ISSN (Online): 2350-0530
ISSN (Print): 2394-3629

International Journal of Research -GRANTHAALAYAH
November 2020, Vol 8(11), 63 – 75
DOI: https://doi.org/10.29121/granthaalayah.v8.i11.2020.1470

DISTRIBUTION DEPENDENT CORRELATIONS: A MATHEMATICAL PRINCIPLE UTILIZED IN PHYSIOLOGY, OR CORRELATION BIAS?

Arne Torbjørn Høstmark *[1]

*[1] Institute of Health and Society, Faculty of Medicine, University of Oslo, Norway, Box 1130 Blindern, 0318 Oslo, Norway

DOI: https://doi.org/10.29121/granthaalayah.v8.i11.2020.1470

Article Type: Research Article

Article Citation: Arne Torbjørn Høstmark. (2020). DISTRIBUTION DEPENDENT CORRELATIONS: A MATHEMATICAL PRINCIPLE UTILIZED IN PHYSIOLOGY, OR CORRELATION BIAS?. International Journal of Research - GRANTHAALAYAH, 8(11), 63-75. https://doi.org/10.29121/granthaalayah.v8.i11.2020.1470

Received Date: 08 September 2020

Accepted Date: 26 November 2020

Keywords:
Correlation Rules
Relative Amounts
Distribution
Biological Regulation

ABSTRACT

In many studies, we may raise the question of whether relative amounts of particular variables are positively or negatively associated, but investigations specifically focusing upon this issue seem hard to find. Previously, we reported some general rules for associations between relative amounts of positive scale variables. The main research question of the present work was: How are correlations between percentages of the same sum brought about? One particular feature of such correlations seemed to be that distributions (ranges) of the variables were crucial for obtaining either positive or negative correlations, and for their strength, suggesting the name Distribution Dependent Correlations (DDC). Certainly, such correlations might cause bias. However, previous findings raise the question of whether DDC might have a physiological relevance as well. In the current work, we extend and systematize theoretical considerations, and show results of computer experiments to test the hypotheses. Finally, we briefly mention a couple of examples from physiology. The results seem to support the idea that true, within-person distributions of the variables are crucial for obtaining positive or negative correlations between their relative amounts, raising the question of whether evolution might utilize DDC to regulate metabolism.

Definitions and Abbreviations

Variability: the width or spread of a distribution, measured e.g. by the range and standard deviation.

Distribution: graph showing the frequency distribution of a variable within a particular range. In this article, we also use distribution when referring to a particular range, a – b, on the scale.

Uniform distribution: every value within the range is equally likely. In this article, we may write "Distribution was from a to b", or "Distributions of A, B, and C were a – b, c – d, and e - f, respectively".

"Low–number variables" have very low numbers relative to "high-number variables".

WBC = White Blood Cells; N = segmented neutrophil leukocytes; L = Lymphocytes; M = Monocytes; E = Eosinophil leukocytes; B = Basophil leukocytes

1. INTRODUCTION

In various studies, we may raise the question of whether relative amounts of particular variables are positively or negatively associated. However, investigations specifically focusing upon this issue seem hard to find, in a literature search. The apparent lack of interest might possibly relate to a methodological concern arising when correlating percentages of the same sum, since significant associations could be obtained mathematically, without e.g. having any biological implications (Pearson, 1897). On the other hand, it may not always be apparent whether positive (negative) associations between percentages of the same sum should be rejected as correlation bias, or rather be considered to have biological relevance. In this context, we previously reported that relative amounts of e.g. particular body fatty acids can be positively or negatively associated as a consequence of their particular *concentration distributions* (range/variability/skewness), suggesting the name *Distribution Dependent Correlations*, DDC (Høstmark and Haug, 2018; 2019a, b; 2020a-c). Furthermore, we raised the question of whether evolution might utilize such correlations as a regulatory mechanism (Høstmark and Haug, 2018; 2020a, c). If so, we should find examples in physiology. Furthermore, since DDC rules are general, they should apply to any unit system in nature. Previously we suggested two general DDC rules (Høstmark and Haug 2019c):

1) If S is the sum of 3 positive scale variables (S = A + B + C), where A and B have low numbers and low variability relative to C, then we might expect a positive %A vs. %B association, and a negative %C vs. %A(%B) association. A decrease (increase) in the variability (range) of A and/or B should improve (make poorer) the %A vs. %B association. In contrast, a narrowing (broadening) of the C range should make poorer (improve) the %A vs. %B association.
2) If A and/or B have very high numbers relative to C, then we should expect a negative %A vs. %B association, irrespective of the ranges of A and B.

The aim of the present work was to further reason about and discuss how Distribution Dependent Correlations are brought about, carry out computer experiments to test hypotheses, and briefly present a couple of examples related to physiology.

2. MATERIALS AND METHODS

Previously (Høstmark and Haug, 2020b), the association between relative amount of arachidonic acid (20:4 n6) and percentage of e.g. eicosapentaenoic acid (EPA, 20:5 n3) was investigated, in chicken lipids. From histograms, the physiological concentration distributions (g/kg wet weight) for the fatty acids were determined. Next the sum (S, g/kg wet weight) of all fatty acids was computed, as well as and the remaining sum (R) when omitting the couple of fatty acids under investigation, thereby apparently obtaining 3 positive scale variables. With these variables, and with surrogate random number variables, generated with the true concentration distributions, computer analyses as described in detail below, were carried out. For the purpose of the present work, the three positive scale variables were named A, B, and C. Previous analyses (Høstmark and Haug, 2020a, b) demonstrated that correlations between e.g. %A and %B depended upon the particular *distribution* (range) of each of the variables involved. Thus, we obtained similar correlation outcomes using the true (measured) values, or random numbers, if the ranges were like the measured ones.

A major part of the present work consists of computer experiments using random numbers to explore further, how distributions of A, B, and C might influence the association between their relative amounts. Thus, A + B + C = S. Dependency between percentages is shown by the equation %A + %B + %C = 100. Using random numbers for A, B, and C, each of which sampled within defined ranges; we studied histograms, scatterplots, and correlations (Spearman’s rho). Computer experiments were performed, to study how *alterations* in the ranges of the random numbers might change associations between %A, %B, and %C. Several repeats were carried out, with new sets of random numbers (n = 200 each time); the general outcome was always the same, but corresponding correlation coefficients and scatterplots varied slightly.

We present the results mainly as scatterplots with correlation coefficients. In most of the computer experiments, the random numbers had *uniform* distribution, but random numbers with *normal* distribution were used as well, however giving qualitatively similar results. We used SPSS 25.0 for the analyses, and for making figures. The significance level was set at $p < 0.05$. We present further details under Results and Discussions.

