SRI International

Technical Note 502 · January 28, 1991

SIMPSON'S PARADOX: A MAXIMUM LIKELIHOOD SOLUTION[1]

Prepared by:

Matthias P. Kläy
Sandoz AG
Praeklinische Forschung, Informatik
Postfach CH-4002 BASEL
Switzerland, Europe

Leonard P. Wesley
Artificial Intelligence Center
SRI International
Menlo Park, California 94025

[1] This work was supported in part by the Swiss National Science Foundation.

INTRODUCTION

Simpson's paradox exemplifies a class of problems that can arise when the logic used to reason about the semantics of propositional sentences does not adequately capture certain dependencies between sentences of interest. This paradox has been known as early as 1903 [YUL03], and has been discussed extensively in the statistical literature [SIM51, DAW79, BLY72, BLY73, CHU42]. The phenomena that typically give rise to Simpson's paradox can occur in cases such as destructive testing (e.g., determining the breaking strength of materials in orthogonal directions), and identifying the composition of complex alloys. It has also been reported to occur in "real life" several times since its discovery [KNA85, WAG82]. One such occurrence received wide attention in 1973 over the appearance of a sex bias in the admission policy for graduate students at the University of Berkely [BIC75]. Given that automated systems will be expected to recognize and cope with the underlying phenomena of this paradox, it is important to develop effective methods for dealing with them, particularly as it impacts the choice of logics that systems must use to reason about real world problems. Only recently, however, has there been any significant indication that Simpson's paradox merits serious attention by the AI community [PEA88].

MOTIVATION

Simpson's paradox is often described within a medical context where a physician is confronted with the task of choosing only one of two incompatible treatments, say A and B, that must be administered to her patient. There are only two outcomes of each treatment, survival and death of the patient. The data from which the physician must make a decision are the survival and death rates of patients receiving either A or B in two distinct cities, say c and $\overline{c}$. We represent the survival and death from receiving treatments A and B by a, $\overline{a}$, b, and $\overline{b}$ respectively. The number of survivals from using treatment A in city c is represented by N_{ac}, the number of deaths from using treatment A in city c is indicated by $N_{\bar{a}c}$, and $N_{\overline{ac}}$ represents the number of deaths in city $\overline{c}$. Similar notation is used for the remaining survival and death rates for treatments A and B in cities c and $\overline{c}$. Suppose the data indicate that $N_{ac} = 50$, $N_{\bar{a}c} = 950$, $N_{bc} = 1000$, $N_{\bar{b}c} = 9000$,

$N_{a\bar{c}} = N_{\bar{a}\bar{c}} = 5000$, $N_{b\bar{c}} = 95$, and $N_{\bar{b}\bar{c}} = 5$.[†] Using traditional analysis methods, we find that with respect to each city separately,

$$\frac{N_{ac}}{N_{ac} + N_{\bar{a}c}} < \frac{N_{bc}}{N_{bc} + N_{\bar{b}c}} \quad \text{and} \quad \frac{N_{a\bar{c}}}{N_{a\bar{c}} + N_{\bar{a}\bar{c}}} < \frac{N_{b\bar{c}}}{N_{b\bar{c}} + N_{\bar{b}\bar{c}}} \quad . \tag{1}$$

A plausible interpretation of this data is that treatment B appears more favorable to the survival of patients in cities c and $\bar{c}$ than treatment A. However, when the data for each city is pooled we find

$$\frac{N_{ac} + N_{a\bar{c}}}{N_{ac} + N_{\bar{a}c} + N_{a\bar{c}} + N_{\bar{a}\bar{c}}} > \frac{N_{bc} + N_{b\bar{c}}}{N_{bc} + N_{\bar{b}c} N_{b\bar{c}} + N_{\bar{b}\bar{c}}} \tag{2}$$

to be paradoxical in that now treatment A appears to be more favorable than treatment B overall.

