Revisiting Statistics and Evidence-Based Medicine: On the Fallacy of the Effect Size Based on Correlation and the Misconception of Contingency Tables

Sergey Roussakow (roussakow@neogalen.org)
Galenic Researches International LLP, London

Research Article

Keywords: Effect size, Pearson's coefficient of correlation, point biserial correlation, mean square contingency coefficient, 2 × 2 table, contingency table

Posted Date: August 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-871875/v3

Revisiting Statistics and Evidence-Based Medicine: On the Fallacy of the Effect Size Based on Correlation and the Misconception of Contingency Tables

Sergey Roussakow,* MD, PhD (0000-0002-2548-895X)

85 Great Portland Street

London, W1W 7LT, United Kingdome

roussakow@neogalen.org

+44 20 3885 0302

Affiliation: Galenic Researches International LLP

85 Great Portland Street

London, W1W 7LT, United Kingdome

Word count: 5,132.

ABSTRACT

BACKGROUND: Evidence-based medicine (EBM) is in crisis, in part due to bad methods, which are understood as misuse of statistics that is considered correct in itself. The study questions the correctness of the basic statistics related to the effect size (ES) based on correlation (CBES). METHODS: Monte Carlo simulation of paired samples, mathematical analysis, conceptual analysis, bias analysis. RESULTS: Correlation and ES are not related. CBES is a fallacy, mainly based on the point biserial correlation (PBC) fallacy and misconception of contingency tables (MCT), which makes no distinction between gross crosstabs (GCTs) and contingency tables (CTs). Misapplication of Pearson's correlation coefficient to point biserial datasets and GCTs gives ES parameters that are not related to correlation. Equations directly expressing ES in terms of correlation coefficient are flawed, since it is impossible without including covariance. Generalization of these fallacies leads to erroneous inferences, conversions, transformations, meta-analyses, and misunderstanding of the nature of correlation. MCT leads to misuse of the relevant statistics and is so ubiquitous that all findings from CTs are suspect. CONCLUSIONS: Two related common statistical misconceptions, CBES and MCT, have been exposed and fixed. These misconceptions are threatening because most of the findings from correlation, paired samples and CTs, including meta-analyses, can be misleading. Since exposing these fallacies casts doubt on the reliability of the statistical foundations of EBM in general, we urgently need to revise them.

KEW WORDS

Effect size, Pearson's coefficient of correlation, point biserial correlation, mean square contingency coefficient, 2 × 2 table, contingency table.

KEY FINDINGS

The study for the first time:

- revealed two related common misconceptions in statistics: the fallacy of effect size based on correlation and the misconception of contingency tables,
- revealed the fallacy of the point biserial correlation coefficient, which is an effect size parameter not related to correlation,
- revealed the similarity between the fallacies of the point biserial correlation coefficient and the Pearson's phi applied to gross crosstabs,
- fixed the fallacies and proposed necessary changes in statistics,
- corrected and redefined some statistical terms,
- questioned the reliability of the statistical foundations of EBM and stated the need to revise them.

Revisiting Statistics and Evidence-Based Medicine: On the Fallacy of the Effect Size Based on Correlation and the Misconception of Contingency Tables

INTRODUCTION

Evidence-based medicine (EBM) is "one of our greatest human creations,"[1, 2] but there is a growing awareness that it is undergoing a crisis,[1, 3, 4, 5, 6] in part due to "bad methods."[1] However, the idea of bad methods comes down to misusing statistics,[1] that is, misusing methods that are correct in themselves.[6] Therefore, believing in the reliability of the statistical foundations is the cornerstone of EBM. Unfortunately, there is cause for concern. This article exposes two common misconceptions in statistics, a fallacy of the effect size based on correlation (CBES) and a related misconception of contingency tables (MCT).

The introduction of the CBES concept in the 1980s is usually credited to Jacob Cohen,[7] although it seems to have been known before. Since then, it has been included in all statistical and meta-analysis manuals and is widely used, especially in psychometrics.[8, 9, 10, 11] The basic equation[10]

[1] $r^2 = \frac{d^2}{d^2+\frac{(n_1+n_0-2)(n_1+n_0)}{n_1 n_0}}$, where d is the effect size (ES) known as Cohen's d, r is the coefficient of bivariate correlation commonly known as Pearson's (product-moment) coefficient of correlation,[12]

given equal groups ($n_0 = n_1 = n$), reduces to

[2] $r^2 = \frac{d^2}{d^2+\frac{4(n-1)}{n}} \cong \frac{d^2}{d^2+4}$, so[13]

[3] $r \cong \frac{d}{\sqrt{d^2+4}}$;

this is the basic formula used by Cohen.[14] The corresponding equation for the dependence of d from r is

[4] $d = \frac{2r}{\sqrt{1-r^2}}\sqrt{\frac{n-1}{n}} \cong \frac{2r}{\sqrt{1-r^2}}$.[9, 14]

Equation [1] is fundamentally flawed and in itself suggests the fallacy of CBES, since it is designed for unequal groups $\{n_1, n_0\}$ (or, more precisely, it allows for inequality of groups), while any correlation statistic, by definition, must refer to paired (equal) groups.

CBES was the weakest place in the Cohen's effect size concept, since the large ES ($d = 0.8$) corresponded to $r = 0.371$, which is a weak correlation according to Pearson. To get around this problem, Cohen had to introduce (or use) the point-biserial correlation and the corresponding "biserial" *r* estimate connected to the raw *r* estimate with a correction factor of 1.253,[15] but even the adjusted "large" r was only 0.465, and still was lower than the Pearson's strong correlation limit of 0.5 (and even more so the modern limit of 0.7[16]). Motivated by this discrepancy, I investigated the causes of the discrepancies.

METHODS

The statistical model used to analyze ES versus correlation is shown in Fig. 1.

Primary dataset	
Y_1	Y_2
61	66
23	30
83	74
29	29
56	47
57	69
88	53
45	49
51	47
28	19

PBS dataset	
X	Y
1	60.9
1	22.7
1	83.2
1	28.9
1	55.8
1	56.6
1	87.6
1	45.5
1	51.3
1	27.8
0	66.5
0	29.9
0	74.2
0	28.9
0	47.5
0	69.4
0	53.0
0	49.3
0	47.2
0	19.1

Parameter	Primary biserial statistics			Parameter	Point biserial statistics		
	Y_1		Y_2		X		Y
μ_1	50	f_l	0.5				
σ_1	20	f_u	1.5				
n_1	10	n_2	10	n_x	20	n_y	20
m_1	52.024	m_2	48.501	m_x	0.500	m_y	50.263
s_1	20.924	s_2	17.426	s_x	0.500	s_y	19.335
cov	291.880			cov_{xy}	0.881		
r	**0.801**			r_p	**0.091**		
Δm	3.523			$\Delta m/cov_{xy}$	**4.000**		
	Unpaired		**Paired**				
df	18	df'	9				
s_p	20.296	s_p'	9.360	s_p	13.676		
s_c	19.335						
t	0.388	t'	0.842				
p	0.702	p'	0.422				
d	**0.174**	**d'**	**0.376**	$\mathbf{d_p}$	-3.639		
$\mathbf{r_d}$	**0.091**	$\mathbf{r_d'}$	**0.195**				
$\boldsymbol{\delta_c}$	**0.182**			r_p/δ_c	**0.500**		
$\boldsymbol{\delta_{rp}}$	**0.174**	$\boldsymbol{\delta_r}$	**2.534**				

Fig. 1. An example of the correlation-based effect size (CBES) analytical model.

Note: μ_1 and σ_1 are the expected mean and standard deviation (SD) of the Y_1; f_l and f_u are the lower and upper bounds, respectively, of the random contingency factor f; n, group size; df, degrees of freedom; m, group mean; s, group SD; cov, covariance (here and below, population (biased) SDs and covariances are used to avoid small sample bias); r, coefficient of correlation (Pearson's r); r_p – point biserial correlation coefficient; Δm, effect magnitude;

S_p, pooled SD; S_c, combined SD; t, t-statistic; p, p-value; d, Cohen's (pooled) effect size; r_d, d-based coefficient of correlation; δ_c, combined effect size; δ_{rp}, point biserial correlation-based effect size; δ_r, actual correlation-based effect size; mark ' (single quotation mark) refers to paired parameters; marks $_1$ and $_2$ refer to samples Y_1 and Y_2, respectively; marks $_x$ and $_y$ refer to samples X and Y, respectively.

Let there be a biserial (bivariate), continuous dataset consisting of two normally distributed paired variables Y_1 and Y_2 (primary dataset), i.e.

[5] $$Y_2 \sim Y_1 \begin{vmatrix} \bar{Y}_1 \sim N(\mu_1, \sigma_1) \\ \bar{Y}_2 \sim N(\mu_2, \sigma_2) \end{vmatrix}.$$

The variable Y_1 is randomly normalized using the Excel NORM.INV[17] function, so that

[6] $y_1 = NORM.INV(\Pr, \mu_1, \sigma_1)$, where μ_1 is the expected mean, σ_1 is the expected standard deviation (SD), Pr is the random probability (probability mass function), where

[7] $Pr = RAND()$, where RAND() is the Excel RAND[18] function providing a random probability of $0 \le Pr \le 1$.