3. RESULTS AND DISCUSSIONS

We first consider how *Distribution Dependent Correlations* might arise, using two lines of reasoning: 1) utilizing the equation of a straight line, and 2) applying the relationship between sum (S) of all variables and their fractions (percentages) of S.

3.1. UTILIZING THE EQUATION OF A STRAIGHT LINE

3.1.1. THEORETICAL CONSIDERATIONS

If S is the sum of many positive scale variables, S = A + B + C + …, we may simplify to S = A + B + R, i.e. %A + %B + %R = 100, or %B = -%A + (100 -%R), where R is the sum of all variables, except A and B. This equation seems to resemble the equation of a straight line, however involving percentage amounts of three unknown variables (A, B, R), each of which with a defined distribution. Accordingly, it is hard to know whether there might be an association between relative amounts of e.g. A and B. However, in two particular conditions we may further simplify the equation by approximations, apparently to involve two variables only. This may be achieved: 1) if the expression (100 - %R) approaches zero, or 2): if %R approaches zero.

3.1.1.1. % R APPROACHING 100

If %R consists of high values (close to 100) and the low-number, corresponding values of %A and %R are such that (100 - %R) > %A, then the equation appears to approach %B = %A, showing a linear positive association between %A and %B. The requirement (100 - %R) > %A is indeed satisfied, sine the remaining value when calculating (100 - %R) would have to be divided between %A and %B. For example, suppose that %R could reach 99%, then the remaining percent would have to be divided between %A and %B. Hence, the slope of the %A vs. %B regression line must be positive. Accordingly, with high %R values relative to A and B percentages, we might expect a positive %B vs. %A association. In this context, we should keep in mind that all variables in the denominator (S) are still there when dealing with percentages of S. Nevertheless, the above simplified relationship between the percentages seemed to work well under this "extreme" condition, as verified by computer experiments, using random numbers (*vide infra*). Conceivably, with very high R-values relative to A (B), a small increase (decrease) in %R, should be accompanied by a compensatory decrease (increase) in %A and %B. Furthermore, it follows from the above reasoning that any change in the ranges of A, B, or R should influence the magnitude of %R, thereby changing the %A vs. %B association. Thus, if the R-range is moved towards higher (lower) values, then the expression (100 - %R) should move closer to (away from) zero, thereby improving (making poorer) the %A vs. %B association. Additionally, since the equation %A + %B + %R = 100 is valid, also a change of A (B) ranges towards lower (higher) values should improve (make poorer) the correlation between %A and %B, accompanied by increased (decreased) values of %R, caused by altering the A (B) ranges.

Above we reasoned that we should expect a positive association between %B and %A, if the expression (100 - %R) approached zero. However, in this case it is inappropriate to approximate the expression to %B = %A, like Y = X. In the latter case, both the abscissa and the ordinate may have any value on the scale, and the Y vs. X graph would have slope = 1. In contrast to this, %B and %A – values are limited by the B and A *ranges*, respectively. A more general approximation would be:

$$\%B_{(p-q)} = [(\%B_{max} - \%B_{min})/(\%A_{max} - \%A_{min})]*\%A_{(r-s)} + z$$

The subscript parentheses indicate ranges of %A and %B, and z = 100 - %R. Thus, z becomes increasingly small as %R increases. The approximated slope value would be:

$$(\%B_{max} - \%B_{min})/(\%A_{max} - \%A_{min})$$

The slope should approach +1, only if ranges of A and B are equal. Additionally, the slope value computed manually based on maximum and minimum values of %A and %B may deviate somewhat from the slope made by the computer, especially with poor scatterplots. To improve readability in the mathematical expressions below, we omit indication of ranges.

3.1.1.2. % R APPROACHING ZERO

Above we reasoned that there should be a positive %B vs. %A association if (100 - %R) > %A. Obviously, this requirement is present also with low values of %R, raising the question of how to explain that the correlation between A and B percentages in this case must be *negative*. The reason is that, when %R is close to zero, the equation approaches: % B = - % A + 100, or %A + %B = 100. Hence, the A and B percentages of S must vary inversely.

3.1.2. COMPUTER EXPERIMENTS TO TEST THE HYPOTHESES

In most of the calculations below, we used random numbers with uniform (rectangular) distribution; however, the correlation outcomes were similar with uniform and normal distribution of the random numbers (results not shown), provided that ranges were equal.

3.1.2.1. %R APPROACHING 100

To obtain high %R values relative to %A(%B), the following ranges were arbitrarily chosen: A 0.1 - 0.2; B 0.3-0.5; R 2 – 20. With these ranges, we generated 200 uniformly distributed random numbers. As shown in Fig. 1 (left panel), %A correlated positively with %B (Spearman's rho = 0.872; %R vs. %A (%B): rho = -0.933 (- 0.986), p < 0.001 for all. Quartiles of the %A, %B, and %R distributions were *0.9, 1.4, 2.3; 2.5, 3.4, 5.8; and 92.0, 95.3, 96.6%*, respectively, i.e. showing very high %R values relative to %A(%B). Equation of the regression line (SE in parentheses) was %B = 2.45 (0.08) *%A + 0.34(0.16). The slope value estimated manually applying the above formula was 2.30. Skewness of %A, %B, and %R histograms were 1.46, 1.41, and -1.39, respectively. Thus, percentages of the low-number variables (A, B) were positively skewed, and percentage of the high-number variable (R) was negatively skewed (further commented below).

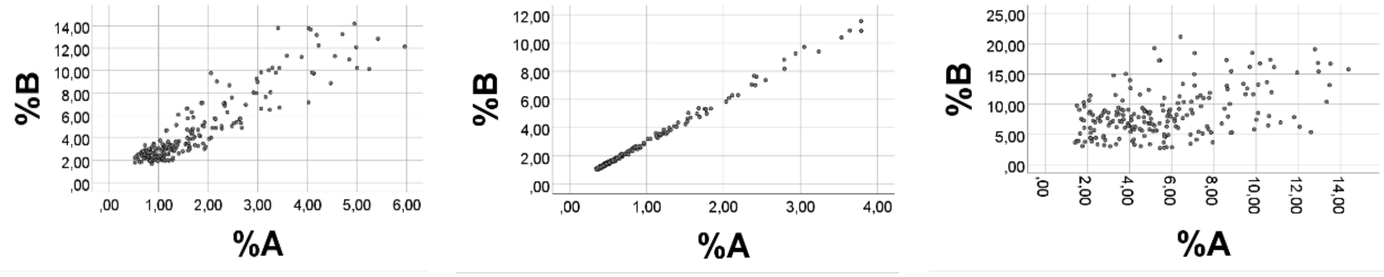

Figure 1: Association between %A and %B, as influenced by changing the ranges of A, B, and R. The figure relates to the equation %A + %B + %R = 100, see text. Uniformly distributed random numbers (n = 200) were used. Left panel; ranges were A 0.1 – 0.2; B 0.3 – 0.5; R 2 – 20; %A vs. %B (Spearman's rho = 0.872, p < 0.001). Middle panel: A 0.10 – 0.11; B 0.30 – 0.33; R 2 – 30; %A vs. %B: rho = 0.995, p <0.001. Right panel: A 0.10 – 0.5; B 0.2 -0.7; R 2 – 8; %A vs. %B: rho = 0.362, p < 0.001.

