PLOS ONE

RESEARCH ARTICLE

# Factoring a 2 x 2 contingency table

**Stanley Luck** *

Science, Technology and Research Institute of Delaware, Wilmington, DE, United States of America

* stan.luck@vectoranalytics.ai

## Abstract

We show that a two-component proportional representation provides the necessary framework to account for the properties of a 2 × 2 contingency table. This corresponds to the factorization of the table as a product of proportion and diagonal row or column sum matrices. The row and column sum invariant measures for proportional variation are obtained. Geometrically, these correspond to displacements of two point vectors in the standard one-simplex, which are reduced to a center-of-mass coordinate representation, $(\delta, \mu) \in \mathbb{R}^2$. Then, effect size measures, such as the odds ratio and relative risk, correspond to different perspective functions for the mapping of $(\delta, \mu)$ to $\mathbb{R}^1$. Furthermore, variations in $\delta$ and $\mu$ will be associated with different cost-benefit trade-offs for a given application. Therefore, pure mathematics alone does not provide the specification of a general form for the perspective function. This implies that the question of the merits of the odds ratio versus relative risk cannot be resolved in a general way. Expressions are obtained for the marginal sum dependence and the relations between various effect size measures, including the simple matching coefficient, odds ratio, relative risk, Yule's $Q$, $\phi$, and Goodman and Kruskal's $\tau_{c|r}$. We also show that Gini information gain ($IG_G$) is equivalent to $\phi^2$ in the classification and regression tree (CART) algorithm. Then, $IG_G$ can yield misleading results due to the dependence on marginal sums. Monte Carlo methods facilitate the detailed specification of stochastic effects in the data acquisition process and provide a practical way to estimate the confidence interval for an effect size.





**Citation:** Luck S (2019) Factoring a 2 x 2 contingency table. PLoS ONE 14(10): e0224460. https://doi.org/10.1371/journal.pone.0224460

**Editor:** Fabio Rapallo, Universita degli Studi di Genova, ITALY

**Received:** April 25, 2019

**Accepted:** October 14, 2019

**Published:** October 25, 2019



**Data Availability Statement:** Data for this project were retrieved from the publicly accessible Nursing Home Compare website, https://data.medicare.gov/data/nursing-home-compare. Copies of the input files for our analysis are available here, https://doi.org/10.6084/m9.figshare.7934960.v1.

**Funding:** The author received no specific funding for this work.

**Competing interests:** The author has declared that no competing interests exist.

## Introduction

In research with contingency tables, the ability to compare experimental results from different studies is essential for studying the dependence between categorical variables and how it is maintained. However, the data acquisition is controlled by sample size parameters that appear as row and column sums for the various categories. Association coefficients that are not adjusted for unbalanced sample size can differ between tables even if the underlying system response is unchanged [1, 2]. The dependence of the $\phi$ coefficient on the margins led to the development of the normalized form, $\phi/\phi_{max}$ [3, 4]. Recently, VanLiere and Rosenberg investigated the allele frequency dependence of the $r^2$ linkage disequilibrium measure [5]; note that $\phi$ and $r$ refer to the same coefficient. Olivier and Bell discussed the limitations of the $\phi$ coefficient and proposed effect size thresholds for the odds ratio because it is a measure that is "not

problematic" [6]. The odds ratio is invariant to scaling of rows or columns, but there is continuing debate on the merits of the odds ratio versus the relative risk [7–10]. Warrens [11] showed that members of the general family of association coefficients that are linear transformations of the simple matching coefficient do not satisfy all three desiderata for a well-behaved coefficient. The lack of consensus on the utility of the many alternative effect size measures [11, 12] led us to consider whether there might be a core set of principles and elementary properties for $2 \times 2$ tables that might broadly apply. In this paper, we review coordinate systems for representing proportional variation in a $2 \times 2$ table, which corresponds to a two-component system of point vectors in the standard one-simplex with two degrees of freedom. Then, we examine the equivalence class of tables induced by an odds ratio. The scaling invariance corresponds to a diagonal symmetry such that an odds ratio does not possess a simple interpretation in terms of proportional effects. We discuss the connections between proportion difference, odds ratio, Yule's *Q*, and relative risk and show that an effect size statistic is more generally regarded as a perspective function, i. e., a linear fractional transformation [13] of proportional variation. A contingency table factors into a product of proportion and diagonal row or column sum matrices. Rows and columns of the proportion matrix correspond to different representations of the relation between categorical variables. Therefore, a $2 \times 2$ table is associated with four different forms of proportional variation. Together, these constitute the full implementation of the Goodman and Kruskal proposal that adjustment for unbalanced sample size is needed in the estimation of effect size [2]. Various forms of stochastic effects can affect a data acquisition process, so a $2 \times 2$ table is associated with a distribution. We discuss the use of Monte Carlo methods as a practical way to simulate a distribution of tables and estimate the confidence interval for an effect size. Finally, our interest in effect size measures developed in the course of plant breeding research at DuPont to identify agriculturally beneficial genetic variation in maize [14]. These studies involved high-dimensional search to assess linkage disequilibrium and genome-wide association (GWAS) in maize populations, including the use of the classification and regression tree (CART) algorithm. An essential step in CART is an exhaustive search over the range of each independent variable for an optimal binary partition of the response data [15, 16]. We show that the Gini information gain is equivalent to $\phi^2$, and we compare their behavior with a scaling invariant effect size measure using a publicly available data set. Satisfactory resolution of these longstanding issues in the application of effect size for statistics would have broad implications for high-dimensional data analysis and machine learning. The main novel contributions of this work are: 1) identification of the correspondence between factoring the $2 \times 2$ table and effect size, 2) identification of the four forms of proportional variation with row or column sum invariance, 3) identification of an effect size measure for a $2 \times 2$ table as a mapping of proportional variation for a two-component system in $\triangle^1 \times \triangle^1$ to $\mathbb{R}^2$, 4) identification of the equivalence between Gini information gain and the $\phi$ coefficient, 5) development of an improved CART association algorithm using a proportional displacement measure with correction for unbalanced sample size for the response.

## 1 Methods

### 1.1 Notation

In this work, we study the connection between odds ratio, proportion and $\phi$ for a $2 \times 2$ table. Our notation for the three required coordinate systems is briefly summarized here. We deviate slightly from convention and use the symbol $\triangle^1$ to designate the standard one-simplex [13] such that the dot product of a vector, $\mathbf{u} \in \triangle^1$, with the one-vector satisfies the condition $\mathbf{u} \cdot \mathbf{1} = 1$. Ratio vectors, $(\alpha, 1)$ and $(\beta, 1)$, with $\alpha, \beta \in \mathbb{R}^1$ are elements of the projective line, $\mathbb{P}^1$. $(\alpha, 1)$

corresponds to the proportion, $p_\alpha = \alpha/(\alpha + 1)$, and the proportion vector, $\mathbf{p}_\alpha = (p_\alpha, 1 - p_\alpha)$, in $\triangle^1$. The subscript for a proportion corresponds to its $\mathbb{P}^1$ coordinate. Similarly, $(\beta, 1)$ corresponds to the proportion vector $\mathbf{p}_\beta = (p_\beta, 1 - p_\beta)$. $(a, b)$, $(c, d)$, $(a, c)$, and $(b, d)$ are vectors in $\mathbb{R}^2$. $(a, b)$ corresponds to the ratio vector, $(a/b, 1)$, in $\mathbb{P}^1$. $(a/b, 1)$ corresponds to the proportion, $p_{a/b} = \frac{a}{b}/\left(\frac{a}{b} + 1\right)$, and the proportion vector, $(p_{a/b}, p_{b/a}) = (p_{a/b}, 1 - p_{a/b})$, in $\triangle^1$. Ratio and proportion vectors are defined in a similar way for the other $\mathbb{R}^2$ vectors. The slightly cumbersome subscript notation is necessary because we are working with proportions for both row space such as '$p_{a/b}$', and column space such as '$p_{a/c}$'. However, in subscripts for marginal sum proportions the division by $N$ is dropped; e. g., $p_{a+c} = (a + c)/N$ where $N = a + b + c + d$. Ratio and proportion vectors are examples of perspective functions of the general form $P(\mathbf{u}, t) = \frac{\mathbf{u}}{t}$ for $\mathbf{u} \in \mathbb{R}^N$, $t \in \mathbb{R}^1$, and $t > 0$ [13]. Another familiar example is normalization by the Euclidean norm, $P(\mathbf{u}, ||\mathbf{u}||) = \frac{1}{||\mathbf{u}||}\mathbf{u}$.

## 1.2 Coordinate systems for proportion and odds ratio

In this section, we discuss coordinate systems for representing binary proportional variation in categorical data analysis. For the point vector $(a, b) \in \mathbb{R}^2$, the ratio corresponds to a linear fractional transformation

$$\begin{aligned} \frac{a}{b} &= \frac{(a + b) + (a - b)}{(a + b) - (a - b)}, \\ &= \frac{1 + \delta_s}{1 - \delta_s}, \end{aligned} \tag{1}$$

where $\delta_s$ is the difference in proportion

$$\begin{aligned} \delta_s &= \frac{a - b}{a + b}, \\ &= \frac{\frac{a}{b} - 1}{\frac{a}{b} + 1}. \end{aligned}$$

The '$s$' designation arises from the connection with the proportional displacement, $\boldsymbol{\delta}_s$, between the pair of vectors $(a, b)$ and $(b, a)$,

$$\begin{aligned} \boldsymbol{\delta}_s &= \frac{1}{a + b}[(a, b) - (b, a)], \\ &= \delta_s(1, -1), \end{aligned} \tag{2}$$

and the correspondence of these vectors to a diagonally 'symmetric' $2 \times 2$ table as described in Section 1.4. We will encounter several expressions of the form Eq (1), indicating that elements of projective geometry [13, 17] provide the framework for the analysis of proportional variation. Consequently, our objective is to identify vector algebraic structures for representing proportional variation in asymmetric $2 \times 2$ tables. They provide the framework for analyzing the relationships between binary proportion, odds ratio, Yule's $Q$, relative risk, and $\phi$.

