TYPE Methods
PUBLISHED 30 January 2023
DOI 10.3389/fsysb.2023.1042156

Check for updates

OPEN ACCESS

EDITED BY
Kristin Tøndel,
Norwegian University of Life Sciences,
Norway

REVIEWED BY
Marissa Renardy,
Applied BioMath, United States
Abel Folch Fortuny,
DSM, Netherlands

*CORRESPONDENCE
Edoardo Saccenti,
✉ edoardo.saccenti@wur.nl

SPECIALTY SECTION
This article was submitted to Data and Model Integration,
a section of the journal
Frontiers in Systems Biology

RECEIVED 12 September 2022
ACCEPTED 09 January 2023
PUBLISHED 30 January 2023

CITATION
Saccenti E (2023), What can go wrong when observations are not independently and identically distributed: A cautionary note on calculating correlations on combined data sets from different experiments or conditions.
Front. Syst. Biol. 3:1042156.
doi: 10.3389/fsysb.2023.1042156

What can go wrong when observations are not independently and identically distributed: A cautionary note on calculating correlations on combined data sets from different experiments or conditions

Edoardo Saccenti*

Laboratory of Systems and Synthetic Biology, Wageningen University and Research, Wageningen, Netherlands

In the scientific literature data analysis results are often presented when samples from different experiments or different conditions, technical replicates or times series are merged to increase the sample size before calculating the correlation coefficient. This way of proceeding violates two basic assumptions underlying the use of the correlation coefficient: sampling from one population and independence of the observations (independence of errors). Since correlations are used to measure and infer associations between biological entities, this has tremendous implications on the reliability of scientific results, as the violation of these assumption leads to wrong and biased results. In this technical note, I review some basic properties of the Pearson's correlation coefficient and illustrate some exemplary problems with simulated and experimental data, taking a didactic approach with the use of supporting graphical examples.

KEYWORDS

covariance, data fusion, data merging, pearson's correlation, repeated measures, spearman's correlation, technical replicates, violation of statistical assumptions

1 Introduction

The Pearson's correlation coefficient (Pearson, 1895; Spearman, 1907) is certainly one of the most popular measures of association used in biology and in the Life Sciences. Unfortunately, it is also one of the most misused. Recently, several papers have been brought to my attention by collaborators in which the sample correlation coefficient is calculated following questionable practices, in particular when data from different experiments or conditions are combined before calculating the correlation. The goal of this cautionary note is to show what happens when the basic assumptions underlying the calculation and the use of the sample correlation are not met.

To set the scene, I start by introducing some notation and by recalling some basic statistical principles. Taken n observations $(x_1, x_2, \ldots, x_n)$ of a variable x and n observations $(y_1, y_2, \ldots, y_n)$ of a variable y, the Pearson's sample correlation coefficient $r^{(xy)}$ (to which I will term "sample correlation" for sake of simplicity) is defined as

$$r^{(xy)} = \frac{C^{(xy)}}{\sqrt{V^{(x)}}\sqrt{V^{(y)}}}, \tag{1}$$

where $V^{(x)}$ is the sample variance[1]

$$V^{(x)} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - M_n^{(x)}\right)^2 \qquad V^{(y)} = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - M_n^{(y)}\right)^2, \tag{2}$$

and $C^{(xy)}$ is the sample covariance:

$$C^{(xy)} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - M_n^{(x)}\right)\left(y_i - M_n^{(y)}\right) \tag{3}$$

with $M_n^{(x)}$ and $M_n^{(y)}$ the sample mean for x and y:

$$M_n^{(x)} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad M_n^{(y)} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{4}$$

Let's now recall now some facts about the sample correlation $r^{(xy)}$ (Eq. 1): it can always be calculated, however its validity and (correct) interpretation (including its statistical significance, as expressed by the associated p-value) rest on several statistical assumptions. I will focus here on the two assumptions stating that the observations used to calculate the correlation must be identically and independently distributed; these formulated as.

A1 All the $(x_1, x_2, \ldots, x_n)$ observations of x (and y) are sampled from the same (normal) distribution.

A2 The $(x_1, x_2, \ldots, x_n)$ observations of x (and y) are independent (independence of errors).

To be of any significance, and to be able to make inference on population parameters, the observations must be randomly selected, that is must be a representative random sample of a larger population. In addition, the relationships between x and y must be linear, since (Eq. 1) cannot account for non-linear relationships. Chapter 32 of Motulsky. (2014) offers a low level yet very precise presentation of all the assumptions underlying the use of the correlation coefficient.

This note deals with the problems arising when assumptions A1 and A2 are violated, that is when observations are non-independently and non-identically distributed. Consequences of the violations of other assumptions, like deviation from normality of the observations, have been discussed elsewhere (Calkins, 1974; Havlicek and Peterson, 1976; Havlicek and Peterson, 1977; Wilcox, 2009). The papers by Schober et al. (2018) and Janse et al. (2021) discuss pitfalls and interpretative problems of correlations.

In what follows, all observations are well-behaved and follow a normal (Gaussian) distribution:

$$x_i \sim N\left(\mu, \nu^2\right) \tag{5}$$

where μ and ν are the population mean and standard deviation (same for variable y).

2 Three research scenarios when correlations are wrongly calculated

I will consider three research scenarios often found in literature in which data is manipulated in some way before the correlation among measured variables is calculated.

RS1 A researcher has two data sets A and B which contain measurements of the variables x and y. Data set A contains n observations of x and y, while data set B contains m observations. A and B are combined into one data set containing $n + m$ observations and the Pearson's sample correlation (Eq. 1) is calculated between x and y. The reasons for merging the two data sets can be different. Data sets A and B can come from different batches measured during an experiment, or even from different experiments. Often the data sets are merged to increase the sample size with the idea (wrong, in this case) of obtaining a more a reliable estimation of the correlation between x and y.

RS2 A researcher has performed an experiment where a large number of variables have been measured over two conditions, on n observations for Condition A and m observation for Condition B, like, for example, in the case of a transcription experiment where thousands of gene expressions have been measured on case and control samples. To reduce the dimensionality of the problem, the researcher restricts the analysis to those genes that are differentially expressed between the two conditions. Moved by the interest of understanding regulatory mechanisms, the researcher decides to build a correlation network using all the measurements (observations) available for those genes (let's call two of such genes x and y) that are differentially expressed.