To test whether the %A vs. %B association *improves* by moving the %R distribution towards higher values, we need to increase %R This was achieved through narrowing ranges of A(B) towards the lower limit, and broadening the R range towards higher values, e.g. A 0.10 – 0.11; B 0.30 – 0.33; R 2 – 30. As anticipated, the %A vs. %B association did improve, as judged from the scatterplot (Fig. 1, middle panel) and correlation coefficient (%B vs. %A: rho = 0.998; %R vs. %A (%B): rho = -0.998 (- 1.000), p < 0.001 for all. Equation of the regression line (SE in parentheses) had changed to: %B = 2.98 (0.02) *%A + 0.02(0.02). Slope computed manually by the above approximation was 3.10. Quartiles of the %A, %B, and %R distributions were 0.5, 0.6, 1.0; 1.3, 1.7, 3.2; and 95.8, 97.7, 98.2%, respectively. Thus, the %R distribution had moved towards higher values, in line with the improved association between %A and %B. Conceivably, the %A and %B distributions had moved towards lower values.

Skewness (SE) of %A, %B, and %R histograms had increased, being 2.19 (0.17), 2.19 (0.17), and -2.19 (0.17), respectively, as compared with 1.46, 1.41, and -1.39, before narrowing.

To test whether the positive %A vs. %B association will *be poorer* by moving the %R distribution towards *lower* values, we need to decrease %R. This was achieved by arbitrarily changing ranges to be A 0.10 – 0.5; B 0.2 -0.7; R 2 – 8. As shown in Fig. 1, right panel, the scatterplot for the %A vs %B association became poorer, as also verified by the correlation coefficient: rho = 0.362 (p<0.001) for the %A vs. %B association. Correlation between %R and %A (%B): rho = -0.773 (- 0.844), p < 0.001 for all. Quartiles of the %A, %B, and %R distributions were: 3.4, 5.1, 7.3; 5.7, 7.9, 10.4; 82.6, 87.5, 89.6%. Thus, the %R distribution had moved towards lower values, in keeping with the observed poorer %A vs. %B correlation. Accordingly, the %A and %B distributions had moved towards higher values. Skewness (SE) of %A, %B, and %R histograms had changed to 0.83 (0.17), 0.94 (0.17), and -0.96 (0.17), respectively, showing attenuated skewness of the distributions of the percentages. Equation of the regression line was %B = 0.60 (0.08) *%A + 5.10 (0.53). With this poor scatterplot, we observed a large difference concerning the computer-calculated slope value of 0.60, and the one obtained manually by the above-approximated formula, i.e. 1.23. We previously explained the relationship between ranges and skewness (Høstmark, 2019d).

3.1.2.2. %R APPROACHING ZERO

Above we argued that, with %R approaching zero, there should be a negative %A vs. %B association, since the equation would approach %B = -%A + 100. The following ranges were changed to: A 1- 5; B 2 - 3; and R 0.10 – 0.15. These ranges gave the following values of %A, %B, and %R quartiles: %A 36.9, 42.8, and 54.2%; %B 43.2, 55.2, and 61.3; %R 1.8, 2.2, 2.6%. Accordingly, the %R distribution had small values compared with those of %A (%B). As shown in Fig. 2, there was – as expected- a strong inverse association between %A and %B (rho = -0.999, p<0.001).

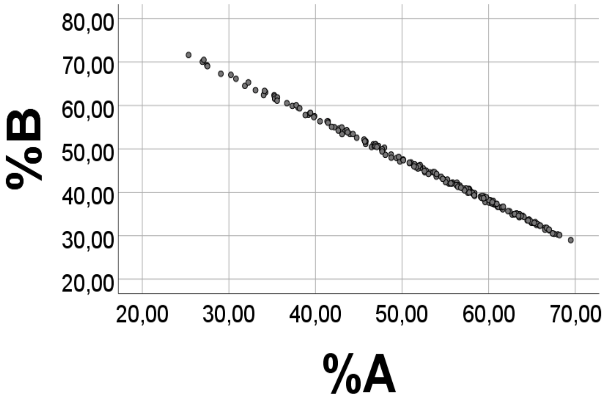

Figure 2: Association between A and B percentages of S, in the equation S = A + B + R, see text. Uniformly distributed RANDOM numbers (n = 200) were used. Ranges were; A: 1 - 5; B: 2 - 3; R: 0.10 – 0.15. %A vs. %B: rho = -0.999, p < 0.001. Equation of the regression line (SE in parentheses): %B = -1.04 (0.002) *%A +99.7(0.11), n =200.

Equation of the regression line was %B = -1.04 (0.002) *%A +99.7(0.11), n =200, i.e. showing a slope close to -1.

Thus, when %R in the equation %B = -%A + (100 -%R) approached zero, then we obtained a negative %A vs. %B association. In additional computer experiments, we changed ranges of the variables in many ways. However, the correlation outcomes were always as predicted above. Thus, decreasing (increasing) the %R range improved (made poorer) the negative %A vs. %B association (not shown), irrespective of whether the %R distribution was moved towards lower (higher) values by altering the R range only, by changing the A and/or B ranges, or by altering ranges of all of the current variables. These experiments seem to support the idea that the above negative association between A and B percentages is a case of *Distribution Dependent Correlations.*

3.2. APPLYING THE RELATIONSHIP BETWEEN SUM (S) AND PERCENTAGES OF S TO EXPLAIN ASSOCIATIONS BETWEEN THE PERCENTAGES

3.2.1. THEORETICAL CONSIDERATIONS

Above it was suggested that, with a combination of two low-number variables (A, B) relative to one high-number variable (R), we might expect a positive association between %A and %B. With this condition, A (B) *percentages* of S should decrease, and %R increase when S increases from lowest to highest value. Accordingly, we should expect a positive correlation between A and B percentages, since both of them are negatively related to S. To explain this outcome further, we omit ranges of the variables, i.e. A + B + R = S. The A, B, and R *fractions* of S are A/S, B/S, and R/S, respectively. Since A and B have low numbers and low ranges relative to S, the A and B *fractions* (percentages) should decrease with increasing S from lowest to highest value within the S range. The R fraction of S is R/S=R/ (A + B + R), or 1/ (1 + z/R), if z =A + B. Since z is small compared with R, the R fraction (and percentage) of S should *increase* with increasing R, and accordingly also with increasing S, because R is the main contributor to S. Thus, S should be positively associated with %R. Since % A and %B are both negatively associated with S, these percentages should be positively associated. Furthermore, the positive association between %R and S explains the negative association between %R and %A (%B). Accordingly, the relationships between S and percentages of S may explain the positive %A vs. %B association, as well as the inverse relationships between %R and %A (%B).

