Measurement 50 (2014) 319–323

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Aggregation of comparisons data and reversal phenomena of metrological interest ☆

Giampaolo E. D'Errico

Istituto di Ricerca Metrologica – INRIM, strada delle cacce, 73, 10135 Torino, Italy

ARTICLE INFO

Article history:
Received 1 October 2013
Received in revised form 7 January 2014
Accepted 10 January 2014
Available online 20 January 2014

Keywords:
Partwise comparisons
Data amalgamation
Yule–Simpson's paradox
Reversal phenomenon
Decision making
Design of experiments.

ABSTRACT

Aggregation of comparisons data to rank experimental results and take decisions is being more and more practiced in diverse areas, spanning over a variety of disciplines including, e.g., quality function deployment in industrial engineering, scientometrics, and recovery rate testing of new medications. Problems in decision making may be accrued from the presence of hidden confounding interactions, spurious relationships, lurking variables at work. An analysis of partitioned datasets is carried-on using contingency tables and conditional probabilities. The focus is on intermediate interpretation of evidence to avoid paradoxical reversal of statistical inference when passing from sub-level data to the global level: to this aim, care in partitioning criteria is needed to balance distribution of partitioned data over successive levels, not to incur statistical dependence. An example of counter-intuitive amalgamation effects – also known as Yule-Simpson's "paradox" – is presented and discussed, showing how to prevent such effects by proper design of experiments.

1. Introduction

Aggregation of data initially collected according to two separated groupings may give raise to a disturbing phenomenon known under the name of Simpson's paradox. However, "paradox" is a wrong label. In most general terms, it may happen that the probabilistic plausibility of a conclusion drawn from inspection of each single dataset might become questioned (and even contradicted) if data are considered as a whole. Given that merging datasets together is obtained after logically consistent operations, the paradoxicality is not the consequence of an inadvertent mistake, but an uncontrolled effect of the intervention of some confounding variable, sometimes a lurking one.

The paradox was so-named after the title of Blyth's paper [1], referring to the paper published in 1951 by Simpson [2], although the phenomenon was previously detected by diverse authors, such as Yule in 1903 [3] (as noted by Good and Mittal [4], this misnaming is a confirmation of the law that every eponymy is wrong). A "paradox" such Yule–Simpson's one can be referred to as a reversal phenomenon. Apart from precedence of names, an analysis of this phenomenon – in the following called the Yule–Simpson's paradox – is required in statistical terms.

The paucity of studies of Simpson's paradox (and of its generalization) – compared to the abundance of realistic examples that can be found in statistical literature – was a ground for complaint in a more recent work [5]. Although improperly defined a paradox, the occurrence of such a situation is the source of authentic dilemmas when an incontrovertible decision is to be taken based on available evidences. This situation may be encountered in a variety of areas, such as:

- medicine, for instance in epidemiology, where the problem is approached in terms of confounding (e.g. [6]) and also by pointing out other paradoxes (e.g. [7]), or in medical research methodology (e.g. [8]);

☆ Based on the oral presentation "Globalising Partition-wide Comparisons" delivered by author at IMEKO TC8/TC23/TC24 Conference *Metrological traceability in the globalization age*, Paris, 6–8 April 2011.

E-mail address: g.derrico@inrim.it

0263-2241/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.measurement.2014.01.014

- psychological science, see e.g. [9];
- insurance ratemaking, see e.g. [10];
- scientometrics (treatment of bibliometric indicators and impact measures), see e.g. [11], and experimental research assessment, see e.g. [12];
- study of probabilistic causality models, see e.g. [13], and of the logic of Simpson's paradox, see e.g. [14].

From a more general point of view, problems with reliability of partwise comparison are addressed by [15], where paradoxes possibly arising in decision analysis are focused with application to engineering design and consumer survey.

The areas pointed-out so far involve measurement results and comparisons [16,17]; this is the reason why the reversal phenomenon is of metrological interest too. The metrologist may be challenged by analyses of results of comparative experiments performed by third-parties (each of them independent of each other): if different experimenters are presenting different groupings of results, e.g. for conformity purpose, a reversal phenomenon may play a role against impartial conclusion, by introducing a difficult-to-detect bias (for some extensions of the idea of bias, see [18]). This will be the central theme of the next Section 2. Section 3 is devoted to point-out concluding remarks.