Proportional normalization of a ratio vector produces a proportion vector

$$\begin{aligned} \frac{1}{\frac{a}{b} + 1}\left(\frac{a}{b}, 1\right) &= \left(\frac{a}{a + b}, \frac{b}{a + b}\right), \\ &= P((a, b), a + b), \end{aligned}$$

which is an element of $\triangle^1$ (Fig 1). Then, a proportion vector has the form $\mathbf{v} = (v_1, 1 - v_1)$, with

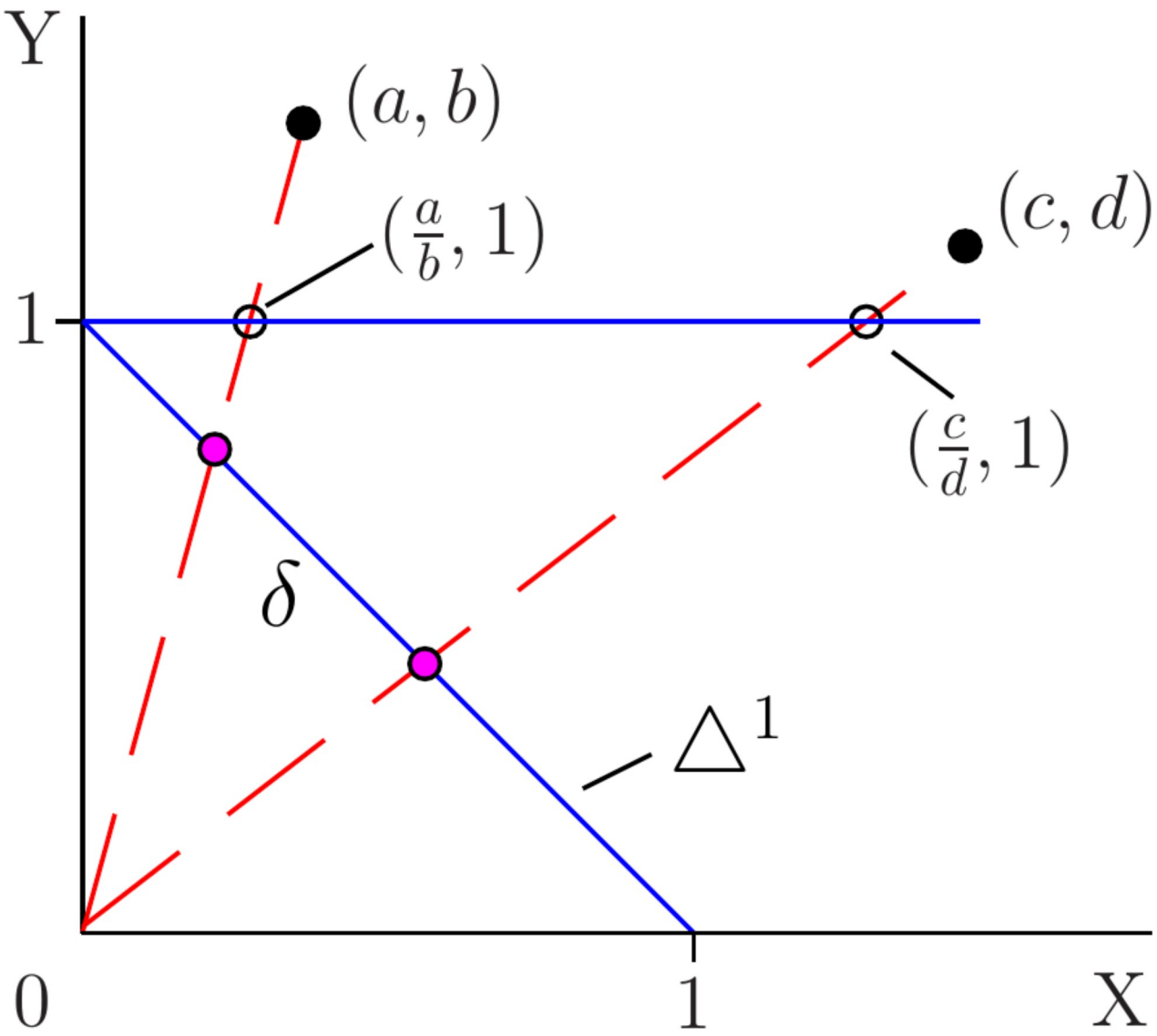


**Fig 1. Coordinates for a two-component binary proportional system.** Proportional variation for vectors, $(a, b)$ and $(c, d)$, is represented either as points, $\left(\frac{a}{b}, 1\right)$ and $\left(\frac{c}{d}, 1\right)$, in $\mathbb{P}^1$, or as points in the standard one-simplex, $\triangle^1$. $\delta$ is the proportional displacement between the vectors. Proportion and ratio are related by a linear fractional transformation, as indicated by the dashed lines.

https://doi.org/10.1371/journal.pone.0224460.g001

derivative $d\mathbf{v} = dv_1(1, -1)$ such that $d\mathbf{v} \cdot \mathbf{1} = 0$ for $0 \leq v_1 \leq 1$. In contrast, the corresponding ratio vector has the form

$$\mathbf{v}' = \left(\frac{v_1}{1 - v_1}, 1\right),$$

with derivative

$$d\mathbf{v}' = \frac{\mathrm{d}v_1}{(1 - v_1)^2}(1, 0).$$

Then, the difference between proportion vectors **u** and **v**

$$\begin{aligned}\boldsymbol{\delta} &= \mathbf{u} - \mathbf{v},\\ &= (u_1 - v_1)(1, -1),\end{aligned}$$

is parameterized by a single parameter, $u_1 - v_1$, and variation in binary proportion corresponds to translation in $\triangle^1$. The difference between ratio vectors is also parameterized by a single parameter,

$$\begin{aligned}\boldsymbol{\delta}' &= \mathbf{u}' - \mathbf{v}',\\ &= \frac{(u_1 - v_1)}{(1 - u_1)(1 - v_1)}(1, 0).\end{aligned}$$

Therefore, $\triangle^1$ and $\mathbb{P}^1$ correspond to different constraints in representing proportional variation. However, the order of categories in a contingency table is arbitrary, and it is not possible to identify a unique category that should serve as the perspective coordinate for a ratio. This introduces ambiguity, as we will see later in the discussion of the odds ratio. On the other hand, in factoring out the effects of marginal sums, the $\triangle^1$ representation provides an important function in the analysis of 2 × 2 tables.

Now, we discuss the representation of a two-component system of binary proportions in $\triangle^1$ and $\mathbb{P}^1$ coordinate systems, and describe intrinsic properties of various effect size measures. The formulae take on a more compact, intuitive form because scaling invariance is built-in. The algebraic intuition gained here helps in comprehending the more cumbersome expressions obtained later using the $\mathbb{R}^2$ representation. The exception is the $\phi$ coefficient, which does not possess a $\triangle^1$ representation due to the lack of scaling invariance (section 1.4). In particular, we discuss properties of the odds ratio, $\omega = \beta/\alpha$, where $\alpha, \beta \geq 0$, corresponding to $(\alpha, 1)$ and $(\beta, 1)$ on the $\mathbb{P}^1$ line, respectively. Then, relative risk is defined as $\rho = p_\beta/p_\alpha$, where $p_\beta = \beta/(\beta + 1)$ and $p_\alpha = \alpha/(\alpha + 1)$. The corresponding proportional basis consists of $\mathbf{p}_\alpha = (p_\alpha, 1 - p_\alpha)$ and $\mathbf{p}_\beta = (p_\beta, 1 - p_\beta)$. Next, we introduce the center-of-mass basis

$$\begin{aligned}\boldsymbol{\delta}_{\beta-\alpha} &= \frac{1}{2}(\mathbf{p}_\beta - \mathbf{p}_\alpha),\\ &= \delta(1, -1),\\ \boldsymbol{\mu}_{\alpha+\beta} &= \frac{1}{2}(\mathbf{p}_\alpha + \mathbf{p}_\beta),\\ &= (\mu, 1 - \mu),\end{aligned}$$

with the parameters $\delta = \frac{p_\beta - p_\alpha}{2}$ and $\mu = \frac{p_\alpha + p_\beta}{2}$; note that the alternative basis $\boldsymbol{\delta}_{\alpha-\beta}$ and $\boldsymbol{\mu}_{\alpha+\beta}$ would also suffice. Then, variation is represented by the two-parameter vector $(\delta, \mu)$, reflecting the fact that there are two degrees of freedom. Using the relations $\alpha = \frac{p_\alpha}{1-p_\alpha}$, $\beta = \frac{p_\beta}{1-p_\beta}$, $p_\alpha = \mu - \delta$ and $p_\beta = \mu + \delta$, we obtain

$$\begin{aligned}\omega &= \frac{p_\beta - p_\alpha p_\beta}{p_\alpha - p_\alpha p_\beta},\\ &= \frac{\delta^2 + \mu(1 - \mu) + \delta}{\delta^2 + \mu(1 - \mu) - \delta}.\end{aligned} \tag{3}$$

Then, we introduce Yule's $Q$ [1] to obtain

$$\begin{aligned} Q &= \frac{\omega - 1}{\omega + 1}, \\ &= \frac{\delta}{\delta^2 + \mu(1-\mu)}. \end{aligned} \tag{4}$$

Similarly, the relative risk is expressed as

$$\begin{aligned} \rho &= \frac{p_\beta}{p_\alpha}, \\ &= \frac{\mu + \delta}{\mu - \delta}, \end{aligned} \tag{5}$$

and the ratio difference is expressed as

$$\begin{aligned} \beta - \alpha &= \frac{p_\beta - p_\alpha}{(1-p_\beta)(1-p_\alpha)}, \\ &= \frac{\delta}{1 + \mu(\mu - 2) - \delta^2}. \end{aligned} \tag{6}$$

Inspection of Eqs (3–6) shows that the odds ratio and relative risk correspond to linear fractional transformations of proportional variation, and an effect size statistic corresponds to a perspective function $P((\delta, \mu), t) = (\delta/t, \mu/t)$, where $t$ is a polynomial function of $\delta$ and $\mu$. However, algebraic considerations alone are not sufficient to explain why a particular form might be preferred for $t$ or to provide operational interpretations for the different perspective normalizations in Eqs (4–6). In his 1912 paper, Yule remarked that the $Q$ coefficient has the merit of possessing a simple form "but the demerit of not possessing an equal simplicity of interpretation" [1]. Given the lack of an interpretation for the different normalizations, we find that Yule's remark also extends to the odds ratio and relative risk. Furthermore, rearranging Eqs (4) and (5) gives the corresponding relations

$$\delta^2 - \frac{\delta}{Q} + \mu(1-\mu) = 0, \tag{7}$$

$$\delta(\rho + 1) - \mu(\rho - 1) = 0, \tag{8}$$

with $0 \leq \mu - \delta \leq 1$ and $0 \leq \mu + \delta \leq 1$. Each of the four forms of proportional variation identified in the section 1.3 satisfies these relations. Thus, there are a range of values of $(\delta, \mu)$ for a fixed value of either $Q$, or $\rho$ (Fig 2). This ambiguity in proportional effects explains why the question of the merits of the odds ratio versus relative risk is still not resolved [18, 19]. A more precise approach would take into account the two-dimensional nature of the proportional variation, which could involve separate thresholds for $\delta$ and $\mu$. In any case, the specification of a perspective function should be based on the assessment of cost-benefit trade-offs for variations in $\delta$ and $\mu$, which will depend on the particular application.

## 1.3 Decomposition of proportional variation for a 2 × 2 contingency table

In this section, the two-component framework is used in the analysis of proportional variation for a 2 × 2 table (Table 1). We are particularly concerned with the confounding effect of the row and column sums in the formulation of association measures [2, 5, 11]. Each marginal sum corresponds to a categorical sample size that is determined by experimental procedure.