Alternatively, Simpson's paradox can be characterized in terms of the following question: if by treatment B favorable to S (i.e., survival) we mean $P(S|B) > P(S)$, B indifferent to S we mean $P(S|B) = P(S)$, and B unfavorable to S we mean $P(S|B) < P(S)$, then is it possible that $P(S|B) < P(S|\bar{B})$, and simultaneously have $P(S|BC) \geq P(S|\bar{B}C)$ and $P(S|B\bar{C}) \geq P(S|\bar{B}\bar{C})$? If so, what treatment is justified by the available data? In an attempt to interpret the data, our physician might reason that if we let p represent the propositional sentence *Treatment is given in city* c,let q represent the sentence *Treatment is given in city* $\bar{c}$, let r represent the sentence *Treatment* B *is more favorable than treatment* A, and given $p \rightarrow \neg q$, $p \rightarrow r$, $q \rightarrow r$, then $p \vee q \rightarrow r$ would seem to follow logically according to the disjunction axiom. However, the pooled data suggests otherwise–i.e., $p \vee q \rightarrow \neg r$ –that this axiom does not hold in this case.

While some might argue that the underlying cause of this paradox is an inappropriate analysis of the data within a classical framework, no classical solution has been offered to date. Others have suggested that the disjunction axiom is simply not sound with respect to the calculus of increased proportions; rather, the calculus

[†] The data for this example is borrowed from [BLY72].

of extreme probabilities is a suitable mathematical framework for coping with this dilemma [PEA88]. Foulis and Randall offer still another view of this paradox based on their investigations of similar phenomena within the physics domain [FOU72, RAN73]. Their work has resulted in the development of a language, called Empirical Logic (EL), that provides the foundation of the generalized maximum likelihood solution discussed in this paper.

A MAXIMUM LIKELIHOOD APPROACH

At the heart of an EL-based formalism to a maximum likelihood solution is the notion of incompatibility. In our medical example, the two treatments are incompatible in the sense that once one treatment has been administered then it is impossible, even in principle, to measure the result and extract a meaningful outcome for the other treatment. A determination is impossible because if the patient is healed, the premise of an illness is no longer valid. If the patient dies, no further measurements are possible. Incompatibility, in this context, does not mean measurements cannot be made simultaneously. It means that the execution of one action (e.g., administering treatment A) precludes, in principle, extracting a meaningful interpretation from the outcome of pursuing the second action (e.g., administering treatment B).

According to Kolmogoroff [KOL33], an experiment is represented by its sample space which consists of all possible outcomes of the experiment, and the collection of probability measures on the sample space. In the presence of incompatible measurements, we have to reflect the fact that there is not simply one grand canonical measurement that produces all outcomes of the experiment, but rather several measurements, each capable of producing only a subset of the outcomes of the experiment [FOU72, RAN73]. In our medical scenario, the collection of all outcomes is $X = \{a, \overline{a}, b, \overline{b}\}$ given the two incompatible measurements $A = \{a, \overline{a}\}$ and $B = \{b, \overline{b}\}$. Thus, a complete description of the sample space consists not only of the outcome set X, but also the collection of measurements $\mathcal{A} = \{A, B\}$.

Although this may appear to be a trivial reformulation of a simple two-sample experiment, from an EL perspective the explicit inclusion of the structure of measurements in the sample space is a true generalization of Kolmogoroff's axiomatic theory of probability. This generalization allows us to distinguish sharply between two representations of the simple two-sample experiment that are sometimes confused. Letting $s =$ *the survival of the patient* and $d =$ *the death of the patient*, one might represent our medical experiment with the Cartesian product $\{T_A, T_B\} \times \{s, d\}$ with outcome set $\{T_As, T_Ad, T_Bs, T_Bd\}$, where T_A means treatment A is applied and T_B means treatment B is applied. This Cartesian product sample space refers to the joint measurement of the random variables *treatment* with outcomes $\{T_A, T_B\}$ and *effect* with outcomes $\{s, d\}$. It describes a one-sample experiment with two measurements that are made simultaneously on one and the same patient, whereas the first sample space represents the two-sample case with incompatible measurements. The two models describe different situations. Their sample spaces are not equivalent, and consequently they generate different sets of probability measures. Moreover the Cartesian product representation blurs the fact that we have to deal with incompatible measurements. An event such as $\{T_Ad, T_Bd\}$, read as *given treatment A or B, the patient dies*, seems to have a dubious ontological status since there exists no real action that could result in both outcomes. The random variable *effect* with outcomes $\{s, d\}$ is not really an independent measurement without any reference to the treatment. We show here that it is the inadmissable representations of the experiment that lie at the heart of Simpson's Paradox.