On the other hand, with two variables (A and B) having high numbers, relative to a third one (R), we should expect a *negative* %A vs. %B association. In this case, sum (S) of the 3 variables would approach S = A + B, i.e. %A + %B =100. Conceivably, when approaching a condition involving two variables only, their relative amounts should be expected to vary inversely. However, is the above "relation to sum" approach useful to explain such correlations? In this case, the A fraction of S would approach A/S = A/ (A + B) = 1/(1 + B/A) and the B-fraction would approach B/(A+B) = 1/(A/B + 1), raising the question of whether the A (B) fraction of S increases or decreases with increasing S-values. A negative %A vs %B association should be expected if S is positively related to one of the fractions, and negatively associated with the other one. To evaluate this question, we have to know the ranges of A and B, and consider whether %A(%B) increases or decreases as S goes from lowest to highest value. The minimum value of the A-fraction would approach 1/ (1 + Bmax/Amin), and the maximum value 1/ (1 + Bmin/Amax). Corresponding S values would be: (Amin + Bmax), and (Amax + Bmin). Accordingly, if (Amax + Bmin) > (Amin + Bmax), then we should expect %A to be positively associated with S. Conversely, if (Amax + Bmin) < (Amin + Bmax), then we should expect a negative S vs. %A relationship. A similar reasoning should apply to S vs. %B. Thus, a negative %A vs. %B association should be expected, if %A (%B) increases (decreases) with increasing S. However, if S is close to be similar with the lowest and highest %A (%B) values, then the "association to sum" approach should not be useful to explain correlations between %A and %B, as exemplified below.

3.2.2. COMPUTER EXPERIMENTS TO TEST THE HYPOTHESES

3.2.2.1. POSITIVE %A VS. %B ASSOCIATIONS

We generated 200 uniformly distributed random numbers with ranges shown in Fig. 1 (left panel), i.e. A: 0.1 - 0.2; B: 0.3 - 0.5; and R: 2 - 20. As expected, S correlated negatively with %A (rho = -0.904) and with %B (rho = -0.949), and positively with %R (rho = 0.962), p<0.001 for all. The finding that S correlated negatively with both of %A and %B might explain the positive %A vs. %B association (rho = 0.872, p < 0.001), see Fig. 1 (left panel).

Next, the *A and B ranges were narrowed* appreciably, and *R broadened*, to be like ranges shown in Fig. 1, middle panel, i.e. A: 0.10 – 0.11; B: 0.30 – 0.33; and R: 2 – 30. With the current values of the A, B, and R fractions of S, we obtain the following approximated values of the percentages: %A = 11/S, and %B = 32/S, or S = 11/%A, and S = 32/%B, suggesting an inverse relationship between S and A (B) percentages. Furthermore, since one particular S-value in this special case corresponds closely to one %A (%B) value only, the S vs. % A (% B) scatterplot should be close to a line, as was observed (Fig. 3). Additionally, from the approximated formulas above, it is seen that a one-unit increase in %A (%B) at low levels of the percentages would be associated with a larger decrease in S than a similar increase at higher levels, suggesting a curvilinear negative S vs. %A (%B) association, with the concave upwards. With the current ranges of the variables, R = S - 0.43. The R fraction of S, i.e. R/S = 1- 0.43/S, showing that

%R and S are positively associated. Additionally, the formulas also imply that the increase in S per unit increase in %R is augmented with increasing values of R (%R). Hence, we should expect a curvilinear positive S vs. %R scatterplot with the concave upwards. Similar conclusions about the curve shapes may be obtained by considering derivatives of the above functions (not shown). A computer experiment with random numbers was in accordance with these considerations (Fig. 3). Correlations were: S vs. %A (%B): rho = -0.998 (-0.998); S vs. %R, rho = 0.999; %A vs. %B, rho = 0.996, p < 0.001 for all (n = 200).

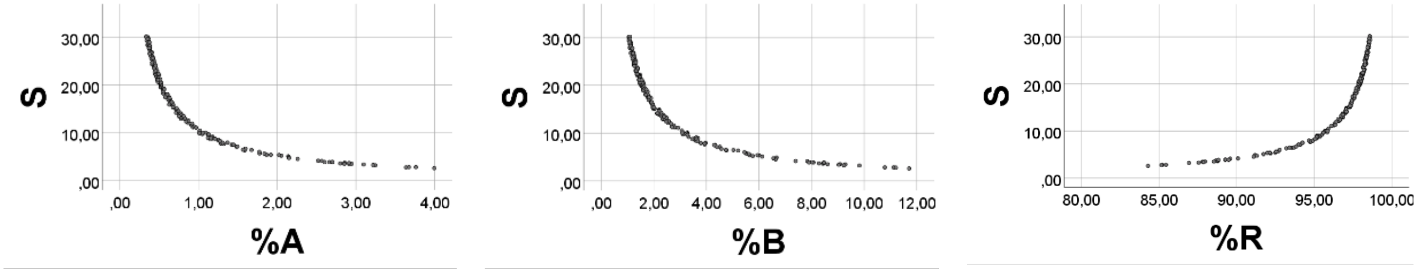

Figure 3: Association between sum (S) of A, B, and R, and their percentages, when *ranges of A and B were narrow* relative to R, i.e. A: 0.10 – 0.11; B: 0.30 – 0.33; and R: 2 – 30. The figure relates to the equation S = A + B + R, see text. Uniformly distributed RANDOM numbers (n = 200) were used. Left and middle panels: S vs. %A (%B): rho = -0.998 (-0.998). Right panel: S vs. %R, rho = 0.999; %A vs. %B, rho = 0.996, p < 0.001 for all (n = 200).

When changing ranges of A, B, and R to be like those shown in Fig. 1, right panel, the scatterplot of the S vs. %A (%B, %R) should be poorer, since in this case the close association between S and percentages of S would be disturbed. The results of a computer test was qualitatively as expected (Fig. 4). S vs. %A (%B) gave rho = - 0. 484 (-0.669), left and middle panels. Also, the S vs. %R scatterplot (Fig. 4, right panel) became poorer; rho = 0.752), p < 0.001 for all.

