [2,*]

[1] Institute of Linguistics, Russian Academy of Sciences, Bolshoi Kislovsky Lane, 1 bld. 1, Russia, 125009, Moscow; monstrator@mail.ru

[2] Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Vavilova St., 32, Russia, 119991, Moscow

* Correspondence: nech99@mail.ru

Abstract: The article proposes a method for determining the *p*-value of correlations in compositional data, i.e., those data that arise as a result of dividing the original values by their sum. Data organized in this way are typical for many fields of knowledge, but there is still no consensus on methods for interpreting correlations in such data. In a space closed by normalizing quantity, correlation coefficients behave differently than under normal conditions: their probabilities of occurrence do not coincide with those inherent in the standard scale of estimates. In the tens of the new millennium, almost all newly emerging methods for estimating correlation in compositional data began to require mandatory log-transformation of the variable values. In the method proposed here there are no log-transformations. We return to the early stages of attempting to solve the problem and rely on negative shifts in correlations in the multinomial distribution. In modeling the data, we use a hybrid method that combines the hypergeometric distribution with the distribution of any other law. During our work on the calculation method, we found that the number of degrees of freedom in compositional data measures discretely only when all the normalizing sums are equal and that it decreases when the sums are not equal, becoming a continuously varying quantity. Estimation of the number of degrees of freedom and the strength of its influence on the magnitude of the shift in the distribution of correlation coefficients is the basis of the proposed method.

Keywords: compositional data; mathematical expectation shift; loss of degrees of freedom; hybrid model

1. Introduction

The emergence in the new millennium of large databases, such as, for example, the Human Microbiome Project, has caused a marked increase in research activity aimed at finding methods for statistical processing of data that have an organization consisting of fractions summarized to a constant (shares to 1, percent to 100, etc.). Following the established tradition in English-language literature, we will call such data compositional, or simply composition, consisting of k parts: $x_1/\sum_{i=1}^{k} x_i + x_2/\sum_{i=1}^{k} x_i + \cdots + x_k/\sum_{i=1}^{k} x_i = y_1 + y_2 + \cdots + y_k = 1$,where $x_i \geq 0$.

Such normalization results in a closed structure generating a special relationship between variance and covariance (*var* and *cov*). Since by all sample size measurements $\sum_{i=1}^{k} y_i = 1$, then $var\left(\sum_{i=1}^{k} y_i\right) = \sum_{i=1}^{k} var(y_i) + 2\sum_{i=1}^{k(k-1)/2} cov(y_i, y_j) = 0$. It follows from this equation that not only $\sum_{i=1}^{k} var(y_i) = -2\sum_{i=1}^{k(k-1)/2} cov(y_i, y_j)$, but in every single case $var(y_i) = -\sum_{i=1}^{k-1} cov(y_i, y_j)$. This was demonstrated in [1], where it was also found that when a composition consists of variables with the same parameters, the zero correlation (hereafter r_0) is not zero, but the value $(1-k)^{-1}$.

A construction for r_0 that takes into account differences between the sizes of variables was independently derived in [2] and [3]. In the latter paper, the derivation is based on multinomial and Dirichlet laws of distribution. Since in such distributions the variance is proportional to the value $E(y_i)(1 - E(y_i))$, where E is the mathematical expectation, Pearson's linear correlation formula by

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0716.v1

substituting the corresponding values of *var* and *cov* turns into the construction $r_0 = -\sqrt{E(y_i)E(y_j)/\left[\left(1-E(y_i)\right)\left(1-E(y_j)\right)\right]}$, or, – through the original values, – into

$$r_0 = -\sqrt{\sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j / \left[\left(N-\sum_{i=1}^{n} x_i\right)\left(N-\sum_{j=1}^{n} x_j\right)\right]}, \quad (1)$$

where N is the sum of all values comprising the composition, and n is the sample size.

However, a year earlier in [2] Eq. (1) was derived in a fundamentally different way, in which from a composition initially represented by a set of variables distributed according to the same law and having equal means and variances, a rearranged composition consisting of variables with different parameters is created by simple summation of different numbers of initial variables.

It follows from derivability of Eg. (1) in this way – and this is emphasized in [2] – that it is related to all distribution laws without exception, but naturally, only if the composition is constructed in accordance with the derivation algorithm, i.e., in such a way in which parametric differences of variables emerge as a result of summation in them different amounts of some initial "quantum" representing some type of distribution.

Although the work [2] has been and still is cited in the relevant literature, the above-mentioned aspect of this work does not seem to have received the attention it deserves until now. This can be inferred indirectly from the fact that further history has been marked by doubts about the validity of Eq. (1) when, in empirical data, differences in variances are not proportional to $E(y_i)(1-E(y_i))$ (as far as we know, this was first raised in [4] (p. 692)).

Indeed, any simulation, if the input variables x_i with different parameters are not generated in it by any of distributions conjugated with the multinomial or if it is not carried out by the summation process described above, will cause the magnitude of the bias of the correlations variation center in the normalized data to disobey formula (1). Of course, the summation process leads to a sort of "binomization" of any initial law, bringing it closer to the law of the normal distribution, but it does not follow from this that the initial variables x_i must exhibit the parameter ratios characteristic of the family of distributions in question. In reality, this ratio can turn out to be anything but it will be determined exclusively by the law of the original "quantum" of the distribution. It follows that when evaluating the applicability of formula (1) to empirical data, the orientation towards the value $E(y_i)(1-E(y_i))$, which almost always (except for specific cases possible with the compound Dirichlet-multinomial version) implies that a variable's variance in the initial data must not exceed the variable's mean, may in many cases be quite wrong.

Since in practice the differences in the parameters of variables in data simulations were not given by summation, disappointment in Eq. (1) resulted in its almost complete neglect after the publication of J. Aitchison's works [5,6], in which a logarithmic transformation of variables was required for the analysis of compositional data. In addition, the postulate of "subcompositional coherence" was introduced, which required to preserve the equality of measurable quantities when the number of constituent proportions in a composition is changed. Obviously, correlation coefficients in normalized data change in such cases inevitably, so in the result J. Aitchison, being faithful to this postulate, simply abandoned the use of correlation analysis, which cannot be said about all his followers. Ignoring J. Aitchison's view of the correlation coefficient uninformativeness, some works based on logarithmic transformations [7–16] do offer their methods for its evaluation. This is not surprising, as the postulate of invariance of values has no theoretical basis and has to do with a problem related not to the specificity of computation but to the selection of material to be processed. Either a researcher makes a mistake in defining the boundaries of the system, or he or she deliberately considers different variants of its possible boundaries, or simply compares relations within the system with those that are formed in one of its subsystems. The values here must necessarily be different, otherwise we would have to assume that, for example, the extinction of the dinosaurs should have had no effect on the nature of interactions between other members of the animal kingdom.

After logarithmic transformations Eq. (1), of course, becomes inapplicable, and the efficiency of methods based on such transformations is seriously doubted by the authors of this paper (for details

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0716.v1

see Section 2). Obviously, by any method of modelling, the equality $var(y_i) = -\sum_{i=1}^{k-1} cov(y_i, y_j)$ will always remain true for composition. For this reason, the authors, who have not seen any intelligible theoretical justification for the inevitability of the logarithmic transformation (and the absence of such not only for the postulate mentioned above, but also for other special prescriptions in the methodology initiated by the works of J. Aitchison, is explicitly stated in [17] (p. 831), also see no serious theoretical grounds for refusing to take the negative shift value determined by formula (1).

As our empirical data from simulations show, Eq. (1) does not give quite accurate values for correlations corresponding to the zero level, which is particularly noticeable for small samples sizes. Therefore, the first of our objectives is to find a method for correcting Eq. (1) that would allow it to satisfactorily match the empirical mean obtained in the distribution of the correlation coefficient calculated between the quantities y_i and y_j. Hereafter, we will refer to the required value as the mathematical expectation shift, or simply the shift, meaning a shift from zero. Finding the shift will provide a starting point for converting asymmetric empirical distributions into distributions with a median of zero, which in turn paves the way for the main task of restoring the ability to interpret the results using standard methods.