RS3 A researcher has measured variable x and y several times on the same subjects, obtaining m measures for each subject. A typical case is when technical replicates are measured for each (or some of the) sample, usually in duo or triplicates, or when time series are acquired. To increase the total sample size, they then decide to combine all the $n \times m$ observations and calculate the correlation over the $n \times m$ observation of x and y.

The first two scenarios RS1 and RS2, albeit different, can be schematized in the same way, as shown in Figure 1.

2.1 Violation of sampling from one population

Scenarios RS1 and RS2 entail the calculation of sample correlations after merging of two (or more) data sets: I wish to illustrate the problems that arise from such a way of operating with a simple simulated example. Figure 2A shows the correlation plot of $n + m = 200$ observations of variables x and y. A positive linear relationship seems to exist between x and y: the sample correlation coefficient is $r^{xy}_{(n+m)} = 0.76$ ($P\text{-}val < 10^{-5}$). Equipped with this rather strong correlation and statistical significance, the researcher may claim association between the two variables and build a story around it, explaining and discussing the biological relevance of it.

The problem becomes evident when we look at how the data presented in Figure 2B has been built. I have proceeded as described in research scenarios RS1 and RS2, merging two data sets

[1] The notation V $^{(x)}$ with the use of superscript (x) may seem, at fist, unpractical. Its utility will become evident in the remaining of the paper when more and different quantities are introduced.

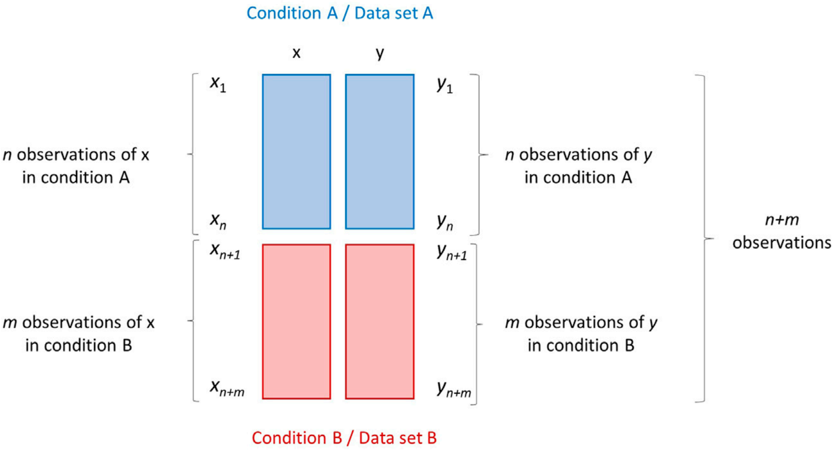
Condition A / Data set A
x
y
x_1
y_1
n observations of x in condition A
n observations of y in condition A
x_n
y_n
x_{n+1}
y_{n+1}
n+m observations
m observations of x in condition B
m observations of y in condition B
x_{n+m}
y_{n+m}
Condition B / Data set B

FIGURE 1
Research scenarios RS1 and RS2. Graphical illustration of two data sets A and B (depicted as blocks of different color) containing measurements/ observations of two variables *x* and *y* measured on two different conditions. Data set A contains *n* observations, while data set B contains *m* observations. The data set obtained by merging A and B row-wise (on top of the other) has dimension ($n + m$) × 2.

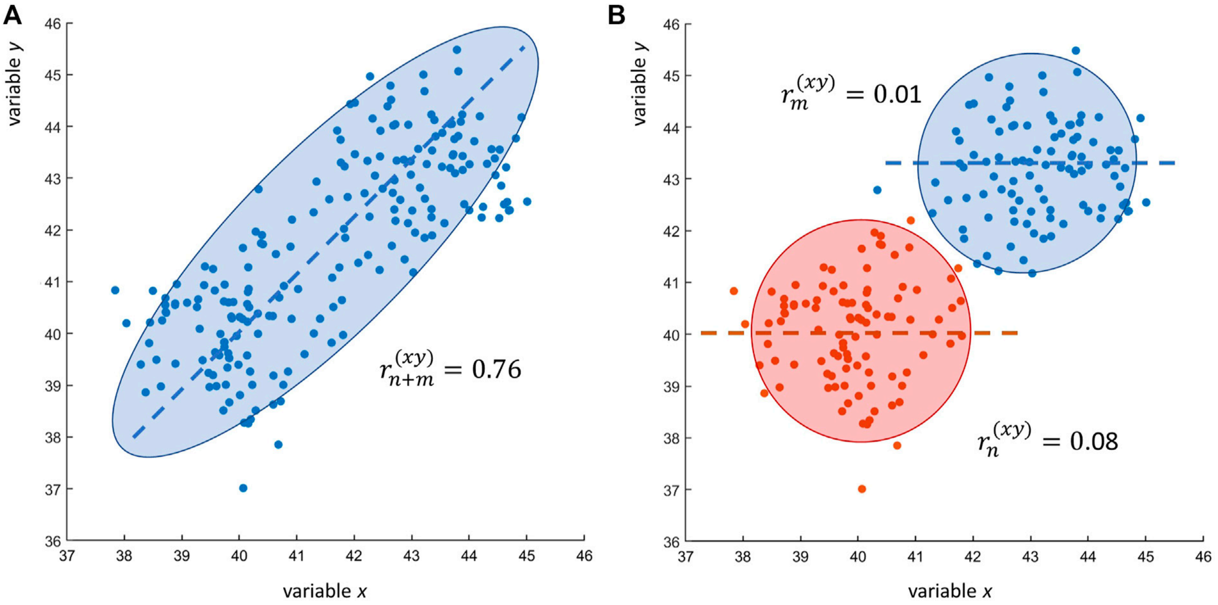

FIGURE 2
Research scenarios RS1 and RS2. **(A)** Scatter plot of $n + m = 200$ observations of two variables *x* and *y*: the Pearson's sample correlation between *x* and *y* is $r_{n+m}^{(xy)} = 0.76$. **(B)** The same scatter plot as in Panel A but with data points color coded to highlight the actual data structure: when taken separately the $n = m = 100$ observations of *x* and *y* (Condition A: blue; condition B: red) are uncorrelated: $r_n^{(xy)} = 0.01$ (data set A) and $r_n^{(xy)} = 0.08$ (data set B). The observed high correlation $r_{n+m}^{(xy)} = 0.76$ is an artifact due to the merging of two data set containing variables coming from two different populations: in this case *x* and *y* come from two independent normal distributions with population means $\mu_A = (43, 43)$ and $\mu_B = (40, 40)$ and unit variance $\nu^2 = 1$.