The dependent paired variable Y_2 is obtained by multiplying the variable Y_1 by a random contingency factor f,

[8] $$y_2 = \begin{cases} r_e > 0: f y_1 \\ r_e < 0: \begin{cases} f y_1 \mid y_1 < \bar{y}_1 \\ \frac{y_1}{f} \mid y_1 \ge \bar{y}_1 \end{cases} \end{cases}$$, where r_e is the expected direction of the correlation coefficient, f is the random contingency factor

[9] $f = f_l + RAND() \times (f_u - f_l)$, where f_l and f_u are the lower and upper bounds of f, respectively, so f will be randomly distributed around the central value $\frac{f_l + f_u}{2}$, and the dependence of f from f_l and f_u will be as follows:

[10] $$f_l < f_u \begin{cases} f_l < 1 \Rightarrow \begin{cases} f_u < 1 \Rightarrow f < 1 \\ f_u = 1 \Rightarrow f \le 1 \\ f_u > 1 \Rightarrow f <> 1 \end{cases} \\ f_l = 1 \Rightarrow \quad f \ge 1 \\ f_l > 1 \Rightarrow \quad f > 1 \end{cases} r^2 < 1$$

$$f_l = f_u \ \{ \quad f = f_l = f_u = const \quad \} r^2 = 1$$

The smaller the difference between f_l and f_u, the stronger the correlation, and $f_l = f_u$ makes the dependence functional (ie, $r^2 = 1$).

[11] $\lim_{f_l \to f_u} r^2 = 1.$

The expected direction of the correlation (r_e, equation [8]) is selectable, so both directions have been tested. The negative correlation model is simplified, so the prefect negative correlation $r = -1$ is never achieved even with $f_l = f_u$, but is nevertheless relevant for assessing the negative correlation effect.

In fact, such complex dependency modeling was not necessary, since the principles of correlation are universal, so all patterns could be demonstrated on unrelated paired samples. This was done to avoid doubts about the adequacy of the model.

The derivative (secondary), heterogeneous (continuous-binary), "point biserial" dataset XY includes a continuous variable Y, which is the combined primary dataset, i.e.

[12] $Y = \{Y_1 \cup Y_2\},$

and a binary variable X, which is the combination index attributing the Y values, i.e.

[13] $x_i = \begin{cases} 1 \mid \mathrm{y_i} \in Y_1 \\ 0 \mid \mathrm{y_i} \in Y_2 \end{cases}.$

Because the model uses a small sample of $n = 10$ for simplicity, this causes small sample bias that distorts the relationships inherent in infinite samples. To avoid the bias, the model defaults to population standard deviations and covariances (biased estimators without Bessel's correction). The option of sample (unbiased) statistics is available in the model given in the Supplement.

The calculated parameters include sample sizes (n), arithmetic means (m), effect magnitudes (Δm), standard deviations (s), covariances (cov), Pearson's correlation coefficients (r), the point biserial correlation (PBC) coefficient (r_p) as the Pearson correlation coefficient for the PBC dataset (XY), degrees of freedom in terms of t-test (df), pooled standard deviations (S_p),

[14] $S_p = \sqrt{\frac{s_1^2 n_1 + s_0^2 n_0}{n_1 + n_0 - 2}}$ for unpaired samples and $S_p = \sqrt{\left(\frac{n}{n-1}\right)\left(\frac{s_1^2 + s_0^2 - 2 r s_1 s_0}{2}\right)}$ for paired samples, where *s* is the population (biased) SD;

t-statistics (t), p-values, combined standard deviations (S_c)

[15] $S_c = \sqrt{\frac{n_1 s_1^2 + n_0 s_0^2 + \frac{n_1 n_0}{n_1 + n_0}(\bar{y}_1 - \bar{y}_0)^2}{n_1 + n_0}}$ (biased estimate) and $S_c = \sqrt{\frac{n_1 s_1^2 + n_0 s_0^2 + \frac{n_1 n_0}{n_1 + n_0}(\bar{y}_1 - \bar{y}_0)^2}{n_1 + n_0 - 1}}$ (unbiased estimate) (the combined SD is the true SD of the combined sample, since the pooled SD is the weighted average of the group SDs, that is, an approximation that depends only on variance and ignores the means difference);

actual effect sizes (AES), including the pooled ES (Cohen's d),

[16] $d = \frac{\Delta \mathrm{m}}{S_p}$,

and combined ES (δ_c),

[17] $\delta_c = \frac{\Delta \mathrm{m}}{S_c}$;

Cohen's effect size-based correlation coefficient (r_d) (Eq. [2]); and correlation-based effect sizes (Eq. [4]). For the purposes of this study, CBES (δ_r) is understood in its direct meaning, that is, refers to the effect size based on the true (primary) correlation r, and the effect size based on the PBC coefficient (r_p) is referred to as the point biserial effect size (PBES) (δ_{r_p}). All effect sizes are directional for better accuracy and visualization. The unpaired primary statistics is used because it corresponds to the PBC statistics which is unpaired (see DISCUSSION).

A Monte Carlo simulation of 10,000 iterations was performed. Then scatter charts were plotted and the coefficients of linear correlation (R^2) calculated for CBES and PBES versus AES. The working model in MS Excel is available in the Supplement.

RESULTS

In conflict with equations [2]–[4] implying functional relationship between correlation and effect size, they are unrelated.

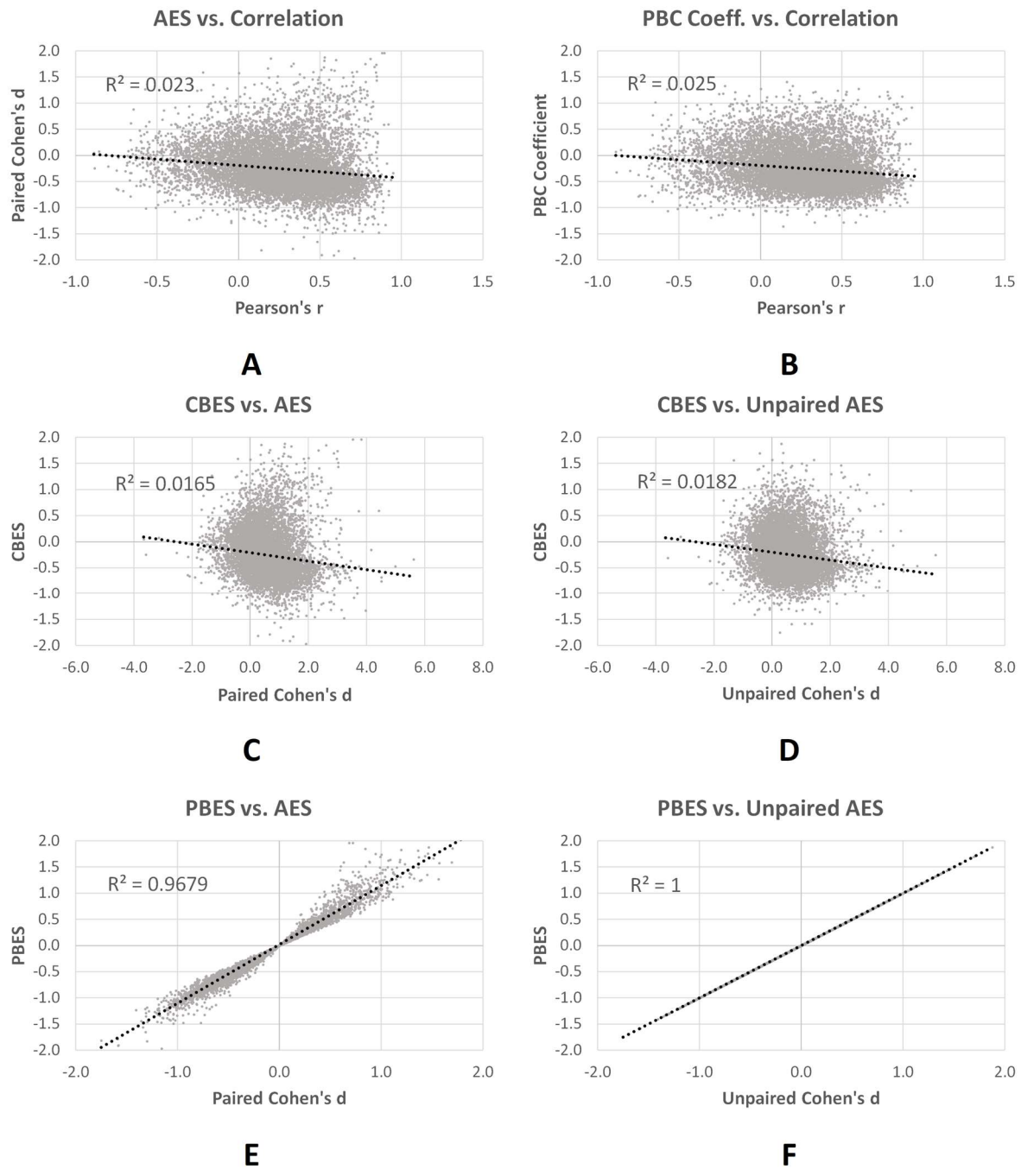


Fig. 2. *Relationships between effect size and correlation (Monte Carlo simulation, 10,000 iterations;* $\mu_1 = 50, \sigma_1 = 20$*,* $f_l = 0, f_u = 3, r_e < 0$*).*