A complete and detailed exposition of the mathematical theory underlying EL and a maximum likelihood solution to Simpson's Paradox can be found in [FOU72, RAN73, KLÄ90]. Here we begin to summarize the essential features of the proposed solution by first defining a maximum likelihood estimator (MLE) over a generalized Kolmogoroff-like sample space. Next we characterize compositions of generalized sample spaces (GSSs) in terms of a two-stage hierarchical experiment that we posit is an appropriate method for viewing our medical example. MLEs are then defined over a hierarchical GSS. Finally, we show how the data used in the above example are reconciled by the proposed MLE solution.

A GSS is a nonempty set $\mathcal{A}$ of nonempty sets where $E, F \in \mathcal{A}$ are operations (e.g., $E =$ administering treatment A and $F =$ administering treatment B) and $x \in E$ is called an outcome from performing operation E. Furthermore $X = \cup\{E | E \in \mathcal{A}\}$ is the collection of all outcomes. If $\mathcal{A} = \{X\}$, i.e., $\mathcal{A}$ contains exactly one operation, then $\mathcal{A}$ is said to be classical in the Kolmogoroff sense. If $\mathcal{A} = \{E,\ F\}$ where $E \cap F = \emptyset$, i.e., E and F are incompatible operations, then $\mathcal{A}$ is considered semiclassical. Clearly $\mathcal{A} = \{A,\ B\}$ with $A = \{a,\ \bar{a}\}$ and $B = \{b,\ \bar{b}\}$ from above is semiclassical. If $\mathcal{A}$ is semiclassical then a probability weight on $\mathcal{A}$ is a map $\omega : X \mapsto [0,\ 1]$ such that for $x \in X$ $\sum_{x \in E} \omega(x) = 1$ for all $E \in \mathcal{A}$. Within an EL framework, the data from the execution of operations are absolute frequencies. The number of times $x \in X$ has been observed is represented by N_x, and the total number of observations on $\mathcal{A}$ is $N = \sum_{x \in X} N_x$.

Where $\Omega(\mathcal{A})$ is the collection of all probability weights on $\mathcal{A}$, a MLE on $\mathcal{A}$ is a mapping $\hat{\omega}$ such that

$$\prod_{x \in X} \hat{\omega}(x)^{N_x} \geq \prod_{x \in X} \omega(x)^{N_x} \tag{3}$$

for all $\omega \in \Omega(\mathcal{A})$. If $\mathcal{A}$ is classical then $\hat{\omega}(x) = N_x/N$ is the expected result. If $\mathcal{A}$ is semiclassical, then $\hat{\omega}(x) = N_x/N_E$, where E is a unique operation in $\mathcal{A}$ with $x \in E$, and $N_E = \sum_{y \in E} N_y$ is the total number of observed outcomes from performing operation E.

Lagrange multipliers are used frequently to solve MLE problems. Maximizing $P = \prod_{x \in X} \omega(x)^{N_x}$ with respect to $\omega \in \Omega(\mathcal{A})$ is equivalent to maximizing the expression $\ell n\ P = \sum_{x \in X} N_x \ell n\ \omega(x)$, subject to the conditions $\sum_{x \in E} \omega(x) - 1 = 0 \equiv f_E$ for all $E \in \mathcal{A}$. By Lagrange-multipliers,

$$-\nabla \ell n P + \sum_{E \in \mathcal{A}} \lambda_E \nabla f_E = 0 \quad ,$$

where differentiation is respect to the variables $\omega(x)$. For each $x \in X$ we get

$$N_x \cdot \frac{1}{\omega(x)} = \sum_{x \in F} \lambda_F \quad ,$$

where the sum is to be taken over all $F \in \mathcal{A}$ such that $x \in F$. We therefore find the solution

$$\hat{\omega}(x) = \frac{N_x}{\sum_{x \in F} \lambda_F} \quad .$$

To determine the coefficients λ, observe that we have for each operation $E \in \mathcal{A}$ an equation

$$1 = \sum_{x \in E} \hat{\omega}(x) = \sum_{x \in E} \left[\frac{N_x}{\sum_{x \in F} \lambda_F} \right] \quad ,$$

where a solution is admissible only if $\sum_{x \in F} \lambda_F > 0$ for all $x \in X$.

