Selvitella *Journal of Statistical Distributions and Applications* (2017) 4:2
DOI 10.1186/s40488-017-0056-5

Journal of Statistical Distributions and Applications

RESEARCH Open Access

The ubiquity of the Simpson's Paradox

Alessandro Selvitella

Correspondence:
aselvite@math.mcmaster.ca
Department of Mathematics and Statistics of McMaster University, 1280 Main Street West, Hamilton, (ON) L8S-4K1, Canada

Abstract

The *Simpson's Paradox* is the phenomenon that appears in some datasets, where subgroups with a common trend (say, all negative trend) show the reverse trend when they are aggregated (say, positive trend). Even if this issue has an elementary mathematical explanation, it has a deep statistical significance. In this paper, we discuss basic examples in arithmetic, geometry, linear algebra, statistics, game theory, gender bias in university admission and election polls, where we describe the appearance or absence of the *Simpson's Paradox*. In the final part, we present our results concerning the occurrence of the *Simpson's Paradox* in Quantum Mechanics with focus on the Quantum Harmonic Oscillator and the Nonlinear Schrödinger Equation. We discuss how likely it is to incur in the *Simpson's Paradox* and give some concrete numerical examples. We conclude with some final comments and possible future directions.

Keywords: Simpson's Paradox, Quantum mechanics, Schrödinger Equation, Prisoner's Dilemma

Mathematics Subject Classification 2000: 35Q55, 34L40, 62H20, 62H17, 62P35

1 Introduction

In 1973, the Associate Dean of the graduate school of the University of California Berkeley worried that the university might be sued for sex bias in the admission process (Bickel et al. 1975). In fact, looking at the admission rates broken down by gender (male or female), we have the following contingency table:

Applicants	Admitted	Deny
Female	1494	2827
Male	3738	4704

The Chi-square statistics for this test has one degrees of freedom with value $\chi^2 = 111.25$ and corresponding p-value basically $= 0$, while the Chi-square statistics with Yates continuity correction for this test has a value of $\chi^2 = 110.849$ and corresponding p-value again approximately 0 (precision order 10^{-26}). A naïve conclusion would be that men were much more successful in admissions than women, which would clear be understood as a bad episode of gender bias. At that point, Prof. P.J.Bickel from the Department of Statistics of Berkeley, was asked to analyze the data.

In a famous paper (Bickel et al. 1975) with E.A.Hammel and J.W.O'Connell, P.J.Bickel studied the problem in detail. Graduate departments have independent admissions procedures and so they are autonomous for taking decisions in the graduate admission process. A further division in subgroups does not find a real counterpart in the structure of Berkeley's system. The analysis of the data, performed department by department, produces the following table:

Dpt	Male applications	Male admissions	Female applications	Female admissions
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	191	28%	393	24%

As Bickel, Hammel and O'Connell say in (Bickel et al. 1975), "The proportion of women applicants tends to be high in departments that are hard to get into and low in those that are easy to get into" and it is even more evident in departments with a large number of applicants. The examination of the aggregate data was showing a misleading pattern of bias against female applicants. However, if the data are properly pooled, and taking into consideration the tendency of women to apply to departments that are more competitive for either genders, there is a small but statistically significant bias in favour of women. The authors concluded that "Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation" (Bickel et al. 1975). This episode is one of the most celebrated real examples of what is called *Simpson's Paradox*: the trend of aggregated data might be reversed in the pooled data.

Note that the *Simpson's Paradox* is not confined to the discrete case, but it can appear also in the continuous case. Even if less famous, we want to mention the following example which has been discussed on the New York Times recently (Norris 2013). Still today, the *Simpson's Paradox* can be a source of confusion and misinterpretation of the data.

An article of the journalist F.Norris (2013) raised the concerns of readers, because of the following apparently paradoxical result. F.Norris analyzed the variation of the US wage over time. Accordingly to the statistics, from 2000 to 2013, the median US wage (adjusted for inflation) has risen of about 1%, if the median is computed on the full sample. However, if the same sample is broken down into four educational subgroups, the median wage (adjusted for inflation) of each subgroup decreased. The percentages of variation for each subgroup are summarized in the following table:

Group	Median change
Total	+0.9%
High School Dropouts	−7.9%
High School Graduates, No college	−4.7%
Some College	−7.6%
Bachelor's or Higher	−1.2%

Here, the reason of the reversal is that the relative sizes of the groups changed greatly over the period considered. In particular, there were more well-educated and so higher wage people in 2013 than in 2000.

In both the cases described above (discrete and continuous, respectively), the variables involved in the paradox are confounded by the presence of another variable (department and level of education, respectively).

The problem of the occurrence of this paradox was considered already in the 19th century. The first author which treated this topic has been Pearson (1899), followed by the contributions of Yule (Yule 1903; Yule and Kendall 1937) and Simpson (1951).

In his paper (Simpson 1951), Simpson considered a 2 × 2 × 2 contingency table with attributes A, B, and C and illustrated the paradox using a heuristic example of clinic patients, divided into a Treatment Group and in a No-Treatment group. The data were examined by gender and showed that both males and females responded favorably to the treatment, with respect to who did not receive the treatment. On the other side, the aggregated data showed an opposite behaviour, since there seemed to not be anymore any association between the use of the treatment and the survival time (see (Goltz and Smith 2010) for more details).

The name "*Simpson's Paradox*" was first used by Blyth (1972). Some authors prefer to not give full credit to Simpson, since he did not discover this phenomenon and to call it *Amalgamation Paradox* or *Yule-Simpson's Effect* instead.

In this paper, we outline that the *Simpson's Paradox* is not confined to statistical problems, but it is ubiquitous in science. We give a series of formal definitions in Section 2. In Section 3, we show the ubiquity of the *Simpson's Paradox* in several areas of technical and social sciences and we also give some examples of its occurrence. In Section 4, we outline our new result on the occurrence of the paradox in the context of Quantum Mechanics, with particular attention posed to the Quantum Harmonic Oscillator and to the Nonlinear Schrödinger Equation. We conclude with a brief discussion on how likely is the *Simpson's Paradox* in Quantum Mechanics (Section 5), with a numerical example (Section 6) and some final comments (Section 7).

Very few papers in the literature treat the *Simpson's Paradox* related to problems in *Quantum Mechanics*. At our knowledge, the only ones avaialable are the fast track communication by Paris (2012), an experimental result by Cialdi and Paris (2015), the preprint by Shi (2012) and a recent paper by the author (Selvitella 2017), which is the first paper that connects the *Simpson's Paradox* to Partial Differential Equations and Infinite Dimensional Dynamical Systems.