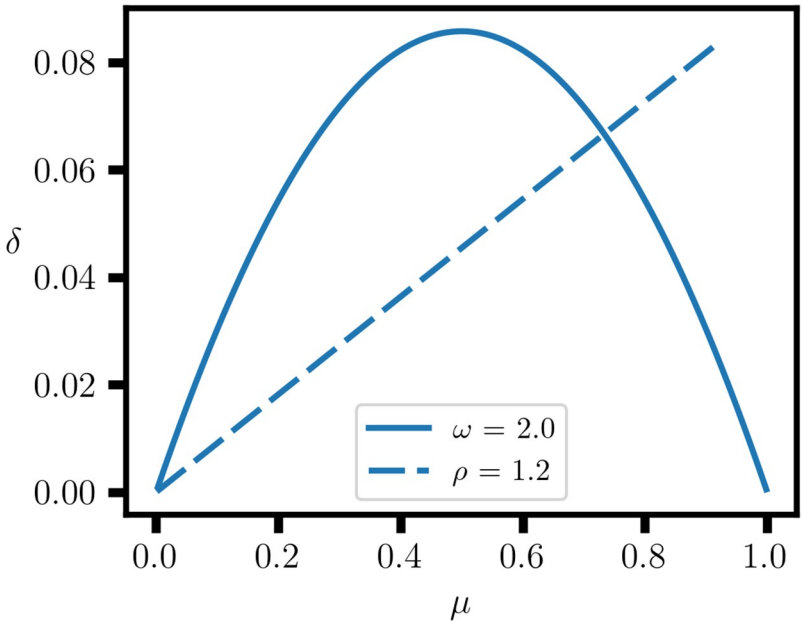


**Fig 2. Center-of-mass coordinates for a two-component biproportional system.** In the $\triangle^1$ representation, the center-of-mass coordinates are $\mu = (p_\alpha + p_\beta)/2$ and $\delta = (p_\beta - p_\alpha)/2$. The proportional variation, $(\mu, \delta)$, for fixed odds ratio, $\omega = 2$, and relative risk, $\rho = 1.2$, are shown. The odds ratio and relative risk are perspective functions of the center-of-mass coordinates.

https://doi.org/10.1371/journal.pone.0224460.g002

Suppose the first row of Table 1 is multiplied by a number $k$ to reflect a change in sample size; then, $(a, b) \mapsto (ka, kb)$. Then, the simple matching coefficient [11], $s_M$, is expressed as

$$s_M = \frac{ka + d}{k(a + b) + c + d},$$

which is not invariant to scaling by $k$. Alternatively, each marginal sum serves as a proportional normalization factor; e. g., $P((a, b), a + b)$. Then, $s_M$ can be expressed as the weighted sum of proportions

$$\begin{aligned} s_M &= \frac{a + c}{a + b + c + d}\frac{a}{a + c} + \frac{b + d}{a + b + c + d}\frac{d}{b + d} \\ &=: x_{a+c}p_{a/c} + x_{b+d}p_{d/b} \end{aligned} \tag{9}$$

**Table 1. The 2 × 2 contingency table.**

| | Column 1 | Column 2 | Row sum |
|---|---|---|---|
| **Row** 1 | a | b | $a + b$ |
| **Row** 2 | c | d | $c + d$ |
| **Column sum** | $a + c$ | $b + d$ | |

Observed counts, $(a, b, c, d)$, of the joint occurrence of categorical events are listed in the table. The sample size parameters for the classifications appear as the row sums, $(a + b, c + d)$, and column sums, $(a + c, b + d)$. The column and row sums are linearly related, and each sum serves as a scale factor for proportion.

https://doi.org/10.1371/journal.pone.0224460.t001

**Table 2. Diagonal scaling invariance of the odds ratio.**

| | Column 1 | Column 2 | Row sum |
|---|---|---|---|
| **Row 1** | $ka$ | $\frac{b}{j}$ | $ka+\frac{b}{j}$ |
| **Row 2** | $jc$ | $\frac{d}{k}$ | $jc+\frac{d}{j}$ |
| **Column sum** | $ka+jc$ | $\frac{b}{j}+\frac{d}{k}$ | |

The odds ratio does not distinguish between rows and columns, $\omega=\frac{a}{b}\frac{d}{c}=\frac{a}{c}\frac{d}{b}$, so the odds ratio is invariant to unitary scaling of the diagonal elements with $j>0, k>0$.

https://doi.org/10.1371/journal.pone.0224460.t002

$$\begin{aligned}&=\frac{a+b}{a+b+c+d}\frac{a}{a+b}+\frac{c+d}{a+b+c+d}\frac{d}{c+d}\\ &=:x_{a+b}p_{a/b}+x_{c+d}p_{d/c}\end{aligned}\tag{10}$$

for columns or rows, respectively. The proportions are invariant to scaling of either rows or columns, but the corresponding weights ($x_i$) are not because the overall sum, $a+b+c+d$, does not distinguish between row or column sums. Therefore, $s_M$ can differ between two tables because of differences in sample size even though the underlying system response properties might be unchanged. Warrens [11] has shown that members of the general family of coefficients that are linear transformations of $s_M$ do not satisfy the criteria for a well-behaved coefficient. As discussed by Goodman and Kruskal [2], dependence on sample size parameters complicates the interpretation of association coefficients. The concepts discussed in this paper support their proposal that normalization to adjust for unbalanced sample sizes is necessary.

The invariance of the odds ratio to scaling of either rows or columns is expressed as

$$\omega\frac{bc}{k}-\frac{ad}{k}=0,\tag{11}$$

$k>0$. This expression remains valid if either $bc\mapsto cb$ or $ad\mapsto da$. Thus, the odds ratio does not distinguish between ratios for rows and columns, $\omega=\frac{a}{b}\frac{d}{c}=\frac{a}{c}\frac{d}{b}$ [18, 20], which introduces ambiguity with respect to proportional effects. Consider the equivalence class of tables obtained by unitary scaling of the diagonal elements ('u-scaling'),

$$\omega\frac{b}{j}\frac{c}{j^{-1}}-\frac{a}{k^{-1}}\frac{d}{k}=0,$$

with $j, k>0$ (Table 2). The two numerical examples of such tables shown in Fig 3 demonstrate that while the odds ratio and $Q$ are invariant, the proportions are not. Furthermore, in the

| Contingency Table<br>a b<br>c d | Column Proportions | $\frac{1}{2}(p_{a/c}-p_{b/d})$ | $\frac{1}{2}(p_{a/c}+p_{b/d})$ | Odds Ratio $\frac{ad}{bc}$ | Q |
|---|---|---|---|---|---|
| **320 40**<br>**15 25** | 0.955 0.615<br>0.045 0.385 | 0.17 | 0.78 | 13.33 | 0.86 |
| **80 20**<br>**30 100** | 0.727 0.167<br>0.273 0.833 | 0.28 | 0.45 | 13.33 | 0.86 |

**Fig 3. Contingency tables with fixed odds ratio.** While the odds ratio, $\omega=\frac{ad}{bc}$, is fixed in these tables, the proportions are not. The Yule $Q$ statistic is also invariant because it is related to $\omega$ by the linear fractional transformation $Q=\frac{\omega-1}{\omega+1}$.

https://doi.org/10.1371/journal.pone.0224460.g003

**Table 3. The Yule symmetric table.**

| | Column 1 | Column 2 | Row sum |
|---|---|---|---|
| **Row 1** | $\sqrt{ad}$ | $\sqrt{bc}$ | $\sqrt{ad}+\sqrt{bc}$ |
| **Row 2** | $\sqrt{bc}$ | $\sqrt{ad}$ | $\sqrt{ad}+\sqrt{bc}$ |
| **Column sum** | $\sqrt{ad}+\sqrt{bc}$ | $\sqrt{ad}+\sqrt{bc}$ | |

The equivalence class for an odds ratio includes this symmetric table obtained by geometric averaging of the diagonal elements; $j=\sqrt{b/c}$ and $k=\sqrt{d/a}$ in Table 2. This results in the equalization of column and row sums and a loss of information.

https://doi.org/10.1371/journal.pone.0224460.t003

special case where $j=\sqrt{b/c}$ and $k=\sqrt{d/a}$, the row and column sums are equalized due to the geometric averaging of the diagonal elements, and the Yule symmetric table (Table 3) is obtained. This table serves as the basis for Yule's $\omega$ coefficient [1], also known as the coefficient of colligation [21]. However, row and column sums are linearly related by a column proportion matrix

$$\begin{pmatrix} a+b \\ c+d \end{pmatrix} = \begin{pmatrix} \frac{a}{a+c} & \frac{b}{b+d} \\ \frac{c}{a+c} & \frac{d}{b+d} \end{pmatrix} \begin{pmatrix} a+c \\ b+d \end{pmatrix}.$$

This linear relation is not preserved by u-scaling because of the mixing of effects between rows and columns (Table 2), so the odds ratio by itself is not suitable as an effect size measure. The linear relation also implies that row and column sums play equal roles as sample size parameters directly or indirectly, and that either rows or columns can be equalized, but not both simultaneously. It is necessary to choose between rows or columns in conditioning a contingency table for unbalanced sample sizes.

A self-consistent representation of proportional variation must account for the scaling invariance of the odds ratio. Therefore, our objective is to obtain a decomposition of the odds ratio in terms of elementary proportions by conditioning for the effect of the marginal sums. Consider scaling of the expression $\omega bc - ad = 0$ by column sums to obtain the fractional representation

$$\omega \frac{b}{n_1(b+d)}\frac{c}{n_2(a+c)} - \frac{a}{n_1(a+c)}\frac{d}{n_2(b+d)} = 0, \tag{12}$$

where $n_1$ and $n_2$ are normalization factors for the subsequent conversion to proportion vectors. Since there are two ways to express the odds ratio as a product of ratios, there are also two ways to group the fractional products to form proportion vectors. The standard grouping is formed from the columns of the table with $n_1 = n_2 = 1$ to obtain the two vectors

$$\left(\frac{a}{a+c}, \frac{c}{a+c}\right), \left(\frac{b}{b+d}, \frac{d}{b+d}\right). \tag{13}$$

However, we also obtain a second pair of vectors formed from the rows with $n_1 = \frac{a}{a+c} + \frac{b}{b+d}$

and $n_2 = \frac{c}{a+c} + \frac{d}{b+d}$ yielding

$$\left( \frac{a}{a + b\frac{a+c}{b+d}}, \frac{b}{b + a\frac{b+d}{a+c}} \right), \left( \frac{c}{c + d\frac{a+c}{b+d}}, \frac{d}{d + c\frac{b+d}{a+c}} \right). \tag{14}$$

The proportions in both Eqs (13) and (14) are invariant to scaling of columns, as required. The second form of proportional variation corresponds to an effect size measure with the normalization needed for experimental work, and has not been previously mentioned in the effect size literature to the best of my knowledge. Proportion vectors invariant to the scaling of rows are obtained in a similar way. A more concise way to obtain the proportion vectors is to observe that a matrix can be factored as a product of a diagonal column sum ($\mathbf{M}_{\text{csum}}$) or a row sum ($\mathbf{M}_{\text{rsum}}$) and proportion matrices, $\mathbf{P}_{\text{csum,c|r}}$ or $\mathbf{P}_{\text{rsum,c|r}}$, respectively.