2. Materials and Methods

Under various experimental conditions, we observe features of the distribution of correlation coefficients, capture these features and on the basis of this we try to create such a mathematical model of the phenomenon under study, which would allow us to predict with satisfactory accuracy the probability of occurrence of any value.

To create observable series of correlation coefficients, we use data simulation methods dictated by such a null hypothesis, which is complicated by an additional requirement for the initial units of experience: all of them, no matter what variable they refer to, must have the same parameters. In this case, the parameter differences appear only at the level of comparison of variables, since they usually consist of a different number of units.

If, for example, we model a composition of three variables with mathematical expectations of 500, 50, and 5 through the discrete uniform distribution, then, following the noted requirement, we cannot simply set them in ranges from 0 to 1000, from 0 to 100, and from 0 to 10, but we can create expectations of 500 and 50 by summing 100 and 10 values in the range from 0 to 10. This modeling method is used by us but is not preferred.

In the basic method, we create an evenly mixed mass from all the quantitatively planned variables and select from it those numbers of units that correspond to the planned values of the summing variable $\sum_{i=1}^{k} x_i$. A sample without replacement will give a multivariate hypergeometric distribution, with replacement – a multinomial distribution. We use the hypergeometric version because it is easily created in Excel. The low variance of these distributions can easily be increased to any level. To do this, it is sufficient to replace the classical urn with balls of different colors representing units of different variables by an urn which, instead of balls giving a value of 1, contains balls with a value varying within a given range and according to a given law. Such hybrid models fit well with reality, in which the indecomposable element of observation does not simply fall into a particular area of space, but also multiplies in it with varying degrees of success.

Although not all laws when inserted into the hypergeometric distribution yields results clearly predictable by our method over the full range of their distribution, as discussed in more detail in Subsection 3.5., the median values in all cases are defined through Eq. (1). Methods based on logarithmic transformations of the variables do not pass the testing under such distributions. For example, one way of proportionality estimation, ρ_P ("proportionality correlation coefficient"), which, according to the authors of this method [9,10], should be considered as a suitable alternative to standard correlation coefficients, gives approximately the same erroneous estimates when comparing logarithms of simply relative values, and estimates with enhanced errors when taking logarithms of dividing relative values (absolute values can also be taken here) by their geometric mean (in terms of log-ratio analysis, this transformation is referred to as clr – centered log-ratio). If the composition consists of variables with the same parameters, this normalized version behaves about the same as

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0716.v1

the standard coefficients. But if the composition contains variables with large and small mathematical expectations, its behavior is inverted with respect to the behavior of the standard coefficients: instead of small negative shifts of zero values from the center of variation large negative shifts occur, and instead of large negative shifts small negative or positive shifts occur. This is a serious drawback of all methods, that use clr-transformation, independent of how the zero-value problem is solved in them.

Especially surprising is the SparCC method [7] often used in microbiological research, also included as part of other algorithms [18]. This method periodically goes beyond ±1 or does not work at all, generating negative variance (again, the larger the difference between the sizes of the variables, the more often this happens, although clr-transformation is not used here).

Among all methods, the "based on symmetric balances" method [15] can be distinguished, which does eliminate the shift and distributes the transformed correlation coefficients on the Student's scale, but does so, again, only when the composition consists of equal variables and there is no loss of degrees of freedom (i.e., all values of $\sum_{i=1}^{k} x_i$ are the same; see below for details). As the differences in the variables increase, so do the differences in the results, quickly reaching catastrophic sizes. Let us show this with a specific example, where the raw data were obtained by summing 50, 30, 10, and 5 standard continuous distributions in 16 independent measurements. This yields fractions with mathematical expectations of 0.526, 0.316, 0.105, and 0.053. The corresponding variables form six pairwise interactions (see row 1 in Table 1). A total of 11110 combinations were obtained, after each of which the standard Pearson's coefficients (hereafter r), our transformed coefficients (see Subsection 3.4.) and, as prescribed by the based on symmetric balances method, also Pearson coefficients, but calculated from the variables transformed by this method, were simultaneously recorded from the normalized data. To demonstrate the differences, it seems sufficient to show the magnitudes of the biases at the median points of the distributions.

Table 1. Comparison of results obtained by using Eqs. (3) and (6), with the results obtained by the "based on symmetric balances" method (n=16).

Proportions of the interacting variables	0.526; 0.316	0.526; 0.105	0.526; 0.053	0.316; 0.105	0.316; 0.053	0.105; 0.053
Average of the distribution Pearson's r	-0.70314	-0.34715	-0.24172	-0.23137	-0.15359	-0.07947
Shift computed by Eq. (3)	-0.70419	-0.35102	-0.24067	-0.22567	-0.15494	-0.07818
Median by Eq. (6)	0.00417	0.00967	-0.00221	-0.00915	0.00163	-0.00284
Median by [14]	0.58669	0.24666	-0.17097	0.17799	-0.19086	-0.29599

Trusting the method of symmetric balances, we would accept the value of 0.5867 observed in the bottom row as significant with a p-value of 0.017, whereas experimental verification shows that the place of this value is in the center of the distribution and its empirical p-value in this experiment is unity, which indicates the complete absence of a relationship. Our method, as seen in the fourth row, eliminates negative shifts by bringing the median value close to zero.

Compared to the methods requiring log-transformations, the method proposed in [19], based on permutations and renormalization, gives considerably more accurate estimates, which, however, may vary markedly depending on the number of variables comprising a composition. In addition, the lack of accounting for the loss of degrees of freedom in this method may lead to an overestimation of the strength of relationship. In our method, the relationship between two selected variables is completely independent of how the rest of the composition space is organized, which can be represented as a single variable, but all relationships in it depend on the structure of the summing variable that determines the number of degrees of freedom.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0716.v1

3. Results

3.1. Loss of Degrees of Freedom

As the empirical averages obtained from data simulations show, Eq. (1) gives a satisfactory approximation, the accuracy of which, however, directly depends on the sample size and on the structure of the summing variable $\sum_{i=1}^{k} x_i$, expressed in the size of distances between its values. Both of these factors are related to the problem of establishing the number of degrees of freedom (hereafter *df*), which in compositional data, as follows from our numerous observations and experiments, are not necessarily discrete quantities of the natural order.

Let us try to show why the value of *df* is not equal to *n-2*, as in the usual estimation of Pearson's linear correlation, but less than or equal to *n-2* when we are dealing with a composition.

Let v_j be the values of the summing variable ranked in ascending order, and $\sum_{j=1}^{n} v_j = N$. Then, in the multivariate hypergeometric distribution, the expectation and variance for the absolute values of x_i can be expressed as $E(x_i)_j = v_j \sum_{i=1}^{n} x_i/N$ and $var(x_i)_j = v_j \sum_{i=1}^{n} x_i/N\,(1 - \sum_{i=1}^{n} x_i/N)(N - v_j)/(N-1)$. It can be seen that for each non-matching value of v_j there will be different values of $E(x_i)_j$ and $var(x_i)_j$. Turning to relative values, the expectation turns into a constant equal to $\sum_{i=1}^{n} x_i/N$. The variance here, being in the interval from zero to one, reverses its direction of growth: $var(y_i)_j = {v_j}^{-1} \sum_{i=1}^{n} x_i/N\,(1 - \sum_{i=1}^{n} x_i/N)(N - v_j)/(N-1)$. With the multinomial distribution, the variance is defined more simply: $var(y_i)_j = {v_j}^{-1} \sum_{i=1}^{n} x_i/N\,(1 - \sum_{i=1}^{n} x_i/N)$.

The difference in variance between the *j*-th and *k*-th members of the sequence here is expressed by the ratio $[v_k(N - v_j)]/[v_j(N - v_k)]$ in without-replacement samples and v_k/v_j in with-replacement samples. Obviously, as *j* increases, the range of variation of the fractions inevitably decreases, which implies that each successive cell of the summing variable makes smaller contributions on average to the procedure for calculating the correlation coefficient than the previous one. But a decreasing contribution entails an increase, not a decrease, in the value of the correlation coefficient, and this corresponds not to $df = n - 2$, but to $df < n - 2$.