Fig. 2 shows that there is no correlation between correlation coefficient and AES ($\mathrm{R}^2 = 0.023$, Fig. 2-A), as well as between CBES and AES, either paired ($\mathrm{R}^2 = 0.0165$, Fig. 2-C) or unpaired ($\mathrm{R}^2 = 0.0182$, Fig. 2-D). An example in Fig. 1 shows that the CBES following the correlation coefficient of $\mathrm{r} = 0.801$ is $\delta_\mathrm{r} = 2.534$, while the paired AES (Cohen's d) is $\mathrm{d}' = 0.376$ and the unpaired one is $\mathrm{d} = 0.174$.

In contrast, PBES perfectly correlates with unpaired AES ($\mathrm{R}^2 = 1.0$, Fig. 2-F) meaning they are the same parameter ($\delta_{\mathrm{r_p}} \equiv \mathrm{d}$), and closely correlates with paired AES ($\mathrm{R}^2 = 0.9679$, Fig. 2-E) – although the example in Fig. 1 shows that the correlation-related difference between them can be very large ($\mathrm{d}' = 0.376$ vs. $\delta_{\mathrm{r_p}} = 0.174$) – but does not relate to the primary (true) correlation ($\mathrm{R}^2 = 0.025$, Fig. 2-B). For example, in Fig. 1 $\mathrm{r_p} = 0.091$, and $\mathrm{r} = 0.801$.

A summary of the relationships between effect size and correlation is as follows:

- Correlation is not related to AES.
- CBES is not related to any AES $\left(\delta_r \neq \left\{\delta_{r_p}, d, d', \delta_c, \delta_c'\right\}\right)$.
- PBES matches the primary unpaired ES:
    - the unpaired pooled ES (Cohen's d) is by equation [4];
    - the primary unpaired combined effect size is twice the PBC coefficient ($\delta_c = 2r_p$):
- PBES does not match the primary paired ES, except for $r = 0$ ($\delta_{r_p} \neq d' \neq \delta_c' \mid r \neq 0$).
- PBC is not related to primary correlation ($r_p \neq r$).

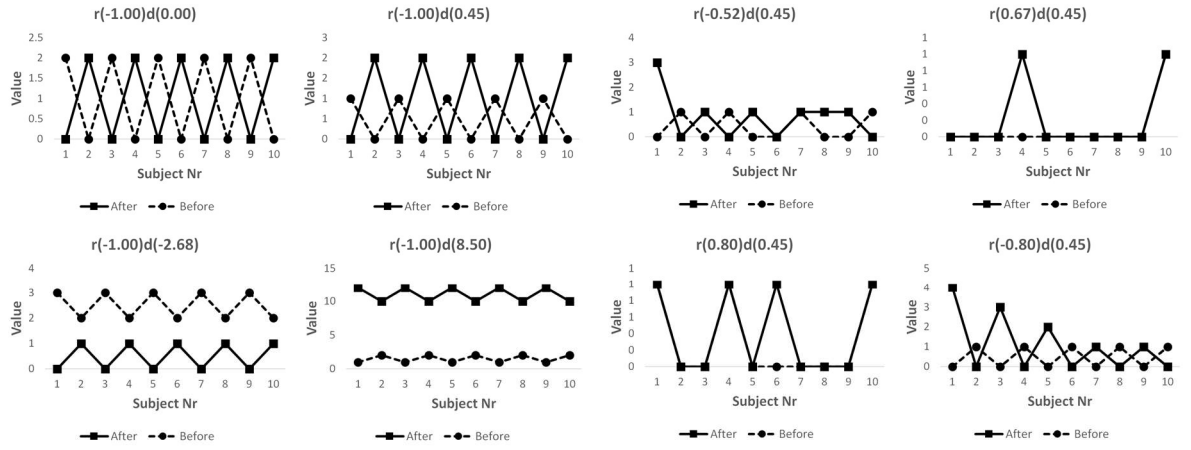
Thus, correlation and CBES are not related to AES, suggesting that equations [2]–[4] are flawed.

DISCUSSION

Although after Cohen's milestone monograph,[7] the relationship between correlation and ES seems apparent and even trivial, it is actually a logical fallacy stemming from the trivial notion that, since they are related to between-group differences, they are interrelated and therefore mutually convertible.

Different nature, independence and inconvertibility of effect size and correlation

Fig. 3 visually demonstrates the independence of effect size and correlation, namely, the possibility of designing any effect size for a fixed correlation (Fig. 3-A)) or any correlation for a fixed effect size (Fig. 3-B).



Pattern		r(-1.00)d(0.00)		r(-1.00)d(0.45)		r(-1.00)d(-2.68)		r(-1.00)d(8.50)	
Group		After	Before	After	Before	After	Before	After	Before
Subject	1	0	2	0	1	0	3	12	1
	2	2	0	2	0	1	2	10	2
	3	0	2	0	1	0	3	12	1
	4	2	0	2	0	1	2	10	2
	5	0	2	0	1	0	3	12	1
	6	2	0	2	0	1	2	10	2
	7	0	2	0	1	0	3	12	1
	8	2	0	2	0	1	2	10	2
	9	0	2	0	1	0	3	12	1
	10	2	0	2	0	1	2	10	2
n		10	10	10	10	10	10	10	10
m		1	1	1.0	0.5	0.5	2.5	11.0	1.5
s		1.05	1.05	1.05	0.53	0.53	0.53	1.05	0.53
Δm		0.00		0.50		-2.00		9.50	
Sp		1.49		1.12		0.75		1.12	
r		-1.00		-1.00		-1.00		-1.00	
d		0.00		0.45		-2.68		8.50	

A

Pattern		r(-0.52)d(0.45)		r(0.67)d(0.45)		r(0.80)d(0.45)		r(-0.80)d(0.45)	
Group		After	Before	After	Before	After	Before	After	Before
Subject	1	3	0	0	0	1	1	4	0
	2	0	1	0	0	0	0	0	1
	3	1	0	0	0	0	0	3	0
	4	0	1	1	0	1	1	0	1
	5	1	0	0	0	0	0	2	0
	6	0	0	0	0	1	0	0	1
	7	1	1	0	0	0	0	1	0
	8	1	0	0	0	0	0	0	1
	9	1	0	0	0	0	0	1	0
	10	0	1	1	1	1	1	0	1
n		10	10	10	10	10	10	10	10
m		0.8	0.4	0.2	0.1	0.4	0.3	1.1	0.5
s		0.92	0.52	0.42	0.32	0.52	0.48	1.45	0.53
Δm		0.40		0.10		0.10		0.60	
Sp		0.89		0.22		0.22		1.34	
r		-0.52		0.67		0.80		-0.80	
d		0.45		0.45		0.45		0.45	

B

Fig. 3. *Visual demonstration of the independence of correlation and effect size. A – different effect sizes for a fixed correlation (r = -1); B – different correlations for a fixed effect size (d = 0.45) – in a paired sample of 10 pairs measured by some parameter before and after an intervention (with the parameter value along the vertical axis).*

r, Pearson's correlation coefficient; d, the paired Cohen's (pooled) effect size [21]; the lines show the parameter changes between the subjects to visualize the association (concordance of changes) between the pairs.

Being a variance-normalized mean difference (Tab. 1), ES is a kind of signal-to-noise ratio that characterizes the relative magnitude of the mean difference regardless of the concordance of changes and, therefore, is applicable to both paired and unpaired samples.

Tab. 1. *Fundamental differences between effect size and correlation.*

	Effect size	**Correlation**
Characteristic	Mean difference normalized to variance	Covariance cleared of variance

	Effect size	Correlation
Essence	Signal-to-noise ratio	Relationship
Effect magnitude	Relevant	Irrelevant
Concordance of changes	Irrelevant	Relevant
Samples	Any (paired and unpaired)	Paired

Correlation, which is covariance cleared of variance, described by Pearson's (product-moment) coefficient of (biserial) correlation

[18] $$r = \frac{cov_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^{n}((x_i - m_X)(y_i - m_Y))}{\sqrt{\sum_{i=1}^{n}(x_i - m_x)^2 \sum_{i=1}^{n}(y_i - m_y)^2}},$$

is a measure of relationship (causation or dependence) that characterizes the concordance of specific (paired) differences (changes) regardless of the magnitude of the difference (Fig. 3-A), and applies only to paired samples. Thus, correlation and effect size are fundamentally different parameters that in principle cannot be reduced to each other.