containing $n = 100$ and $m = 100$ observation of *x* and *y* and then I have calculated the correlation between *x* and *y*. The reality is that *x* and *y* are not correlated at all: when the two data sets (conditions) are considered separately, the correlation between *x* and *y* is zero, since the generating mechanism of the data shown in Figure 2A is the following:

$$\begin{aligned}(x_{1<i<n}, y_{1<i<n}) &\sim N(40, 1)\\ (x_{n+1<i<n+m}, y_{n+1<i<n+m}) &\sim N(43, 1),\end{aligned} \tag{6}$$

which generates variables *x* and *y* that are independent and uncorrelated. The second scenario RS2 is very often encountered in papers dealing with the analysis of very large omics data sets. This way of proceeding is also problematic. In fact, when the researcher looks for differentially expressed genes, (or for metabolites with different concentrations), they perform some statistical test to compare the observed means of variable *x* and *y* in condition A *versus* condition B (in this case a *t*-test, for instance), testing the Null hypothesis (similar considerations hold for variable *y*):

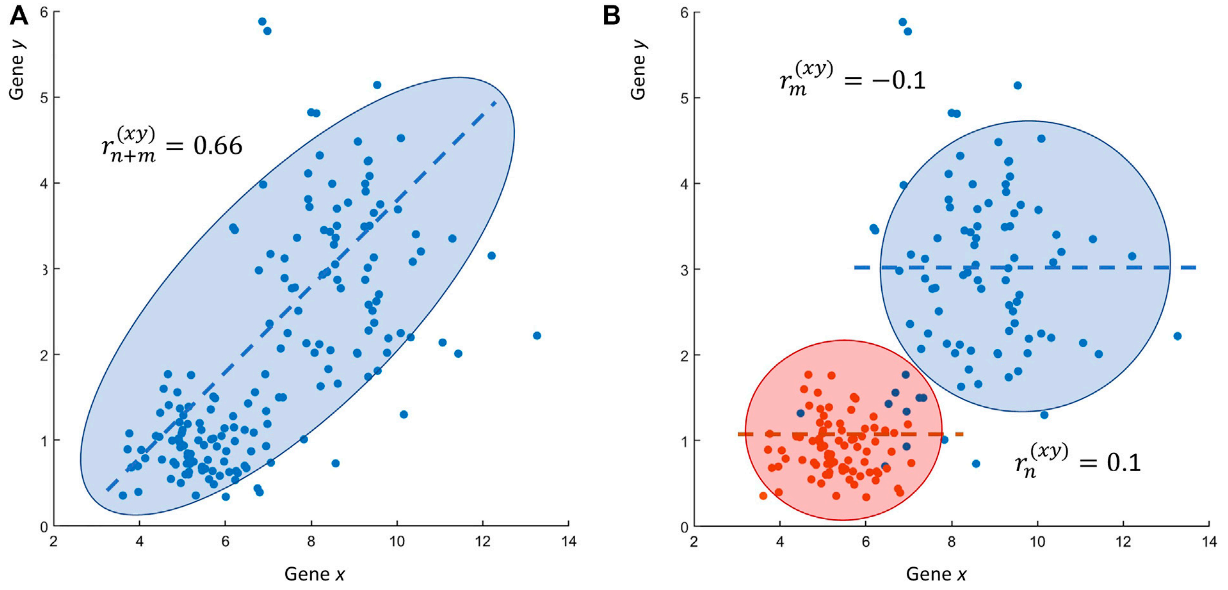

FIGURE 3
Real life research scenario RS2. **(A)** Scatter plot of the expression values of two genes *x* and *y*: the Pearson's sample correlation between gene *x* and *y* is $r_{n+m}^{(xy)} = 0.66$. **(B)** The same scatter plot as in Panel A but with gene expression color coded to highlight the actual data structure: when taken separately the expression of genes *x* and *y* (Condition A: blue; condition B: red) are uncorrelated: $r_m^{(xy)} = -0.1$ and $r_n^{(xy)} = 0.1$, respectively. The observed correlation $r_{n+m}^{(xy)} = 0.66$ is an artifact due to the merging of the expression profile of two genes that have been found to be differentially expressed between condition A and B (case and controls) (P-val $< 10^{-5}$) violating, in this way, the assumption of sampling from the same distribution. Data from (Li et al., 2014). Note that there is nothing wrong with analysis of the data in the original publication: here the data has been only used to build an illustrative example.

$$\mathrm{H}_0\colon \mu_A^{(x)} = \mu_B^{(x)}, \tag{7}$$

against the alternative

$$\mathrm{H}_1\colon \mu_A^{(x)} \neq \mu_B^{(x)}. \tag{8}$$

There is of course no problem using the *t*-test to find genes that are differentially expressed (even if more powerful approaches have been introduced for this type of data). The problem arises when the differentially expressed genes are used to compute correlations. Selecting the variables for which H_0 is rejected is the equivalent of selecting variables for which the distribution of *x* is different between two conditions. Stated in other words, by doing so the researcher is looking specifically for those variable that violates the assumption of sampling from one distribution!

If the reader thinks that these are just simulated numerical examples, it is not complicated to show that such problematic situations can be easily encountered when using real-life experimental data. Figure 3 shows a case similar to the one given in Figure 2, this time obtained using data from a transcriptomic study: the expression profiles of two genes *x* and *y*, measured at two different conditions in a case-control scenario, are uncorrelated ($r_n^{(xy)} = 0.1$, $r_m^{(xy)} = -0.1$) when the two conditions are considered separately Figure 3A. However, they become correlated ($r_{n+m}^{(xy)}$, $P\text{-}val < 10^{-5}$) (Figure 3B) if the correlation is taken over all the observations combined, i.e., when the two data sets are combined.