These results show that positive associations between percentages in the current conditions are *distribution dependent* ones, and that they may be explained by simple algebraic approaches.

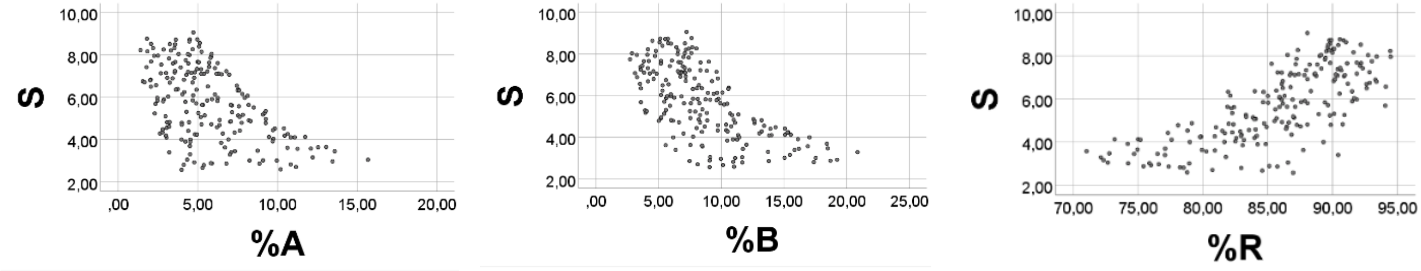

Figure 4: Association between the sum (S) of A, B, and R, and their percentages of S, when ranges of A and B were broadened. The figure relates to the equation S = A + B + R, see text. Uniformly distributed RANDOM numbers (n = 200) were used. Ranges were: A: 0.10 – 0.5; B: 0.2 -0.7; R: 2 – 8. Left and middle panels: S vs. %A (%B): rho =-0.484 (-0.669). Right panel: S vs. %R, rho = 0.752, p < 0.001 for all (n = 200).

3.2.2.2. NEGATIVE %A VS. %B ASSOCIATIONS

3.2.2.2.1. DIFFERENT RANGES OF A AND B

We first considered two high-number variables (A, B) relative to R; if ranges of A and B were ***different***. One example of this condition could be: A 1- 5; B 2 - 3; and R 0.10 – 0.15. The theoretical minimum and maximum values of the A fraction in this case would approach A/(A+B) = 1/ (1 + 3) = 0.25, and 5/ (5 + 2) = 0.71, respectively. Similarly, minimum and maximum values of the B fraction would be 2/ (5 +2) = 0.29, and 3/ (1 +3) =0.75. Corresponding minimum (maximum) S values would be: 4 (7) and 7 (4). Therefore, we should expect a positive association between S and the A-fraction (percentage), and a negative one between S and the B-fraction (percentage). Accordingly, A and

B percentages should be negatively associated. A computer test showed these values: Spearman's rho for S vs %A (%B): 0.875 (-0.864); %A vs. %B: rho =-0.999, p<0.001 for all; n = 200. The extrapolated scatterplot (not shown) did not cross the %A axis at exactly 100%, or the %B axis at 0%, due to the approximations done. Thus, with the current A and B ranges, it seems that the "relation to sum" approach might apply to explain also some negative correlations between percentages of the same sum.

3.2.2.2.2. EQUAL RANGES OF A AND B

We next tested the suggested conditions expected to give *negative correlations between %A and %B*, i.e. two high-number variables (A, B) relative to R; if ranges of A and B, and where ranges were ***equa****l.* The following ranges were chosen: 1- 10 for both of A and B, and 0.1 – 0.15 for R, giving the following approximated minimum and maximum values of the A and B fractions of S: 1/(1 + B/A) = 1/(1 + 10/1) = 0.09, and 1/(1 + 1/10) = 0.91, respectively, i.e. %A was running from approximately 9 to 91%. However, the corresponding S values would be 11 for both of the "extreme" %A (%B) values. We might, accordingly, expect a poor association between S and %A (%B). A computer experiment with uniformly distributed random numbers (n = 200) showed a strong inverse relationship between %A and %B (rho = -0.999, p<0.001). However, S did not correlate significantly with %A (rho = -0.055, p = 0.443) or %B (rho = 0.085, p = 0.232), see scatterplot in Fig. 5.

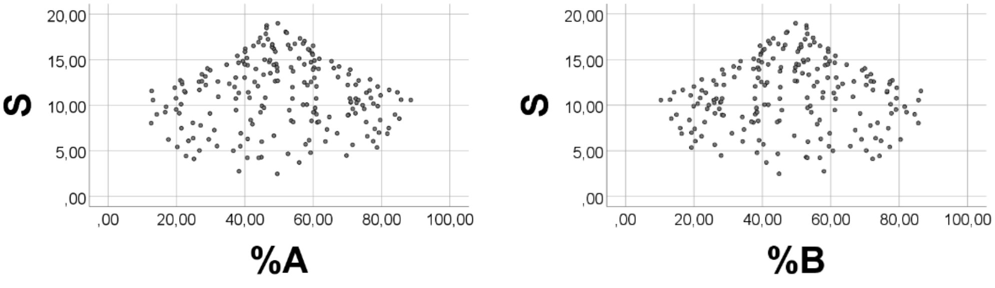

Figure 5: Scatterplot of S vs. %A (left panel) and %B (right panel) when ranges of A and B were both 1 – 10, and R 0.10 – 0.15. The figure relates to the equation %B = -%A + (100 -%R), see text.

This example suggests that, with *similar* or close to similar A (B) ranges, we should expect poor correlations between S and A (B) percentages, since approximately similar S-values are found at the minimum and maximum vales of %A (%B). This suggestion was corroborated in computer experiments (results not shown). Accordingly, under the current conditions, the "relation to sum" approach does *not* seem useful to explain the observed strong negative %A vs. %B correlation (rho =0.999, p<0.001, n=200). However, also in the current example, the negative correlation is well explained by the equation %B = -%A + (100 - %R), which may be roughly approximated to %B = -%A + 100, when %R values are small.

4. TURNING POINT

In the general equation %A + %B + %C =100, i.e. %B = -%A + (100 - %C), increasingly higher (lower) values of %C is expected to promote a positive (negative)%A vs. %B association. Therefore, a *Turning Point* should be found where a positive (negative) correlation between A and B percentages turns to become negative (positive), in response to progressively altering ranges of the variables. Accordingly, close to the Turning Point we should probably not find a significant correlation between A and B percentages. We previously reported this outcome in computer experiments (Høstmark, 2019c).