Variance (covariance cleared from correlation) is a pure measure of sample variability, correlation (covariance cleared from variance) is a pure measure of the association or concordance of paired sample changes, and covariance is a complex parameter that combines variability and association (Tab. 2).

Tab. 2. *Properties of variance, covariance, and correlation.*

	Variance	Covariance	Correlation
Variability	Yes	Yes	No
Association	No	Yes	Yes
Direction	No	Yes	Yes

These parameters are interconnected, so that any one of them can only be expressed in terms of the other two (Eq. [18]). Given paired samples, we can express ES in terms of variance and correlation, or variance and covariance, but never just correlation:

[19] $d = \frac{\Delta m}{S_p} \underset{n_X = n_Y = n}{\Longrightarrow} d = \frac{\Delta m}{\sqrt{\frac{s'^2_X + s'^2_Y - 2rs'_X s'_Y}{2}}} = \frac{\Delta m}{\sqrt{\frac{s'^2_X + s'^2_Y - 2cov'_{XY}}{2}}}$, where *s'* and *cov'* are the sample (n - 1) statistics.

Given equal variances, we can express ES in terms of covariance and correlation as follows:

[20] $r = \frac{cov'_{XY}}{s'_X s'_Y} \underset{s'_X = s'_Y = s'}{\Longrightarrow} r = \frac{cov'_{XY}}{s'^2} \Rightarrow s' = \sqrt{\frac{cov'_{XY}}{r}}$,

[21] $S_p = \sqrt{\frac{s'^2_X + s'^2_{Y0} - 2rs'_X s'_Y}{2}} \underset{s'_X = s'_Y = s'}{\Longrightarrow} s'\sqrt{1 - r} = \sqrt{cov'_{XY}\frac{1-r}{r}}$,

[22] $d = \frac{\Delta m}{S_p} \underset{s'_X = s'_Y = s'}{\underset{n_X = n_Y = n}{\Longrightarrow}} d = \frac{\Delta m}{s'\sqrt{1-r}} = \frac{\Delta m}{\sqrt{cov'_{XY}\frac{1-r}{r}}}$.

Thus, ES cannot be expressed in terms of correlation alone and must necessarily also include covariance and/or variance. Therefore, ES and correlation are mutually inconvertible, so equations [2]–[4] and the CBES concept in general are fundamentally wrong, and the observed independence of the effect size from correlation (Fig. 1-2) is natural.

Point Biserial Correlation Fallacy

The main reason why the independence of correlation and ES remains hidden is the so-called "Point Biserial Correlation" (PBC). Since Cohen used it to justify the interconvertibility of ES and correlation by introducing (or using) the relation [2],[14] it has been considered a proof of the validity of CBES. In fact, the PBC is a multilevel fallacy.

The fallacy is in the fact that PBC is not a correlation. It seems that Karl Pearson would be extremely surprised if he knew that his correlation coefficient for homogeneous variables would be applied for heterogeneous (binary and continuous), as this changes its nature, and it is no longer a correlation coefficient. On Fig. 1, it is easy to see that the subjects in the point biserial dataset are no longer paired as in the primary dataset, so PBC loses the correlation value and is unpaired by default. So what is PBC?

The PBC coefficient is the Pearson correlation coefficient applied to point biserial dataset, that is, the ratio of the covariance of the variables X and Y to the product of their standard deviations (Eq. [18]). Given paired samples, i.e. if $N = 2n$, so that the number of indices 1 and 0 are equal, the mean and SD of the binary variable X are 0.5, that is, are constants:

[23] $$\left.\begin{array}{l}\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=0.5\\ s_X=\sqrt{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N}}=0.5\end{array}\right| \{N=2n, \sum_{i=1}^{n}x_i=n\,, \sum_{j=n+1}^{N=2n}x_j=0\}.$$

Since the Y variable is the combined primary Y_1 and Y_2 variables (Eq. [12]), its SD is the combined standard deviation of the Y_1 and Y_2 primary samples (Eq. [15]).

[24] $s_Y = S_c(Y_1, Y_2)$

The XY covariance is

[25] $cov_{XY}=\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{N}$.

Given equation [23], there are n of $[x_i-\bar{x}=1-0.5=0.5]$ and n of $\left[x_j-\bar{x}=0-0.5=-0.5\right]$. Therefore, the total sum in [25] can be expressed as the sum of two equal subsamples of *n* elements each corresponding to the initial samples Y_1 and Y_2, so following equation [23],

[26] $cov_{XY}=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y}) \underset{[23]}{\Longrightarrow} \frac{1}{2n}\left(\sum_{i=1}^{n}0.5(y_{1i}-\bar{y})+\sum_{j=n+1}^{2n}-0.5\left(y_{2j}-\bar{y}\right)\right)=\frac{1}{4n}\left(\sum_{i=1}^{n}(y_{1i}-\bar{y})-\sum_{j=1}^{n}\left(y_{2j}-\bar{y}\right)\right)=\frac{1}{4n}\left(\sum_{i=1}^{n}y_{1i}-\sum_{j=1}^{n}y_{2j}\right)=\frac{1}{4n}(\bar{y}_1 n-\bar{y}_2 n)=\frac{1}{4}(\bar{y}_1-\bar{y}_2)$,

i.e., the covariance of the point biserial dataset (XY) is reduced to a quarter of the primary effect value (see Fig. 1) and is independent of correlation (as expected) or even variance (beyond expectations). Then, the PBC is

[27] $r_p=\frac{\bar{y}_1-\bar{y}_2}{2s_Y} \underset{[24]}{\Longrightarrow} \frac{\overline{y_1}-\overline{y_2}}{2S_c}=\frac{1}{2}\delta_c$, where δ_c is the unpaired primary combined effect size.

Since, as mentioned above, the *Y* sample is naturally unpaired when pooled, r_p is independent of correlation and depends only on the primary variance.

In the case of using sample statistics, the relationships [26]-[27] change in the sample size-dependent manner due to Bessel's correction:

[28] $cov_{XY}=\frac{n_{XY}}{4(n_{XY}-1)}(\bar{y}_1-\bar{y}_2)$, where $n_{XY}=n_X=n_Y$

[29] $r_p=\frac{1}{2}\sqrt{\frac{n_{XY}}{n_{XY}-1}}\,\delta_c$,

so that, eg, for the case of primary $n = 10$ (ie, $n_{XY} = 20$), $cov_{XY} = \frac{\Delta m}{3.8}$ and $r_p \approx 0.513\ \delta_c$.

In the case of unequal samples (this is a false assumption since it means unpaired samples, but it is usual in the conventional CBES concept (eg, see equation [1] and PBC),

[30] $r_p = \delta_c \sqrt{\frac{n_1 n_2}{(n_1+n_2)^2}}$, which reduces to equation [27] with equal samples.

The PBC coefficient can be then transformed into Cohen's d as follows:

[31] $S_c = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2 + \frac{n_1 n_2}{n_1+n_2}(\overline{y_1}-\overline{y_2})^2}{n_1+n_0}} \underset{n_1=n_2=n}{\Longrightarrow} S_c = \sqrt{\frac{s_1^2+s_2^2}{2} + \frac{(\overline{y_1}-\overline{y_2})^2}{4}}$,

[32] $S_p = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_0-2}} \underset{n_1=n_2=n}{\Longrightarrow} \sqrt{\frac{n}{n-1}\ \frac{s_1^2+s_2^2}{2}}$,

then, following equation [27],

[33] $r_p = \frac{1}{2}\delta_c = \frac{\overline{y_1}-\overline{y_2}}{2\sqrt{\frac{s_1^2+s_2^2}{2}+\frac{(\overline{y_1}-\overline{y_2})^2}{4}}} = \frac{\frac{\overline{y_1}-\overline{y_2}}{S_p}}{\frac{2\sqrt{\frac{s_1^2+s_2^2}{2}+\frac{(\overline{y_1}-\overline{y_2})^2}{4}}}{S_p}} = \frac{d}{2\sqrt{\frac{s_1^2+s_2^2}{2S_p^2}+\frac{(\overline{y_1}-\overline{y_2})^2}{4S_p^2}}} =$

$\frac{d}{2\sqrt{\frac{s_1^2+s_2^2}{2\left(\frac{n}{n-1}\frac{s_1^2+s_2^2}{2}\right)}+\left(\frac{d}{2}\right)^2}} = \frac{d}{\sqrt{d^2+4\frac{n-1}{n}}}$,

which is the known equation [2].