2.2 A closer look to the mathematics of the problem

This section presents a mathematical explanation of what observed in Figures 2, 3. Let's define the overall mean $M_{m+n}^{(x)}$ over all the $(x_1, x_2, \ldots x_{n+m})$ observations of *x* (identical formulas hold for *y*)

$$M_{n+m}^{(x)} = \frac{1}{n+m}\sum_{i=1}^{n+m} x_i, \tag{9}$$

and the partial means $M_n^{(x)}$ and $M_m^{(x)}$ over the $(x_1, x_2, \ldots x_n)$ and $(x_{n+1}, x_{n+2}, \ldots x_{n+m})$, as

$$M_n^{(x)} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{10}$$

$$M_m^{(x)} = \frac{1}{m}\sum_{i=n+1}^{n+m} x_i, \tag{11}$$

with similar definition for variable *y*. The overall variance $S_{n+m}^{(x)}$ of x_{n+m} is given by (Chan et al., 1982)

$$V_{n+m}^{(x)} = \frac{1}{n+m-1}\sum_{i=1}^{n+m}\left(x_i - M_{m+n}^{(x)}\right)^2, \tag{12}$$

and the partial variances $S_n^{(x)}$ and $S_m^{(x)}$ by

$$\begin{aligned} V_n^{(x)} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - M_n^{(x)}\right)^2 \\ V_m^{(x)} &= \frac{1}{m-1}\sum_{i=n+1}^{n+m}\left(x_i - M_m^{(x)}\right)^2. \end{aligned} \tag{13}$$

The total variance $S_{n+m}^{(x)}$ taken over all the observations of *x* can be expressed as function of the partial variances (Eq. 13) of the subsets

$$V_{n+m}^{(x)} = \frac{1}{n+m-1}\left[(n-1)V_n^{(x)} + (m-1)V_m^{(x)} + \frac{nm}{n+m}\left(M_n^{(x)} - M_m^{(x)}\right)^2\right]. \tag{14}$$

The covariance $C_{m+n}^{(xy)}$ between the $n + m$ observations of x and y can be expressed in term of the partial covariance $C_n^{(xy)}$ and $C_m^{(xy)}$ between the n (and m) observations of x and y:

$$C_{n+m}^{(xy)} = \frac{1}{n+m-1}\Bigg[(n-1)C_n^{(xy)} + (m-1)C_m^{(xy)} + \frac{nm}{n+m}\left(M_n^{(x)} - M_m^{(x)}\right)\left(M_n^{(y)} - M_m^{(y)}\right)\Bigg]. \tag{15}$$

By combining (Eq. 15) with (Eqs 13, 14) it is possible to obtain the sample correlation $r_{n+m}^{(xy)}$ between $n + m$ observations x and y as a function of the correlation $r_n^{(x)}$ and $r_m^{(x)}$ calculated on the two subsets (Hayes, 2012):

$$\begin{aligned} r_{n+m}^{(xy)} &= \frac{C_{n+m}^{(xy)}}{\sqrt{V_{n+m}^{(x)}}\sqrt{V_{n+m}^{(y)}}} \\ &= \frac{1}{n+m-1}\frac{(n-1)C_n^{(xy)} + (m-1)C_m^{(xy)} + \frac{nm}{n+m}\left(M_n^{(x)} - M_m^{(x)}\right)\left(M_n^{(y)} - M_m^{(y)}\right)}{\sqrt{V_{n+m}^{(x)}}\sqrt{V_{n+m}^{(y)}}} \\ &= \frac{n-1}{n+m-1}r_n^{(xy)} + \frac{m-1}{n+m-1}r_m^{(xy)} \\ &+ \frac{1}{n+m-1}\frac{nm}{n+m}\frac{\left(M_n^{(x)} - M_m^{(x)}\right)\left(M_n^{(y)} - M_m^{(y)}\right)}{\sqrt{V_{n+m}^{(x)}}\sqrt{V_{n+m}^{(y)}}}. \end{aligned} \tag{16}$$

Eq. 16 shows that the correlation between x and y taken over the full data set is a weighted sum of the correlations $r_n^{(xy)}$ and $r_m^{(xy)}$ taken over the two subsets plus and additional term

$$\Delta r_{n+m}^{(xy)} = \frac{1}{n+m-1}\frac{nm}{n+m}\frac{\left(M_n^{(x)} - M_m^{(x)}\right)\left(M_n^{(y)} - M_m^{(y)}\right)}{\sqrt{V_{n+m}^{(x)}}\sqrt{V_{n+m}^{(y)}}}. \tag{17}$$

The term $\Delta r_{n+m}^{(xy)}$ does not depends on the correlation between x and y, but only on the difference between the mean value of x and y in the two sub sets. As a consequence, even if $r_n^{(xy)}$ and $r_m^{(xy)}$ are zero, the correlation taken over the merged data set is different from zero if x and y have not the same average $M_n^{(x)}$ and $M_m^{(x)}$ and $M_n^{(y)}$ and $M_m^{(y)}$ in the two sub sets. This is exactly what happens in the examples shown in Figures 2, 3. Working out the calculations for data in Figure 2 we have:

$$n = m = 100 \tag{18}$$

$$\begin{aligned} &M_n^{(x)} = 43.2 \quad M_n^{(y)} = 43.2 \quad M_m^{(x)} = 39.9 \quad M_m^{(y)} = 40.1 \\ &V_n^{(x)} = 0.92 \quad S_n^{(y)} = 0.86 \quad V_m^{(x)} = 0.77 \quad V_m^{(y)} = 0.96 \\ &M_{n+m}^{(x)} = 41.5 \quad M_{n+m}^{(x)} = 41.6 \\ &V_{n+m}^{(x)} = 3.5 \quad S_{n+m}^{(x)} = 3.3 \\ &r_n^{(xy)} = 0.08 \quad r_m^{(xy)} = 0.01 \\ &r_{n+m}^{(xy)} = 0.76. \end{aligned} \tag{19}$$

These numerical results are consistent with those we obtained for the synthetic data sets, where uncorrelated observations of x and y were generated according to:

$$\begin{aligned} (x_1, x_2, \dots x_n) &\sim \mathcal{N}(40, 1) & (y_1, y_2, \dots y_n) &\sim \mathcal{N}(40, 1) \\ (x_{n+1}, x_{n+2}, \dots x_{n+m}) &\sim \mathcal{N}(43, 1) & (y_{n+1}, y_{n+2}, \dots y_{n+m}) &\sim \mathcal{N}(43, 1), \end{aligned} \tag{20}$$

with $n = m = 100$. Eq. 16 shows that the contrary is also possible: two variables can be correlated in two different data sets but uncorrelated when the correlation is taken over the two merged data sets: this is shown in Figure 4.