As simulations show, any increase in the distance between neighboring values of the summing variable leads to an increase in the variance of the empirical distribution of the correlation coefficients calculated between the variables of the normalized data. We speak precisely about the distances between values and not about the variance of the summing variable as a whole, as it is due to outliers not always commensurate with the loss of *df*. Therefore, when estimating the number of *df*, it is methodologically more correct to take into account not the final variance, but its stepwise changes in the ranked sequence. This principle is implemented in the construction proposed below, which is derived mainly empirically:

$$df_0 = \left\{\frac{n-2}{n}\right\}\left(1 + \sum_{j=2}^{n} (v_1/v_j)^{\sqrt{0.5V_j\left(v_j/v_{j-1}\right)^{\ln([j-1]/[1+\ln(j-1)])}}}\right) \quad (2)$$

In this construction, $V_j = j\sigma_j/(1 + \sum_{j=2}^{n} (v_1/v_j)^{1/\sqrt{2\pi}})$ is the current coefficient of variation, which is expressed without percentage conversion and calculated stepwise over the parallel sequence $(v_1/v_j)^{1/\sqrt{2\pi}}$. The zero in the index at *df* is due to the fact that as the correlation coefficient values move away from zero the loss decreases somewhat, as discussed in more detail in Subsection 3.3.

So far the values of df_0 obtained by this formula agree quite well with empirical data. For instance, if we generate a summing variable using the construction v_j = 60+15(*j*-1), at *n*=6 we obtain df_0=3.64 (9.07% loss), at *n*=15 – df_0=10.16 (21.8% loss), at *n*=45 – df_0=24.70 (42.6% loss), at *n*=102 – df_0=41.96 (58.04% loss). For transformation results of the distributions obtained by simulation on these summing variables, see Subsection 3.4., Table 4, rows 8, 11, 14, and 16.

When comparing the values of $[v_k(N - v_j)]/[v_j(N - v_k)]$ and v_k/v_j it is seen that the hypergeometric distribution gives higher variance differences than the multinomial and those of "binomized" which are not created by without-replacement samples from the urn with a fixed

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0716.v1

number of "quanta". It is to be expected that the loss of *df* when without-replacement samples are created will also be higher.

However, the difference does not appear to be very large, as it is not yet detectable even in cases leading to very high *df* losses. Table 4 in rows 15a and 15b (see Subsection 3.4.) shows the results obtained by summing the discrete uniform distribution with values 0 and 1 ($var(x_i)_j = E(x_i)_j/2$ and $var(y_i)_j = E(y_i)[1 - E(y_i)]/[2E(v_j)]$; the values of the summing variable here are varying quantities, so v_j is replaced by $E(v_j)$) and by summing the continuous uniform distribution between 0 and 1 ($var(x_i)_j = E(x_i)_j/6$ and $var(y_i)_j = E(y_i)[1 - E(y_i)]/[6E(v_j)]$). Row 15 shows the results of without-replacement sampling with the same parameters. In these three cases *N* and *E*(*N*) are 32004, v_1 and $E(v_1)$ are 252, v_n and $E(v_n)$ are 10080. Given the noted differences in ratios of variances, for rows 15a and 15b we obtain a 40-fold difference between the variances in the first and last cells of the summing variable, and for row 15 a 57.93-fold difference, which is clearly confirmed experimentally. Despite the nearly 1.5-fold gap, there is no significance of the differences in the values of deviations from the expected Student values.

3.2. Correcting for the Magnitude of the Shift by Degrees of Freedom

While observing the nature of deviations of empirical averages from expected by formula (1), it was found that their trajectory has some similarity with the trajectory characterizing deviations of sample correlation coefficients from values peculiar to the general population. To correct for these deviations, R. Fisher [20] derived the formula $\rho = r[1 + (1 - r^2)/(2n)]$, which was later refined [21] into $\rho = r[1 + (1 - r^2)/(2n - 6)]$. Since the resemblance to our case here is approximate, from R. Fisher's formula is borrowed only the general scheme of construction of the correction coefficient, which has eventually grown to the following cumbersome construction:

$$\mu = r_0 \left[1 + \frac{(1 - r_0^2)}{2df_0^{\,[1+(n-2)^{-(n-1)}](n-1)^{-2}}} \left\{ \frac{(n-2)[1 - \ln(df_0/[n-2])/(n-2)]]^{-1}}{n - 1 - (n-2)/(n^2/(n-1)\,[2 - |r_0|^{(n-2)/(n+2)}] + \ln[2n])} \right\} \right]^{-1} \quad (3)$$

This construction, as well as Eg. (2), has been derived empirically, by attempts to create such a mathematical model of the studied object, which with satisfactory accuracy would agree with the whole array of experimental data. The abundance of sample size values is due to the desire to cover limiting cases up to n=3 and $df \leq 1$. If we do not take into account small n up to n=7, as well as cases with very large loss of degrees of freedom, the formula can be simplified to $\mu = r_0/[1 + (1 - r_0^2)/(2df_0)]$.

The tests carried out at this moment provide only hypothetical possibilities to identify the directions of errors of the shift calculation method presented in (3). The number of empirical series obtained in the tests is 2191. Each series consists of more than 10000 units (the longest series have more than a million units). All series differ either in the size of the shift or in the peculiarities of the structure of *the* summing variable, i.e., in the size of the values [1-df_0/(n-2)] (loss of degrees of freedom), or in the sparsity coefficient *S* (see Subsection 3.4.). The test results are shown in Table 2.

Table 2. Estimating the *p*-value of the shift magnitude.

	sampling characteristics	series number	arithmetic mean and its *p*-value		*p*-value of distribution by the χ^2 test
1	The whole aggregate, except rows 19 and 20	1846	-0.0032	0.8861	0.7951
	Hypergeometric distribution				
2	n=3	92	0.0209	0.8513	0.6635
3	n=4	92	-0.0192	0.8639	0.7514
4	n=5-6	81	0.0273	0.7728	0.2512
5	n=7-9	158	0.0038	0.9635	0.3912
6	n=10-11	545	0.0043	0.9203	0.5628

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0716.v1

7	n=15, 18-20, 25-26	226	-0.0461	0.4525	0.2388
8	n=35, 42-45	366	-0.0138	0.7669	0.3044
9	n=60, 62	161	-0.0062	0.9361	0.6522
10	n=86, 96	194	0.0049	0.9448	0.5997
11	$s_{\bar{r}} < 0.001$	251	0.0141	0.8188	0.3395
12	$s_{\bar{r}} < 0.00017$ (series >1000000)	9	-0.1396	0.6390	--------
13	$\|r_0\| > 0.5$	24	0.0257	0.9079	--------
14	$0.5 > \|r_0\| > 0.3$	119	-0.0285	0.7683	0.1481
15	$\|r_0\| < 0.05$	231	-0.0450	0.4795	0.4839
16	$[1 - df_0/(n-2)] = 0$	527	-0.0066	0.8747	0.7066
17	$0 < [1 - df_0/(n-2)] < 0.2$	954	0.0052	0.8671	0.5537
18	$0.2 < [1 - df_0/(n-2)] < 0.55$	330	-0.0408	0.4468	0.2352
19	$[1 - df_0/(n-2)] > 0.55$	184	0.1990	**0.0150**	**0.0001**
20	Outliers at the beginning of the variable $\sum_{i=1}^{k} x_i$	161	-0.1897	**0.0087**	**0.0499**
	Hybrid model distributions				
21	The whole aggregate, except row 22	527	0.0038	0.9296	0.9539
22	exp ($5 < x < 10$) and x^A with $A > 3$	199	0.2855	**$1.51*10^{-6}$**	**$1.60*10^{-5}$**

The distribution of the value $(\bar{r} - \mu)/s_{\bar{r}}$, where $\bar{r}$ is the arithmetic mean of the series of correlation coefficients and $s_{\bar{r}}$ is its statistical error, was evaluated in Table 2. In calculating the χ^2 values the theoretical values were determined by the probability density function of Student's t-distribution.