2. Statistical analysis with applicability to metrology

Let random events be represented by set-theoretical variables a, b and c; let prime denote the operation of set complementation (in terms of logical connectives, its counterpart is negation), let product denote the operation of set intersection (its counterpart is conjunction, here denoted by &) and let the sign + denote the set-theoretic union (logical counterpart, disjunction). If P stands for a probability measure on such sets, the following system of equations can be written in terms of conditional probabilities:

$$P(a|b) \geqslant P(a|b'), \tag{1a}$$

$$P(a|bc) \leqslant P(a|b'c), \tag{1b}$$

$$P(a|bc') \leqslant P(a|b'c'). \tag{1c}$$

There is no conflict within system of Eq. (1), where each inequality sign can be replaced by its opposite one: however, if at least one of them is strict inequality, the situation can be described as positive Simpson's reversal (as negative Simpson's reversal, if opposite signs replace inequalities in the above system).

The sets b and b', which appear in all of three equations above, are dichotomic subsets: the label "dichotomic" is a short for denoting that two such subsets define an exhaustive partition of their whole universe, say B, into two mutually exclusive subsets, i.e., such that their union is $b + b' = B$ and their intersection bb' is $\varnothing$ (the empty set). Similarly, c and c' are dichotomic with respect to their universe, say C. They are introduced in Eqs. (1b) and (1c), whereas they do not appear in Eq. (1a): while they are remaining latent within Eq. (1a), c and c' play a confounding role. That role becomes explicit through Eqs. (1b) and (1c). Note that $bc + bc' = b$ and $b'c + b'c' = b'$.

Let b stand for patient treated with a new drug and b' for patient treated with placebo, where the group of patients under treatment and the control group are equally numerous (1000 each); let c stand for male patients and c' for female patients. Suppose all inequality signs in the system of Eq. (1) are limited to strict inequalities: thus, if a stands for success (recovery) rate, the interpretation of Eq. (1a) is that the new drug is effective (compared to placebo), regardless patients' gender; on the contrary, Eqs. (1b) and (1c) state that the drug is worst than placebo for both males – Eq. (1b) – and females – Eq. (1c).

A possible realization of this scenario is shown by a numerical example in next tables. From the global dataset reported in Table 1, the success rate of the new drug, administered to patients regardless their gender, indicates it is superior to placebo effects. However, if the global dataset is partitioned according to the patients' gender (the population under treatment and control is composed of two subpopulations of 1000 males and 1000 females), placebo outperforms drug in both subpopulations. A "paradox" therefore would emerge, since an inference about drug efficacy, although based on results from the same experiment, is prone to Simpson's reversal, the final decision depending on which table – whether Table 1 or Table 2 – is taken into account.

If probability is interpreted in term of relative frequency, data in Tables 1 and in 2 lead to a translation of the system of Eq. (1) into the following system of Eq. (2), where; in terms of conditional recovery rates, Eq. (2a) corresponds to Table 1, Eqs. (2b) and (2c) correspond to Table 2:

$$P(\text{success}|\text{drug}) = 50\% > P(\text{success}|\text{placebo}) = 40\%, \tag{2a}$$

$$P(\text{success}|\text{drug} \,\&\, \text{male}) = 60\% < P(\text{success}|\text{placebo} \,\&\, \text{male}) = 70\%, \tag{2b}$$

Table 1
Global dataset (n, l, m, k, N, M, L, K: see Table 2).

Treatment	a: Success count; rate	a': Failure count	Total
b: Drug	$(n + l) = 500$; 50%	500	$(N + L) = 1000$
b': Placebo	$(m + k) = 400$; 40%	600	$(M + K) = 1000$
Total	900	1100	2000

Table 2
Dataset from Table 1 partitioned according patient's gender.