Eq. 16 also explains why the Pearson's sample correlation is so sensitive to outliers, to the point that one single outlier is sufficient to pull a zero correlation to 1. Having one outlier is the equivalent of having one additional data set (or condition/class/group) with just m = 1 observations. As a result, Eq. 16 simplifies to

$$r_{n+1}^{(xy)} = \frac{n-1}{n}r_n^{(xy)} + \frac{1}{n+1}\frac{\left(M_n^{(x)} - x_{n+1}\right)\left(M_n^{(y)} - y_{n+1}\right)}{\sqrt{V_{n+1}^{(x)}}\sqrt{V_{n+1}^{(y)}}}. \tag{21}$$

If the outlying observation (x_{n+1}, y_{n+1}) is very distant from the average of the other n observations, the resulting correlation can be severely inflated, as shown in Figure 5A. The reader could argue that the use of Spearman's rank correlation (Spearman, 1904) would have avoided this problem, since Spearman's correlation is less sensitive to outliers. This is certainly true, as shown in Figure 5B: the presence of an outlier, even if really far from the average of all other n, not outlying, observations, does not affect the correlation. However, this argument holds true only in case of a few outliers: if the number m of outliers increases, also the Spearman's correlation increases, albeit less dramatically than in the case of Pearson's correlation, but still to a significant extent, leading to claim the existence of a correlation between x and y when x and y are not correlated at all. This is shown in Figures 6A, B. The agreement between Pearson's and Spearman's indexes increases when the number of outliers increase since the Spearman's index is not robust against a large number of outliers. As shown in the plot, when there are one to two outliers, the Pearson's coefficient gives a large correlation while the Spearman's not; and as the number of outliers increase, both agree on a large correlation.

2.3 Violation of independence of observations

The third scenario RS3 pertains the violation of the assumption that the observations (i.e., the samples on which variables x and y are measured), are independent, i.e., independence of errors. Working with repeated measures is the most striking case of non-independent observations, since the same subject is measured more times. This scenario is graphically illustrated in Figure 7.

What are the consequences of the violation of the independence of the samples? This depends on the type of dependence present between the observations. Take for instance the case when technical replicates are used (wrongly!) to increase the sample size. In this case more observations (measurements) are available of x and y on the same sample(s), with the model (a similar equation holds for y)

$$x_{i,j} = \mu_x + \epsilon_j, \tag{22}$$

where $x_{i,j}$ is the jth replicate of observation ith of x and ϵ_j is the replication error. An example of $n = 25$ observations of x and y, each with three replicates, is shown in Figure 8A: if only one observation per subject is taken ($n = 25$), the correlation between x and y is $r^{(xy)} = 0.91$, while if all $n + m = 25 + 2 \times 25 = 25 \times 3 = 75$ observations are taken, the correlation is $r^{(xy)} = 0.76$. In general, using replicates considering them as independent observation will lower the value of the correlation coefficient. Repeated measures must be handled carefully with special approaches: there is ample literature on this topic, see for instance (Bakdash and Marusich, 2017) and reference therein.

Dependence of observations can also arise because of reasons that are out of the control of the experimenter, like in presence of correlated measurement noise, where data can be modeled as

$$x_i = \mu_x + \epsilon_{x_i} + \phi_x \tag{23}$$

$$y_i = \mu_y + \epsilon_{y_i} + \phi_y, \tag{24}$$

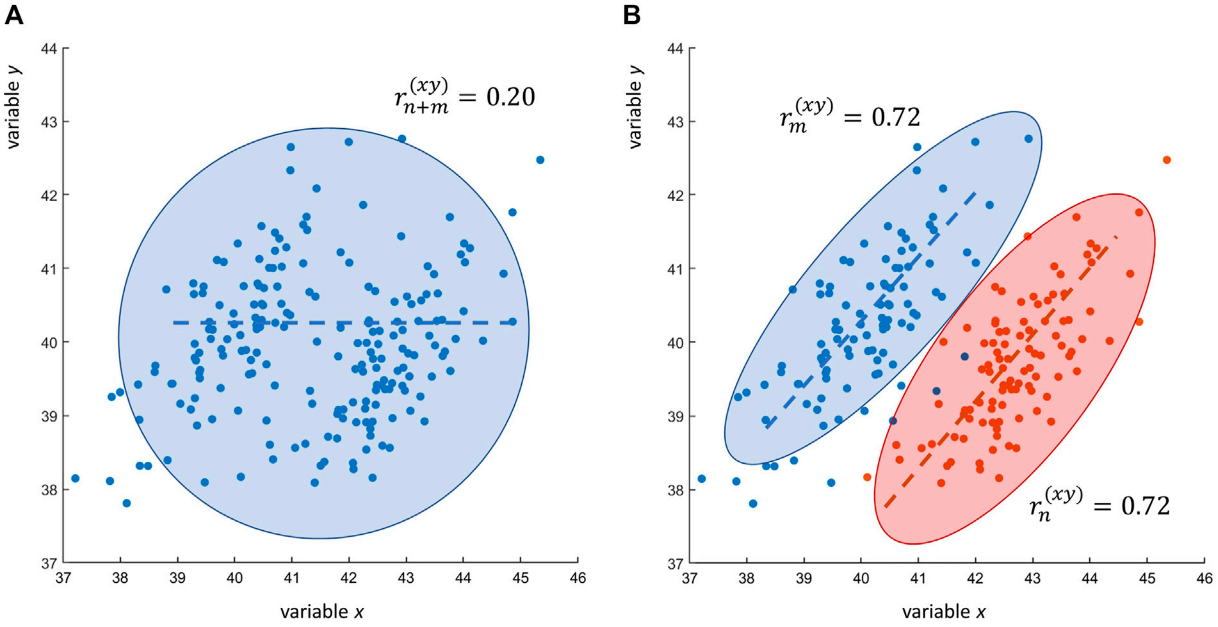

FIGURE 4
Research scenarios RS1 and RS2. **(A)** Scatter plot of $n + m = 200$ observations of two variables x and y: the Pearson's sample correlation between x and y is $r_{n+m}^{(xy)} = 0.20$. **(B)** The same scatter plot as in Panel A but with data points color coded to highlight the actual data structure: when taken separately the $n = m = 100$ observations of x and y (data set A, blue; data set B: red) are correlated: $r_{n}^{(xy)} = 0.72$ (data set A) and $r_{n}^{(xy)} = 0.72$ (data set B). The low observed correlation $r_{n+m}^{(xy)} = 0.20$ is an artifact due to the merging of two data set containing variables coming from different populations: in this case x and y come from two independent normal distribution with population means $\mu_A = (43, 40)$ and $\mu_B = (40, 40)$ and unit variance.