2 Measures of amalgamation

In this section, we give the definition and some popular examples of *Measures of Amalgamation*. For more details, we refer to (Good and Mittal 1987).

2.1 Definitions

First, we define the *Process of Amalgamation* of contingency tables $\mathbf{t}_i, i = 1, \dots, n$.

Definition 1 *Let* $\mathbf{t}_i = [a_i, b_i; c_i, d_i]$, $i = 1, \dots, l$ *be* 2×2 ***contingency tables*** *corresponding to the i-th of l mutually exclusive sub-populations, with* $a_i b_i c_i d_i \neq 0$. *Let* $N_i = a_i + b_i + c_i + d_i$ *denote the sample size for the i-th sub-population and let* $N = N_1 + \dots + N_l$ *be the total sample size of the population. If the n tables are added together, the process is called* ***Amalgamation****. We obtain a table* $\mathbf{T} := [A, B; C, D] := \left[\Sigma_{i=1}^{l} a_i, \Sigma_{i=1}^{l} b_i, \Sigma_{i=1}^{l} c_i, \Sigma_{i=1}^{l} d_i\right]$, *where* $A + B + C + D = N$.

After having amalgamated a group of contingency tables, we can define the *Measure of Amalgamation.*

Definition 2 *A function* $\alpha : M_{p\times p} \to \mathbb{R}$ *is called* ***Measure of Amalgamation****.*

Given the definition of *Measure of Amalgamation*, we can formally define the *Simpson's Paradox.*

Definition 3 *We say that the* ***Simpson's Paradox*** *occurs for the* ***Measure of Amalgamation*** α *if*

$$\max_i \alpha(\mathbf{t}_i) < \alpha(\mathbf{T}) or \min_i \alpha(\mathbf{t}_i) > \alpha(\mathbf{T}),$$

with α *defined on the set of contingency tables and real valued, as in Definition 2.*

We fix some terminology that we are going to use in the list of examples below in the context of contingency tables (see (Good and Mittal 1987)). Sampling Procedure *I*, called also *Tetranomial Sampling*, is performed when we sample at random from a population. Sampling Procedure II_R (respectively II_C), called also *Product-Binomial Sampling*, is performed when the row totals (respectively columns) is fixed and we sample until this marginal totals are reached. Sampling Procedure *III* controls both row and column totals.

2.2 Examples

Consider the contingency table $\mathbf{t} = [a, b; c, d]$, given by

$\mathbf{t} =$

	S	not S
T	a	b
not T	c	d

The following are popular examples of *Measures of Amalgamation* (see (Good and Mittal 1987)).

- The *Pierce's measure*:

  $$\pi_{Pearce}(\mathbf{t}) = \frac{a}{a+b} - \frac{c}{c+d}.$$

  Under *Tetranomial Sampling* and *Product-Binomial Sampling* with row fixed, this measure becomes

  $$\pi_{Pearce} = P(S|T) - P(S|\bar{T}).$$

  It compares the probability of an effect S under treatment and the probability of an effect S without any treatment (row categories are considered to be the "causes" of the column categories).
- The *Yule's measure* is given by the formula:

  $$\pi_{Yule}(\mathbf{t}) = \frac{ad - bc}{N^2}.$$

  It compares a/N with respect to its expected value under independence of rows and columns. In fact:

  $$\pi_{Yule}(\mathbf{t}) = \frac{ad - bc}{N^2} = \frac{a}{N} - \frac{(a+b)(a+c)}{N^2} = P(S \cap T) - P(S)P(T),$$

  since $N = a + b + c + d$.

- The *Odds Ratio* is probably the most popular one:

$$\pi_{Odds}(\mathbf{t}) = \frac{ad}{bc}.$$

The *Odds Ratio* is the ratio between the probability of success and the probability of failure, given a treatment or a no-treatment. In fact

$$\pi_{Odds}(\mathbf{t}) = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{\frac{a/(a+b)}{b/(a+b)}}{\frac{c/(c+d)}{d/(c+d)}} = \frac{P(S|T)/P(\bar{S}|T)}{P(S|\bar{T})/P(\bar{S}|\bar{T})}.$$

- The *Weight of Evidence* is given by:

$$\pi_{Weight_C}(\mathbf{t}) = \log\frac{a(b+d)}{b(a+c)}.$$

Under *Tetranomial Sampling* or column fixed *Product-Binomial Sampling*, the *Weight of Evidence* represents the logarithm of the *Bayes factor* in favour of *S*, knowing that the treatment was *T*, namely:

$$\pi_{Weight_C} = \log\frac{P(T|S)}{P(T|\bar{S})}.$$

- The *Causal Propensity*:

$$\pi_{Causal}(\mathbf{t}) = \log\frac{d(a+b)}{b(c+d)},$$

under *Tetranomial Sampling* or *Product-Binomial Sampling* with row fixed, represents the propensity of T causing S rather than $\bar{S}$:

$$\pi_{Causal}(\mathbf{t}) = \log\frac{P(\bar{S}|\bar{T})}{P(\bar{S}|T)}.$$

3 The *Simpson's Paradox* appears not just in statistics

In this section, we give very basic examples of the appearance of the *Simpson's Paradox* in fields different from statistics. In particular, we give examples in arithmetic, geometry, statistics, linear algebra, game theory and election polls.

- **Arithmetic**: There exist quadruplets $a_1, b_1, c_1, d_1 > 0$ and $a_2, b_2, c_2, d_2 > 0$ such that $a_1/b_1 > c_1/d_1$ and $a_2/b_2 > c_2/d_2$ but $(a_1 + a_2)/(b_1 + b_2) < (c_1 + c_2)/(d_1 + d_2)$. Example: $(a_1, b_1, c_1, d_1) = (2, 8, 1, 5)$ and $(a_2, b_2, c_2, d_2) = (4, 5, 6, 8)$. In this case, the *Measure of Amalgamation* is given by:

$$\pi(\mathbf{t}) = \frac{a}{b} - \frac{c}{d} = \frac{ad - bc}{bd}.$$

If we consider the contingency tables

$$\mathbf{t}_1 = [a_1, b_1; c_1, d_1]$$

and

$$\mathbf{t}_2 = [a_2, b_2; c_2, d_2]$$

and the amalgamated one:

$$\mathbf{T} = [a_1 + a_2, b_1 + b_2; c_1 + c_2, d_1 + d_2],$$

we have that:

$$\max_{i=1,2} \pi(\mathbf{a_i}) < 0 < \pi(\mathbf{T})$$

and so we have the *Simpson's Paradox*, accordingly to Definition 3.