The motivations of the samples by the values of n, df_0 and r_0 hardly require any special explanations: these are the parameters on which the magnitude of the shift depends. The motivation of the samples by the magnitude of the error is also quite obvious: the smaller this value is, the more pronounced the errors of the method should be.

Rows 19 and 20 in Table 2 (see Section 6 for outliers at the beginning of the summing variable) present excluded data that show two opposing trends in the errors of the method. Although without these data, the reliability picture looks quite good, this likely indicates only that the errors observed in the marked rows simply have not yet had a chance to appear in the overall data set due to the insufficient number of trials.

However, practical calculations are only marginally affected by errors of this magnitude. In order to make deviations of ±0.2, as in the cases described in rows 19 and 20, close to zero, the shifts obtained by Eq. (3) should be reduced or increased on the average by about 0.0008. The transformed correlation coefficients, if their values are far from zero, will change by an appreciably smaller amount.

The hybrid models tested to date give a broadly similar picture, which we have considered unnecessary to detail here, confining ourselves to the two rows. The reasons for placing a certain part of the data in a separate row 22 will become clear after reading the contents of subsection 3.5.

3.3. Transformation of Correlation Coefficients

Let us agree that r without an index further will denote not the Pearson correlation coefficient in general, but namely the one calculated from the normalized data of the composition. In [3] the estimation of the difference between r and r_0 is done by Fisher's z-transformation: the value $(z[r] - z[r_0])\sqrt{n-3}$ gives the value of the t-criterion by which the p-value of the difference is estimated, and the formula $(z[r] - z[r_0])/\sqrt{1 + (z[r] - z[r_0])^2}$ can be used to calculate the correlation coefficient corresponding to this criterion. However, such a procedure with expected values over |0.5| begins to noticeably deviate its results from the expected curve in the direction of their decreasing, bringing the deviation to approximately -0.08 at values from |0.92| up to |0.95|. Although this phenomenon is relevant only in the absence or insignificant losses df, which, if they are

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0716.v1

ignored, can easily lead to overcoming such deficits, this overestimated "bend" of the distribution curve still seems to the authors much more difficult to correct than the problems generated by the elementary construction $(r-\mu)/(1-r\mu)$ (here the value of r_0 is replaced by the value of μ corrected by (3)).

When using this construction, two phenomena immediately become apparent: first, the larger the shift and the smaller the number of pairwise comparisons, the more the medians of the transformed distributions shift from zero to negative values, and second, as the differences between the values of the summing variable increase, the deviations from the values expected by the Student increase. The magnitudes of the deviations in both cases can be very substantial, but there are fundamental differences in their direction: in the absence of variation of the summing values and/or in their relatively weak variation, the negative tail has consistently overestimated values of the transformed coefficients, while the positive tail has the opposite; in strong variation the values in both tails overestimate, indirectly indicating the loss of degrees of freedom. The experimental data illustrating these phenomena are shown in the table below.

The rows of Table 3 show the deviations from the expected by Student values observed in the transformed distributions, each with 11110 correlation coefficients. Deviations of transformations by means of the construction $(r-\mu)/(1-r\mu)$ are highlighted in color, below them deviations of transformations by means of Eq. (6) are located (see Subsection 3.4.). Shown are the sizes of the deviations from the expected values at the points corresponding to the empirical p-values (p) indicated in the top row. For example, p=0.4 in the distribution of 11110 units corresponds to places with ordinal numbers 2222 (0.4 multiplied by 5555) and 8889 (11110-2222+1). In the empirical distribution reflected in item 2, we have initial values -0.8761 and -0.4369 at the indicated locations, which after transformation by the formula $(r-\mu)/(1-r\mu)$ acquire values -0.5385 and +0.2807, and -0.4293 and +0.4167 after transformation by Eq. (6). For n=6, the expected value of the correlation coefficient at p=0.4 is 0.4257. If we subtract this value from the transformed values taken as module, we obtain the deviation values reflected in item 2 in the left and right columns of 0.4. But if we do the same over the whole series, we get a misunderstanding of the direction and magnitude of the deviations in the immediate vicinity of zero. Therefore in the negative tail the transformed empirical value is subtracted from the expected value, and starting from point p=1 the opposite is done. This allows the deviations to retain the correct sign and the correct distance between the empirical and expected values throughout.

Table 3. Dependence of the pattern of distribution on the shift magnitude and degrees of freedom.

	p / *n*; *df*₀; *μ*	**0.0025**	**0.025**	**0.4**	**0.7**	**0.9**	**1**	**0.9**	**0.7**	**0.4**	**0.025**	**0.0025**
1	4; 1,93	0.001	0.007	0.112	0.166	0.190	**-0.205**	-0.206	-0.201	-0.167	-0.019	-0.004
	μ=-0.440	-0.000	-0.005	-0.002	-0.006	-0.009	**0.004**	0.004	0.002	-0.005	-0.005	-0.004
2	6; 4	0.018	0.031	0.113	0.140	0.149	**-0.156**	-0.155	-0.156	-0.145	-0.056	-0.015
	μ=-0.639	0.009	-0.003	0.004	0.005	0.002	**-0.004**	-0.003	-0.005	-0.009	-0.005	0.002
3	20; 18	0.023	0.036	0.047	0.050	0.054	**-0.054**	-0.052	-0.053	-0.054	-0.042	-0.036
	μ=-0.917	-0.005	0.001	-0.001	-0.001	0.002	**-0.001**	0.000	-0.001	-0.004	-0.003	-0.005
4	62; 17.46	0.235	0.205	0.101	0.052	0.020	**-0.003**	0.014	0.048	0.101	0.230	0.269
	μ=-0.053	-0.010	-0.006	0.003	0.003	0.001	**0.000**	0.002	0.004	0.006	0.006	0.008
5	62; 58.18	-0.005	0.003	0.003	0.002	0.001	**-0.001**	-0.002	-0.002	-0.002	0.002	-0.003
	μ=-0.051	-0.006	0.002	0.002	0.001	0.001	**0.000**	-0.001	-0.002	-0.002	-0.000	-0.008

The first three items in Table 3 show distributions with large shifts. It can be seen that for df_0=18, despite a much larger shift than for df_0=4 and df_0=1.93, the deviations in the highlighted row are markedly reduced. Items 4 and 5 demonstrate how uninformative the variance values of the summing variable can be for estimating df_0. In item 4, its series is constructed using the formula $v_j = 20 + 4j + \sum_{i=0}^{j} j_i$ ($j \in \{0, 1, \ldots, n\}$): 20, 25, 31, …, 2026, 2090, 2155. The coefficient of variation (V) here is 82.1%. In item 5, the series consists of 60 values each of 60, to which outliers of 1000 and 2000 are

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0716.v1

added, and this, despite the anomalous variance (V=253.5%), leads to a barely perceptible decrease in df_0 for such a series length. As can be seen from comparing these items, without taking the loss of df into account, one can be fundamentally wrong in estimating correlation coefficients.

As can be seen from the numerator of the construction $(r-\mu)/(1-r\mu)$ and the deviations in the median point reflected in column p=1, the values that coincide with the empirical mean move away from the median as n and df decrease. This is markedly corrected by multiplying the shift by $1+(1-r^2)/df_0$. However, if we estimate the correlations on the standard scale with df=n-2, this will give only some semblance of symmetry, but not at all a relief from large deviations, which will disappear only in the nearest vicinity of zero. If, as in the case illustrated in item 4, we use for the estimation values corresponding to 17.46 degrees of freedom instead of 60, this will lead to another kind of distortion: deviations from the expected values will gradually begin to grow minus instead of plus. This tendency is clearly visible when the losses of df are very large, as in this case. Since nothing similar is observed in cases where the normalizing denominator is a constant and thus df=n-2, it can be concluded that as the value of the correlation coefficient increases, the loss of df gradually decreases. For this reason, the following function for the change of df is introduced, which is necessary for a correct p-value determination:

$$df = df_0\left\{\left(\frac{n-2-df_0}{n-2}\right)^{\sqrt{(n-2)/df_0}}+1\right\}^{|\rho|^{2([n-2]/df_0)^{1/4}\left(2+|\rho|^2(n-2)/df_0\right)/5\,df_0/(n-2)}}, (4)$$

where $\rho=(r-\mu[1+(1-r^2)/df_0])/(1-r\mu)$ is an aid construction used for recursive correction.