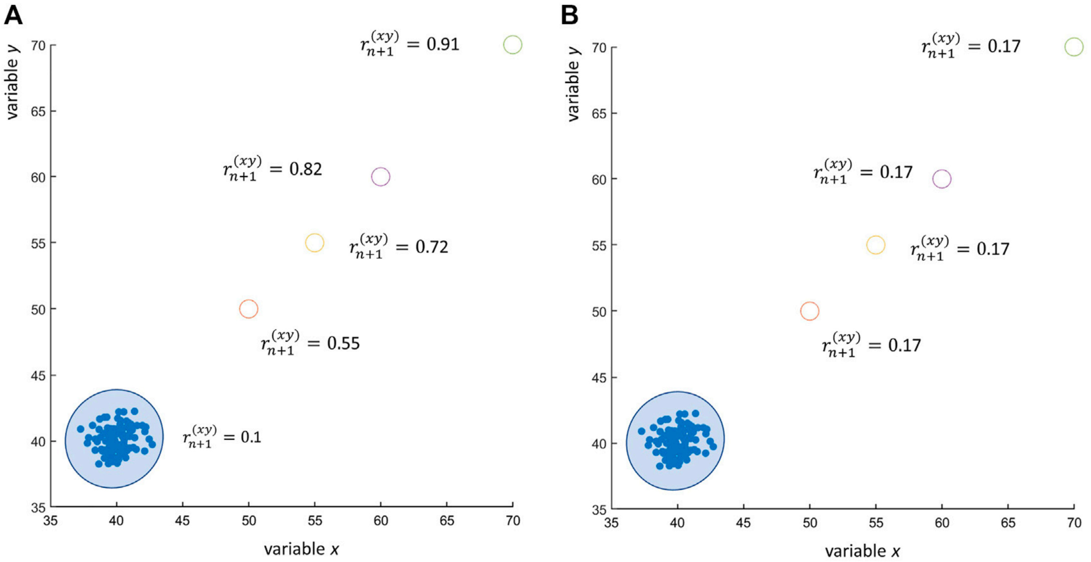

FIGURE 5
Effect of outliers on the correlation coefficient: 100 observations of two uncorrelated variables x and $y \sim N(40, 1)$ (observed Pearson's correlation $r_{n}^{(xy)} = 0.1$) to which an outlier observation x_{n+1}, y_{n+1}, with increasing distance from the sample mean (open dots ○), is added. **(A)** The Pearson's correlation $r_{n+1}^{(xy)}$ calculated on the $n + 1$ observations (including the outlier x_{n+1}, y_{n+1}) increases with the distance of the outlier x_{n+1}, y_{n+1} from the sample mean, consistently with Eq. 21. **(B)** The Spearman's correlation coefficient is not sensitive to the distance of the outlier from the sample mean. However, this is not true if $m \geq 2$ outliers are present, as shown in Figure 6.

where ϕ_x and ϕ_y are correlated error terms normally distributed with zero mean and given error variance-covariance. The presence of correlated error can induce correlation between two variables that are originally uncorrelated, as shown in Figure 8B. The effect of correlated and uncorrelated error on the Pearson correlation coefficient is discussed in (Saccenti et al., 2020).

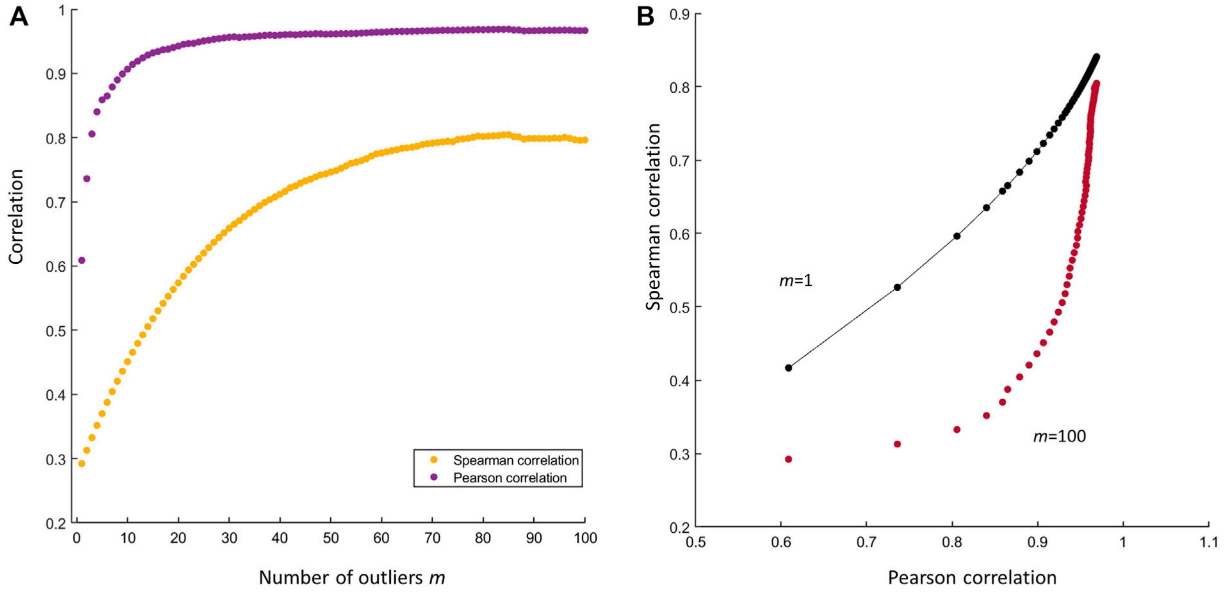

FIGURE 6
(A) Effect of the number of outliers on the Pearson's and Spearman's correlation. Both indexes increase with the number *m* of outliers **(B)** Scatter plot of the Spearman's index against the Pearson's in presence of *m* = 1 outlier (black dots) and in presence of *m* = 100 outliers (purple dots). The difference between the two indexes is larger when *m* = 1 since the Spearman's correlation is robust to the presence of that outlier. If the number of the outlier is large (*m* = 100) the difference between the two indexes becomes increasingly smaller since the Spearman index is also affected by the outliers: both indexes record an inflated correlation.