- **Geometry**: Even if a vector v_1 has a smaller slope than another vector w_1, and a vector v_2 has a smaller slope than a vector w_2, the sum of the two vectors $v_1 + v_2$ can have a larger slope than the sum of the two vectors $w_1 + w_2$. Example: take $w_1 = (a_1, b_1)$, $v_1 = (c_1, d_1)$, $w_2 = (a_2, b_2)$, $v_2 = (c_2, d_2)$. The same *Measure of Amalgamation* of the previous example makes the game here as well.
- **Statistics**: A positive/negative trend of two separate subgroups might reverse when the subgroups are combined in one single group. This happens in both the discrete and continuous case. We gave examples of this in the introduction, with the Berkeley Gender Bias (discrete) case and the "time vs US wage" case (continuous).
- **Linear algebra** There exists $A_1, A_2 \in Mat_{n\times n}$ such that

$$det(A_1) > 0, \quad det(A_2) > 0, \quad but \quad det(A_1 + A_2) < 0.$$

Consider for example $A_1 = \mathbf{t}_1$ and $A_2 = \mathbf{t}_2$, as above.

- **Game theory**: The *Prisoner's Dilemma* shows why two players A and B might decide to not cooperate, even if it appears that, for both of them, it is more convenient to cooperate. If both A and B cooperate, they both receive a reward p_1. If B does not cooperate while A cooperates, then B receives p_2, while A receives p_3. Similarly, if viceversa. If both A and B do not cooperate, their payoffs are going to be p_4. To get the *Simpson's Paradox*, the following must hold:

$$p_4 = a_1/b_1 > p_2 = c_1/d_1 > p_3 = a_2/b_2 > p_1 = c_2/d_2.$$

Here $p_3 > p_1$ and $p_4 > p_2$, and $p_4 > p_3$ and $p_2 > p_1$ imply that it is better to not cooperate for both A and B both given the fact that the other player does or does not cooperate (*Nash Equilibrium*). Note that, if we use these quadruplets for the table of rewards, we get for the rewards of player A:

Rewards for A	B cooperates	B does not
A cooperates	p_1	p_3
A does not	p_2	p_4

and for the rewards of player B:

Rewards for B	B cooperates	B does not
A cooperates	p_1	p_2
A does not	p_3	p_4

Using the values in our examples, we get for the rewards of player A:

Rewards for A	B cooperates	B does not
A cooperates	0.75	0.8
A does not	0.2	0.25

and for the rewards of player B:

Rewards for B	B cooperates	B does not
A cooperates	0.75	0.2
A does not	0.8	0.25

Note that this implies that both players A and B are pushed, for personal convenience, to not cooperate, independently of what the other player does, but end up getting a worse reward than if they would have both cooperated. In fact, the *amalgamated* contingency table, gives:

Rewards for A+B	B cooperates	B does not
A cooperates	1.5	1
A does not	1	0.5

that prizes the decision of cooperation. The *Measure of Amalgamation* considered here can be thought in the form of an *Utility Function*, such as:

$$U_A(a,b) = p_1ab + p_3a(1-b) + p_2b(1-a) + p_4(1-a)(1-b)$$

and

$$U_B(a,b) = p_1ab + p_2a(1-b) + p_3b(1-a) + p_4(1-a)(1-b).$$

Here $a = 1$, means that A cooperates, while $a = 0$ means that A does not. Similarly for B. Note that, under the conditions on p_1, p_2, p_3 and p_4 mentioned above, the Utility is bigger for the choice of not cooperation for both A and B, given any decision taken by the other player. In fact,

$$p_1 = U_A(1,1) < U_A(0,1) = p_2$$

and

$$p_3 = U_A(1,0) < U_A(0,0) = p_4$$

and analogously for U_B. However, when we combine the utilities, we get *Utility Function*

$$U_{A+B}(a,b) = 2p_1ab + (p_2+p_3)a(1-b) + (p_3+p_2)b(1-a) + 2p_4(1-a)(1-b).$$

This utility is always bigger for cooperation, if we require $2p_4 < p_2 + p_3 < 2p_1$, as we chose in our example. In fact:

$$2p_4 = U_{A+B}(0,0) < U_{A+B}(1,0) = p_2 + p_3 = U_{A+B}(0,1) < 2p_1 = U_{A+B}(1,1).$$

In this way, we have restated the *Prisoner's Dilemma* in the context of the *Simpson's Paradox*.

- **Election polls**: Suppose candidates T and C run for elections in two states $State_1$ and $State_2$. Suppose that candidate T and C receive in $State_1$ a percentage of votes:

$$\%votes\ for\ T = \frac{a}{b} > 1 - \frac{a}{b} = \%votes\ for\ C$$

and that candidate T and C receive in $State_2$ a percentage of votes:

$$\%votes\ for\ T = \frac{2}{d} > 1 - \frac{c}{d} = \%votes\ for\ C.$$

Is it possible that overall candidate C receives a higher percentage of votes? Clearly, this is not possible because $\frac{a}{b} > 1 - \frac{a}{b}$ implies $a > 0.5b$ and $\frac{c}{d} > 1 - \frac{c}{d}$ implies $c > 0.5d$ and so

$$0.5b + 0.5d < a + c,$$

which implies

$$\frac{a+c}{b+d} > 0.5$$

and so

$$\frac{a+c}{b+d} > 1 - \frac{a+c}{b+d}.$$

In this case, we do not have any paradox and this is related to the fact that there is an extra constraint on the construction of the contingency table. Note that since the set of real numbers for which these inequalities hold is an open set, the inclusion of a not strong third candidate will not change the situation. What happens if the third candidate is as strong as T and C?

4 The *Simpson's Paradox* in quantum mechanics

In this section, we turn our attention to a novel result of us (Selvitella 2017) concerning the occurrence of the *Simpson's Paradox* in Quantum Mechanics. In particular, we show how we can detect an unintuitive behaviour in the interaction between solitary wave solutions in the case of the Quantum Harmonic Oscillator and the Nonlinear Schrödinger Equation. We start with the Quantum Harmonic Oscillator.