$$\begin{aligned} \begin{pmatrix} a & b \\ c & d \end{pmatrix} &= \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix} \begin{pmatrix} \frac{a}{n_1(a+c)} & \frac{b}{n_1(b+d)} \\ \frac{c}{n_2(a+c)} & \frac{d}{n_2(b+d)} \end{pmatrix} \begin{pmatrix} a+c & 0 \\ 0 & b+d \end{pmatrix}, \\ &= \mathbf{N}_{\text{csum,c|r}} \mathbf{P}_{\text{csum,c|r}} \mathbf{M}_{\text{csum}}, \end{aligned} \tag{15}$$

$$\begin{aligned} &= \begin{pmatrix} a+b & 0 \\ 0 & c+d \end{pmatrix} \begin{pmatrix} \frac{a}{n_1(a+b)} & \frac{b}{n_2(a+b)} \\ \frac{c}{n_1(c+d)} & \frac{d}{n_2(c+d)} \end{pmatrix} \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}, \\ &= \mathbf{M}_{\text{rsum}} \mathbf{P}_{\text{rsum,c|r}} \mathbf{N}_{\text{rsum,c|r}}. \end{aligned} \tag{16}$$

The $\mathbf{N}_{\text{csum,c|r}}$ and $\mathbf{N}_{\text{rsum,c|r}}$ proportion normalization factors provide the different scaling structures (Eq 12) needed for column and row proportion matrices, which correspond to different projective representations of the relationship between variables (Fig 4). The standard protocol is to equalize the marginal sums for the response or dependent variable, and calculate the response effect size for variation of the treatment or independent variable. Depending on whether the response variable is listed in columns or rows, the corresponding representation would be either $\mathbf{P}_{\text{csum,r}}$ or $\mathbf{P}_{\text{rsum,c}}$, respectively. Examples of corresponding proportion difference measures, $\delta_{c,a-c}$ and $\delta_{r,a-b}$, are also shown in Fig 4. Our subscript notation is explained by the following example,

$$\begin{aligned} \delta_{r,a-b} &= \frac{1}{1 + \frac{c}{a}\frac{a+b}{c+d}} - \frac{1}{1 + \frac{d}{b}\frac{a+b}{c+d}}, \\ &= \frac{a}{a + c\frac{a+b}{c+d}} - \frac{b}{b + d\frac{a+b}{c+d}}. \end{aligned}$$

Thus, $\delta_{r,a-b}$ corresponds to the difference between '$a$' and '$b$' elements of the $\mathbf{P}_{\text{rsum,c}}$ proportion matrix. Then, calculation of an effect size requires the specification of a perspective function for mapping the relevant $(\delta, \mu)$ vector to $\mathbb{R}^1$ (Section 1.2). Proper practice also requires that an effect size estimate must be qualified by a confidence interval (Section 1.5).

| Scaling Invariance | Proportion Matrix | Proportion Difference |
|---|---|---|
| Rows | $\mathbf{P}_{\mathrm{rsum,r}} = \begin{pmatrix} \frac{1}{1+\frac{b}{a}} & \frac{1}{1+\frac{a}{b}} \\ \frac{1}{1+\frac{d}{c}} & \frac{1}{1+\frac{c}{d}} \end{pmatrix}$ | $\delta_{r,a-c} = \frac{1}{1+\frac{b}{a}} - \frac{1}{1+\frac{d}{c}}$ |
| | $\mathbf{P}_{\mathrm{rsum,c}} = \begin{pmatrix} \frac{1}{1+\frac{c}{a}\frac{a+b}{c+d}} & \frac{1}{1+\frac{d}{b}\frac{a+b}{c+d}} \\ \frac{1}{1+\frac{a}{c}\frac{c+d}{a+b}} & \frac{1}{1+\frac{b}{d}\frac{c+d}{a+b}} \end{pmatrix}$ | $\delta_{r,a-b} = \frac{1}{1+\frac{c}{a}\frac{a+b}{c+d}} - \frac{1}{1+\frac{d}{b}\frac{a+b}{c+d}}$ |
| Columns | $\mathbf{P}_{\mathrm{csum,r}} = \begin{pmatrix} \frac{1}{1+\frac{b}{a}\frac{a+c}{b+d}} & \frac{1}{1+\frac{a}{b}\frac{b+d}{a+c}} \\ \frac{1}{1+\frac{d}{c}\frac{a+c}{b+d}} & \frac{1}{1+\frac{c}{d}\frac{b+d}{a+c}} \end{pmatrix}$ | $\delta_{c,a-c} = \frac{1}{1+\frac{b}{a}\frac{a+c}{b+d}} - \frac{1}{1+\frac{d}{c}\frac{a+c}{b+d}}$ |
| | $\mathbf{P}_{\mathrm{csum,c}} = \begin{pmatrix} \frac{1}{1+\frac{c}{a}} & \frac{1}{1+\frac{d}{b}} \\ \frac{1}{1+\frac{a}{c}} & \frac{1}{1+\frac{b}{d}} \end{pmatrix}$ | $\delta_{c,a-b} = \frac{1}{1+\frac{c}{a}} - \frac{1}{1+\frac{d}{b}}$ |

**Fig 4. Four forms of proportional variation for a 2 × 2 table.** Separate proportion matrices are obtained in factoring a 2 × 2 matrix for scaling by the column sum (csum) or the row sum (rsum). Columns and rows of a proportion matrix correspond to different representations of the relationship between categorical variables.

https://doi.org/10.1371/journal.pone.0224460.g004

## 1.4 The $\phi$ coefficient

In this section, we discuss why $\phi$ does not serve as a well-behaved effect size measure and further explain the connection between $\delta_s$ and diagonally symmetric 2 × 2 tables. The $\phi$ coefficient is of particular importance in GWAS because it serves as a standard measure of linkage disequilibrium between molecular markers [3, 5]. The popularity of $\phi$ is due to its correspondence with Pearson's correlation coefficient. Binary {0, 1} representations are invoked for the categorical variables, then the correlation coefficient formula is applied to obtain

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, \tag{17}$$

which is also often referred to as '*r*'. However, the limitations of $\phi$ as an association measure

are well known [3, 5, 6, 11, 22]. Alternatively, $\phi$ is obtained from the relation with Pearson's chi-squared statistic, $\chi^2 = (a + b + c + d)\phi^2$ [3, 23], which also averages over rows and columns resulting in confounding effects. Introducing the ratio product for marginal sums,

$$\omega_M = \frac{(a+b)(c+d)}{(a+c)(b+d)}, \tag{18}$$

$\phi$ can be written as the row sum factorization

$$\begin{aligned}\phi &= \sqrt{\omega_M}\frac{ad-bc}{(a+b)(c+d)},\\ &= \sqrt{\omega_M}\left(\frac{a}{a+b}-\frac{c}{c+d}\right),\\ &= \frac{1}{\sqrt{(a+c)(b+d)}}\left(\sqrt{\frac{c+d}{a+b}}a-\sqrt{\frac{a+b}{c+d}}c\right),\end{aligned}$$

which corresponds to the scaling of $\mathbf{P}_{\text{rsum,r}}$

$$\sqrt{\omega_M}\begin{pmatrix}\frac{a}{a+b} & \frac{b}{a+b}\\ \frac{c}{c+d} & \frac{d}{c+d}\end{pmatrix} = \frac{1}{\sqrt{(a+c)(b+d)}}\begin{pmatrix}\sqrt{\frac{c+d}{a+b}}a & \sqrt{\frac{c+d}{a+b}}b\\ \sqrt{\frac{a+b}{c+d}}c & \sqrt{\frac{a+b}{c+d}}d\end{pmatrix}.$$

Therefore, $\phi$ corresponds to u-scaling of the $2 \times 2$ table with $j = \sqrt{\frac{a+b}{c+d}}$ and $k = \sqrt{\frac{c+d}{a+b}}$. Alternatively, the column sum factorization for $\phi$ is

$$\begin{aligned}\phi &= \frac{1}{\sqrt{\omega_M}}\left(\frac{a}{a+c}-\frac{b}{b+d}\right),\\ &= \frac{1}{\sqrt{(a+b)(c+d)}}\left(\sqrt{\frac{b+d}{a+c}}a-\sqrt{\frac{a+c}{b+d}}b\right),\end{aligned}$$

which corresponds to the u-scaling of the $2 \times 2$ table with $j = k = \sqrt{\frac{b+d}{a+c}}$. The following factorizations also hold:

$$\phi = \delta_{r,a-b}\frac{\left(a + c\frac{a+b}{c+d}\right)\left(b\frac{c+d}{a+b}+d\right)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, \tag{19}$$

$$=: \delta_{r,a-b}M_{r,a-b}, \tag{20}$$

$$= \delta_{c,a-c}\frac{\left(a + b\frac{a+c}{b+d}\right)\left(c\frac{b+d}{a+c}+d\right)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, \tag{21}$$

$$=: \delta_{c,a-c}M_{c,a-c}. \tag{22}$$

Consequently, each proportion difference, $\delta_i$, is associated with a factorization $\phi = M_i\delta_i$, where $M_i$ depends on marginal sums. Therefore, $\phi$ corresponds to a weighted average of the $\delta_i$. The multiplication of row and column sums together in each $M_i$ has a compounding effect because the sums are not independent.

Consider a diagonally symmetric 2 × 2 table with $d = a$ and $c = b$ in Table 1, and equal row and column sums. Then, Eq (12) becomes

$$\frac{\omega b^2 - a^2}{(a+b)^2} = 0,$$

which corresponds to the proportion vectors $\frac{1}{a+b}(a, b)$ and $\frac{1}{a+b}(b, a)$, and proportion difference $\delta_s$ (Eq 2). Since $\frac{1}{a+b}[(a, b) + (b, a)] = (1, 1)$, the $\triangle^1$ coordinates are $\left(\delta_s, \frac{1}{2}\right)$, so there is only one degree of freedom. Thus, there is a correspondence between 2 × 1 tables [23] and diagonally symmetric 2 × 2 tables. However, $M_i = 1$ for diagonally symmetric tables, and Eq (17) simplifies to give $\phi_s = \delta_s$. Thus, $\delta_s$ and $\phi_s$ are equivalent measures of proportional variation. Conversely, the $\delta_i$ in Fig 4 can be regarded as constituting an extension of $\phi_s$ to asymmetric tables. The $\phi$ coefficient per se does not account for the loss of symmetry when $M_i \ll 1$, because it does not distinguish between the $\delta_i$. However, when $M_i \approxeq 1$ the four expressions collapse into one or nearly so, and the values of $\phi$ and $\delta_i$ will be approximately the same. This includes the case where either $b = c = 0$ or $a = d = 0$ resulting in a diagonal 2 × 2 table. The connection with $\phi$ suggests that Cohen's recommendations of effect sizes of 0.1, 0.3 and 0.5 for small, medium, and large effects, respectively, for $\phi$ [6, 24] can also be invoked for the various forms of $\delta_i$, but this assumes that the $\mu_i$ coordinate is irrelevant.