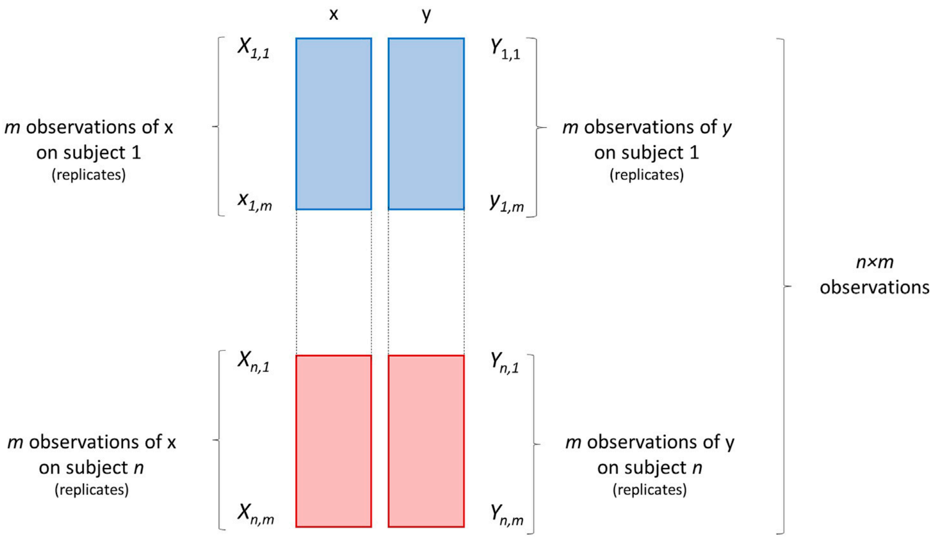

FIGURE 7
Graphical illustration of a data set containing *m* repeated (or replicated) measurements/observations of two variables *x* and *y* on *n* subjects. Each block of repeated measurements is depicted as a block of different color.

Another commonly seen error is the estimation of the correlation using time series data to increase the sample size, which is another violation of the assumption of independent observations. This type of data also needs to be handled carefully, since it is very easy to obtain misleading results: a classic reference on this topic is (Yule, 1926).

3 Plot the data rather than blindly trusting correlation values

The examples discussed in the previous section should (hopefully) suggest that plotting the data is a critical step for the analysis, understanding and interpretation of correlations: visual exploration

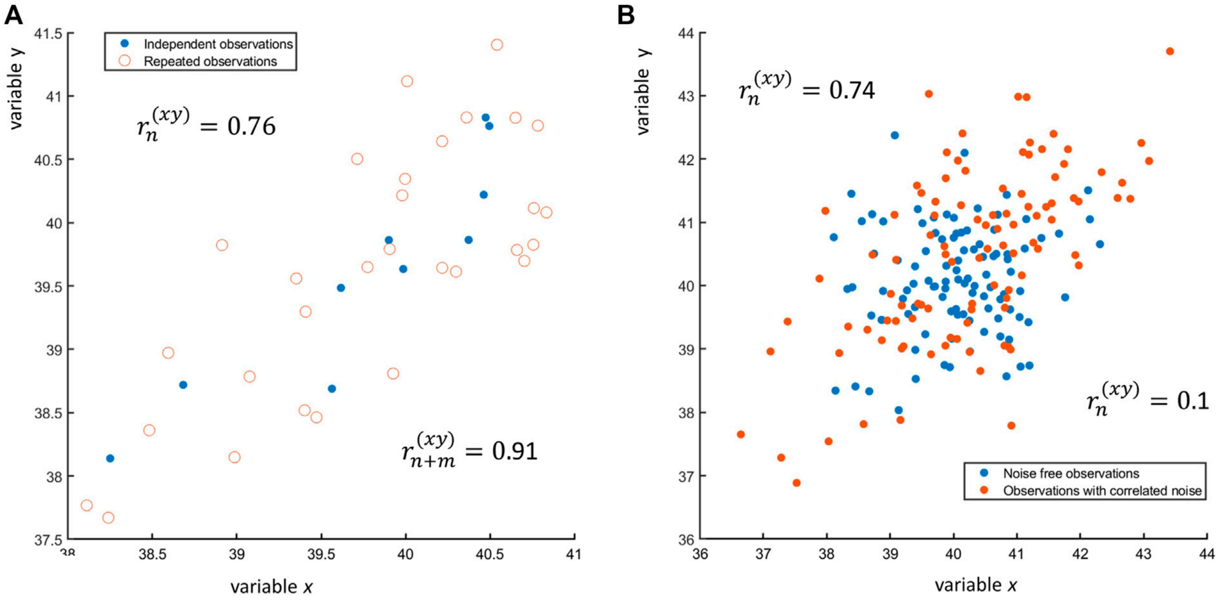

FIGURE 8
(A) Scatter plot of n = 10 observations of two variables x and y (blue dots ●) and of three replicates for each observation (red open circles ○). The sample correlation between the n = 10 observations is $r_n^{(xy)}$ = 0.91. **(B)** if the replicates are considered using a total of n + m = 30 observations, the sample correlation reduces to $r_{n+m}^{(xy)}$ = 0.76.

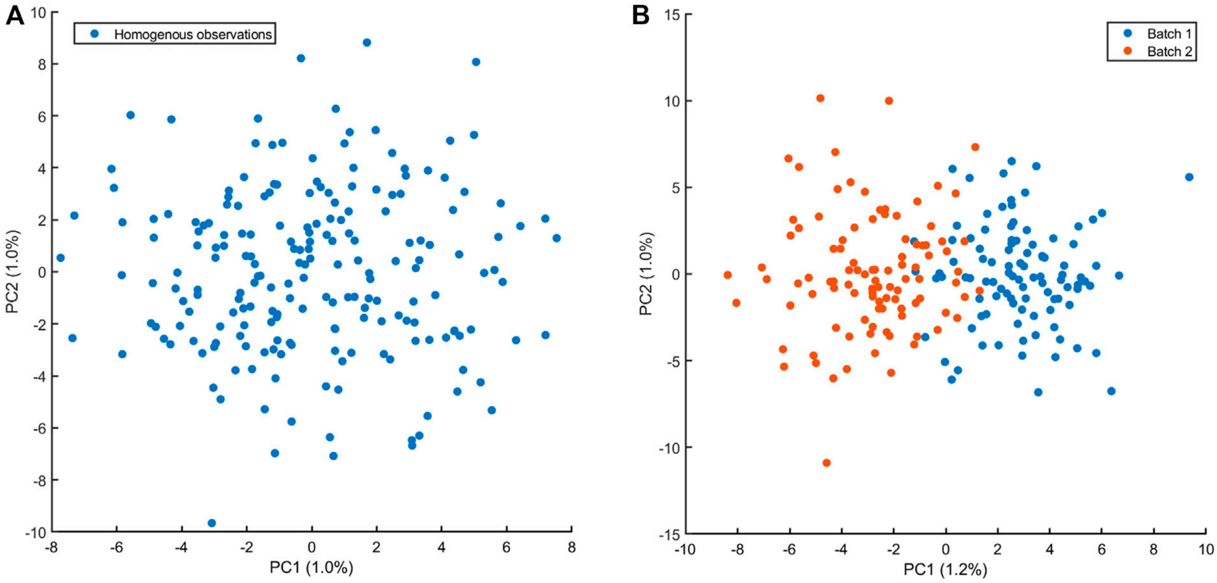

FIGURE 9
Principal component analysis score plot of n = 200 observations of p = 1,000 variables. **(A)** Observations are sampled from the same normal multivariate distributions: variables are uncorrelated with unit variance and mean 40. There is no structure in the data and no outliers are visible. **(B)** Observations sampled in presence of batch effect affecting 500 variables which are sampled from a normal distribution with mean 40.2. The batch effect is clearly visible in the PCA score plot, suggesting a violation the assumption from one distribution. Caution must be taken when correlations are calculated!