4.1 The quantum harmonic oscillator

We consider the following Linear Schrödinger Equation in the presence of a Harmonic Potential:

$$i\hbar\frac{\partial}{\partial t}\psi(t,\mathbf{x}) = -\frac{\hbar^2}{2m}\Delta_{\mathbf{x}}\psi(t,\mathbf{x}) + \frac{1}{2}m\omega^2|\mathbf{x}|^2\psi(t,\mathbf{x}). \tag{1}$$

Here $i = \sqrt{-1}$ is the *complex unit*, $\hbar$ is the *Planck constant*, m represents the *mass* of a particle, ω is the *angular velocity* and $(t,\mathbf{x}) \in (0,+\infty) \times \mathbf{R}^n$. There exists a solution of Eq. (1) in the form

$$\psi(t,x) = u(\mathbf{x} - \mathbf{x}(t))e^{i\left[\mathbf{x}\cdot\mathbf{v}(t)+\gamma(t)+\frac{\omega t}{2}\right]} \tag{2}$$

with the following conditions on $u(\mathbf{x})$, $\mathbf{x}(t)$, $\mathbf{v}(t)$ and $\gamma(t)$:

- the *profile* $u(\mathbf{x})$ for $\mathbf{x} \in \mathbf{R}^n$ satisfies the equation
$$-\frac{\hbar^2}{2m}\Delta_{\mathbf{x}}u(\mathbf{x}) + \frac{1}{2}m\omega^2|\mathbf{x}|^2u(\mathbf{x}) + \frac{\omega}{2}u(\mathbf{x}) = 0; \tag{3}$$
- the *position vector* $\mathbf{x}(t)$ and the *velocity vector* $\mathbf{v}(t)$ satisfy the following system of ODEs:
$$\begin{cases} \dot{\mathbf{x}}(t) = \frac{\hbar}{m}\mathbf{v}(t), \\ \dot{\mathbf{v}}(t) = -\frac{m}{\hbar}\omega^2\mathbf{x}(t); \end{cases} \tag{4}$$
- the *complex phase* $\gamma(t)$ is such that

$$\dot{\gamma}(t) = \frac{1}{\hbar}\mathcal{L}(\mathbf{x}(t), \dot{\mathbf{x}}(t); m, \omega),$$

where $\mathcal{L}(\mathbf{x}(t), \dot{\mathbf{x}}(t); m, \omega) := \frac{1}{2}m|\dot{\mathbf{x}}(t)|^2 - \frac{1}{2}m\omega^2|\mathbf{x}(t)|^2$ is the *Lagrangian* of the system of ODEs (4). For why this is true, we refer to (Berezin and Shubin 1991) and (Selvitella 2017).

4.2 The nonlinear Schrödinger Equation

In the rescaled variables $m = 1$ and $\hbar = \frac{1}{2}$, the Nonlinear Schrödinger Equation takes the following form:

$$\begin{cases} i\frac{\partial}{\partial t}\psi(t,\mathbf{x}) = -\Delta_{\mathbf{x}}\psi(t,x) - |\psi(t,\mathbf{x})|^{p-1}\psi(t,\mathbf{x}), \\ \psi(0,\mathbf{x}) = \psi_0(\mathbf{x}), \end{cases}$$

Here, $n \geq 1$ and $1 < p < 1 + \frac{4}{n}$ is the L^2-*subcritical exponent.* There exist solutions, called ***solitons***, of the form $\psi(t,\mathbf{x}) = e^{i\omega t}Q_\omega(\mathbf{x})$ with $\omega > 0$ and where $Q_\omega \in H^1(\mathbf{R}^n)$ is a solution of

$$\Delta Q_\omega + Q_\omega^p = \omega Q_\omega, \quad Q_\omega > 0. \tag{5}$$

These solutions Q_ω can be computed explicitly in dimension $n = 1$ and take the form

$$Q_\omega(x) = \omega^{\frac{1}{p-1}} \left(\frac{p+1}{2\cosh^2\left(\frac{p-1}{2}\omega^{\frac{1}{2}}x\right)} \right)^{p-1} .$$

In any dimension $n \geq 1$, the solitons which minimize the so called *Energy Functional*

$$E[Q_\omega] := \frac{1}{2}\int_{\mathbf{R}^n} d\mathbf{x}|\nabla Q_\omega|^2 + \frac{\omega}{2}\int_{\mathbf{R}^n} d\mathbf{x}|Q_\omega|^2 - \frac{1}{p+1}\int_{\mathbf{R}^n} d\mathbf{x}|Q_\omega|^{p+1}$$

are called ***ground states***. These solutions are radially symmetric for $n > 1$ (in fact, they are even for $n = 1$), exponentially decaying and unique up to symmetries (see (Berestycki and Lions 1983; Berestycki et al. 1981; Gidas et al. 1979; Kwong 1989)).

4.3 The main theorems

In Quantum Mechanics and in the context of the Schrödinger Equation, there is a very natural *Measure of Amalgamation*, given by the $L^2(\mathbf{R}^n)$ inner product.

Definition 4 *Consider two solutions* $\psi(t,\mathbf{x})$ *and* $\phi(t,\mathbf{x})$ *of Eq.* (1). *The Measure of Amalgamation between* $\psi(t,\mathbf{x})$ *and* $\phi(t,\mathbf{x})$ *is given by the* $L^2(\mathbf{R}^n)$ *inner product:*

$$Cov(\psi(t,\cdot),\phi(t,\cdot)) :=< \psi(t,\cdot),\phi(t,\cdot)) >_{L^2(\mathbf{R}^n)} .$$

Using the $L^2(\mathbf{R}^n)$ inner product, we can show that, for the Quantum Harmonic Oscillator, there exist quadruplets of solitons, which exhibit the *Simpson's Paradox.*

Theorem 1 ***[Existence of the Simpson's Paradox]*** *Consider Eq.* (1) *for every spatial dimension* $n \geq 1$*. Then, for every* $m > 0$ *and* $\omega > 0$*, there exists a set of parameters* $(x_i(t), \gamma_i(t), v_i(t))$ *with* $i = 1, \ldots, 4$*, such that the following is true. If we consider an initial datum of the form* $\psi(0,x) = \Sigma_{i=1}^4 \psi_i(0,x)$ *with* $\psi_i(0,x)$ *such that*

$$\psi_i(t,x) = \left(\frac{m\omega}{\pi\hbar}\right)^{1/4} e^{i[x\cdot v_i(t)+\gamma_i(t)+\frac{\omega t}{2}]}e^{-\frac{m\omega}{2\hbar}|x-x_i(t)|^2},$$

then the Simpson's Paradox occurs in the following cases.