Treatment	a: Success count; rate	a': Failure count	Total
c, Males			
b: Drug	$n = 450$; 60%	300	$N = 750$
b': Placebo	$m = 175$; 70%	75	$M = 250$
Subtotal	625	375	1000
c', Females			
b: Drug	$l = 50$; 20%	200	$L = 250$
b': Placebo	$k = 225$; 30%	525	$K = 750$
Subtotal	275	725	1000

$$P(\text{success}|\text{drug \& female}) = 20\% < P(\text{success}|\text{placebo \& female}) = 30\%. \tag{2c}$$

The paradoxical situation could be managed taking into account the goal of the experiment and the criteria of data stratification. In the above scenario – following the line of reasoning depicted in [19] –, the experiment might be aimed at deciding whether the new drug is approved or not. If both Tables 1 and 2 are available, a prudential safe-oriented decision should be not to admit the new drug in the pharmaceutical market. However, if the decision maker is not acquainted with Table 2 – or perhaps results according to stratified data are purposely unreported (after all, the patient's gender can be known to the experimenter before administering the drug) – approval of the new drug could not be excluded on the base of results reported in Table 1 only. Suppose instead that Tables 1 and 2 are referred to a diverse scenario, such that: b and b' now represent black and white varieties of a plant, respectively; c and c' represent long-stemmed and short-stemmed varieties, respectively; a stands for yield rate. Although numerical results are the same of the drug experiment above, the decision maker might favor the white variety, thus giving priority to the yield rate, regardless the stem length, that can be considered a more or less negligible accident (contrary to gender in the medical example).

A compact explanation of Yule–Simpson's paradox is exposed in [20]; more discussion with application to numerical examples can be found e.g. in [10]. In the following, the statistical machinery that produces such paradoxes is reverse-engineered in its elemental components. Let us reconsider the system of Eq. (1) noting that a can be 2-partitioned (dichotomized) into $a = ac + ac'$ (with $(ac)(ac') = \varnothing$): thus, adding probabilities conditioned on b yields $P(a|b) = P(ac|b) + P(ac'|b)$.

By definition of conditional probability, $P(ac|b) = P(abc)/P(b)$, $P(ac'|b) = P(abc')/P(b)$; moreover $P(abc) = P(a|bc)P(bc)$, $P(abc') = P(a|bc')P(bc')$; $P(bc) = P(c|b)P(b)$, $P(bc') = P(c'|b)P(b)$ (similarly for $P(a|b') = P(ac|b') + P(ac'|b')$. Thus:

$$\begin{aligned} P(a|b) &= P(abc)/P(b) + P(abc')/P(b) \\ &= P(a|bc)P(c|b) + P(a|bc')P(c'|b), \end{aligned} \tag{3a}$$

$$P(a|b') = P(a|b'c)P(c|b') + P(a|b'c')P(c'|b'). \tag{3b}$$

According to Eqs. (2b) and (2c), it may be the case that $P(a|bc) < P(a|b'c)$, $P(a|bc') < P(a|b'c')$ and, in the same time, according to Eq. (2a), $P(a|b) > P(a|b')$. Since Eqs. (3a) and (3b) are weighted means (with weights $P(c|b)$, $P(c'|b)$ in Eq. (3a) and weights $P(c|b')$, $P(c'|b')$ in Eq. (3b)), this kind of reversal may happen – unless b and c are statistically independent (relatively to the probability measure P) – with no reason to invoke any paradox. In fact, given integers n, N, m, M, l, L, k, and K such that $\frac{n}{N} < \frac{m}{M}$ and $\frac{l}{L} < \frac{k}{K}$, it may be true that $\frac{n+l}{N+L} > \frac{m+k}{M+K}$ (see [21] for an elegant geometrical illustration). For example, taking data from Tables 1 and 2:

$$\frac{450+50}{750+250} = 50\% > \frac{175+225}{250+750} = 40\%, \tag{4a}$$

$$\frac{450}{750} < \frac{175}{250}, \tag{4b}$$

$$\frac{50}{250} < \frac{225}{750}. \tag{4c}$$

In fact, independence would impose $P(c|b) = P(c|b')$ and $P(c'|b) = P(c'|b')$, thus mathematically preventing any possibility of reversal at all (absence of confounding variable); in terms of ratios, independence can be translated into conditions:

$$\frac{n}{N} = \frac{l}{L} \quad \text{and} \quad \frac{m}{M} = \frac{k}{K}. \tag{5}$$