of scatter plots like those shown in Figures 2–4, can easily reveal the presence of outliers and data structures that can point to violation of the assumption of sampling from one population or independence of observations (although the latter can be tricky to spot). In the case of multivariate data when $p \gg 2$ variables are measured, plotting and visual exploration of all possible correlation plots is usually not a feasible approach, since the number of plots increases as $\frac{1}{2}p(p-1)$. In this case, Principal Component Analysis (PCA) (Pearson, 1901; Hotelling, 1933; Jolliffe, 2002) is an extremely valuable tool since it can be used to reduce the dimensionality of high-dimensional data and can highlight the presence of outliers and of (unwanted) data structure that hampers the calculation of the correlation coefficient. An example is given in Figure 9, where the PCA plot of a simulated data set without and with data structure (in this case a batch effect) is shown.

4 Conclusion

In this technical note, I have shown some of the consequences of neglecting the assumptions of sampling from one population and independence of observations when calculating the Pearson's correlation coefficient. I illustrated cases of the violation of these assumptions that originate when data sets coming from different experiments or pertaining different experimental conditions or in presence of batch effects are merged before the calculation of the correlation coefficient. It is shown that this way of proceeding will result in inflation or deflation of correlations: inflation or deflation of correlations. In both cases wrong inference will be made, leading to believe that a correlation exists when it does not, or that a correlation does not exist when it actually does. Similar problems arise when correlations are taken over repeated measures or time series.

The hope is that the reader, after having read and meditated the examples, will be able to recognize those situations where the calculation of the sample correlation is not allowed because of the violations of fundamental statistical assumptions.

Data availability statement

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Author contributions

ES is the sole author, having conceived, written, and edited this manuscript.

Funding

This study has received funding from Netherlands Organisation for Health Research and Development (ZonMW) through the PERMIT project (Personalized Medicine in Infections: from Systems Biomedicine and Immunometabolism to Precision Diagnosis and Stratification Permitting Individualized Therapies, project number 456008002) under the PerMed Joint Transnational call JTC 2018 (Research projects on personalised medicine—smart combination of pre-clinical and clinical research with data and ICT solutions).

Acknowledgments

Robert van Vorstenbosch is acknowledged for fruitful comments on the manuscript.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bakdash, J. Z., and Marusich, L. R. (2017). Repeated measures correlation. *Front. Psychol.* 8, 456. doi:10.3389/fpsyg.2017.00456

Calkins, D. S. (1974). Some effects of non-normal distribution shape on the magnitude of the pearson product moment correlation coefficient. *Rev. Interam. Psicol.* 8, 261–288.

Chan, T. F., Golub, G. H., and LeVeque, R. J. (1982). "Updating formulae and a pairwise algorithm for computing sample variances," in *COMPSTAT 1982 5th symposium held at toulouse 1982* (Berlin: Springer), 30–41.

Havlicek, L. L., and Peterson, N. L. (1977). Effect of the violation of assumptions upon significance levels of the pearson r. *Psychol. Bull.* 84, 373–377. doi:10.1037/0033-2909.84.2.373

Havlicek, L. L., and Peterson, N. L. (1976). Robustness of the pearson correlation against violations of assumptions. *Percept. Mot. Ski.* 43, 1319–1334. doi:10.2466/pms.1976.43.3f.1319

Hayes, K. (2012). Updating formulae for the sample covariance and correlation. *Teach. Statistics Int. J. Teach.* 34, 65–67. doi:10.1111/j.1467-9639.2011.00491.x

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 24, 417–441. doi:10.1037/h0071325

Janse, R. J., Hoekstra, T., Jager, K. J., Zoccali, C., Tripepi, G., Dekker, F. W., et al. (2021). Conducting correlation analysis: Important limitations and pitfalls. *Clin. Kidney J.* 14, 2332–2337. doi:10.1093/ckj/sfab085

Jolliffe, I. T. (2002). *Principal component analysis*. Berlin: Springer.

Li, B., Tsoi, L. C., Swindell, W. R., Gudjonsson, J. E., Tejasvi, T., Johnston, A., et al. (2014). Transcriptome analysis of psoriasis in a large case–control sample: Rna-seq provides insights into disease mechanisms. *J. Investigative Dermatology* 134, 1828–1838. doi:10.1038/jid.2014.28

Motulsky, H. (2014). *Intuitive biostatistics: A nonmathematical guide to statistical thinking*. USA: Oxford University Press.

Pearson, K. (1901). LIII. *On lines and planes of closest fit to systems of points in space. Lond. Edinb. Dublin philosophical Mag. J. Sci.* 2, 559–572. doi:10.1080/14786440109462720

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proc. R. Soc. Lond.* 58, 240–242.

Saccenti, E., Hendriks, M. H., and Smilde, A. K. (2020). Corruption of the pearson correlation coefficient by measurement error and its estimation, bias, and correction under different error models. *Sci. Rep.* 10, 438–519. doi:10.1038/s41598-019-57247-4

Schober, P., Boer, C., and Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesth. Analgesia* 126, 1763–1768. doi:10.1213/ANE.0000000000002864

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *Am. J. Psychol.* 18, 161–169. doi:10.2307/1412408

Spearman, C. (1904). Measurement of association, part ii. correction of 'systematic deviations. *Am. J. Psychol.* 15, 88–101.

Wilcox, R. R. (2009). Comparing pearson correlations: Dealing with heteroscedasticity and nonnormality. *Commun. Statistics-Simulation Comput.* 38, 2220–2234. doi:10.1080/03610910903289151

Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between time-series?–a study in sampling and the nature of time-series. *J. R. Stat. Soc.* 89, 1–63. doi:10.2307/2341482