Eventually, attempts to align the distribution trajectory resulted in the original construction being overgrown with a multitude of corrective functions, which are presented separately:

$$r_{df}=\frac{r+|\mu|^{\alpha}[1+\beta\,(1-r^2)/df]}{\gamma-\mu\{(1-r^2)/df\,|\mu|^{\delta}+r\}}\varepsilon\,, (5)$$

where $\alpha=\left(1+\rho^3(1-|R|)^2[1+\mu]^{\ln(df_0)}\right)^{1-r^2}$,

$$\beta=\left\{1+\left([n-2]/\left[2^{1/(1+|\mu|^{-\ln|\mu|})}n+1\right]\right)^2/\ln(n\ln[n-\ln(n)])\right\}^{1-|\rho|^{0.5}},$$

$$\gamma=(1-S)^{(1+\mu^3)\left(1-|\rho|^{0.5(1-\rho^{1/3})}\right)},\ \delta=1-|r|^{-2n^{\ln(n)}\ln|\mu|},$$

$$\varepsilon=\left(1-S_{ij}\rho^3\left[1-\sqrt{|\rho|}\right]\right)^{1-|\rho|^{(n-2)/(2df_0)}}.$$

The function α, where R is a critical value at the significance level of 0.001 (but not at df, but at n-2, since the final step brings the correlation coefficient to that level, as discussed in Section 6), lowers values after the main "bend" of the normal distribution curve has passed, with negligible effect on low values. The function β, in contrast, affects values at the center of the distribution, but has negligible effect on high values.

The indices S and S_{ij} that can be seen in the functions γ and ε require a separate explanation. They are related to a special factor which in certain cases has a rather noticeable impact on the magnitudes of the deviations of the empirical values from the expected ones. This factor in microbiome studies is commonly referred to as sparsity [22]. Let us demonstrate the effects of this factor using a simple, but highly exaggerated example. At n=100 there are two variables, each consisting of 20 units. The maximum possible negative value here is -0.25, and it occurs only when all forty units are distributed in different forty cells. Obviously, a simulation-generated empirical series here, despite the negative arithmetic mean, will have a pronounced slope towards positive values.

It is unlikely that anyone would consider such data in a real study. Nevertheless, due to the fact that when a certain threshold is reached in the relationship between n, df_0 and the size of the original variables, the trend outlined in this example still becomes noticeable, special coefficients defining the functions γ and ε are introduced to correct it: $S=\sum_{j=1}^{n}\left(nv_j10^d\right)^{-1}$ is the sparsity coefficient on the summing variable (as the formula shows, it is the inverse of the harmonic mean; d here is the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0716.v1

corresponding number of decimal places, which is necessary to eliminate fractional values if the original data consist of such), and $S_{ij} = th\{0.5[S/E(y_i) + S/E(y_j)](n-2)/df_0\}$ is a coefficient taking into account the degree of sparsity of the matched pairs of variables (*th* is the hyperbolic tangent).

So far our experimental data show that the deviation character specific to low-density cases – the minus throughout the negative tail turns into a minus of approximately the same magnitude in the positive tail, but not all the way through, but around the expected value 0.4 turns positive – starts to become noticeable when $[S/E(y_i) + S/E(y_j)]$ exceeds 1 (very rough estimate).

3.4. Final Transformation and Method Errors

If the normalizing denominator of the composition is a constant, then *df*=*n*-2 throughout the distribution of correlation coefficient values. In all other cases *df*<*n*-2, so the results are processed in two steps. First, the empirical correlation coefficient transformed by (5) is estimated on a scale corresponding to the value of *df* calculated by (4). Then through the obtained *p*-value the value of the *t*-criterion for *df*=*n*-2 is found, and then the calculation by (6) completes the conversion procedure:

$$r_{n-2} = t/\sqrt{n-2+t^2} \quad (6)$$

The *p*-values corresponding to fractional degrees of freedom are determined by the formula $p = p_{n_{wh}} - (df - n_{wh})(p_{n_{wh}} - p_{n_{wh}+1})$, where n_{wh} is the integer (whole) part of the constituent *df*. For example, for r_{df}=0.5491 with *df*=3.7890 we get $p = p_3 - (3.7890 - 3)(p_3 - p_4) = 0.3378 - 0.7890(0.3378 - 0.2592) = 0.2758$. As can be seen, this is a crude linear approximation, which, however, gives appreciable deviations only at very small *n*.

In general terms the errors of the method can be estimated from the data presented in Table 4.

Table 4. Deviations from the expected values.

	p *n; df*₀; loss	**0.001**	**0.01**	**0.05**	**0.3**	**0.7**	**1**	**0.7**	**0.3**	**0.05**	**0.01**	**0.001**
	(I) $[1 - df_0/(n-2)] = 0$											
1	3; 1; **0.0**	0.000	0.000	-0.000	-0.002	0.003	**-0.001**	0.001	0.002	-0.001	-0.000	0.000
2	4; 2; **0.0**	0.000	-0.001	-0.001	-0.002	-0.002	**0.000**	0.002	0.000	0.001	-0.001	-0.000
3	7; 5; **0.0**	-0.000	0.000	-0.001	0.000	0.001	**0.000**	0.000	0.000	0.000	-0.000	-0.001
4	42; 40; **0.0**	-0.001	-0.003	-0.001	0.000	0.000	**0.000**	0.000	0.001	-0.000	-0.001	-0.002
5	96; 94; **0.0**	0.002	0.001	0.001	0.001	0.001	**-0.000**	-0.000	-0.000	0.001	0.001	0.000
	(II) $0 < [1 - df_0/(n-2)] < 0.2$											
6	62; 58.18; **0.030**	0.002	0.001	0.001	0.001	0.000	**-0.000**	0.000	0.000	0.001	0.000	0.001
7	96; 86.55; **0.079**	0.002	0.001	0.001	0.000	0.000	**-0.000**	-0.000	-0.000	0.000	0.000	0.001
8	6; 3.64; **0.091**	-0.002	-0.003	-0.004	-0.005	-0.003	**0.000**	-0.004	-0.007	-0.006	-0.003	-0.002
9	42; 35.44; **0.114**	-0.000	-0.000	0.000	0.001	0.000	**-0.000**	-0.000	-0.001	-0.001	-0.001	-0.000
10	10; 7.04; **0.120**	-0.002	-0.001	0.002	0.000	-0.001	**0.000**	-0.001	-0.003	-0.004	-0.002	-0.002
	(III) $0.2 < [1 - df_0/(n-2)] < 0.55$											
11	15; 10.16; **0.218**	-0.002	-0.005	-0.005	-0.005	-0.004	**0.001**	-0.002	-0.005	-0.008	-0.008	-0.008
12	15; 8.84; **0.320**	-0.002	0.001	0.003	0.000	-0.000	**0.001**	-0.001	-0.001	-0.001	-0.000	-0.002
13	96; 54.61; **0.419**	-0.011	-0.007	-0.005	-0.003	-0.001	**-0.000**	-0.001	-0.003	-0.005	-0.006	-0.007
14	45; 24.70; **0.426**	-0.012	-0.008	-0.007	-0.005	-0.002	**0.000**	-0.001	-0.004	-0.007	-0.008	-0.010
15	10; 3.96; **0.505**	-0.010	-0.001	0.001	-0.001	-0.002	**0.001**	-0.001	0.002	0.003	0.001	-0.012
15a	*10; 3.96;* ***0.505***	-0.006	-0.002	0.001	0.001	-0.001	**0.000**	-0.000	0.002	0.002	0.002	-0.005
15b	*10; 3.96;* ***0.505***	-0.010	-0.006	-0.001	-0.002	-0.002	**0.001**	0.001	0.001	0.000	0.001	-0.009
	(IV) $[1 - df_0/(n-2)] > 0.55$											
16	102; 41.96; **0.580**	-0.006	-0.006	-0.005	-0.003	-0.001	**0.000**	-0.001	-0.003	-0.005	-0.006	-0.005
17	14; 4.38; **0.635**	-0.009	0.002	0.006	0.007	0.004	**0.000**	0.003	0.004	0.005	-0.001	-0.007
18	18; 4.81; **0.699**	-0.011	-0.001	0.001	0.005	0.004	**0.001**	0.004	0.003	-0.002	-0.002	-0.006
19	62; 17.46; **0.709**	-0.002	0.000	0.003	0.005	0.002	**-0.000**	0.002	0.004	0.003	0.002	0.003