## 1.5 Confidence interval for proportional effects

Each step of a data acquisition process is subject to stochastic effects, and data quality can vary between data sets. Therefore, the specification of a confidence interval (CI) for the effect size is an integral part of data analysis [25, 26]. A contingency table for experimental data is associated with a distribution of tables, $\mathcal{P}(\theta)$, and corresponding distributions for the effect size. The specification of $\mathcal{P}(\theta)$ must be based on a realistic assessment of all sources of error and uncertainty to form an error model for the data, $\mathcal{E}(\theta)$. For binary variables, a common approach is to estimate variance from a binomial distribution; the normal distribution is a useful approximation for large sample sizes. Then, estimating the CI for an effect size requires a propagation of error calculation, which is often not straightforward. Analytical approaches for estimating confidence intervals for ratios [27, 28], proportion and difference of two proportions [29, 30], correlation coefficients [31, 32], and odds ratios [9] are already quite involved. Fractional transformation, the bounded range, and the discrete properties of an effect size for proportional variation introduce complications that make it difficult to obtain convenient expressions for error propagation. Alternatively, Monte Carlo (MC) methods [33, 34] provide a more practical approach to estimate confidence intervals for quantities such as $\delta_{r,b-a}$ and $\delta_{c,c-a}$. In an MC simulation, a 2 × 2 MC table is obtained by generating the $N = a + b + c + d$ events by making random draws according to specified sample proportions [9] and $\mathcal{E}(\theta)$. A set of MC tables is obtained by repeating the sampling process many times; MC distributions are formed for proportions and effect size from the MC tables. Many MC runs are performed, collecting the relevant statistics for each MC distribution, including the mean, median, variance, and histogram. Finally, the degree of convergence for the MC simulation is estimated from the statistics for the MC runs. Fig 5A and 5C shows constrained MC simulations with fixed column sums $n_1 = a + c$ and $n_2 = b + d$ and sampling proportions $\frac{1}{a+c}(a, c)$ and $\frac{1}{b+d}(b, d)$, respectively. Fig 5B and 5D shows greater internal scatter because only the overall sum, $N$, is fixed, with corresponding sampling proportions $\frac{1}{N}(a, b, c, d)$. Even though the underlying distributions are discrete, the $\pm 2\sigma$ interval for a normal distribution serves as a good approximation for the $\delta_{c,c-a}$ confidence interval in this example. More generally, the distribution of effect size is asymmetric which would be represented by separate confidence intervals for positive and negative deviation from

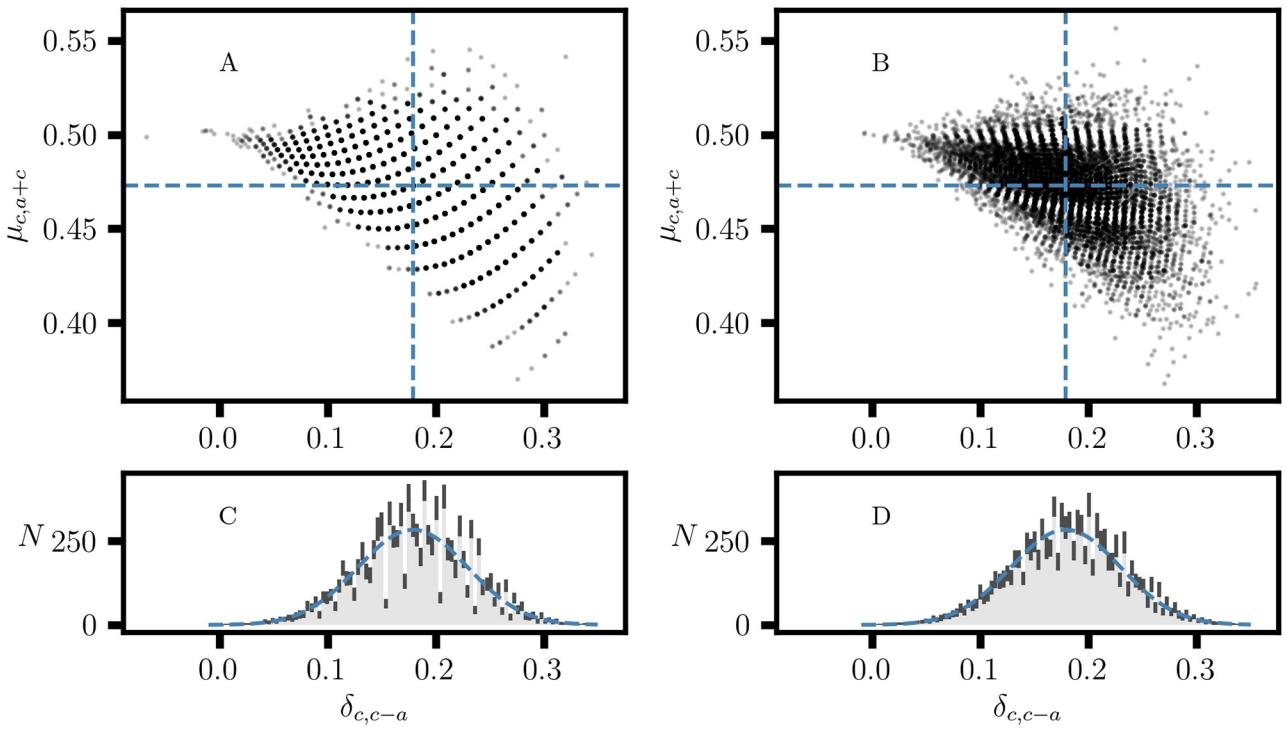


**Fig 5. Two sets of constrained Monte Carlo (MC) simulations of the distribution of proportional variation, ($\delta_{c,c-a}$, $\mu_{c,c+a}$), for a 2 × 2 table with *a*, *b*, *c*, *d* = [10, 30, 30, 20].** A,C: MC with fixed column sums, $n_1 = a + c$ and $n_2 = b + d$. B,D: MC with fixed overall sum $N = a + b + c + d$. A,B: Data for 10000 MC tables. The dashed lines indicate the expected values, $\delta_{c,c-a} = 0.18$ and $\mu_{c,a+c} = 0.473$. C,D: Each histogram is the mean of 64 MC runs with 10000 MC tables per run. Each whisker is the ±2 standard deviation interval. The normal distribution, $\mu = 0.18$ and $\sigma = 0.0506$, is shown as a dashed curve.

https://doi.org/10.1371/journal.pone.0224460.g005

the median. The advantage of the MC method is that the simulation can accommodate a detailed specification of $\mathcal{E}(\theta)$, including heteroscedasticity [25, 35] and correction for attenuation from misclassification [35, 36]. This capability is essential in accounting for the effects of instrumental and other operational factors on the quality of data produced by a data acquisition system.

## 1.6 Decomposition of proportional effects for an *r* × *c* table

A table with more than two rows or columns is commonly referred to as an $r \times c$ table. The matrix factorization (Eqs 15 & 16) extends in a straightforward way to produce the $r \times c$ proportion matrices. For independent and dependent variables with $r$ and $c$ categories, respectively, proportional variation is represented as $r$ points in the standard $\triangle^{c-1}$ simplex, with $r(c-1)$ degrees of freedom. Various multicategorical association measures have been proposed for $r \times c$ tables. However, we choose Cramer's $V^2$ [37, 38] as an example to illustrate the difficulties. $V^2$ is defined as a normalization of Pearson's $\chi^2$ such that $\chi^2 = n(q-1)V^2$, where $n$ is the total event count and $q = min(r, c)$. $V$ is equivalent to $\phi$ for $2 \times 2$ tables. Similarly, it is straightforward to show that Goodman and Kruskal's $\tau_c$ and $\tau_r$ [37] are both equivalent to $\phi^2$ for $2 \times 2$ tables. These equivalences confirm that Pearson's $\chi^2$, $V^2$, $\tau_c$ and $\tau_r$ are composite statistical quantities that average over alternative forms of variation and are therefore subject to ambiguous interpretation. The $\mathbb{R}^{r(c-1)} \mapsto \mathbb{R}^1$ mappings consist of multidimensional sums and products across rows and columns, resulting in confounding effects because of dependence between them.

In the absence of an engineering or functional model, the specification of a vector basis for proportional variation for an $r \times c$ table is not a well-posed problem [39]; i. e., there isn't a unique solution. This constitutes a fundamental limitation for the formulation of an effect size measure. Consider a two-component proportional system represented by vectors, $\mathbf{u}, \mathbf{v} \in \triangle^N$ with $N > 1$, and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{N+1}$. The two default center-of-mass vectors are $\boldsymbol{\mu} = (\mathbf{u} + \mathbf{v})/2$, and $\boldsymbol{\delta} = (\mathbf{u} - \mathbf{v})/2$. However, there isn't a standard procedure for choosing the additional $2N - 2$ vectors needed to form a complete basis. Alternatively, a single coordinate or a sum of coordinates could serve as the basis for estimating an effect size. This corresponds to choosing a $\triangle^1 \times \triangle^1$ subspace for the representation of proportional variation; e. g., $\delta = (u_i + u_j) - (v_i + v_j)$, with $\{(u_i + u_j, 1 - u_i - u_j), (v_i + v_j, 1 - v_i - v_j) \in \triangle^1\}$. A representation of the $2N$ degrees of freedom for a two-component $\triangle^N \times \triangle^N$ system would require the specification of $N$ $2 \times 2$ tables. Therefore, the $2 \times 2$ table serves an elementary role in the decomposition of multiproportional variation due to the minimal properties of $\triangle^1$. The recommended approach is to adopt a multidimensional representation of proportional variation and "reduce any multiple-level or multiple-variable relationship to a set of two-variable relationships" [25]. Similar advice has been given for avoiding the compounding effect of the ANOVA null hypothesis, to break down "complicated hypotheses into smaller, more easily understood pieces" [40]. Ways in which an $r \times c$ table might be partitioned and marginalized have been described by Kateri [41]. The objective is to construct a set of $2 \times 2$ tables that encompass relevant forms of proportional variation for the particular application. This multidimensional representation should be combined with the specification of cost-benefit trade-offs in assessing the effect size for proportional variation. In the next section, we discuss the use of $2 \times 2$ tables in the CART algorithm. However, high-dimensional search is still a developing area [42, 43], and a detailed assessment of the pros and cons for various approaches is beyond the scope of this paper.