With %A + %B + %C = 100, the *Turning Point* between e.g. the A and B percentages may also be related to ***skewness*** of the distributions of the A, B, and C percentages (Høstmark, 2019d). Thus, high negative (positive) skewness of the %C histogram is associated with strong positive (negative) %A vs. %B correlation (Fig. 6), see also the examples above. As explained previously, varying skewness of the histograms may be related to differences in *ranges* of the variables (Høstmark, 2019c; 2019d). In Fig. 6, skewness of %C was plotted against rho for %A vs. %B. With uniformly distributed random numbers, giving altogether 49 particular combinations of A, B, C ranges (n = 200 for each of the points), we were able to produce a scatterplot appearing like a mirror image of a sigmoidal curve (Fig. 6). Scatterplot shapes like that shown in Figure 6 could as well be made for %A skewness (abscissa) against the ordinate being rho for %B vs. %C correlation (or for %B skewness as related to rho for %A vs. %C), not shown. This outcome was similar if using random numbers with uniform and normal distributions (Høstmark, 2019d).

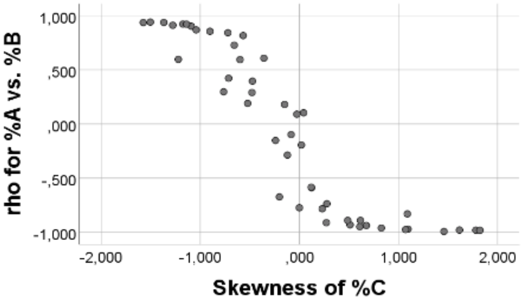

Figure 6: Association between skewness and correlations. With reference to the equation %A + %B + %C = 100, or %B = - %A + (100 - %C), see text, skewness of the %C histogram was plotted against Spearman's rho for the correlation between percentages of the remaining two variables (A and B). The figure was made using uniformly distributed random numbers of A, B, and C. Each of the 49 points was computed based on 200 random number "cases" (i.e. total N = 9800), each "200-set" was computed with particular ranges for A, B, and C. From Høstmark (2019d). Copyright: Høstmark, AT.

5. DISTRIBUTION DEPENDENT CORRELATIONS: A MATHEMATICAL PRINCIPLE CAUSING BIAS, OR A REGULATORY MECHANISM IN PHYSIOLOGY?

The present work strongly suggests that *variability* is crucial for obtaining correlations (positive as well as negative) between percentages of the same sum. True within-subject variability is related to various periods of time, e.g. month, week, or day, and is largely governed by genetics. However, also external factors, such as diet, physical activity, and environment in general could influence variability. All of these factors compose the true biological variability. In addition, the distribution (range) of a biological variable could be strongly influenced by various types of random and systematic errors, e.g. related to sampling, storage, measurements, and to information bias. Conceivably, between-subject variability should be greater than the within-subject one, due to between-subject variability of DNA *per se*, and differences in epigenetic influences, such as DNA methylation and histone modification.

It is beyond the scope of this article to discuss the many types of error in physiological research. However, the many causes of variability do seem to be an argument in favor of considering Distribution Dependent Correlations (DDC) as a mathematical artifact, when it comes to possible physiological interpretations of such correlations. On the other hand, the mathematical principle of DDC could offer an excellent tool to regulate metabolism, raising the question of whether evolution might have utilized this principle. The two examples below apparently seem in favor of this latter idea. *Thus, by determining the within-person variability, i.e. where on the scale the variables are placed, it follows from the DDC rules described in this article, whether relative amounts will be positively or negatively associated, or not correlated at all.* In other words, evolution could govern associations between percentages of the same sum, through regulating within-person distributions of variables. Since the mathematical rules giving *Distribution Dependent Correlations* are general ones, they might apply to any unit systems in nature. The idea that true, within-person variability may be involved as a potent regulatory factor in biology seems to be a novel one. Below, we show in brief two examples from physiology.

6. EXAMPLES FROM PHYSIOLOGY

6.1. EXAMPLE 1: CORRELATIONS BETWEEN FATTY ACID PERCENTAGES, AS OBTAINED IN A DIET TRIAL IN CHICKENS

We recently reported that relative amounts of fatty acids that are precursors of eicosanoids (docosanoids) were positively associated in breast muscle lipids of chickens (Høstmark and Haug, 2018; 2020b-c). In this case, the concentration *distributions* of the various fatty acids were crucial for obtaining the correlations. For example, relative amount of arachidonic acid (AA, 20:4 n6) was shown to correlate positively with percentage eicosapentaenoic acid (EPA, 20:5 n3), and with some other eicosanoid precursor fatty acid percentages (Høstmark and Haug, 2020b). All of these fatty acids were low-number ones, with low variability, relative to sum of the remaining fatty acids. Since AA and EPA derived eicosanoids have opposing cellular effects (Mayes, 2000; Baker, 1990; Gogus and Smith, 2010), it was suggested that the positive association between %AA and %EPA might possibly serve to ensure a proper balance between the metabolic effects of these powerful metabolites. Surprisingly at the time, the positive correlations between eicosanoid precursor percentages could be well reproduced when random numbers were used in lieu of the true values of the fatty acids, provided that the random numbers were sampled with the true ranges (Høstmark and Haug, 2018; 2019a). Additionally, minor changes in the ranges had major effects on the correlation outcomes, suggesting that the correlations were *distribution dependent* ones. Thus, by orchestrating where on the scale the concentrations of various fatty acids are placed, evolutionary mechanisms might achieve that some correlations between relative amounts must be positive whereas others must be negative, due to the described mathematical principles of *Distribution Dependent Correlations.*

In breast muscle lipids of chickens, the equation %AA= -%EPA + (100 - %ALA) was considered (Høstmark and Haug, 2019e), where ALA = α-linolenic acid (18:3 n3). This equation is of the type previously described in this article: %B = -%A + (100 -%R). Two of the fatty acids (AA and EPA) are low-number ones relative to ALA (Høstmark and Haug, 2019e). Using random numbers, the "ALA" range was *hypothetically* increased, from a very narrow range, while keeping the true ranges of AA and EPA. By progressively moving the %ALA distribution towards higher values, the negative %AA vs. %EPA association was attenuated, to eventually pass through a *Turning Point* (Fig. 7); thereafter, a negative %EPA vs. %AA correlation turned to become increasingly positive, in response to further increasing %ALA values (Høstmark and Haug, 2019e). In this experiment, the Turning Point between positive and negative %EPA vs. %AA correlations was attained as the 1st, 2nd, and 3rd quartiles of the %ALA distribution were approximately 28, 30, and 38%, respectively.