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0716.v1

20	32; 5.13; **0.829**	-0.003	0.004	0.008	0.009	0.006	**-0.001**	0.004	0.005	0.001	0.005	0.001
(V) Outliers at the beginning of the variable $\sum_{i=1}^{k} x_i$												
21	28; 15.12; **0.419**	-0.025	-0.026	-0.021	-0.015	-0.007	**0.001**	-0.005	-0.015	-0.023	-0.026	-0.026
22	45; 24.58; **0.428**	-0.037	-0.036	-0.029	-0.016	-0.006	**0.000**	-0.006	-0.016	-0.028	-0.032	-0.029
23	62; 34.99; **0.417**	-0.048	-0.043	-0.034	-0.019	-0.007	**0.000**	-0.007	-0.019	-0.033	-0.040	-0.035

The magnitudes of the deviations are defined in Table 4 in the same way as in Table 3, but they do not correspond here to a single distribution, but represent the average of the deviations fixed in 21 series of empirical coefficients generated in the one common matrix. Each series differs in the size of μ and S_{ij}, but averaging over all series seems justified, as Eq. (5) converts both small- and large-shift distributions without apparent difference. The data here, as in the shift deviation analysis in Table 2, are divided into four groups that correspond to the levels of loss of *df*. All examples in rows 6-20 are ranked in order of increasing magnitude of loss. The cases with outliers at the beginning of the summing variable are placed in a special category.

In rows 15, 15a and 15b the results of distributions obtained by different simulation methods are compared (see Section 3).

When reviewing the data in Table 4, one can conclude that method errors, expressed in the size of the deviations of the empirical values from the expected values, tend to increase as the losses of *df* increase. However, it is obvious that this trend is broken in places. The differences in these misfits are often such that it is not easy to attribute them to statistical bias (compare, for example, rows 8 and 16 with their neighbors). This inconsistency is primarily due to the weakness of the method in estimating the distances between the values of the summing variable – especially in the initial part of its ranked row.

Typical cases in which the noted disadvantage is most pronounced are shown in subsection V of the table. A clear overestimation of the loss value, leading to underestimation of the correlation indicators, occurs when the summing variable is led by a value (or a very small number of them relative to the total length of the series) that is strongly inferior to the subsequent values. In the example in row 21, the summing variable begins from values of 20 and 35, followed by a series of 26 values that are obtained using the formula v_i = 60+5(i-1). In the example in row 22, the summing variable begins from a value of 10, followed by a series starting with a value of 32 with relatively small distances between values. In row 23, the values of 10 and 30 precede the 60 values equal to 60. *P*-values for deviations in the median point here are 0.0109, 0.4362 and 0.7515 respectively, which would seem to suggest that df_0 is too small only in the first case. However, the large negative deviations in these distributions will disappear if we increase the df_0 to 17, 29, and 45, respectively. This would make the median values 0.00055 (p=0.1701), -0.00021 (p=0.5100) and -0.00041 (p=0.1201). Since one indicator in the center lost significance of the deviation from zero and the others did not get it, there can hardly be any doubt that the problem here is, after all, overestimation of losses rather than some oddities in the distribution.

Thus, without any risk of error exceeding ±0.02 with respect to the standard Student's estimator, the method proposed above seems to the authors to be usable for any losses of *df*, but only with a principal caveat: there should be no outliers at the head of the summing variable. If so, one should expect underestimation of the correlations to be greater than -0.02. As for outliers at the end of the row, however large they may be, they will not reduce more than their own unit of freedom.

3.5. Hybrid Models

In practice, if it is not a simulation, the analysis is carried out on already formed systems that have passed a certain historical or evolutionary path. If the units under study are regarded as having equal potential of possibilities, then the total space of the system at the first stage of its formation is filled with uniform distribution, which leads to the variations according to the hypergeometric law when dividing into variables. If the equality of possibilities is preserved, which can be considered a natural condition of the null hypothesis, then the evolution of the system should proceed with equal

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0716.v1

reproduction parameters of the initial units in all variables. Thus, when modelling the stage of evolution, the initial unit can be directly specified as a reproducing unit.

A caveat should be made here, however, that not every reproduction law and/or range of variation of the initial unit yields results close to the standard scale of estimation, as can partly be seen from the data in Table 5.

Table 5. Deviations from the expected values in hybrid models.

	p / *n*; df_0; insert	0.001	0.01	0.05	0.5	1	0.5	0.05	0.01	0.001
	(I) Binomial and normal distributions									
1	11; 9; B (10, 0.5)	0.001	0.000	0.001	0.002	-0.001	0.000	0.001	0.001	0.003
2	32; 30; B (10, 0.5)	-0.000	-0.003	-0.002	0.001	-0.000	-0.001	-0.003	-0.002	-0.006
3	7; 5; B (10, 0.5)	-0.001	-0.005	-0.005	0.002	0.001	-0.002	0.001	0.001	0.003
4	10;5.85; B (100, 0.5)	-0.000	-0.002	-0.006	-0.007	0.001	-0.005	-0.010	-0.008	-0.006
5	11; 9; N (650, 10^4)	0.003	-0.001	-0.002	0.002	-0.002	-0.000	-0.001	0.001	0.002
6	32; 30; N (50, 64)	-0.008	-0.002	0.001	0.001	0.000	0.000	-0.002	-0.003	0.001
7	42; 40; N (50, 64)	0.004	-0.000	0.000	0.001	-0.000	0.000	-0.002	-0.003	0.003
8	65; 63; B (100, 0.5)	-0.001	-0.003	-0.002	0.001	-0.001	-0.000	0.002	0.003	0.002
8a	$\|\mu\| < 0.05$	-0.006	-0.006	-0.003	0.001	-0.002	-0.001	0.004	0.006	0.009
8b	$\|\mu\| > 0.05$	0.005	-0.000	-0.001	0.000	-0.001	0.001	-0.001	-0.002	-0.007
	(II) Uniform and power-law distributions									
9	32; 30; DU (0, 10000)	0.003	-0.002	-0.001	0.001	-0.000	-0.000	-0.001	-0.001	-0.003
10	44; 35.53; DU (0, 1000)	-0.003	-0.004	-0.003	-0.000	-0.001	-0.001	-0.000	0.000	0.007
11	32; 30; [CU (0, 1000)]^2	-0.008	-0.006	-0.005	-0.001	-0.000	-0.002	-0.005	-0.006	-0.007
12	10; 3.96; [CU (0, 1000)]^2	-0.011	-0.004	0.001	0.002	-0.000	0.002	0.004	-0.000	-0.010
13	32; 30; [DU (0,10)]^4	-0.016	-0.010	-0.008	-0.000	-0.001	-0.001	-0.000	-0.001	0.004
14	32; 30; [CU (0,1000)]^4	-0.016	-0.013	-0.008	-0.001	-0.001	-0.002	-0.001	0.001	-0.001
15	32; 30; [DU (0,10)]^6	-0.018	-0.014	-0.011	-0.000	-0.002	-0.002	0.001	0.003	0.009
16	10; 5.85; 1	0.000	-0.005	-0.008	-0.006	0.001	-0.006	-0.006	-0.004	-0.004
17	10; 5.85; DU (0, 10000)	-0.006	-0.009	-0.009	-0.007	0.003	-0.006	-0.008	-0.009	-0.007
18	10; 5.85; [DU (0, 10000)]^2	-0.005	-0.006	-0.010	-0.007	0.002	-0.007	-0.007	-0.011	-0.008
19	10; 5.85; [DU (0, 10000)]^3	-0.003	-0.013	-0.012	-0.008	0.002	-0.005	-0.008	-0.006	-0.006
20	10; 5.85; [DU (0, 10000)]^4	-0.011	-0.013	-0.016	-0.010	0.002	-0.007	-0.009	-0.011	-0.004
21	65; 63; [DU (0, 10000)]^{10}	-0.053	-0.043	-0.029	-0.006	-0.003	-0.007	-0.008	-0.006	-0.004
22	65; 63; [DU (0, 10)]^{10}	-0.048	-0.038	-0.026	-0.004	-0.003	-0.006	-0.004	-0.002	0.009
	(III) Exponential function distributions									
23	32; 30; 2^($\sum_1^{10}$ CU (0, 1))	-0.011	-0.005	-0.006	-0.001	0.000	-0.001	-0.005	-0.005	0.001
24	32; 30; 2^($\sum_1^{20}$ CU (0, 1))	-0.018	-0.020	-0.016	-0.003	-0.001	-0.003	-0.005	-0.004	-0.003
25	32; 30; 2^CU (0, 10)	-0.026	-0.018	-0.012	-0.002	-0.001	-0.004	-0.004	-0.003	0.004
26	32; 30; exp ($\sum_1^{10}$ CU (0, 1))	-0.026	-0.020	-0.014	-0.002	-0.002	-0.002	-0.003	-0.002	0.005
27	32; 30; exp (CU (0, 7))	-0.026	-0.021	-0.012	-0.001	-0.001	-0.003	-0.002	-0.002	0.003
28	32; 30; exp (CU (0, 10))	-0.034	-0.024	-0.018	-0.002	-0.002	-0.004	-0.001	0.002	0.009
29	65; 63; exp (CU (0, 100))	-0.221	-0.184	-0.142	-0.048	-0.002	-0.047	-0.081	-0.071	-0.053