## 1.7 Gini information gain and $\phi^2$

In this section, we examine connections between effect size and information gain (IG) measures used in standard implementations of the CART algorithm. CART creates a binary decision tree by the recursive partitioning of the association between response and independent variables [44–46]. Each node of the tree corresponds to a binary partition of the range of an independent variable. Each terminal node is a classification identified by a unique combination of intervals of the independent variables. In standard implementations, the partition parameters for a node are determined by maximizing IG for the response variable in an exhaustive search of associations over all independent variables. In each iteration, the set of statistics obtained for the binary partitions of an independent variable constitutes a CART association graph. Our objective is to compare CART graphs for effect sizes including IG. To simplify the discussion, we consider the case where the response variable is binary. Then, the data for a partition correspond to a $2 \times 2$ table [47]. Then, IG is defined as the parent node impurity, $I(S)$, minus the weighted impurities for the subnodes $I(S_1)$ and $I(S_2)$,

$$\mathrm{IG}(S_1, S_2) = I(S) - x_1 I(S_1) - x_2 I(S_2), \tag{23}$$

where the weight factor is $x_i = \frac{n_i}{n}$, $n_i$ is the number of elements in node $S_i$, and $n = n_1 + n_2$. Two popular impurity measures are the entropy, $E = -\Sigma p_j \ln p_j$, and Gini impurity, $G = 1 - \sum p_j^2$, where $p_j$ is the proportion of class '$j$' items in a set [16]. For a binary proportion vector, $(p_{m/n}, p_{n/m}) = \frac{1}{m+n}(m, n)$, and the Gini impurity becomes $G(p_{m/n}, p_{n/m}) = 2 p_{m/n}\, p_{n/m}$. However, the $x_i$ are subject to the same limitations as the weight factors for $s_M$ (Eqs 9 & 10), and both $\mathrm{IG}_E$ and $\mathrm{IG}_G$ depend on the marginal sums. More concretely, we show that $\mathrm{IG}_G$ and $\phi^2$ are equivalent in CART. Let the rows and columns of Table 1 correspond to the subnodes and categories

**Table 4. Classification tree partitions for NHC 'short-stay rehospitalized' data.**

| Impurity Measure | Split Value | $S_1$ N, (1 star, 5 star) | $S_2$ N, (1 star, 5 star) |
|---|---|---|---|
| $IG_G$, $IG_E$ | 22.0 | 1966, (0.41, 0.59) | 2077, (0.58, 0.42) |
| $\delta_{c,a-c}$, lower | 13.3 | 301, (0.30, 0.70) | 3742, (0.51, 0.49) |
| $\delta_{c,a-c}$, upper | 32.6 | 3874, (0.49, 0.51) | 169, (0.71, 0.29) |

CART association split value, sample size (N), and (1-star, 5-star) proportions for subnodes $S_1$ and $S_2$, for Gini ($IG_G$) and entropy ($IG_E$) information gain, and proportion difference ($\delta_{c,a-c}$).

https://doi.org/10.1371/journal.pone.0224460.t004

for the response variable, respectively. Then, $G(S)$ for the parent node is

$$G(S) = 2\frac{(a+c)(b+d)}{(a+b+c+d)^2}.$$

$G(S_1)$ and $G(S_2)$ are calculated from proportions for the row vectors $(a, b)$ and $(c, d)$, respectively. Then,

$$\begin{aligned} IG_G(S_1, S_2) &= \frac{2}{a+b+c+d}\left[\frac{(a+c)(b+d)}{a+b+c+d} - \frac{ab}{a+b} - \frac{cd}{c+d}\right], \\ &= G(S)\phi^2, \end{aligned} \tag{24}$$

with substitution of the $\phi$ coefficient from Eq (17). Since $G(S)$ is a constant for binary partitions at a parent node, we conclude that $IG_G$ is equivalent to $\phi^2$. This confirms that $IG_G$ depends on marginal sums due to the $x_i$, in which the normalization factor $N = a + b + c + d$ does not distinguish between rows and columns. Information gain measures of the form Eq (23) will be subject to this limitation, including $IG_E$. It is known that $IG_E$ and $IG_G$ yield very similar results in CART [48], which confirms that $IG_E$ is subject to dependence on marginal sums (Table 4). The limitations of $IG_G$ raise the question of whether the column sum invariant $\delta_{c,a-c}$ statistic might be more appropriate for CART, which we consider in the next section.

## 2 Data analysis and results

### 2.1 Data preparation

The Centers for Medicare and Medicaid (CMS) conduct regular inspections of nursing homes to assess compliance with regulations and survey residents to assess the quality of patient care. The CMS quality measures data and Five-Star rating assignments are publicly available from the Nursing Home Compare (NHC) website [49]. The analysis of NHC data is an important problem in itself [50–52] and is the subject of our ongoing work [53]. Nursing homes are dynamic systems where the measurement of performance is essential for managing cost, but this constitutes a complex problem for which there is not a unique or 'best' solution. The challenge is to develop data analysis methods that can help identify public health criteria for classifying the quality of patient care in nursing homes, or some approximation thereof. However, in this work our interest is limited to the comparison of CART association graphs for effect size measures. First Quarter, 2018 NHC data for eighteen quality measures were retrieved, selecting only those nursing homes with either a 1 star or 5 star overall rating, corresponding to 1394 and 2649 nursing homes, respectively. Selecting '1 star, 5 star' rating data creates a binary response data set, which is convenient for our purpose; otherwise, data for all five ratings would be included in the CART analysis. The distributions of NHC 'Percentage of short-stay residents who were rehospitalized after a nursing home admission' (Rehospitalized)

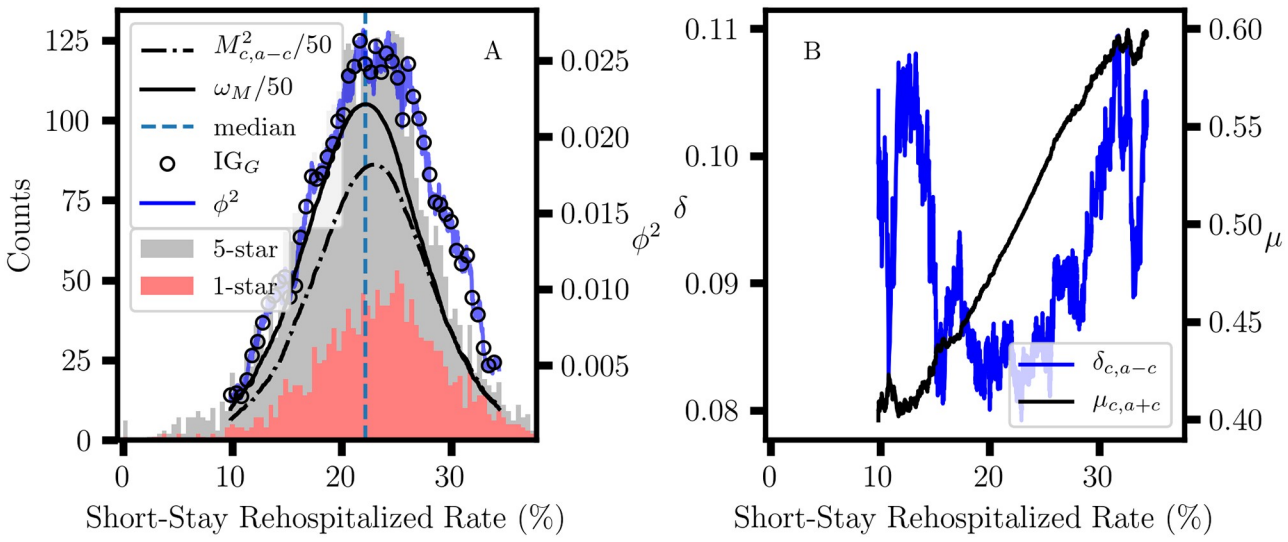


**Fig 6. CART association graph.** A: Stacked histograms for First Quarter, 2018 'Percentage of short-stay residents who were rehospitalized after a nursing home admission' data for nursing homes with a 1 star or 5 star overall rating; the dashed line is the median value. CART associations between Nursing Home Compare 'Rehospitalized' quality measure and '1 star, 5 star' overall rating for $IG_G$ and $\phi^2$ are also shown. $IG_G$ was scaled to match $\phi^2$. Both $M^2_{c,a-c}$ and $\omega_M$ were scaled by 1/50. B: Column scaling invariant center-of-mass coordinates, $(\delta_{c,a-c}, \mu_{c,a+c})$, for the two-component proportional variation in the standard one-simplex, $\triangle^1$.

https://doi.org/10.1371/journal.pone.0224460.g006

quality measure data for 1 star and 5 star overall ratings are broad and largely overlap (Fig 6A). This result implies that the $M_i$ for the corresponding contingency tables will tend to be much less than 1, as required for our demonstration.

## 2.2 Effect size in CART

In demonstrating the marginal sum dependence of various effect size measures, we must choose an elementary contingency table analysis problem. CART analysis for a binary response variable (bCART) is well suited for this purpose. In searching for an optimal binary partition of an independent variable, bCART generates a set of $2 \times 2$ tables where the sample sizes, $n_1$ and $n_2$, of the two subnodes vary over almost the entire range of the fixed sum $N = n_1 + n_2$; a minimum size is usually specified because a partition where either of the subnodes is too small is not informative. We let the rows and columns of Table 1 correspond to the two subnodes and the '1 star, 5 star' rating for the response variable, respectively. Effect size results for a bCART scan for association between the Rehospitalized quality measure and NHC '1 star, 5 star' overall rating are shown in Fig 6. The exact match between $IG_G$ and $\phi^2$ (Fig 6A) is consistent with Eq (24) because $G(S)$ is constant. The parabolic variation of $\phi^2$ is explained by Eq (21) because the variation in the marginal sum factor, $M_{c,a-c}$, outweighs the much smaller variation in the proportional effect, $\delta_{c,a-c}$ (Fig 6B). The parabolic variation of $M^2_{c,a-c}$ is in turn explained by the approximate similarity with $\omega_M$. Replacing each marginal sum in Eq (18) by the corresponding proportion yields

$$\begin{aligned} \omega_M &= \frac{p_{a+b}p_{c+d}}{p_{a+c}p_{b+d}}, \\ &= \frac{p_{a+b}(1-p_{a+b})}{p_{a+c}(1-p_{a+c})}. \end{aligned}$$

The denominator corresponds to the binomial variance for the parent set, which is constant.