Thus, the PBC coefficient is not a correlation, so even its name is misleading: it is essentially another ES parameter, which is equal to half the primary combined effect size, refers to independent samples, and has nothing to do with the actual correlation. Using PBC to justify CBES is an intellectual falsification, where the true primary correlation coefficient is substituted by a point biserial pseudo-correlation, which is in fact a different calculation of the primary effect size (i.e. identical to ES), which results in the informal fallacy of confusing identity and dependency. This manipulation allows to bypass the problem of independence of correlation and effect size and gives the impression that CBES is valid and ES is a relationship measure.

The PBC-related CBES fallacy is multi-layered:

- Since PBES is ostensibly "correlation based," it provokes the calculation of effect size based on the true correlation; in this case, the resulting CBES is false and does not match any AES (Fig. 1, 2BC).
- Even using the PBC coefficient in equation [4], the resulting PBES matches the unpaired AES, and not the paired one as it should, so the PBES is still misleading; if correlation is strong, the error can be striking (eg, $r_p = 0.174$ vs. the actual $d' = 0.376$ in Fig. 1; see also Tab. 4).
- Since CBES concept directly relates ES to correlation, a misbelief arises that ES is a relationship measure (see "Summary, origin and significance of the CBES fallacy and MCT").
- Because the PBC coefficient applies to a point biserial dataset where dichotomous variable numbers may be unequal (ie, $n(1) \neq n(0)$), it allows PBES to be used for unequal (unpaired) samples under the guise of pairwise statistic (see equation [1][10] as an example of this fallacy).
- Using PBC for inferential statistics is either misleading or unnecessary. In the case of paired samples, PBC is misleading since returns unpaired statistics instead of the expected paired one. In the case of unpaired samples, PBC is not necessary since it is easier and more accurate to apply significance testing directly to the primary samples.

Thus, PBC is fundamentally misleading and should be void. The point biserial dataset is an unnecessary artificial construct that we never see in reality and is only used to falsely justify CBES. Some statisticians understand that CBES only refers to PBES, but consider it pairwise and still consider PBC to be a correlation; others sincerely believe that CBES refers to actual correlation, as stated in all manuals,[7, 10, 11] so everyone is misled. The PBC-based inferential statistics is an unnecessary surrogate.

Misconception of Contingency Tables

Another reason why the inconvertibility of correlation and ES remains hidden, and probably the true cause of the misconception, is the misconception of contingency tables (MCT). Currently, any crosstab is considered contingency table by default,[19, 20] resulting in the severe fallacy shown in Fig. 4.

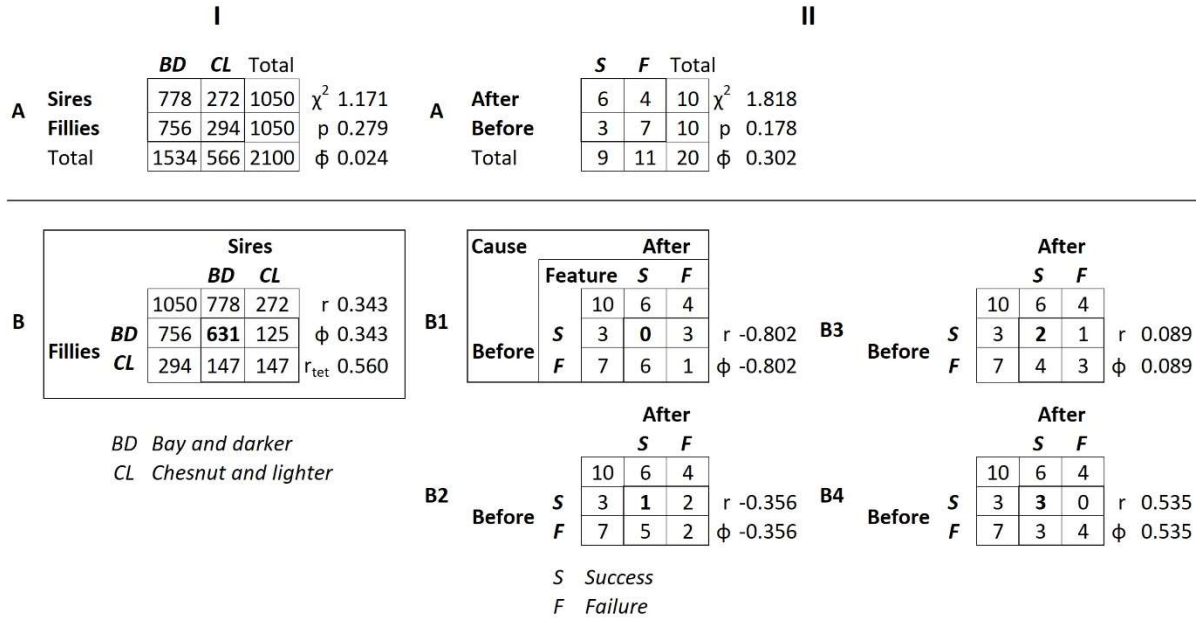


Fig. 4*. Example of 2 × 2 tables: I, Pearson's example;[21] II, simulated example; A, gross (parent, master) crosstab; B, contingency (child) tables.* r*, Pearson's* r*;* φ*, Pearson's* φ *;* $\bar{\varphi}$*, gross Pearson's Phi;* χ^2*, Pearson's chi-square statistic; p, p-value.*

Table IIA is a 2 × 2 matrix (crosstab) of two binary variables, intervention (Before–After) and outcome (Success–Failure). It is a gross crosstab that reports that 3 out of 10 subjects were a success before intervention and 6 out of 10 after the intervention. This table gives information on the effect but does not contain the information on the association of the variables. To display the contingency information, it must be converted to a contingency table IIB, where the cells of the gross (parent, master) table become the marginal statistics of the contingency (child) table, and a new 2 × 2 matrix appears that relates to the contingency (association). For example, in the table IIB2, of three before-after pairs who were a success before intervention, one remained a success after the intervention (SS pair) and the other two changed to a failure (SF). Likewise, of seven failure pairs, two remained a failure (FF), and five changed to a success (FS). Therefore, 3 pairs have kept their outcomes, while 7 changed them.

For this table, we can calculate the Pearson mean square contingency coefficient φ (Pearson's Phi)[21]

[34] $\varphi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$, where $\frac{a \mid b}{c \mid d}$ are the corresponding cells of the 2 × 2 matrix,

as well as any other measures of association,[22, 23, 24, 25, 26, 27] to reveal a negative association (ϕ=-0.356) between the variables in terms of the outcomes (i.e., the groups change their outcomes discordantly).

Thus, we come to new definitions of crosstabs:

- Categorical crosstab is an n × m matrix that displays the mutual frequency distribution of two categorical variables having n and m categories, respectively.
- Binary crosstab (BCT) is a 2 × 2 matrix that displays the mutual frequency distribution of two binary variables.
- Gross crosstab (GCT) is a BCT that displays the mutual frequency distribution of two different binary variables, one of which is paired (i.e., the options are interrelated).
- Contingency table (CT) is a BCT that displays the frequency distribution of the feature pairs against the featured paired cause, so that the marginal statistics of CT match the cells of the parent GCT for these paired binary variables (Fig. *4*-B1).

The differences between BCT, GCT and CT are summarized in Tab. 3.

Tab. 3. *Differences between binary crosstabs, gross crosstabs and contingency tables.*

	Binary crosstab	**Gross crosstab**	**Contingency table**
Number of levels	1		2
Unit	Subject		Pair
Sample	Unpaired	Paired	
Mutual relationship	Unrelated	Master table	Child table
Reduces to contingency tables	Not applicable	Yes	No
Restores to the gross table	Not applicable	No	Yes
Significance of differences	Unpaired tests		Paired tests
Association measures	Inapplicable		Applicable
Related linear dataset	Secondary (point biserial)		Primary

CT has a double-decker design (Fig. 4-IIB1),[28] where the first level is formed by the feature binary variable, and the second by the causal binary variable, and counts pairs. GCT and BCT

are single-decker (Fig. 4-IIA) and count subjects. Each parent GBC can be reduced to one or several child CTs with different associations, where the number of CTs is equal to the minimum GCT cell value plus one. For example, in the example II (Fig. 4), four association options are available with the SS values of 0, 1, 2 and 3 (B1-4). Accordingly, any CT can be restored to the parent GCT (Fig. 4-I). BCT refers to unpaired samples, so it cannot be converted to CT. Unpaired significance tests (Pearson's chi-square, Fisher's exact test, etc.) are used with BCTs and GCTs, but give wrong results when applied to CTs, so the CTs-based significance testing requires paired tests (e.g., McNemar's tests). Thus, CTs, GSTs and BCTs are different in nature.

The example of thoroughbred racehorses[21] (Fig. 4-IB), which Karl Pearson used when introducing his Phi, is a double-decker CT, from which the parent GCT can be easily restored (Fig. 4-IA). When applied to CTs, Pearson's Phi obtains its original meaning of the "mean square contingency coefficient" ϕ, which is another calculation of the Pearson's r for two binary variables (Fig. 4-B). The confusion is in the fact that, due to the similarity of 2 × 2 crosstabs, Pearson's Phi can be also applied to GCT, but, due to the mentioned difference in nature of the tables, this results in a completely different parameter. Pearson's Phi applied to GCT has nothing to do with correlation (contingency), it is a size-independent derivative of the chi-squared statistic

[35] $\overline{\varphi} = \sqrt{\frac{\chi^2}{N}}$, where χ^2 is the Pearson chi square statistic, N is the sample size.