The data in Table 5 is presented in the same way as in Tables 3 and 4, only instead of the loss of *df*, the law and the insertion parameters are given here. The abbreviations here are as follows: B – binomial, N – normal, CU and DU – continuous and discrete uniform distributions.

The results falling within the ±0.02 range can be observed in rows 1-20 and 23-24. Here hybrids with inserts containing binomial, normal, uniform and quadratic power distributions behave indistinguishably or almost indistinguishably from the simple hypergeometric distribution. Deviations increase with increasing degree in power functions, and with increasing base value and range of variation of the argument in exponential functions. The results look somewhat better if the argument of an exponential function varies according to the binomial law or a law "binomized" by

 doi:10.20944/preprints202305.0716.v1

summing up standard continuous distributions. In power functions the range of variation of the argument does not seem to play any role (cf. e.g., rows 13 and 14, 21 and 22).

It is not difficult to see that the asymmetry of the method error distribution increases as the deviations increase, which is clearly seen in the negative slope of the negative tail and less clearly in the positive slope of the positive tail. This is primarily a consequence of the redundancy of the zero values. The sparsity coefficient *S* presented in the previous subsection loses its already weak efficiency here, as with increasing mean values the discrete nature of the variation does not disappear. For example, row 8 contains a distribution where there are two variables consisting of 50 units each, which at the given parameters yields instead of the mean 0.7692 (50/65) the mathematical expectation 38.46 (50^2/65). Meanwhile, the expected number of zeros does not change, averaging about 30 ($65[64/65]^{50}$=29.94). This means that almost half of the sample size, when calculating covariance, always gives deviations from the mean exclusively with a plus sign. In rows 8a and 8b, the overall distribution is divided into 12 series with smaller and 9 with larger shifts. The difference in the deviations, although not very striking, is still significant.

According to the algorithm for calculating the sparsity coefficient, it should have decreased here by a factor of 50 respectively to the mathematical expectation of the insertion. However, here, as in all other distributions presented in Table 5, the coefficient remains as it would be in the case of an ordinary insertion with a constant value of 1, which is motivated by the preservation of the discrete nature of the variation. In real data, of course, we do not know anything about any inserts, but similar distributions, where values varying over a large range alternate with many zeros, are typical, for example, in studies of the microbiome. In such cases, it is probably better to focus on coefficients of variation: if about a third of the variables in the raw data give values above 50%, it is better to take the value 0.02 instead of the small estimated *S* value. By the way, ignoring this coefficient completely (i.e., setting it to 0) would by no means lead to disastrous consequences. In the example under consideration, of course, the relationship between the two small variables noted above would be the most affected, where corresponding to the significance levels of Table 5 deviations of -0.028, -0.009, -0.008, 0.001, -0.002, 0.001, 0.011, 0.010, 0.008 would take the values -0.034, -0.014, -0.012, -0.000, -0.002, 0.000, 0.010, 0.011, 0.010.

As can be seen, most of the distributions presented in Table 5 have no loss of degrees of freedom. Clearly, the varying quantities cannot give the same values for all sums, but we have identical mathematical expectations throughout the sample size, which are taken into calculation. If we calculated directly, some part of *df* would inevitably be lost in each individual case. We would end up with a picture with slightly increased negative deviations, which already prevail over the positive ones. It is reasonable to assume that the expected decrease of *df* is somehow compensated here. If we look at row 30 with an exponent whose argument varies continuously and uniformly up to 100, where both tails have an abnormal negative growth with a characteristic dominance of the negative tail, it is hard to resist the assumption that the additional variation of the insertion leads to an increase in *df*.

In most cases the inequality of sizes of the areas under study is an objective reality, and in these cases taking into account the loss of *df* is necessary. But sometimes data representing samples taken from spaces of the same size, but with a varying number of objects detected on them, are subjected to analysis. As an instance again, samples taken for microbiome studies are appropriate. In such cases, if the same mathematical expectation is assumed for all samples, taking into account the loss of *df* might be redundant. However, we are not sure about this, as we cannot yet answer the question about the existence of the effect of additional variation in objective rather than experimental reality.

4. Discussion

The method proposed in this article for converting the correlation coefficients calculated from proportions yields results that appear with about the same probability as the Pearson correlation coefficients corresponding to them in magnitude, calculated from compositionally unrelated normally distributed variables.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2023 doi:10.20944/preprints202305.0716.v1

This can be stated with certainty not only in the case where the experimental units in the hypergeometric model are real units. This is also true for a certain range of hybrid models in which the researcher-selected law of reproduction of the original unit is built into the hypergeometric distribution. Certainly, the question of the limits of applicability of our method requires careful further investigation. Nevertheless, since the laws discussed in subsection 3.5. are the basic laws reflecting the variability of the real world, it can be assumed that our method is applicable to many compositions of real data.

Naturally, the question immediately arises, in which real, not experimental, data one can postulate the presence of multiplying units with equal initial parameters and whether the null hypothesis, in which inequality is regarded as a factor generating non-random relationships, can be used to evaluate data in which a comparison of variance values with high probability indicates the absence of equality. If the variance in a hybrid simulation is large enough, replication will produce very different "individual portraits" of the composition each time. The main indicator of the initial equality of distribution units – the constancy of the relation between the variance and the mathematical expectation – will change freely by several times for both small and large variances, if the sample size is not large enough. Under such conditions, estimating the probability of initial equality in real experience data is an extremely difficult task.

Let us simulate an explicit inequality situation by three discrete uniform distributions (0, 2000), (10, 40), and (10, 40) at n=16. The mean values of the Pearson correlations here is -0.954 for the two relationships of the large variable with the small variables and 0.829 for the relationship between the small variables. Eq. (5) converts these values to -0.755 and 0.836. If we focus on the fact that all values of the summing variable have the same mathematical expectation, which, according to our method, excludes loss of degrees of freedom, then Eq. (6) will give the same result with p-values of 0.0007 and 0.000055. In our experimental reality with this anomalous variance ratio, the values of the summing variable vary enormously and the number df_0 varies freely from about 5 to 11. Suppose that in a real experiment we were studying environmental niches, then the disparity of size would be an objective reality and we would get, say, a value of df_0=8. This would lower the p-values to 0.0065 and 0.0009, and Eq. (6), returning the usual value of df_0=14, would lower the coefficients to -0.649 and 0.744. But in this series of results these coefficients are at the median point, so their empirical p-values are 1.