From this experiment, it would appear that the *Turning Point* is achieved when ALA range is between 0.1 – 0.4 and 0.1 – 0.3 g/kg; the measured physiological ALA concentration (g/kg) being (mean ± SD) 0.53±0.32 g/kg (Høstmark and Haug, 2019e). These results with *hypothetical* random numbers for ALA raises the question of whether true ALA levels may attain so low levels that the positive association between relative amounts of EPA and AA becomes seriously disturbed, eventually leading to a negative relationship between %EPA and %AA. If so, we *hypothetically* might expect metabolic disturbances related to an imbalance between eicosanoids derived from AA and EPA. We do not know, however, whether such conditions do exist. Anyhow, the above calculations illustrate a potentially strong effect of ALA concentration upon the relationship between relative amounts of EPA and AA.

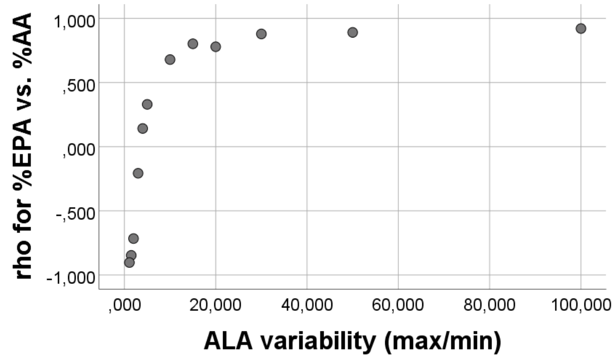

Figure 7: Scatterplot showing the association between ALA variability (expressed as the maximum value divided by the minimum value) and Spearman's rho for the association between %EPA and %AA. The figure relates to the equation % EPA + %AA + %ALA =100, or %AA = - %EPA + (100 - %ALA), see text. For all points on the figure, we generated 163 RANDOM numbers with uniform distribution, sampled within the true (measured) concentration range, i.e. for EPA 0.13 – 0.24 g/kg, and for AA 0.25 – 0.42 g/kg. For ALA we used the following 12 *hypothetical* ranges: 0.1 – 10.0; 0.1 – 5.0; 0.1 – 3; 0.1 – 2.0; 0.1 – 1.5; 0.1 – 1.0, 0.1 – 0.5; 0.1 – 0.4; 0.1 – 0. 3, 0.1 – 0. 2; 0.1- 0.15 1, 5; and 0.1 – 0.11. p<0.001 for all correlation coefficients, except for those close to the Turning Point between positive and negative rho – values. From Høstmark and Haug (2019e). Copyright: Høstmark, A.T.

6.2. EXAMPLE 2. ASSOCIATION BETWEEN RELATIVE AMOUNTS OF WHITE BLOOD CELL (WBC) COUNTS

Searching for possible *negative* associations between percentages of the same sum, in physiology, the counts of WBC seemed a possible candidate. Indeed, counts of segmented neutrophils (N) and lymphocytes (L) are high-number variables relative to sum of the remaining white blood cells (R). The equation %N = -%L + (100 -%R) would approach %N = -%L + 100 with small %R values relative to %N and %L. Hence, we should expect an inverse relationship between %N and %L. Our previous random number analyses (Høstmark and Haug, 2018;2019b) showed that a negative association would prevail until reaching the *Turning Point*. Thus, a negative %N vs. %L association should be expected, using random numbers generated on the basis of reported (Lacher et al., 2012) WBC values. If so, the associations should change, in response to altering ranges, as discussed above. The results shown in Fig. 8 were as anticipated. Thus, relative amounts of random numbers, used to represent the true values of N and L, showed a strong negative correlation in each gender. The correlation outcome was qualitatively the same, irrespective of whether the random numbers were generated based upon the reported (Lacher et al., 2012) *within*-person or the *between*-person values (Spearman's rho being at least -0.9 in each of the conditions, p < 0.001, n = 200).

The N/L ratio has been used as a risk factor, e.g. for atherosclerosis (Meng et al., 2018), and COVID-19 infections (Liu et al., 2020). This ratio is equal to the ratio between N and L percentages, since %N and %L in each subject are computed from the same sum. Thus, the N and L percentages of total WBC allow calculation of the N/L ratio, as well as evaluating whether the relative amounts of N and L correlate.

However, the N/L ratio per se does not provide sufficient information to evaluate whether N and L percentages of total white blood cell count are correlated. However, from the above equations of the regression lines for %N vs. %L (based upon random numbers), it is seen that the slopes are not far from -1, which would be the slope when computing %N and %L from the sum of N and L only, i.e. %N = -%L +100. If so, rho for %N vs. %L should be equal to -1.000, irrespective of the N and L ranges, and the extrapolated regression line should theoretically cross axes at exactly 100%. An improved approximation of the equation would be:

$\%N_{(p-q)} = -(\%N_{max} - \%N_{min})/(\%L_{max} - \%L_{min}) * \%L_{(r-s)} + (100 - \%R_{(t-u)})$

In this equation, R = M + E + B, and subscript parentheses indicate ranges of the percentages.

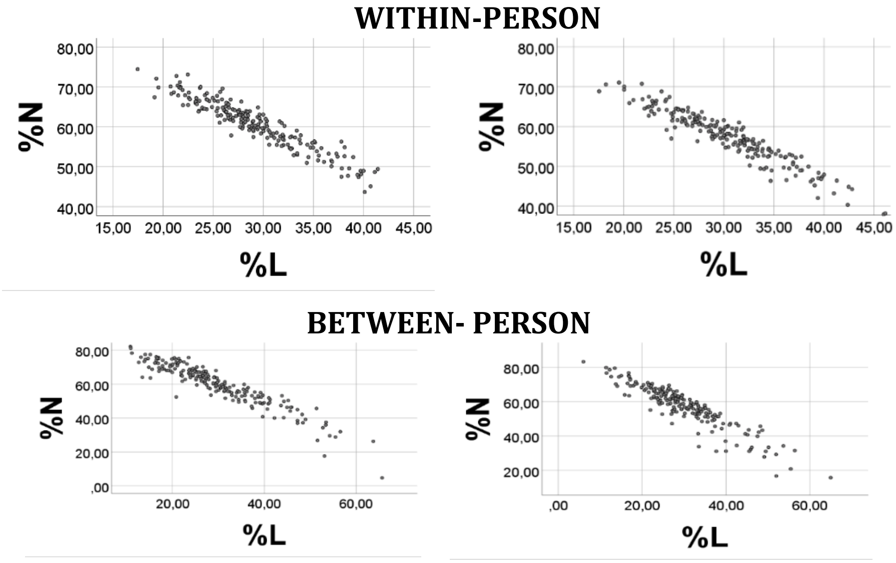

Figure 8: Association between relative amounts of RANDOM numbers in lieu of true values of segmented neutrophil (N) counts, and lymphocyte (L) counts (10^3/μL). Random numbers with normal distribution were generated based on reported mean (SD) values of human blood cells [14]. All of the negative associations were highly significant (Spearman's rho at least equal to - 0.9, p < 0.001, n = 200). Note differences in scale concerning within-person (top) and between-person (bottom) panels, due to differences in variability. Equations of the regression lines were for women, within-person: %N = - 1.16 (0.03)*%L + 94.6 (0.8); for men, within - person: %N = -1.16 (0.03)*%L + 92.6 (0.8); for women, between-person: %N = -1.07 (0.03)*%L + 91.5 (0.9); and for men, between-person: %N = -1.12 (0.04)*%L + 90.7 (1.1). From Høstmark AT (2020d). Copyright: Høstmark, A.T.