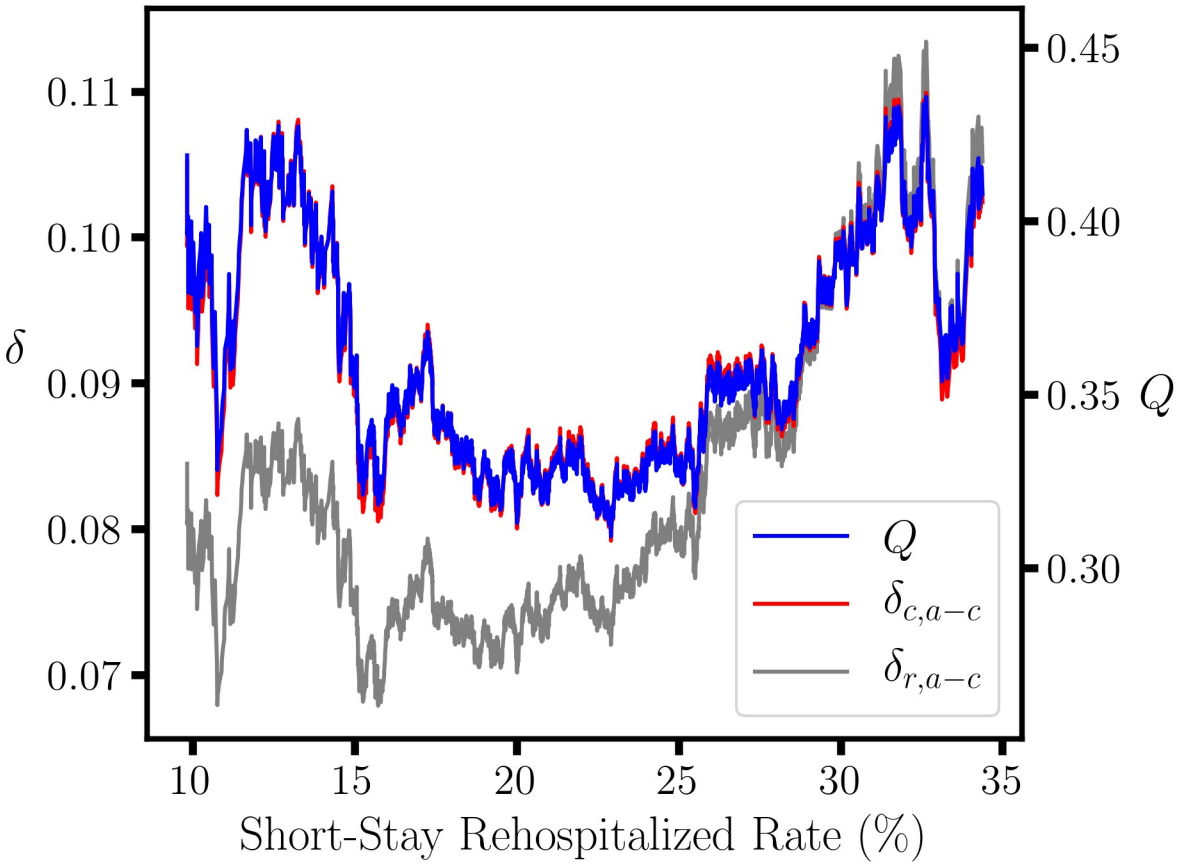


**Fig 7. Scaling invariant effect size statistics for CART.** Yule's $Q$ and $\delta_{c,c-a}$ yield similar results in the CART association between the First Quarter, 2018 Nursing Home Compare 'Rehospitalized' data and '1 star, 5 star' overall rating. $\delta_{c,c-a}$ and $\delta_{r,c-a}$ are the column and row scaling invariant proportion differences, respectively. $Q$ is invariant to scaling of either columns or rows.

https://doi.org/10.1371/journal.pone.0224460.g007

The numerator corresponds to the binomial variance for subnode size proportions, $(a + b)$: $(c + d)$, so $\omega_M$ has a maximum when $a + b = c + d$, which coincides with the median Rehospitalized value. Consequently, the parabolic dependence of $\phi^2$, with the maximum near the median value, largely reflects the variation in the subnode sample size instead of '1 star, 5 star' composition. In contrast, $\delta_{c,a-c}$ is column sum invariant and yields very similar results to Yule's $Q$, which is invariant to scaling of either rows or columns (Fig 7); the correlation is higher than 0.99 for 16 NHC quality measures, and the lowest is 0.91. Note that this similarity does not represent a special relation with $Q$ and results from the numerical properties of $\delta_{c,a-c}$ and $\mu_{c,a+c}$ for these data (Eq 4). The lower correlation ($r = 0.78$) between $\delta_{c,a-c}$ and $\delta_{r,a-c}$ confirms that different forms of proportional variation can be distinguished; $\delta_{r,a-c}$ also measures the difference in subnode composition but is row sum invariant. The U-shaped $\delta_{c,a-c}$ association graph has two maxima, so there are two possible CART partitions (Table 4). The relatively small subnode with Rehospitalized below 13.3% is enriched in the 5 star rating, corresponding to better than average patient care. Above 32.6%, the patient care is worse than average because it is associated with enrichment of the 1 star rating. The middle range from 13.3-32.6% includes the majority of nursing homes with average performance. In comparison, $IG_G$ and $IG_E$ produce subnodes that are nearly equal in size and with much lower degrees of enrichment in the '1 star, 5 star' proportions. Thus, $\delta_{c,a-c}$ is more effective than $IG_G$ and $IG_E$ in identifying partitions that correspond to a difference in the '1 star, 5 star' composition.

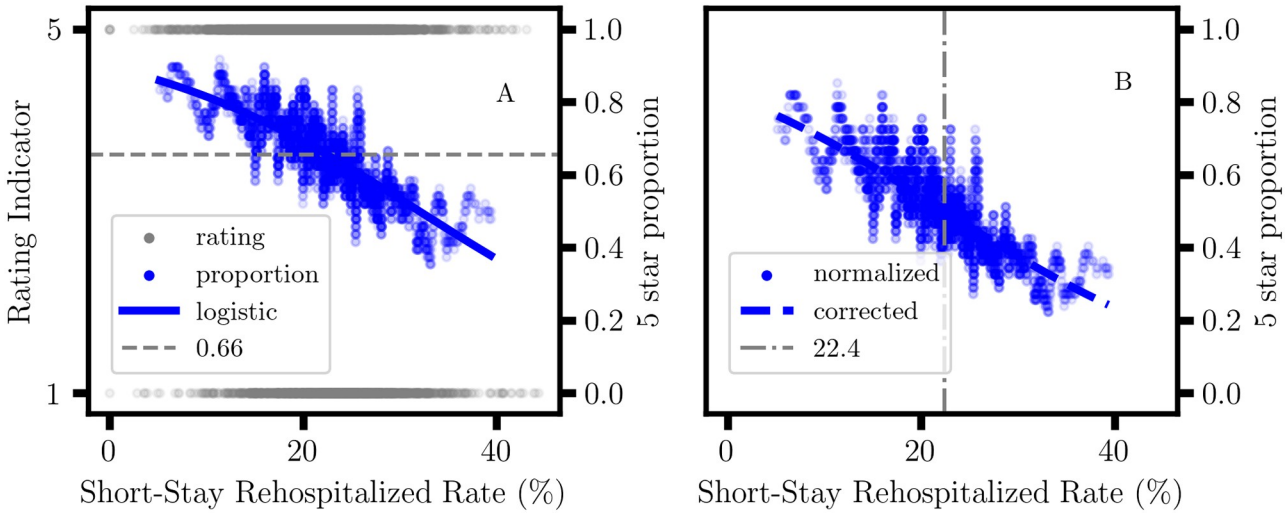


**Fig 8. Sample size effects in logistic regression.** A: Logistic model for Nursing Home Compare 'Rehospitalized' data and '1 star, 5 star' overall rating. The moving average '5 star' proportion is included for reference. The '5 star' sample size proportion, $\frac{n_5}{n_1+n_5} = 0.66$, is shown as a horizontal line; $n_1$ = 1394, and $n_5$ = 2649. B: Normalized '5 star' proportion adjusted for unequal sample sizes, and the adjusted logistic curve. The midpoint value, $x_0$ = 22.4, for the logistic curve is shown as a vertical line; the 'Rehospitalized' median is 22.2.

https://doi.org/10.1371/journal.pone.0224460.g008

The logistic regression method provides a graphical view of the effect of sample size parameters on proportional variation in categorical data analysis (Fig 8A). The '1 star, 5 star' rating data were analyzed using the LogisticRegression function in the scikit-learn library with the 'lbfgs' solver [54]. A moving average of the '5 star' rating proportion is included in the graph as a reference for the logistic curve. The normalized '5 star' proportion adjusted for inequality in the '1 star, 5 star' sample sizes and the corresponding adjusted logistic curve are shown in (Fig 8B). The variation in proportion confirms that the left and right tails of the Rehospitalized distribution correspond to nursing homes with above and below average performance, respectively, consistent with the CART association results. The logistic model for the '5 star' proportion, $y = \frac{c_5}{c_1+c_5}$, is usually expressed as

$$y = \frac{1}{1 + e^{-(a+bx)}}, \tag{25}$$

where parameters, $(a, b)$, are determined from the curve fit. The adjustment for the logistic curve was obtained using the change in coordinates

$$a = -bx_0 - \ln\left(\frac{n_1}{n_5}\right),$$

where $n_1$ and $n_5$ are the sample sizes for the 1 star and 5 star ratings in the data set, respectively. Substitution into Eq (25) yields

$$y = \frac{1}{1 + \frac{n_1}{n_5} e^{-b(x-x_0)}}$$

such that $y(x_0) = \frac{n_5}{n_1+n_5}$. In a data set where $n_1 = n_5$, $y(x_0) = 1/2$, and $x_0$ correspond to the midpoint value for the logistic curve. Then, there are two sample-size-independent parameters, $b$ and $x_0$.

## 3 Discussion

The renewed warnings from the statistics community about the limitations of statistical significance methodology has created a perplexing situation, given that there is a wide range of opinion on the underlying causes and solutions [55, 56]. Claims have also been made about effect size [25, 26, 57] as a better alternative, but the lack of consensus on the utility of commonly used association coefficients, such as the odds ratio [8, 10], the simple matching coefficient and $\phi$ [5, 11], hinders development of this approach. In this paper, we describe a rigorous framework for representing proportional variation in a $2 \times 2$ table, which helps in resolving the marginal sum dependence problem for association coefficients. We show that a $2 \times 2$ table is associated with four forms of proportional variation resulting from the factorization as a product of proportion and diagonal row or column sum matrices. Association coefficients, such as $\phi$, the odds ratio, and the simple matching coefficient, which do not distinguish between rows or columns, correspond to averages of proportional effects and lack clear interpretation. The two-component structure implies that there are two degrees of freedom corresponding to the displacement of two point vectors in the standard one-simplex, $\triangle^1$. An effect size measure then requires the specification of a perspective function of the center-of-mass coordinates, $(\delta, \mu)$, which is potentially unique for each application because of differences in cost-benefit trade-offs. In practice, classification problems vary widely in difficulty depending on the degree of overlap between the underlying distributions. Fisher's irises data set [58] is an example of a classification problem for well separated distributions, where different association coefficients achieve similar results because of degeneracy, particularly when the $2 \times 2$ table is diagonally symmetric or the effects are highly correlated. Conversely, differences in performance between association coefficients are best observed when the underlying distributions overlap. We also show that both Gini and entropy information gain are subject to dependence on marginal sums, which degrades the performance of the CART algorithm. Alternatively, the proportion difference with marginal sum invariance for the response variable provides a significant improvement in the performance of the CART algorithm. We conclude that the results in this paper demonstrate that equalization of either row or column sums of a $2 \times 2$ table serves as a correction for unbalanced sample sizes, as suggested by Goodman and Kruskal [2].