- *In the stationary case, namely when $v_i(t) = 0$ and $x_i(t) = x_i$ for every t; both when $\gamma_i = \gamma_j$ for every $1 \leq i,j \leq 4$ and when $\gamma_i \neq \gamma_j$ $1 \leq i,j \leq 4$, $i \neq j$.*
- *In the non-stationary case: if there exists $t_0 \in \mathbf{R}$ such that the Simpson's Paradox occurs at t_0, then the Simpson's Paradox occurs at any t_1 with $t_1 \neq t_0$.*

Remark 1 *As we can see from Theorem 1, the occurrence of the Simpson's Paradox in the case of the Quantum Harmonic Oscillator is determined by the initial datum and so we can say that it is persistent under the flow of the Quantum Harmonic Oscillator.*

Once we have proved the existence, we want to address the question of how robust this phenomenon is, namely if nearby a quadruplet of solitons, we can find plenty of quadruplets of solitons for which the paradox occurs. We have that the set of parameters for which the paradox occurs contains open sets.

Theorem 2 ***[Stability of the Simpson's Paradox]*** *Suppose that there exists a set of parameters $(x_i(t), \gamma_i(t), v_i(t))$ for $i = 1, \ldots, 4$ such that the Simpson's Paradox occurs in the stationary case. Then, there exists $r > 0$ such that, for every $(\tilde{x}_i(t), \tilde{\gamma}_i(t), \tilde{v}_i(t))$ for $i = 1, \ldots, 4$ inside $B_r((x_i(t), \gamma_i(t), v_i(t)), \quad i = 1, \ldots, 4)$, the Simpson's Paradox still occurs for initial data as above. Moreover, if the Simpson's Paradox occurs for a $\psi(t,x)$ at a certain time $t = \tilde{t}$, then there exists an open ball in $\Sigma := L^2(\mathbf{R}^n, dx) \cap L^2(\mathbf{R}^n, |x|^2 dx)$ such that the Simpson's Paradox still occurs for any $\bar{\psi}(t,x) = \psi(t,x) + w(t,x)$ with $w(t,\cdot) \in \Sigma$ and the same time $t = \tilde{t}$.*

Now, we can discuss the nonlinear case.

Theorem 3 ***[Nonlinear case]*** *Consider the nonlinear Schrödinger Equation in dimension $n = 1$, with $1 < p < 5$ (L^2-subcritical exponent). Then, there exist an initial datum $\psi_0(x)$, in the form of a superposition of solitons (see (Martel and Merle 2006)), for which there exists $t = \tilde{t}_1 \gg 1$ where the Simpson's Paradox occurs and $t = \tilde{t}_2 \gg 1$ where the Simpson's Paradox does not occur.*

Remark 2 *In striking contrast with the Quantum harmonic Oscillator, for the Nonlinear Schrödinger Equation, the Simpson's Paradox is not anymore persistent, but it is intermittent. In fact, we can detect it for large times but it appears and disappears indefinitely.*

Proof For the complete proofs of these theorems, we refer to (Selvitella 2017), while for a brief sketch of the proof of Theorem 1 in the stationary case, we refer to the upcoming Section 5. □

5 How likely is the Simpson's Paradox in quantum mechanics?

An important question is: "How likely is the Simpson's Paradox?". It is in fact interesting to quantify, in some way, the chances that one has to run into the paradox.

In the case of $2 \times 2 \times l$ contingency tables with $l \geq 2$, Pavlides and Perlman (2009) address the problem and, among the other things, they prove the following.

Suppose that a contingency table consists of a factor A with two levels, a factor B with other 2 levels and a third factor C with $l \geq 2$-levels. Then, the array of cell probabilities $\mathbf{p}$ lies on the Simplex

$$\mathcal{S}_{4l} := \left\{ \mathbf{p} | \, p_i \geq 0, \ \forall i = 1, \ldots, 4l; \ \Sigma_{i=1}^{4l} p_i = 1 \right\}.$$

Endow $\mathcal{S}_{4l}$ with the *Dirichlet Distribution on* $\mathcal{S}_{4l}$, denoted by $D_{4l}(\alpha)$ and denote with $\pi_l(\alpha)$ the probability of having the *Simpson's Paradox* under $D_{4l}(\alpha)$. Pavlides and Perlman proved in (Pavlides and Perlman 2009) that $\pi_2(1) = \frac{1}{60}$ and conjectured that for every $\alpha > 0$, there exists $h(\alpha) > 0$ such that

$$\pi_l(\alpha) \simeq \pi_2(\alpha) \times e^{-h(\alpha)\left(\frac{l}{2}-1\right)}, \quad l = 2, 3, \ldots.$$

A similar question can be asked in the case of the *Quantum Harmonic Oscillator* and the *Nonlinear Schrödinger Equation*. In the constructions developed in (Selvitella 2017), we aimed just at finding one single choice of the parameters which gives the *Simpson's Paradox* and we did it mainly with a perturbative method. But how large is (and in which sense it is large) the set of parameters which gives the *Simpson's Paradox*?

To investigate a little bit further this issue, we briefly sketch the proof of Theorem 1, at least in the stationary case and deduce from it a preliminary result on the likelihood of occurrence of the *Simpson's Paradox*.

Consider two *moving solitons* of the form:

$$\psi_i(t,x) = \left(\frac{m\omega}{\pi\hbar}\right)^{1/4} e^{i\left[x \cdot v_i(t) + \gamma_i(t) + \frac{\omega t}{2}\right]} e^{-\frac{m\omega}{2\hbar}|x - x_i(t)|^2},$$

and

$$\psi_j(t,x) = \left(\frac{m\omega}{\pi\hbar}\right)^{1/4} e^{i\left[x \cdot v_j(t) + \gamma_j(t) + \frac{\omega t}{2}\right]} e^{-\frac{m\omega}{2\hbar}|x - x_j(t)|^2},$$

for $1 \leq i \leq j \leq 4$ and with $\mathbf{x}(t)$, $\mathbf{v}(t)$ and $\gamma(t)$ as in Subsection 4.1.