Suppose $\mathcal{A} = \{E,\ F,\ G\}$ where $E = \{a, b\}$, $F = \{b, c\}$, $G = \{c, d\}$, and $X = \{a, b, c, d\}$. Given the data N_a, N_b, N_c, N_d and $N =$ *the total number of observed outcomes*, using Lagrange-multipliers we find the solution for the coefficients to be

$$\lambda_E = N \cdot \frac{N_a}{N_a + N_b}, \quad \lambda_G = N \cdot \frac{N_d}{N_c + N_b}, \quad \text{and}$$

$$\lambda_F = N \cdot \left[\frac{N_b}{N_b + N_d} - \frac{N_a}{N_a + N_c} \right] \quad .$$

and we find for our given $\mathcal{A} = \{E, F, G\}$ the maximum likelihood estimators to be

$$\hat{\omega}(a) = \hat{\omega}(c) = \frac{N_a + N_c}{N} \quad \text{and} \quad \hat{\omega}(b) = \hat{\omega}(d) = \frac{N_b + N_d}{N} \quad .$$

If $N_a = 0$ then the system of equations is consistent if and only if $N_c = 0$, in which case $\hat{\omega}(a) = \hat{\omega}(c) = 0$ and $\hat{\omega}(b) = \hat{\omega}(d) = 1$. A symmetric argument holds when $N_b = 0$.

SIMPSON'S PARADOX REVISITED

A viable method our physician might use to reason about which treatment to give her patient is to first view her task within the context of a two-stage hierarchical experiment. A typical two-stage experiment involves executing an operation E from a semiclassical sample space $\mathcal{A}$, then depending on the outcome $x \in E$ executing an operation F_x from a semiclassical sample space $\mathcal{B}_x$. The idea is that the first stage of this experiment requires the physician to ask for the patient's city of residence. The sample space for this question is $\mathcal{C} = \{\{c, \overline{c}\}\}$. Then depending on the outcome of the patient's answer, the second stage of the experiment involves selecting the treatment, represented by the sample space $\mathcal{A} = \{A, B\}$ with $A =$

$\{a, \overline{a}\}$ and $B = \{b, \overline{b}\}$. The complete experiment is then represented by the forward operational product $\overrightarrow{\mathcal{CA}}$. However, we have yet to give the essential steps of how to construct a MLE over $\overrightarrow{\mathcal{CA}}$ from the MLEs for each $\mathcal{C}$ and $\mathcal{A}$ separately. The details of these steps are given in [KLÄ90].

For each $x \in X$ let $\mathcal{B}_x$ be a semiclassical sample space with outcome set Y_x and let xF denote the set of all pairs (x, y) with $y \in F$; then a two-stage experiment has the general form

$$\bigcup_{x \in E} xF_x$$

where $E \in \mathcal{A}$ and $F_x \in \mathcal{B}_x$ for all $x \in E$. Let $\omega_{\mathcal{A}} \in \Omega(\mathcal{A})$ and for each $x \in X$ let $\omega_x \in \Omega(\mathcal{B}_x)$; then a probability weight on $\overrightarrow{\mathcal{AB}}$ is a map

$$\omega(xy) := \omega_{\mathcal{A}}(x) \cdot \omega_x(y)$$

where the notation xy denotes the ordered pair (x, y) for all $x \in X$ and all $y \in Y_x$. The absolute frequencies observed over $\omega_x \in \Omega(\mathcal{B}_x)$ are of the form $(N_{xy})_{x \in X, y \in Y_x}$. We use $N_x := \sum_{y \in Y_x} N_{xy}$ to denote the number of observed outcomes $x \in X$.