By reducing equation [34] to GCT,

[36] $\overline{\varphi} = \frac{p_1 - p_0}{2\sqrt{pq}} = \frac{p_1 - p_0}{2S_c} = \frac{1}{2}\delta_c$, where p is the proportion of events, q is the proportion of no events,

which is a binary analogue of the continuous equation [27]. Thus, $\overline{\varphi}$ is an unpaired ES parameter and, like its continuous counterpart, it can be transformed into the unpaired binary Cohen's d by the equation [4]:

[37] $d = \frac{2\overline{\varphi}}{\sqrt{1-\overline{\varphi}^2}}\sqrt{\frac{N-2}{N}}$.[29]

So, the term Pearson's Phi (ϕ), or "mean square contingency coefficient," should only be applied to CTs. When applied to GCTs, equations [34] and [35] give the "mean-square effect

size" and should be denoted as the gross Pearson Phi ($\overline{\varphi}$). Fig. 4 shows that $\overline{\varphi}$, which is an effect size parameter (part A), has nothing to do with φ (part B), which is a correlation parameter.

The similarity between GCT and PBC is explained by the fact that GCT can be derived from both primary and the binary point biserial dataset (data binary vs. combination binary), which is similar to the continuous point biserial dataset (continuous vs. combination binary), while CT stems only from the primary dataset. As a result, the patterns revealed in the continuous domain (Fig. 1) are valid for the binary domain too (Fig. 5).

Prim	
Y_1	Y_2
0	0
1	0
0	1
1	0
0	0
1	0
0	1
0	0
0	0
1	1

PB	
X	Y
1	**0**
1	**1**
1	**0**
1	**1**
1	**0**
1	**1**
1	**0**
1	**0**
1	**0**
1	**1**
0	0
0	0
0	1
0	0
0	0
0	0
0	0
0	1
0	0
0	0
0	1

Par	Primary		Point biserial	
	Y_1	Y_2	X	Y
n	10	10	20	20
m	0.400	0.300	0.500	0.350
s	0.490	0.458	0.500	0.477
cov	-0.020		cov_{xy}	0.025
r	**-0.089**		r_p	**0.105**
Δm	0.100		$\Delta m/cov_{xy}$	4.000
	Unpaired	**Paired**		
df	18	9		
sp	0.500	0.522	s_p	0.489
sc	0.477			
t	0.447	0.429		
p	0.660	0.678		
d	**0.200**	**0.192**	$\mathbf{d_p}$	0.307
$\mathbf{r_d}$	**0.105**	**0.101**		
δ_c	**0.210**		r_p/δ_c	**0.50**
$\boldsymbol{\delta_{rp}}$	**0.200**	**-0.170**		

GCT (PB)			
	1	**0**	
X	4	6	10
Y	3	7	10
	7	13	20
φ	**0.105**		
df	18		
χ^2	0.220		
$p_{\chi 2}$	0.639		
t	0.447		
p	0.660		
d	0.200		

CT (Prim)			
		Y	
		1	**0**
X	**1**	4	1	3
	0	6	2	4
		10	3	7
	-0.089		
	9		
	0.200		
	0.655		
	0.429		
	0.678		
	0.192		

Fig. 5. *The relationship of conventional (linear) binary data sets and cross-tabulations. Contingency table (CT) is derived from the primary binary dataset (Prim), the gross crosstab (GCT) – from the point biserial dataset (PB). Unpaired chi-square statistics is by Pearson's test, paired – by McNemar's test. $p_{\chi 2}$ is the chi-square p-value.*

Correlation dependence of the paired significance of differences

The last source of the CBES fallacy is the known dependence of the paired significance of differences (SOD) from correlation: the larger the correlation, the stronger the SOD. Since

SOD depends on the ES, it seems natural that correlation and ES are mutually related. However, this is another fallacy, the nature of which is parsed in Fig. 6.

The paired SOD depends on t-statistics

[38] $t = d\sqrt{\frac{n}{2}}$,[10]

so is determined by the effect size d, which, as shown above (Eqs. [19] and [22]), depends on both variance and covariance (correlation). As shown in Fig. 6A, it is the variance that determines the basic zero-correlation ES (δ_1 and δ_2) and the corresponding chi-square ($\chi^2_{1_0}$ and $\chi^2_{2_0}$). With the fixed variance, changing correlation does indeed change the ES and χ^2 (SOD). So, in an imaginary set of CTs with different associations for each GCT (Fig. 6B), there is a functional dependence of ES from correlation (Fig. 6C),

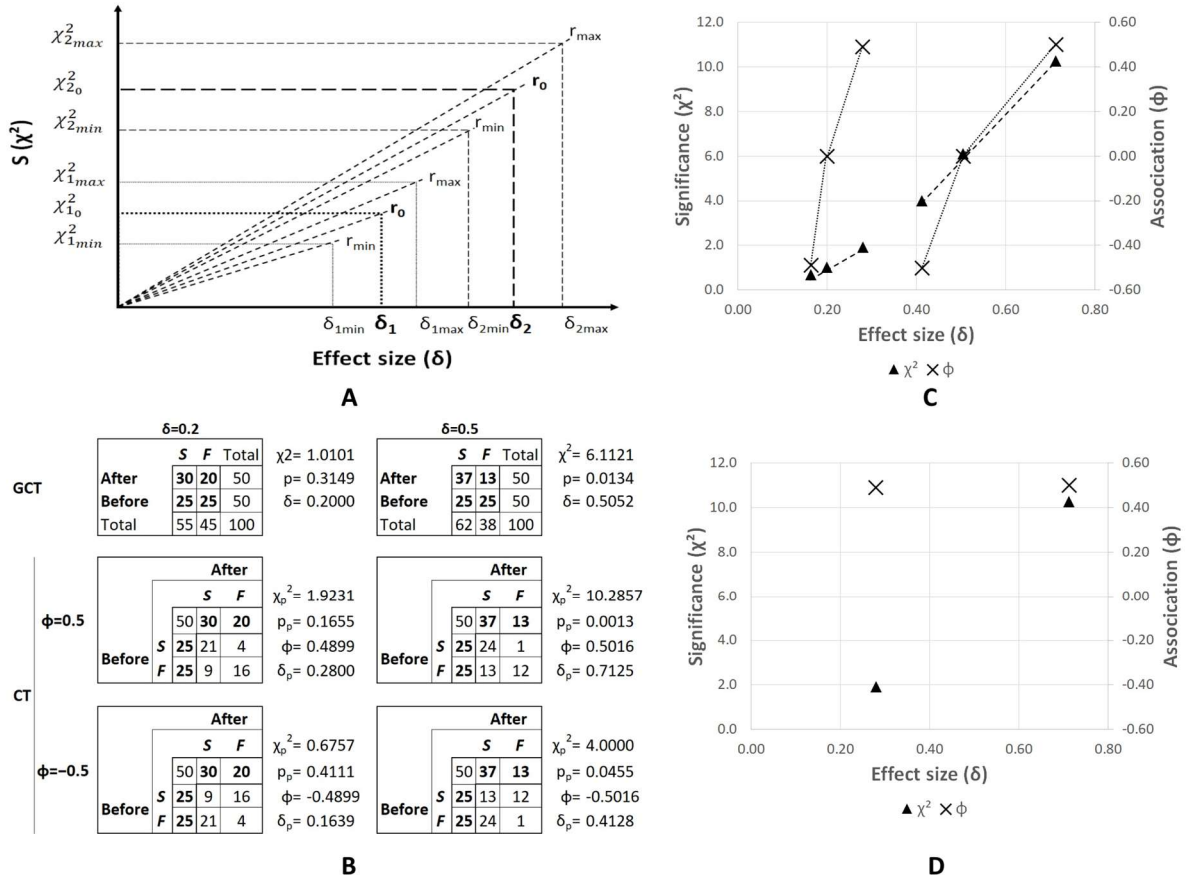


Fig. 6. *Explanation of the relationship between correlation and effect size and significance of differences: A, schematic representation of the fallacy; B-D, an example: B, original tables; C-D, significance and association vs. AES: C, within-GCT (multiple associations); D, between-GCT (one association per GCT).*

However, in real life the dependency disappears since we are always dealing with one association per GCT (Fig. 6D), and variance that fundamentally determines SOD is not related to correlation (Tab. 2). An example case is shown in Fig. 6B-D. Two equal-sized GCTs with different zero-correlation ESs of 0.2 and 0.5 allow several associations (Fig. 6B), and within every GCT, the relationship between the correlation and ES / significance is functional (Fig. 6C). However, in pairwise comparison of actual CTs of different GCTs, the relationship disappears (Fig. 2), since the same correlations (here, $\varphi \approx 0.5$ or $\varphi \approx -0.5$) correspond to different ESs (δ_p of 0.28 vs. 0.7125 at $\varphi \approx 0.5$ and 0.1639 vs. 0.4128 at $\varphi \approx -0.5$, Fig. 6D). In fact, correlation only modulates (Eq. [22]) the pure variance dependence of the unpaired ES ($d = \frac{\Delta m}{s'}$) and the related SOD, but never makes correlation and ES directly interdependent, as CBES suggests.