Consider the case in which, for every $t \in \mathbf{R}$, one has that $x_k(t) = x_k$, for every $k = 1, \ldots, N$ independent of time. It has been proven in (Selvitella 2017) (Proposition 3.3) that the *Covariance* between any of these two solitons is given by:

$$Cov(\psi_i(t,x), \psi_j(t,x)) = \frac{1}{2}\cos(\gamma_i - \gamma_j)\left[\frac{\hbar}{m\omega} - \frac{1}{2}|x_i - x_j|^2\right] e^{-\frac{m\omega}{4\hbar}|x_i - x_j|^2} \qquad (6)$$

Therefore, the proof of Theorem 1 in the stationary case reduces to the problem of finding parameters such that the *Simpson's Paradox* occurs, namely such that

$$Cov(\psi_1(t,x), \psi_2(t,x)) > 0,$$

$$Cov(\psi_3(t,x), \psi_4(t,x)) > 0$$

but

$$Cov(\psi_1(t,x) + \psi_3(t,x), \psi_2(t,x) + \psi_4(t,x)) < 0$$

or viceversa,

$$Cov(\psi_1(t,x), \psi_2(t,x)) < 0,$$

$$Cov(\psi_3(t,x), \psi_4(t,x)) < 0$$

but

$$Cov(\psi_1(t,x) + \psi_3(t,x), \psi_2(t,x) + \psi_4(t,x)) > 0.$$

Now, we define

$$L_{ij}^2 := \frac{m\omega}{2\hbar}|x_i - x_j|^2$$

so that $Cov(\psi_i(t,x),\psi_j(t,x))$ can be rewritten in the following way:

$$Cov(\psi_i(t,x),\psi_j(t,x)) = \frac{\hbar}{2m\omega}\cos(\gamma_i - \gamma_j)\left[1 - L_{ij}^2\right]e^{-\frac{1}{2}L_{ij}^2}.$$

In the following discussion, we treat only the case $\gamma_i = \gamma_j$, for every $i,j = 1,\ldots,4$.

We can restate our hypotheses and thesis in the following way: we suppose that $0 < L_{12} < 1$ and $0 < L_{34} < 1$ and we want to quantify "how many" admissible choices of $0 < L_{12} < 1$ and $0 < L_{34} < 1$, L_{23} and L_{14} there are such that

$$\left[1 - L_{12}^2\right]e^{-\frac{1}{2}L_{12}^2} + \left[1 - L_{23}^2\right]e^{-\frac{1}{2}L_{23}^2} + \left[1 - L_{34}^2\right]e^{-\frac{1}{2}L_{34}^2} + \left[1 - L_{14}^2\right]e^{-\frac{1}{2}L_{14}^2} < 0.$$

Remark 3 *Note that the defining condition for the occurrence of the Simpson's Paradox are all inequalities which is a hint of the fact that the Simpson's Paradox occurs in a open set of the correct topology (see Theorem 2 and (Selvitella 2017)).*

Since we are in dimension $n = 1$, we can choose $x_1 < x_2 < x_3 < x_4$. This implies that $L_{14} = L_{12} + L_{23} + L_{34}$ and so that we have to find an admissible choice of $0 < L_{12} < 1$ and $0 < L_{34} < 1$ and L_{23} such that

$$\begin{aligned}&\left[1 - L_{12}^2\right]e^{-\frac{1}{2}L_{12}^2} + \left[1 - L_{23}^2\right]e^{-\frac{1}{2}L_{23}^2} + \\ &+ \left[1 - L_{34}^2\right]e^{-\frac{1}{2}L_{34}^2} + \left[1 - (L_{12} + L_{23} + L_{34})^2\right]e^{-\frac{1}{2}(L_{12}+L_{23}+L_{34})^2} < 0.\end{aligned}$$

Now, if we define $X := L_{12}$, $Y := L_{34}$ and $Z := L_{23}$, we get that the *Simpson's Paradox* occurs when the following are satisfied:

$$\begin{cases} 0 < X < 1 \\ 0 < Y < 1 \\ \left[1 - X^2\right]e^{-\frac{1}{2}X^2} + \left[1 - Y^2\right]e^{-\frac{1}{2}Y^2} + \left[1 - Z^2\right]e^{-\frac{1}{2}Z^2} + \left[1 - (X + Y + Z)^2\right]e^{-\frac{1}{2}(X+Y+Z)^2} < 0. \end{cases}$$

Figure 1 focuses on a small region of the parameters' space with $0 < X, Y < 1$ and represents the surface which discriminates between where the paradox occurs and when it does not.

Note that, when one of the coordinates (for example Z) becomes larger and larger, the paradox occurs more and more rarely. In fact, the condition

$$\left[1 - X^2\right]e^{-\frac{1}{2}X^2} + \left[1 - Y^2\right]e^{-\frac{1}{2}Y^2} + \left[1 - Z^2\right]e^{-\frac{1}{2}Z^2} + \left[1 - (X + Y + Z)^2\right]e^{-\frac{1}{2}(X+Y+Z)^2} < 0$$

for big Z reduces to

$$\left[1 - X^2\right]e^{-\frac{1}{2}X^2} + \left[1 - Y^2\right]e^{-\frac{1}{2}Y^2} < 0$$

which is incompatible with

$$0 < X < 1, and\, 0 < Y < 1.$$

Figure 2 explains this last sentence visually.

We have decided to test the inequality $f(X,Y,Z) < 0$ over a grid of $n \times n \times n$ values with $n = 1000$ in the parallelepiped $(X,Y,Z) \in [0,1] \times [0,1] \times [0,4]$ and we discovered that about $1.2 * 10^{-4}$ of the times (0.012%) the inequality is satisfied. Note that the choice of the uniform distribution on $[0,1] \times [0,1] \times [0,4]$ has been made because for $Z > 4$ the