With $\omega_{\mathcal{A}} \in \Omega(\mathcal{A})$ and $\omega_x \in \Omega(\mathcal{B}_x)$, by hypothesis

$$\prod_{x \in X} \hat{\omega}_{\mathcal{A}}(x)^{N_x} \geq \prod_{x \in X} \omega_{\mathcal{A}}(x)^{N_x} \text{ and } \prod_{y \in Y_x} \hat{\omega}_x(y)^{N_{xy}} \geq \prod_{y \in Y_x} \omega_x(y)^{N_{xy}} \quad .$$

It has been shown in [KLÄ90] that from the above inequalities the following maximum likelihood estimator for $\overrightarrow{\mathcal{AB}}$ can be derived

$$\prod_{x \in X, y \in Y_x} \Big[\hat{\omega}_{\mathcal{A}}(x) \cdot \hat{\omega}_x(y)\Big]^{N_{xy}} \geq \prod_{x \in X, y \in Y_x} \Big[\omega_{\mathcal{A}}(x) \cdot \omega_x(y)\Big]^{N_{xy}} \quad .$$

Given $\mathcal{A} = \{\{a, \overline{a}\}, \{b, \overline{b}\}\}$ and $\mathcal{C} = \{\{c, \overline{c}\}\}$ for our two-stage hierarchical medical experiment, we find using the above MLE techniques that

$$\hat{\omega}_{\mathcal{A}}(a) = \frac{N_a}{N_a + N_{\overline{a}}}, \quad \hat{\omega}_{\mathcal{A}}(\overline{a}) = \frac{N_{\overline{a}}}{N_a + N_{\overline{a}}} \ , \quad \text{and}$$

$$\hat{\omega}_{\mathcal{A}}(b) = \frac{N_b}{N_b + N_{\overline{b}}}, \quad \hat{\omega}_{\mathcal{A}}(\overline{b}) = \frac{N_{\overline{b}}}{N_b + N_{\overline{b}}} \quad .$$

Now consider $\overrightarrow{c\mathcal{A}}$ which has the outcomes $\{ca, c\overline{a}, cb, c\overline{b}, \overline{c}a, \overline{c}\overline{a}, \ \overline{c}b, \overline{c}\overline{b}\}$ and contains the operations $\{ca, c\overline{a}, \overline{c}a, \overline{c}\overline{a}\}$, $\{ca, c\overline{a}, \overline{c}b, \overline{c}\overline{b}\}$, $\{cb, c\overline{b}, \overline{c}a, \overline{c}\overline{a}\}$, $\{cb, c\overline{b}, \overline{c}b, \overline{c}\overline{b}\}$. Notice that this set of operations is not semiclassical because of overlapping operations. If we let N_c and $N_{\overline{c}}$ denote the frequencies of observing c and $\overline{c}$ and let $N = N_c + N_{\overline{c}}$, we find the MLE over $\overrightarrow{c\mathcal{A}}$ to be

$$\hat{\omega}(xy) = \frac{N_x}{N} \cdot \hat{\omega}_{\mathcal{A}}(y) \quad ,$$

where $x \in \{c, \overline{c}\}$ and $y \in \{a, \overline{a}, b, \overline{b}\}$.

Now imagine a patient arrives at our physician's office and is diagnosed as having an illness that is treatable with either A or B. The patient is then asked for his city of origin; based on the answer a treatment will be selected and the outcome observed. This experiment is represented by $\overrightarrow{c\mathcal{A}}$. Using the same data as [BLY72] we find that

$$\hat{\omega}(ca) = 0.0261 \leq \hat{\omega}(cb) = 0.0521 \quad \text{and}$$
$$\hat{\omega}(\overline{c}a) = 0.2393 \leq \hat{\omega}(\overline{c}b) = 0.4547 \quad , \tag{4}$$

and again we judge treatment B to be better than treatment A in each city. The results in Equation 4 were derived by multiplying the results in Equation 1 by the proportion $\frac{N_c}{N}$ of patients in city c, respectively by proportion $\frac{N_{\overline{c}}}{N}$ of patients in city $\overline{c}$. To find the marginal probability weight on $\mathcal{A}$ we sum the probabilities for both cities. That is, for any probability weight ω on $\overrightarrow{c\mathcal{A}}$, we define the marginal probability weight $\omega_{\mathcal{C}}(y) = \omega(cy) + \omega(\overline{c}y)$ for all $y \in \{a, \overline{a}, b, \overline{b}\}$. This result is a true probability weight on the sample space $\mathcal{A}$ alone, and it is derived by averaging over the factor $\mathcal{C}$. To visit the numbers above once again, we find that

$$\hat{\omega}_{\mathcal{C}}(a) = 0.2654 < \hat{\omega}_{\mathcal{C}}(b) = 0.5068 \quad .$$