Thus, the essence of the SOD-based CBES fallacy is in the perception of the influence of the correlation coefficient on ES and SOD "given fixed variance" as a universal dependence, i.e. ignoring this fundamental reservation. The main misconception stemming from this fallacy is a widespread tendency to draw conclusions about relationship based on the SOD (statistical inference-based correlation fallacy) and vice versa (correlation-based statistical inferencing fallacy) (see examples in the next section).

Moreover, in fact, the paired SOD gives another proof of the CBES fallacy, since the equation [4] nullifies ES at zero correlation – and, accordingly, nullifies the SOD ($p = 1$, no difference) by zeroing the t-statistics (Eq. [38]) – while in reality, zero correlation simply reduces the paired SOD to unpaired, so a correct equation should reduce the paired effect size to unpaired, but not nullify it.

Summary, origin and significance of the CBES fallacy and MCT

Thus, the CBES fallacy is that ES has nothing to do with correlation, so the CBES concept is fundamentally wrong. Applying Pearson's r to point biserial datasets and Pearson's Phi to GCTs gives effect size rather than correlation coefficient, and the fallacy is that these parameters are misleadingly considered to be correlations. As a result, CBES is fundamentally flawed: if it is calculated from PBC coefficient or gross phi, it matches the unpaired AES, but not the paired AES it is intended to match, and if it is calculated from the actual correlation coefficient (Pearson's r or phi), then it does not match any AES. Generalization of these false equations [2]–[4] leads to erroneous inferences, conversions,

transformations,[30, 31] meta-analyses,[10, 11] and misunderstanding of the nature of correlation.[7]

It is difficult to say how long statistics have been captive to these misconceptions. The CBES fallacy seems took shape in the early 80s with the introduction of the idea of effect size[7] and meta-analysis.[8, 32] The MCT seems to go back to Karl Pearson himself, who (or someone of his team) boldly calculated φ for all 2 × 2 tables, including GCTs.[33] However, that time this was no more than a misapplication, as the difference between GST and CT was intuitively understood, though never clearly articulated. Then this understanding gradually blurred over time (Cramer in 1946[34] still correctly applied Pearson's Phi to CTs, while Cohen in 1988[7] was already in the misconception), and nowadays the MCT dominates. In a sense, the MCT is due to the term "2 × 2 table," which makes no distinction between GCTs and CTs and facilitates the misuse of Pearson's Phi.

Despite the apparent inadequacy, the CBES concept has never been questioned, is included in all guidelines,[9-11] equations [2]–[4] are commonly used for calculating ES and related conversions,[30, 31] and even for the correlation-based definition of the ES,[10] so it is a common misconception.

The same applies to the MCT, which is a series of related misconceptions. Typically, it looks like treating GCTs, or even unpaired BCTs, as CTs, misleadingly attributing to them the ability to assess the association of the variables, leading to misapplication of association statistics (e. g., applying association statistics to GCTs) and independence statistics (e. g., applying Pearson's chi-square to CTs). Fundamentally, this misconception stems from three fallacies: confusion of effect and association, as discussed above for CBES; misunderstanding of the mutual relationship between GCTs and CTs as parent and child tables, leading to the belief that CTs simply arise in a single specific form;[28] and lack of understanding of the pairwise nature of association (if samples are not paired (i.e., the pairs are not intrinsically bound), they can be resorted in any order and any association is random (corresponds to a certain random order), therefore meaningless). Finally, Pearson's Phi is often calculated using equation [35] for the gross Pearson Phi,[35] which is non-directional, so it is not a measure of association that requires equation [34].

An example of the consequences of these misconceptions is shown in Fig. 7. Section I presents example CTs (Table 1–3) taken from a credible source.[28] Sections II and III include

the corresponding GCTs and CTs, respectively, obtained by adjusting the original tables. Only Table 1 was indeed a two-decker CT, obtained by applying two binary variables (opinions on death penalty and gun registration) to the same subject (paired samples). However, the confusion of association and effect led to the misleading conclusion that "P value is 0.0232 … suggests that there is an association … ." In fact, there was no association (ϕ=-0.061, Table 1-III), and the conclusion is a statistical error caused by MCT.

Table 1

I

Gun registration	*Favor*	*Oppose*	Total	
		Death penalty		
Favor	784	236	1020	χ^2= 5.1503
Oppose	311	66	377	p= 0.0232
Total	1095	302	1397	

Table 2

Smoker	*Yes*	*No*	Total	
	Lung cancer			
Yes	647	622	1269	χ^2= 22.044
No	2	27	29	p= 2.7E-06
Total	649	649	1298	

Table 3

Treatment	*Yes*	*No*	Total
	Depression improved?		
Pramipexole	8	4	12
Placebo	2	8	10
Total	10	12	22

II

	Favor	*Oppose*	Total	
Death penalty	1095	302	1397	χ^2= 10.944
Gun registration	1020	377	1397	p= 0.0009
Total	2115	679	2794	δ= 0.146

	LC	*NLC*	Total	
SM	647	622	1269	χ^2= 22.044
NSM	2	27	29	p= 2.7E-06
Total	649	649	1298	δ= 0.263

III

			Death penalty		
			Favor	*Oppose*	
		1397	1095	302	χ^2_p= 10.283
Gun regi	*Favor*	1020	784	236	p_p= 0.0013
stration	*Oppose*	377	311	66	ϕ= -0.061

Fig. 7. *An example of the misconception of 2 × 2 tables: [28] I, original tables; II, corrected gross tables; III, corrected contingency tables. LC, Lung Cancer; NLC, No Lung Cancer; SM, Smoker, NSM, Non-Smoker.*

In addition, Pearson's chi-square was misapplied to the CT, so that the reported p-value of 0.0232 (χ^2=5.15) is incorrect, and the actual p-value is 0.0013 (χ^2=10.283). Tab. 4 presents a more illustrative example of the degree of possible error caused by misapplication of significance tests based on the example in Fig. 4II.

Tab. 4. *Significance and association error due to misconception of contingency tables (Fig.* ***4****II).*

		Pearson's test		**McNemar's test**	
Table	**ϕ**	**χ^2**	**p**	**χ^2**	**p**
A	0.302	1.818	0.178	0.143	0.705
B1	-0.802	6.429	0.011	1.000	0.317
B2	-0.356	1.270	0.260	1.286	0.257

Table	ϕ	Pearson's test		McNemar's test	
		χ^2	p	χ^2	p
B3	0.089	0.079	0.778	1.800	0.180
B4	0.535	2.857	0.091	3.000	0.083

For example, misapplication of Pearson's chi-square to CT B1 would show a significant difference ($p = 0.011$), while the actual difference (McNemar's test) is not significant ($p = 0.317$). The same applies to association: a strong negative association in the CT B1 ($\varphi = -0.802$) would be misjudged as a weak positive association ($\varphi = 0.302$), if based on the GCT (A). This is how MCT contributes to data dredging (p-hacking).

Other examples in Fig. 7 are not CTs since count subjects, not pairs. A pseudo-two-decker design of Table 2-I is misleading because when the count is not based on pairs, the Yes–No options, misclassified as "a single binary variable," actually represent different variables "Lung cancer" (Lung cancer – No lung cancer) and "Smoking" (Smoker – Non-Smoker). It is also not a GCT since the samples, albeit equal-sized, are unpaired.[36] Thus, this is a BCT (Table 2-II) that cannot be reduced to CT since any association in the BCT is random, therefore misleading. Finally, Table 3-I is a typical BCT with unequal groups that technically cannot be converted to CT, so its pseudo-two-decker design is simply anecdotal. All of these examples misjudge association based on statistical inference.

The example in Fig. 7 shows that MCT is threatening, because much of the findings obtained from CTs can be misleading. The misconception is widespread: the idea of CT is typically explained using GCTs;[11, 28, 37] BCTs are misleadingly referred to as CTs;[19, 28] CTs are often (mostly?) pseudo CTs;[28] and even true CTs are still misjudged in terms of significance testing[28] and association.[35] With all of the above, there seems to be no publication on CTs unaffected by the misconception and therefore not misleading.

Required changes in statistics

Given the above, the following changes should be made to the corpus of statistics:

- The CBES concept should be abolished as flawed and misleading.