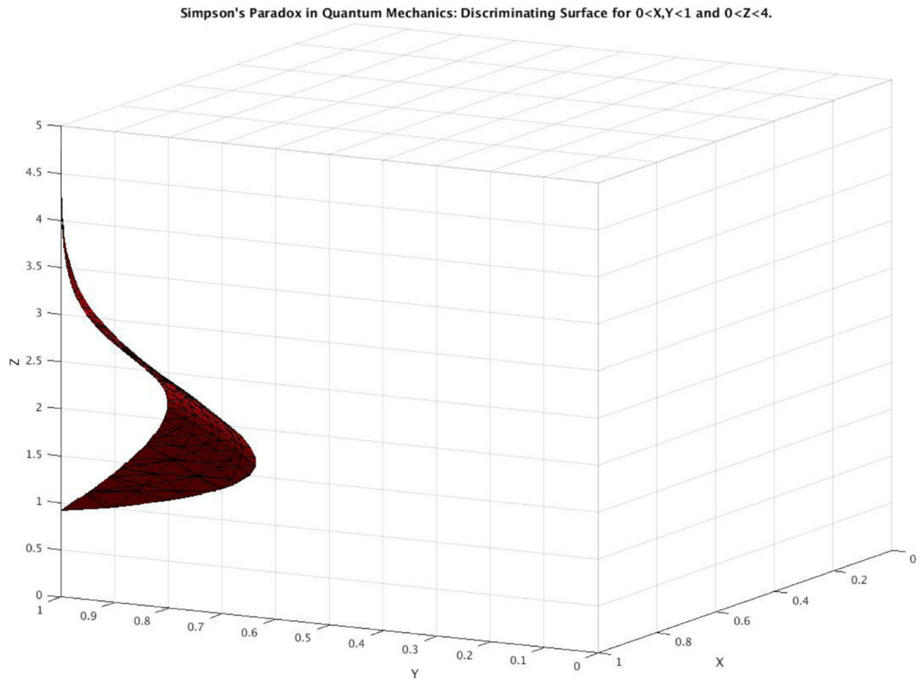

Fig. 1 Surface discriminating between the region of parameters where the Simpson's Paradox occurs and does not occur

Simpson's Paradox's region is almost null (Fig. 2) and because already $0 < X, Y < 1$. This result deserves further investigation. For reproducibility purposes, we give the Matlab Code that we used for the analysis:

```
syms X Y Z
fun=@(X,Y,Z)((1-X.^2).*exp(-X.^2/2)+(1-Y.^2).*exp(-Y.^2/2)
+(1-Z.^2).*exp(-Z.^2/2)+(1-(X+Y+Z).^2).*exp(-(X+Y+Z).^2/2));

n=1000;
S=zeros(n);
m=zeros(n);
x=0:1/n:1;
y=0:1/n:1;
%since the max of this function is 4
z=0:1/n:4;
SP=0;
%syms t
%[X,Y]=meshgrid(0:0.1:1,0:0.1:1);
for i= 1:n+1
    for j=1:n+1
        for k=1:n+1
            if fun(x(i),y(j),z(k))<0
     SP=SP+1;
            else
           SP=SP+0;
    end
    end
end
```

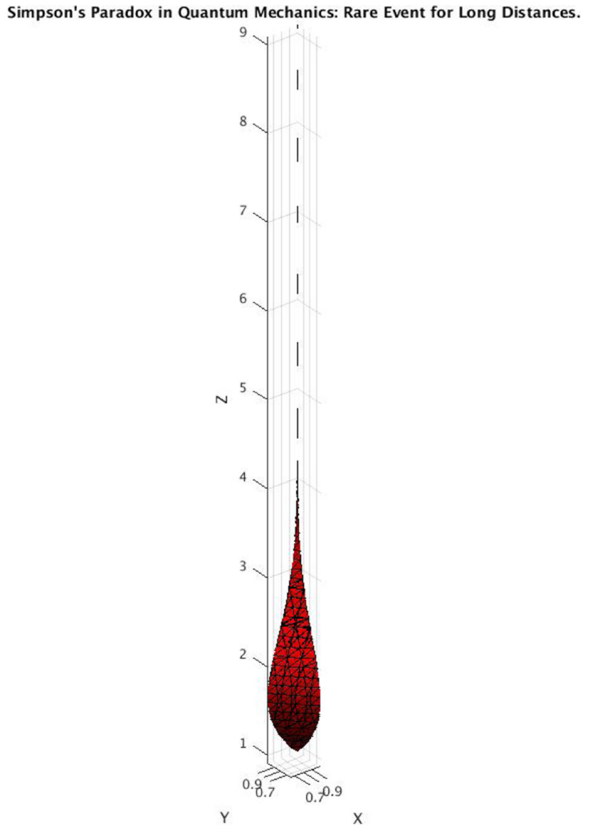

Fig. 2 The likelihood of occurrence of the *Simpson's Paradox* decreases to zero as one of the distances between the particles increases indefinitely

```
SP # Number of occasions in which the Simpson's Paradox occurs
SP/(n+1)^3 # Percentage of of occasions in which the Simpson's
Paradox occurs
```

6 Some numerical examples

For illustration purposes, we give some numerical examples of cases in which the *Simpson's Paradox* occurs and on which it does not. We find interesting to give to each parameters their true physical value.

Consider the *Planck Constant*

$$\hbar = \frac{h}{2\pi} = \frac{1}{2\pi} * 6.62607004 * 10^{-34} m^2 kg/s = 1.0545718 * 10^{-34} m^2 kg/s,$$

the *Mass of an Electron*

$$m = 9.10938356 * 10^{-31} kg$$

with frequency of revolution

$$f = 6.6 * 10^{15} s^{-1}$$

and angular velocity

$$\omega = 2\pi f = 4.1469023 * 10^{16} s^{-1}.$$

Note that the quantity

$$L_{ij}^2 := \frac{m\omega}{2\hbar}|x_i - x_j|^2$$

that we defined and used in Section 5 for the sketch of the proof of the stationary case of Theorem 1, is dimensionless and it is a fundamental quantity.

We choose L_{12}^2, L_{34}^2 and L_{23}^2 which are all around 1. Note that this implies the following about the distance between the particles:

$$\begin{aligned} 1 \simeq L_{ij}^2 &= \frac{m\omega}{2\hbar}|x_i - x_j|^2 \\ &= \frac{9.10938356 * 10^{-31} * 4.1469023 * 10^{16}}{1.0545718 * 10^{-34}}|x_i - x_j|^2 \simeq 3.582091 * 10^{20}|x_i - x_j|^2. \end{aligned}$$