Treatment B still appears to be better and the paradox has been resolved. It is shown in [KLÄ90] that the construction of $\omega_{\mathcal{C}}$, in general, can never display the paradoxical behavior of the typical calculations whatever the data. We characterize the difference between the classical analysis methods of such data and an EL-based analysis in that the former ignores the information contained in the partition of the data according to their origin, whereas the latter averages over this information.

To summarize, we have shown that by taking into account explicitly the mechanisms of how the data is collected and the interdependence of such data, inconsistencies of the typical calculations are avoided.

Acknowledgments: We are deeply indepted to Charles H. Randall, whose death in the Spring of 1987 prematurely terminated his part in his joint work with David J. Foulis on Simpson's Paradox. One author, Matthias P. Kläy was a Visiting Scholar at the Institute for Mathematical Studies in the Social Sciences, Stanford University, where most of this paper was written. Consequently, we would like to express our appreciation to Professor Patrick Suppes for his hospitality and to both Professor Suppes and Professor Nancy Cartwright for many stimulating discussions.

REFERENCES

[BIC75] Bickel, P.J., Hammel, E.A., O'Connell, J.W., "Sex Bias in Graduate Admissions: Data from Berkely," *Science*, **187**, pp. 398-404, (1975).

[BLY72] Blyth, Colin R., "On Simpson's Paradox and the Surething Principle," *Journal of the American Statistical Association,* (Theory and Methods Section) Vol. 67, No. 338, pp. 364-366, (June 1972).

[BLY73] Blyth, Colin R., "Simpson's Paradox and Mutually Favorable Events," *Journal of the American Statistical Association,* (Theory and Methods Section) Vol. 68, No. 343, pg. 746, (September 1973).

[CHU42] Chung, K-L., "On Mutually Favorable Events," *Annals of Mathematical Statistics,* Vol. 13, pp. 338-349, (1942).

[DAW79] Dawid, A.P., "Conditional Independence in Statistical Theory," *Journal of The Royal Statistical Society*, Series A 41 (No. 1), pp. 1-31, (1979).

[FOU72] Foulis, D.J., Randall, C.H., "Operational Statistics I: Basic Concepts," *Journal of Mathematical Physics*, **13**, pp. 1667-1675, (1972).

[KLÄ90] Kläy, M.P., Foulis, D.J., "Maximum Likelihood Estimation on Generalized Sample Spaces: An Alternative Resolution of Simpson's Paradox," to appear in *Foundations of Physics*, (1990).

[KNA85] Knapp, T.R., "Instances of Simpson's Paradox," *The College Mathematics Journal*, Vol. 16, No. 3, pp. 209-211, (1985).

[KOL33] Kolmogoroff, A., *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer-Verlag, Berlin (1933).

[PEA88] Pearl, J., *Probabilistic Reasoning In Intelligent Systems: Networks of Plausible Inference,* Morgan Kaufmann Publishers Inc., San Mateo, CA., pp. 493-497, (1988).

[RAN73] Randall, C.H., Foulis, D.J., "Operational Statistics II: Manuals of Operations and their Logics," *Journal of Mathematical Physics*, **14**, pp. 1472-1480, (1973).

[SIM51] Simpson, E.H., "The Interpretation of Interaction in Contingency Tables," *Journal of The Royal Statistical Society*, Series B 13, pp. 238-241, (1951).

[WAG82] Wagner, C.H., "Simpson's Paradox in Real Life," *The American Statistician*, Vol. 36, No. 1, pp. 46-48, (1982).

[YUL03] Yule, G.U., "Notes on The Theory of Association of Attributes in Statistics," *Biometrika-II*, (2), pp. 121-134, (February 1903).