Suppose we would obtain a similar picture by recording the spatial distribution of three closely related species competing for the same resources. The high variance in dispersal of individuals of the evolutionarily more successful species would be consistent with the fact that they tend to the most resource-rich locations, leaving the "outcasts" who have lost out in the intraspecific competition, on scarce territories. In the positive association of the two small species, we would see a usual consequence of natural selection, which preserved those species (formerly subspecies that evolved from "outcasts") that learned to cooperate.

Obviously, somewhere in the evolutionary origins of this system of interactions, all individuals were potentially equal in their ability to reproduce and choose their habitat. But at the time of observation, the equality had disappeared. Should we reject such data if we know for a fact that correlations at such ratios of variances arise by themselves without any evolutionary process? Here a counter question arises: could such variances in real, rather than experimental, situations have arisen by chance? We are inclined to the negative answer: such anomalous ratios between variances can arise only as a result of a directed choice of a qualitatively better space. This entails the displacement of competitors as well as other causal relationships. Obviously, the decision of whether to accept or reject must be made on a case-by-case basis by the researcher, and good knowledge of the system being analyzed will contribute to the correct choice. However, well known systems are unlikely to be subjected to such an analysis, so we believe that data with anomalous variance should still be better accepted with great caution and some disbelief than rejected altogether, remembering to transform the original correlation coefficient values using the method proposed in this article.

Because of the authors' desire to minimize the errors in various extreme cases (very large losses of df, highly sparse variables, small sample sizes, close to -1 shift), the method as a whole turned out to be very cumbersome. But such cases (with the exception of sparse variables, the abundance of

which haunts microbiome analyses) are very unlikely in real experiments. Therefore, up to the loss of about half of the original number of degrees of freedom, one can use simplified versions of Eqs. (3) and (5) without much risk of obtaining a noticeable decrease in accuracy: $\mu = r_0/[1 + (1 - r_0^2)/(2df_0)]$ and $r_{df} = \frac{r-\mu[1+(1-r^2)/df]}{1-\mu[r-\mu(1-r^2)/df]}$.

It should be added that one can also manage without normalization through the sum of components, replacing it with a partial correlation coefficient, which eliminates the influence of summing variable on the coefficients calculated from the absolute data. Our observations so far show, however, that this is not the best choice, since the partial correlation method behaves normally as long as the values of the summing variable are distributed relatively normally. Of course, the partial correlation coefficient does not get rid of all subsequent procedures, but simply replaces the coefficient calculated from the relative data exceeding it slightly on average, which is exactly the difference of one degree of freedom added by this calculation method.

In one particular case, we can say with a high degree of certainty that the data of the real study are quite consistent with the condition of initial equality of distributed units. These data, in short, represent the distributions of various phonetic features over the semantic space of the language system. The analysis of these data provides interesting results concerning the early stages of the evolution of language. These results are soon to be published and will be available for review and further discussions.

Author Contributions: Conceptualization, Y.V.M.; methodology, Y.V.M.; software, Y.V.M.; investigation, Y.V.M. and Y.D.N.; writing—original draft preparation, Y.V.M.; writing—review and editing, Y.V.M., Y.D.N.; supervision, Y.D.N.; project administration, Y.D.N.; funding acquisition, Y.D.N. Authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Program of Fundamental Research in the Russian Federation for the 2021–2030 period (project No. 121052600299-1).

Acknowledgments: We thank Professor V.I. Galkin of the Physics Department of Moscow State University for useful discussion of the problem and valuable remarks.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Chayes, F. On correlation between variables of constant sum. *Jour. Geophysical Res*. **1960**, Volume 65, 12, 4185–4193.
2. Sarmanov, O. V. O lozhnoy korrelyatsii mezhdu sluchaynymi velichinami [On spurious correlation between random variables]. *Trudy Matematicheskogo instituta imeni V. A. Steklova* **1961**, Volume 64, 173–184 (in Russian).
3. Mosimann, J. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika* **1962**, Volume 49, 1/2, 65–82.
4. Chayes, F.; Kruskal, W. An approximate statistical test for correlations between proportions. *Jour. Geol*. **1966**, Volume 74, 692–702.
5. Aitchison, J. A new approach to null correlations of proportions. *Mathematical Geology* **1981**, Volume 13, 2, 175–189.
6. Aitchison, J. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B* (*Methodological*) **1982**, Volume 44, 2, 139–177.
7. Friedman, J.; Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol*. – **2012**, Volume 8, 9, 1002687
8. Available at: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002687.
9. Fang, H.; Huang, C.; Zhao, H.; Deng, M. CCLasso: correlation inference for *compositional data* through Lasso. *Bioinformatics* **2015**, Volume 31, 19, 3172-3180.
10. Available at: https://academic.oup.com/bioinformatics/article/31/19/3172/211784.

11. Lovell, D.; Pawlowsky-Glahn, V.; Egozcue, J. J.; Bähler, J. Proportionality: a valid alternative to correlation for relative data. *PLoS Comput. Biol*. **2015**, Volume 11, 3, 1004075
12. Available at: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004075.
13. Lovell, D.; Chua, Xin-Yi; McGrath, A. Counts: an outstanding challenge for log-ratio analysis of compositional data in the molecular biosciences. *NAR Genomics and Bioinformatics* **2020**, Volume 2, 2, 5859926
14. Available at: https://academic.oup.com/nargab/article/2/2/lqaa040/5859926?login=false
15. Kurtzt, Z. D.; Müller, C. L.; Miraldi, E. R.; Littmann, D. R.; Blaser, M. J.; Bonneau, R. A. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol*. **2015**, Volume 11, 5, 1004226
16. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4423992/.
17. Ban, Y.; An, L.; Jiang, H. Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics* **2015**, Volume 31, 20, 3322-3329.
18. Available at: https://academic.oup.com/bioinformatics/article/31/20/3322/195785?login=false
19. Erb, I.; Notredame, C. How should we measure proportionality on relative gene expression data? *Theory in Biosciences* **2016**, Volume 135, 1-2, 21-36.
20. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4870310/.
21. Schwager, E.; Mallick, H.; Ventz, S.; Huttenhower, C. A Bayesian method for detecting pairwise associations in compositional data. *PLoS Comput. Biol*. **2017**, Volume 13, 11, 1005852
22. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5706738/.
23. Kynčlová, P.; Hron, K.; Filzmoser, P. Correlation between compositional parts based on symmetric balances. *Mathematical Geosciences* **2017**, Volume 49, 6, 777-796.
24. Available at: https:// www.researchgate.net/publication/312081950
25. Yoon, G.; Gaynanova, I.; Müller, C. L. Microbial networks in SPRING – semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. *Front. Genet.* **2019**, Volume 10, 00516
26. Available at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00516/full#B36
27. Egozcue, J. J. Reply to "On the Harker Variation Diagrams; . . ." by J.A. Cortés. *Mathematical Geosciences* **2009**, Volume 41, 829-834.
28. Shaffer, M.; Thurimella, K.; Sterrett, J. D.; Lozupone, C. A. SCNIC: Sparse correlation network investigation for compositional data. *Molecular Ecology Resources* **2023**, Volume 23, 1, 312-325.
29. Available at: https:// onlinelibrary.wiley.com/doi/10.1111/1755-0998.13704
30. Faust, K.; Sathirapongsasuti, J. F.; Izard, J.; Segata, N.; Gevers, D.; Raes, J.; Huttenhower, C. Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol*. **2012**, Volume 8, 7, 1002606
31. Available at: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002606
32. Fisher, R. A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **1915**, Volume 10, 4, 507-521.
33. Olkin, I.; Pratt, J.W. Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics* **1958**, Volume 29, 201-211.
34. Lutz, K. C.; Jiang, S.; Neugent, M. L.; De Nisco, N. J.; Zhan, X.; Li, Q. A survey of statistical methods for microbiome data analysis. *Front. Appl. Math. Stat.* **2022**, Volume 8, 884810
35. Available at: https://www.frontiersin.org/articles/10.3389/fams.2022.884810/full