- The concept of point biserial correlation, as well as any statistics on point biserial datasets, should be discarded as misleading and unnecessary.
- New definitions for BCTs, GCTs and CTs should be adopted, the term "2 × 2 table" should be avoided as confusing.
- Pearson's mean square contingency coefficient (φ) should only be applied to CTs; it is an association measure that does not apply to meta-analysis.
- Pearson's Phi applied to GCTs should be referred to as the "mean-square effect size" ($\bar{\varphi}$); it should not be used for estimating association.
- Equations [2]–[4] and all related equations linking effect size to correlation / association are misleading and should be used in the correct sense as conversions between different effect sizes measures.
- Unpaired significance tests should only apply to BCTs and GCTs; paired tests should be used for significance testing with CTs.
- In no case should the results of significance tests be used to assess the association between variables and vice versa.
- All conclusions and inferences based on the CBES, as well as all conversions and transformations based on CBES, should be revised.
- All meta-analyses based on CBES should be revised.
- All findings and conclusions based on CTs should be revised.
- The relevant chapters in statistical and meta-analysis manuals should be revised.

CONCLUSIONS

This article exposes two common misconceptions in statistics, the correlation-based effect size (CBES) fallacy and the misconception of contingency tables (MCT), which have been around for over 70 years and go unnoticed. The CBES concept falsely suggests effect size and correlation coefficient to be functionally dependent and directly interconvertible. MCT arises from the confusion of gross crosstabs (GCTs) and contingency tables (CTs), which leads to the CBES fallacy due to the confusion of Pearson's phi and gross Pearson's phi, and the misuse of the relevant inferential statistics.

In fact, correlation is not related to effect size, so CBES is fundamentally false. The CBES fallacy is due to the point biserial correlation (PBC) fallacy, since the PBC coefficient is an

effect size parameter rather than correlation; and the related MCT, where the mean-square contingency coefficient (Pearson's phi) of CTs turns unto the mean-square effect size (gross Pearson's phi) when misapplied to GCTs. CBES is anyway misleading: if it is calculated from PBC coefficient or gross Pearson's phi, it matches the unpaired actual effect size (AES) and does not match the paired AES it is intended to match; and if it is calculated from the actual correlation, then it does not correspond to any AES. Moreover, the PBC coefficient and the gross Pearson's phi do not allow estimation of relationship, since they are not correlation (association) measures, and true correlation measures do not allow statistical inference, as the CBES concept suggests. Generalization of these fallacies leads to erroneous inferences, conversions, transformations, meta-analyses, and misunderstanding of the nature of correlation. The fallacies are so ubiquitous, old, and rooted that virtually all pairwise statistics since the 1980s (probably even the 1960s) are suspect.

The presence of such serious fallacies in the very foundations of statistics, which are considered long known and unshakable, cast doubt on the reliability of the statistical foundations of EBM in general. If statistics is corrupted, then the problems of EBM are deeper than it is believed, because they are not limited to the misuse of statistics but extends to the bad statistics itself. However, this can be a problem and a solution at the same time, as many of the EBM problems can be caused by flawed statistics and resolved by fixing these flaws. That is why we urgently need to revise the statistical foundations of EBM. This article fixes the misconceptions of CBES and contingency tables.

DATA SHARING

An Excel model described in Methods is available in the Supplement. Extra models are available by emailing SR or at http://neogalen.org/projects?mode=view&ret_mode=folder&post_id=4160304&folder_id=220280806

DECLARATIONS

Abbreviations

AES	actual effect size
BCT	binary cross-tabulation

CBES	effect size based on correlation; correlation-based effect size
ES	effect size
PBC	point biserial correlation
PBES	point biserial correlation-based effect size
SD	standard deviation
SOD	significance of differences

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

All data generated or analyzed during this study are included in this published article (and its supplementary information files).

Competing interests

The author declares no conflicts of interest.

Funding

No funding.

Authors' contributions

The sole author is the only contributor.

Acknowledgements

I am grateful to all the editors and reviewers who have made invaluable contributions to the improvement of this article.

Authors Information (optional)

No information.

REFERENCES

[1] Horton R. What is medicine's 5 sigma? Lancet 2015;385(9976):P1380.

[2] Kamerow D. Milestones, tombstones, and sex education. BMJ 2007;334:0-a.

[3] Ioannidis JPA. Why most published research findings are false? PLOS Medicine 2005;2(8):e124.

[4] Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. The Lancet 2009;374(9683):86–9.

[5] Macleod MB, Michie S, Roberts I, et al. Biomedical research: increasing value, reducing waste. Lancet 2014;383(9912):101-4.

[6] Ioannidis JPA. Evidence-based medicine has been hijacked: a report to David Sackett. J Clin Epidemiol 2016;73:82-6.

[7] Cohen J. Statistical Power Analysis for the Behavioral Sciences. Routledge 1988:75-144.

[8] Rosenthal R. Meta-analysis procedures for social research. Sage: Beverly Hills 1984.

[9] Cooper H, Hedges LV. The Handbook of Research Synthesis, Volume 236. Russell Sage Foundation 1994:238.

[10] Hedges LV, Olkin I. Statistical methods for meta-analysis. Orlando: Academic Press Inc 1985:77.

[11] Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. John Wiley and Sons 2009:41-43.

[12] Pearson K. Notes on regression and inheritance in the case of two parents. Proc of the Royal Society of London 1895;58:240–2.

[13] Borenstein (2009), p.48, formula 7.7, 7.8.

[14] Cohen (1988), p. 23, formula 2.2.6.

[15] Cohen (1988), p. 82, formula 3.2.1.

[16] Evans JD. Straightforward Statistics for the Behavioral Sciences. Brooks/Cole Publishing, Pacific Grove 1996.

[17] NORM.INV function. [cited 2022 Jul 18] https://support.microsoft.com/en-us/office/norm-inv-function-54b30935-fee7-493c-bedb-2278a9db7e13

[18] RAND function. [cited 2022 Jul 18] https://support.microsoft.com/en-us/office/rand-function-4cbfa695-8869-4788-8d90-021ea9f5be73#:~:text=If%20you%20want%20to%20use,you%20with%20just%20a%20value

[19] Lauritzen SL. Lectures on Contingency Tables. Electronic edition, 1979-2002. [cited 2021 Jul 14] http://www.stats.ox.ac.uk/~steffen/papers/cont.pdf

[20] Everett BS, Skrondal A. The Cambridge dictionary of statistics, 4th ed. Cambridge University Press 2010.

[21] Pearson K. Mathematical Contributions to the Theory of Evolution. VII. On the Correlation of Characters not Quantitatively Measurable. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 1900;195:1-47.

[22] Yule GU. On the Association of Attributes in Statistics: With Illustrations from the Material of the Childhood Society, &c. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 1900;194:257–319.

[23] Cohen J. A coefficient of agreement for nominal scales. Educat Psychol Measurement 1960;20(1):37–46.

[24] Goodman LA, Kruskal WH. Measures of Association for Cross Classifications. J Amer Stat Assoc 1954;49(268):732–64.

[25] Goodman LA, Kruskal WH. Measures of Association for Cross Classifications. II: Further Discussion and References. J Amer Stat Assoc 1959;54(285):123–63.

[26] Goodman LA, Kruskal WH. Measures of Association for Cross Classifications III: Approximate Sampling Theory. J Amer Stat Assoc 1963;58(302):310–64.

[27] Goodman LA, Kruskal WH. Measures of Association for Cross Classifications, IV: Simplification of Asymptotic Variances. J Amer Stat Assoc 1972;67(338):415–21.

[28] Liu L, Berger VW. Two by Two Contingency Tables. In: Encyclopedia of Statistics in Behavioral Science (Eds: BS Everitt & DC Howell). John Wiley & Sons, Ltd, Chichester 2005;4:2076–81.

[29] Fleiss JL. Measures of effect size for categorical data. In: Cooper H, Hedges LV (Eds.). The handbook of research synthesis. New York: Russell Sage Foundation 1994:245–60.

[30] Olivier J, May WL, Bell ML. Relative effect sizes for measures of risk. Commun Stat Theory Methods 2017;46(14):6774-8.

[31] Olivier J, Bell ML. Effect Sizes for 2×2 Contingency Tables. PLoS One 2013;8(3):e58777.

[32] Glass GV, McGraw B, Smith ML. Meta-analysis in social research. Sage: Beverly Hills 1981.

[33] Ekström J. The Phi-coefficient, the Tetrachoric Correlation Coefficient, and the Pearson-Yule Debate. UCLA 2011 [cited 2021 Jun 15] https://escholarship.org/uc/item/7qp4604r..

[34] Cramér H. Mathematical Methods of Statistics. Princeton: Princeton University Press 1946:282.

[35] Everitt BS, Skrondal A. The Cambridge dictionary of statistics, 4th ed. Cambridge University Press 2010:325.

[36] Doll R, Hill AB. Smoking and carcinoma of the lung; preliminary report. BMJ 1950;2(4682),739–48.

[37] Contingency table. Wikipedia [cited 2021 Jun 14] https://en.wikipedia.org/wiki/Contingency_table

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CBESModel.v16.xlsb