This implies that

$$|x_i - x_j| \simeq 5.2836213 * 10^{-11}m.$$

Recall that the *Bohr Radius*, which represents approximately the most probable distance between the center of a nuclide and the electron in a hydrogen atom in its ground state, is

$$r_{Bohr} = 5.2917721067 * 10^{-11}m$$

We choose $L_{12}^2 = 1 - \epsilon_1^2, L_{34}^2 = 1 - \epsilon_2^2$ and $L_{23}^2 = 1 + \delta^2$ with $\epsilon_1 \ll 1, \epsilon_2 \ll 1$. The following R code produces and example of the paradox in our case:

```
x=1-10^(-10); #L_{12}^2<1--> Positive Correlation
y=1-10^(-10); #L_{34}^2<1--> Positive Correlation
z=1+10^(-5); #L_{23}^2
(1-x^2)*exp(-x^2/2)+(1-y^2)*exp(-y^2/2)+(1-z^2)*exp(-z^2/2)
+(1-(x+y+z)^2)*exp(-(x+y+z)^2/2)
#Reversal Condition <--> Negative Correlation
[1] -0.0888821
```

Of course, there are cases in which the *Simpson's Paradox* does not occur, like

```
x=1-10^(-1); #L_{12}^2<1--> Positive Correlation
y=1-10^(-1); #L_{34}^2<1--> Positive Correlation
z=1+10^(-5); #L_{23}^2
(1-x^2)*exp(-x^2/2)+(1-y^2)*exp(-y^2/2)+(1-z^2)*exp(-z^2/2)
+(1-(x+y+z)^2)*exp(-(x+y+z)^2/2)
#Reversal Condition not satisfied <--> Positive Correlation
[1] 0.1177287
```

7 Discussion

In this paper, we discussed the *Simpson's Paradox* in several settings. In particular, we gave basic examples in arithmetic, geometry, linear algebra, statistics, game theory, gender bias in university admission and election polls, where we described the appearance or absence of the *Simpson's Paradox*. Then, we moved to the presentation of our recent results on the occurrence of the *Simpson's Paradox* in Quantum Mechanics with focus on the Quantum Harmonic Oscillator and the Nonlinear Schrödinger Equation (Selvitella 2017). We discussed the likelihood of the occurrence of the *Simpson's Paradox* and we gave some numerical examples in which the *Simpson's Paradox* occurs and some numerical examples in which the *Simpson's Paradox* does not occur. This depends on in which parameter regions we are. Several problems remain to be addressed. An extended investigation of the question "How likely is the *Simpson's Paradox* in Quantum Mechanics?" is appropriate. In particular, it would be interesting to construct and put a more suitable probability measure on the set of parameters and quantify the likelihood of the *Simpson's Paradox* even further.

Acknowledgements
I thank my family and Linda Forson for their constant support. I thank my supervisor Prof. Narayanaswamy Balakrishnan for his constant help and useful comments. Many thanks to the referees and the editors, whose suggestions improved the quality of the manuscript. This research has been supported by CANSSI and Central Michigan University in occasion of the ICOSDA 2016. The author has no competing interests in the manuscript. I dedicated this paper to the memory of Zia Vera and Zio Salvatore.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 15 December 2016 Accepted: 14 March 2017
Published online: 28 March 2017

References

Bickel, PJ, Hammel, EA, O'Connell, JW: Sex Bias in Graduate Admissions: Data From Berkeley. Science. **187**(4175), 398–404 (1975)

Berestycki, H, Lions, PL: Nonlinear Scalar field equations. Arch. Rational Mech. Anal. **82**(3), 313–345 (1983)

Berestycki, H, Lions, PL, Peletier, LA: An ODE approach to the existence of positive solutions for semilinear problems in $\mathbf{R}^N$. Indiana Univ. Math. J. **30**(1), 141–157 (1981)

Berezin, FA, Shubin, MA: The Schrödinger Equation. Translated from the 1983 Russian edition by Yu. Rajabov, DA Leĭtes and NA Sakharova and revised by Shubin. With contributions by G. L. Litvinov and Leĭtes. Mathematics and its Applications (Soviet Series), 66. Kluwer Academic Publishers Group, Dordrecht (1991). ISBN:0-7923-1218-X 81-01 (35J10 35P05 46N50 47F05 47N50)

Blyth, CR: On Simpson's Paradox and the Sure-Thing Principle. J. Am. Stat. Assoc. **67**(338), 364–366 (1972)

Cialdi, S, Paris, MGA: The data aggregation problem in quantum hypothesis testing. Eur. Phys. J. D. **69**, 7 (2015). doi:10.1140/epjd/e2014-50425-7

Good, IJ, Mittal, Y: The Amalgamation and Geometry of Two-by-Two Contingency Tables. Ann. Stat. **15**(2), 694–711 (1987)

Gidas, B, Ni, WM, Nirenberg, L: Symmetry and related properties via the maximum principle. Comm. Math. Phys. **68**, 209–243 (1979)

Goltz, HH, Smith, ML: Yule-Simpson's Paradox in Research. Pract. Assess. Res. Eval. **15**(15), 1–9 (2010)

Kwong, MK: Uniqueness of positive solutions of $\Delta u - u + u^p = 0$ in $\mathbf{R}^n$. Arch. Rational Mech. Anal. **105**(3), 243–366 (1989)

Martel, Y, Merle, F: Multi solitary waves for the nonlinear Schrödinger Equations. Ann. I. H. Poincaré. **23**(6), 849–864 (2006)

Norris, F: Can Every Group Be Worse Than Average? Yes (2013). https://economix.blogs.nytimes.com/2013/05/01/can-every-group-be-worse-than-average-yes/. Accessed 1 May 2013

Paris, MGA: Two quantum Simpson's Paradoxes. J. Phys. A. **45**, 132001 (2012)

Pavlides, MG, Perlman, MD: How likely is Simpson's Paradox? Am. Stat. **63**, 226–233 (2009)

Pearson, K, Lee, A, Bramley-Moore, L: Genetic (reproductive) selection: Inheritance of fertility in man, and of fecundity in thoroughbred racehorses. Phil. Trans. R. Soc. A. **192**, 257–330 (1899)

Selvitella, A: The Simpson's Paradox in quantum mechanics. J. Math. Phys. **58**(3), 37 (2017). 032101

Shi, Y: Quantum Simpson's Paradox and High Order Bell-Tsileron Inequalities (2012). preprint available at arxiv.org/pdf/1203.2675

Simpson, EH: The Interpretation of Interaction in Contingency Tables. J. R. Stat. Soc. Ser. B. **13**, 238–241 (1951)

Yule, GU: Notes on the Theory of Association of Attributes in Statistics. Biometrika. **2**(2), 121–134 (1903)

Yule, GU, Kendall, MG: An Introduction to the Theory of Statistics. Griffin, London (1937)