Thus, distribution per se (place on the scale/range/variability/skewness) of all types of WBC should influence the negative correlation between N and L percentages. Since M provides about 0.5, E about 0.2, and B about 0.04 (*10^3/µl) of R (Lacher et al., 2012), the order of potency for influencing the %N vs. %L correlation should be: M > E > B, i.e. an apparent small effect of B relative to that of M and E. The current %R values were far from zero, as indicated by quartiles of the computed (random number) %R distribution, being about 10, 12, and 15%, respectively (histogram not shown). Still, the strong negative %N vs. %L association prevailed (Fig. 8). According to the above reasoning, we would anticipate improved (poorer) %N vs. %L association (scatterplots) with decreasing (increasing) values of %R, and this outcome was observed in computer experiments [Høstmark, 2020d). *Obviously, it is not justified to relate our findings to health and disease, since random numbers were used to replace the true values, and effects of large alterations in the ranges were studied.* Nevertheless, the results demonstrate a powerful influence of changing variability upon the negative association between percentages of the same sum, raising the question of whether evolution might utilize the mathematical principles of DDC to obtain an inverse %N vs. %L relationship. Thus, with WBC subgroups directed to particular places on the scale, the relative amounts of N and L must be inversely related, according to the presented mathematical rules.

7. CONCLUSIONS

The results of the present work suggest that *distribution* (range, variability, place on the scale, skewness) of positive scale variables determines whether their relative amounts correlate positively, negatively, or not at all. We accordingly suggest the name *Distribution Dependent Correlations* (DDC). Such correlations may be understood through simple algebraic considerations. The many types of variability could make it hard to detect and appreciate true biological within-person DDC in various studies, and DDC could cause bias. On the other hand, evolution might utilize the mathematical principles of DDC to regulate metabolism, as suggested by the presented examples. Thus, by directing variables to particular places on the scale, evolution might ensure that relative amounts of some variables must become positively associated, whereas percentages of others are negatively correlated. Since DDC rules are general, they should apply to any unit system in nature.

SOURCES OF FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTEREST

The author have declared that no competing interests exist.

ACKNOWLEDGMENT

None.

REFERENCES

[1] Pearson K. Mathematical contributions to the theory of evolution - On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London,1897, 60: 489-496.

[2] Høstmark, A.T., Haug, A. The fatty acid distribution per se explains why percentages of eicosapentaenoic acid (20:5 n3) and arachidonic acid (20:4 n6) are positively associated: a novel regulatory mechanism? Journal of Nutrition and Diet Supplements, 2018, 2(1): 103

[3] Høstmark, A.T, Haug, A. Associations between %AA (20:4 n6) and percentages of EPA (20:5 n3), DPA (22:5 n3), and DHA (22:6 n3) are distribution dependent in breast muscle lipids of chickens. Journal of Nutrition and Diet Supplements, 2019a, 3(1): 103

[4] Høstmark, A.T., Haug, A. The inverse association between relative abundances of oleic acid and arachidonic acid: a case of distribution dependent regulation? Lipids in Health and Disease, 2019b, 18:123

[5] Høstmark, A.T., Haug, A. Relative amounts of eicosanoid and docosanoid precursor fatty acids are positively associated: a distribution dependent regulation. Journal of Nutrition and Food Processing, 2020a, 3; DOI:10.31579/2637-8914/022

[6] Høstmark, A.T., Haug, A. Associations between %AA (20:4 n6) and relative amounts of other body fatty acids. Journal of Nutrition and Food Processing, 2020b, 3(2); DOI:10.31579/2637-8914/024

[7] Høstmark, A.T., Haug A. Distribution dependent and cluster regulation of associations between body fatty acid percentages, as observed in chickens. Journal of Nutrition and Food Processing, 2020c, 3(2); DOI:10.31579/2637-8914/025

[8] Høstmark, A.T. Association between percentages of scale variables, as related to distributions. Journal of Nutrition and Diet Supplements, 2019c, 3(1)104

[9] Høstmark, A.T. . Body fatty acids, nutrition, and health: Is skewness of distributions a mediator of correlations? Journal of Nutrition and Food Processing, 2019d, 2(1); DOI: 10.31579/2637-8914/009

[10] Mayes P.A. “Metabolism of unsaturated fatty acids and eicosanoids”, in Harper’s Biochemistry 25th Ed., R.K. Murray, D.K. Granner, P.A. Mayes, V.W. Rodwell, Eds. New York: McGraw-Hill, 2000, pp. 250-258.

[11] Baker, R.R.. The eicosanoids: a historical overview. Clinical Biochemistry, 1990, 23:455-458.

[12] Gogus, U., Smith, C. n-3 Omega fatty acids: a review of current knowledge. International Journal of Food Science &. Technology, 2010, 45:417–436.

[13] Høstmark, A.T., Haug A. Alpha Linolenic Acid variability influences the positive association between % Eicosapentaenoic acid and % Arachidonic acid in chicken lipids. Journal of Nutrition and Food Processing, 2019e, 2(2); DOI: 10.31579/2637-8914/016

[14] Lacher, D.A., Barletta, J., Hughes, J.P. Biological variation of hematology tests based on the 1999 - 2002 National Health and Nutrition Examination Survey, 2012, National Health Statistics Reports 54 (USA).

[15] Meng, L-B., Yu, Z-M., Guo, P., Wang Q-Q., Qi, R-M., Shan, M-J., Lv, J., Gong, T. Neutrophils and neutrophil-lymphocyte ratio: Inflammatory markers associated with intimal-media thickness of atherosclerosis. Thrombosis Research, 2018, 170: 45-52.

[16] Liu, Y., Du, X., Chen, J., Luo, M., Chen, L., Zhao, Y. Neutrophil-to-lymphocyte ratio as an independent risk factor for mortality in hospitalized patients with COVID-19. Journal of Infection, 2020, 81: e6-e12; DOI: https://doi.org/10.1016/j.jinf.2020.04.002

[17] Høstmark, A.T. Association between relative amounts of white blood cell counts: a case of Distribution Dependent Correlations. Journal of Nutrition and Food Processing, 2020d, 3(2); DOI:10.31579/2637-8914/028.