## Acknowledgments

It is a pleasure to acknowledge helpful discussions and suggestions from many colleagues in the DuPont Genetic Discovery group, particularly Ada Ching, Antoni Rafalski, and Scott Tingey. I also thank my colleagues at the Science, Technology and Research Institute of Delaware for their support, and Open Data Delaware for supporting the development of the NursingHomeMeasures.com website.

## Author Contributions

**Conceptualization:** Stanley Luck.

**Formal analysis:** Stanley Luck.

**Methodology:** Stanley Luck.

**Software:** Stanley Luck.

**Visualization:** Stanley Luck.

**Writing – original draft:** Stanley Luck.

**Writing – review & editing:** Stanley Luck.

## References


1. Yule GU. On the Methods of Measuring Association Between Two Attributes. Journal of the Royal Statistical Society. 1912; 75(6):579–652. https://doi.org/10.2307/2340126
2. Goodman LA, Kruskal WH. Measures of Association for Cross Classifications. J Amer Statis Assoc. 1954; 49:732–764. https://doi.org/10.1080/01621459.1954.10501231
3. Hedrick P. Gametic disequilibrium measures: proceed with caution. Genetics. 1987; 341:331–341.
4. Davenport EC, El-Sanhurry NA. Phi/Phimax: Review and Synthesis. Educational and Psychological Measurement. 1991; 51(4):821–828. https://doi.org/10.1177/0013164491051004003
5. VanLiere JM, Rosenberg NA. Mathematical properties of the $r^2$ measure of linkage disequilibrium. Theoretical Population Biology. 2008; 74(1):130–137. https://doi.org/10.1016/j.tpb.2008.05.006 PMID: 18572214
6. Olivier J, Bell ML. Effect Sizes for 2 × 2 Contingency Tables. PLoS ONE. 2013; 8(3):e58777. https://doi.org/10.1371/journal.pone.0058777 PMID: 23505560
7. Haddock CK, Rindskopf D, Shadish WR. Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. Psychological Methods. 1998; 3(3):339–353. https://doi.org/10.1037/1082-989X.3.3.339
8. Kraemer HC. Reconsidering the odds ratio as a measure of 2 × 2 association in a population. Statistics in Medicine. 2004; 23(2):257–270. https://doi.org/10.1002/sim.1714 PMID: 14716727
9. Ruxton GD, Neuhäuser M. Review of alternative approaches to calculation of a confidence interval for the odds ratio of a 2 × 2 contingency table. Methods in Ecology and Evolution. 2013; 4(1):9–13. https://doi.org/10.1111/j.2041-210x.2012.00250.x
10. Grant RL. Converting an odds ratio to a range of plausible relative risks for better communication of research findings. BMJ. 2014; 348(jan24 1):f7450–f7450. https://doi.org/10.1136/bmj.f7450 PMID: 24464277
11. Warrens MJ. On Association Coefficients for 2 × 2 Tables and Properties That Do Not Depend on the Marginal Distributions. Psychometrika. 2008; 73(4):777–789. https://doi.org/10.1007/s11336-008-9070-3 PMID: 20046834
12. Hubálek Z. Coefficients of Association and Similarity, Based on Binary (Presence-Absense) Data: An Evaluation. Biological Reviews. 1982; 57(4):669–689. https://doi.org/10.1111/j.1469-185X.1982.tb00376.x
13. Boyd SP, Vandenberghe L. Convex optimization. New York, NY: Cambridge University Press; 2004.
14. Beló A, Zheng P, Luck S, Shen B, Meyer DJ, Li B, et al. Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. Molecular Genetics and Genomics. 2008; 279(1):1–10. https://doi.org/10.1007/s00438-007-0289-y
15. Loh WY. Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011; 1(1):14–23.
16. Krzywinski M, Altman N. Points of Significance: Classification and regression trees. Nature Methods. 2017; 14(8):757–758. https://doi.org/10.1038/nmeth.4370
17. Reid M, Szendröi B. Geometry and Topology. New York: Cambridge University Press; 2005.
18. Bland JM, Altman DG. Statistics Notes: The odds ratio. BMJ. 2000; 320(7247):1468–1468. https://doi.org/10.1136/bmj.320.7247.1468 PMID: 10827061
19. Newcombe RG. A deficiency of the odds ratio as a measure of effect size. Statistics in Medicine. 2006; 25(24):4235–4240. https://doi.org/10.1002/sim.2683 PMID: 16927451
20. Sistrom CL, Garvan CW. Proportions, Odds, and Risk. Radiology. 2004; 230(1):12–19. https://doi.org/10.1148/radiol.2301031028 PMID: 14695382
21. Pearson K, Heron D. On Theories of Association. Biometrika. 1913; 9:159–315. https://doi.org/10.2307/2331805
22. Zysno PV. The modification of the phi-coefficient reducing its dependence on the marginal distributions. Methods of Psychological Research. 1997; 2(1):41–53.
23. Richardson JT. The analysis of 2 × 1 and 2 × 2 contingency tables: an historical review. Statistical Methods in Medical Research. 1994; 3(2):107–133. https://doi.org/10.1177/096228029400300202 PMID: 7952428
24. Cohen J. A power primer. Psychological Bulletin. 1992; 112(1):155–159. https://doi.org/10.1037//0033-2909.112.1.155 PMID: 19565683

25. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological reviews of the Cambridge Philosophical Society. 2007; 82(4):591–605. https://doi.org/10.1111/j.1469-185X.2007.00027.x PMID: 17944619

26. Cumming G. Understanding The New Statistics. New York, NY: Routledge; 2012.

27. Marsaglia G. Ratios of Normal Variables. Journal of Statistical Software. 2006; 16(4):1–10. https://doi.org/10.18637/jss.v016.i04

28. von Luxburg U, Franz VH. A Geometric Approach to Confidence Sets for Ratios: Fieller's Theorem, Generalizations, and Bootstrap. Statistica Sinica. 2009; 19:1095–1117.

29. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. Statistics in Medicine. 1998; 17(8):873–890. https://doi.org/10.1002/(sici)1097-0258(19980430)17:8<873::aid-sim779>3.0.co;2-i PMID: 9595617

30. Agresti A. Dealing with discreteness: making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact. Statistical Methods in Medical Research. 2003; 12(1):3–21. https://doi.org/10.1191/0962280203sm311ra PMID: 12617505

31. Banik S, Kibria BM. Confidence Intervals for the Population Correlation Coefficient $\rho$. International Journal of Statistics in Medical Research. 2016; 5(2):99–111. https://doi.org/10.6000/1929-6029.2016.05.02.4

32. Bishara AJ, Hittner JB. Confidence intervals for correlations when data are not normal. Behavior Research Methods. 2017; 49(1):294–309. https://doi.org/10.3758/s13428-016-0702-8 PMID: 26822671

33. Bevington PR, Robinson DK. Data Reduction and Error Analysis for the Physical Sciences. 3rd ed. New York, NY: McGraw-Hill; 2003.

34. Kroese DP, Brereton T, Taimre T, Botev ZI. Why the Monte Carlo method is so important today. Wiley Interdisciplinary Reviews: Computational Statistics. 2014; 6(6):386–392. https://doi.org/10.1002/wics.1314

35. Buonaccorsi JP. Measurement error: models, methods, and applications. Boca Raton: Chapman and Hall/CRC; 2010.

36. Höfler M. The effect of misclassification on the estimation of association: a review. International Journal of Methods in Psychiatric Research. 2005; 14(2):92–101. https://doi.org/10.1002/mpr.20

37. Berry KJ, Johnston JE, Mielke PW. A Measure of Effect Size for $R \times C$ Contingency Tables. Psychological Reports. 2006; 99(1):251–256. https://doi.org/10.2466/pr0.99.1.251-256 PMID: 17037476

38. Thomson G, Single RM. Conditional Asymmetric Linkage Disequilibrium (ALD): Extending the Biallelic $r^2$ Measure. Genetics. 2014; 198(1):321–331. https://doi.org/10.1534/genetics.114.165266 PMID: 25023400

39. Logan JD. Applied Mathematics. 2nd ed. New York, NY: John Wiley & Sons, Inc.; 1997.

40. Casella G, Berger R. Statistical Inference. 2nd ed. Pacific Grove, CA: Duxbury; 2002.

41. Kateri M. Contingency Table Analysis. New York, NY: Springer New York; 2014.

42. Kettenring JR. Coping with high dimensionality in massive datasets. Wiley Interdisciplinary Reviews: Computational Statistics. 2011; 3(2):95–103. https://doi.org/10.1002/wics.141

43. Coveney PV, Dougherty ER, Highfield RR. Big data need big theory too. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2016; 374(2080):20160153. https://doi.org/10.1098/rsta.2016.0153

44. Duda RO, Hart PE, Stork DG. Pattern classification. Wiley; 2001.

45. de Ville B. Decision trees. Wiley Interdisciplinary Reviews: Computational Statistics. 2013; 5(6):448–455. https://doi.org/10.1002/wics.1278

46. Loh WY. Fifty Years of Classification and Regression Trees. International Statistical Review. 2014; 82(3):329–348. https://doi.org/10.1111/insr.12016

47. Mingers J. An empirical comparison of selection measures for decision-tree induction. Machine Learning. 1989; 3(4):319–342. https://doi.org/10.1023/A:1022645801436

48. Krzywinski M, Altman N. Error bars. Nature Methods. 2013; 10(10):921–922. https://doi.org/10.1038/nmeth.2659 PMID: 24161969

49. Nursing Home Compare datasets; 2018. Available from: https://data.medicare.gov/data/nursing-home-compare.

50. Quartararo M, Glasziou P, Kerr CB. Classification Trees for Decision Making in Long-Term Care. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences. 1995; 50A(6):M298–M302. https://doi.org/10.1093/gerona/50A.6.M298

51. Alexander GL. An analysis of nursing home quality measures and staffing. Quality management in health care. 2008; 17(3):242–51. https://doi.org/10.1097/01.QMH.0000326729.78331.c5 PMID: 18641507

52. Raju D, Su X, Patrician PA, Loan LA, McCarthy MS. Exploring factors associated with pressure ulcers: A data mining approach. International Journal of Nursing Studies. 2015; 52(1):102–111. https://doi.org/10.1016/j.ijnurstu.2014.08.002 PMID: 25192963

53. Nursing Home Quality Measures; 2019. Available from: https://nursinghomemeasures.com/.

54. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12(Oct):2825–2830.

55. Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician. 2016; 70(2):129–133. https://doi.org/10.1080/00031305.2016.1154108

56. Leek J, McShane BB, Gelman A, Colquhoun D, Nuijten MB, Goodman SN. Five ways to fix statistics. Nature. 2017; 551(7682):557–559. https://doi.org/10.1038/d41586-017-07522-z PMID: 29189798

57. Grissom RJ, Kim JJ. Effect Sizes for Research. 2nd ed. New York, NY: Routledge; 2011.

58. Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics. 1936; 7(2):179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x