If the variable under comparative analysis (in the medical example, success of drug vs. placebo) is independent on the potential confounding variable (the patient's' gender) – i.e., if Eq. (5) holds –subpopulations data fulfil the following relationship too:

$$\frac{m}{M} - \frac{n}{N} = \frac{k}{K} - \frac{l}{L} = \rho, \tag{6}$$

It can be easily verified [10] that when Eq. (6) holds it is also true that:

$$\frac{m+k}{M+K} - \frac{n+l}{N+L} = \rho. \tag{7}$$

In such a scenario, there is no reversal and no change at all (conditions stated in Eqs. (6) and (7) are also known as collapsibility in contingency tables [23]): analyses of stratified or aggregated data lead exactly to the same statistical inference using the criterion of comparing treatments success rates.

Therefore, attention should be focused on whether statistical dependence arises from introduction of a potentially confounding variable used to partition global data. For instance, statistical dependence is introduced when data are unequally distributed over partitions as it happens in Table 2, where $\frac{m}{M} = 70\%$, $\frac{n}{N} = 60\%$, $\frac{k}{K} = 30\%$, $\frac{l}{L} = 20\%$, and $\rho = 10\%$: in this case Eq. (6) holds, whereas conditions of Eq. (5), namely independence, does not hold. This may happen if data are acquired in experiments where not all involved variables are under the control of the experimenter: in the medical example above, the patient's gender is uncontrolled variable. (For more about the crucial role of independence and conditional independence in statistical inference and other areas of statistics, see e.g. [22]).

On the other hand, if the experimenter has control over all involved variables, paradoxes may be avoided by implementing a balanced experiment design: in this case, it is up to the experimenter that the global population be proportionally distributed over subpopulations being investigated in details. An experiment is balanced if:

$$\frac{N}{M} = \frac{L}{K} \tag{8}$$

For example, in the medical scenario depicted above, Eq. (8) translates into imposing the condition that the ratio of number of patients treated with drug to the number of patients of the same gender treated with placebo be the same for both male and female patients. A special instance of proportional distribution is obtained when both $N = L$ and $M = K$, a fully balanced experiment. In a balanced experiment, Eq. (7) follows – see [10] for algebraic passages – from Eqs. (6) and (8).

A metrologist might be faced to a similar dilemma, the horns of which could be positioned in between the two scenarios where the drug recovery rate and the plant yield rate were the decision criteria, respectively. Requests of assessments in the sectors of health, food and nutrition, ecology may bring challenging questions: procedures involving comparisons, traceability, verification and decision making urge caution. A paradigmatic situation is recently reported in [24]: the producer of a drug that failed to show efficacy against placebo after two trials (performed according to two mutually unrelated procedures), asked the regulatory agency – who, being aware of the expected reversal of result interpretation, refused the option – the permission of pooling both trials in view of a global analysis of comparative drug/placebo performance.

From a broader metrological point of view, a seemingly paradoxical situation may arise, e.g., in the framework of conformity assessment. Guidelines how to conduct conformity assessment are provided by standard [25]; here a potential situation is described where a comparative assessment is supposed in view of a decision based on rating two alternative production processes in terms of product quality. Let the quality of process A and process B be compared by inspection of respective products, say A-products and B-products, sampled at random. Let inspections be performed according to two successive measurements campaigns: let iA (iB, respectively) denote the first campaign with application to process A (B, respectively) and similarly let iiA (iiB) denote the second one. Results of iA are compared to results of iB, and results of iiA are compared to those of iiB. It may happen that, for both campaigns, the percentage of A-products satisfying the specified requirements is sensibly greater than the corresponding percentage of B-products. However, in case the two measurement campaigns are designed distinctly from each other, rather than as a whole, they are prone to lack of overall control. In this case a reversal phenomenon might occur if the inspection results are merged all together: inference based on merged data could be biased due to uncontrolled statistical dependence introduced with merging, and/or to uncontrolled presence of unbalanced distributions of sampled products over their populations.

3. Concluding remarks

A reversal phenomenon such as the so called Simpson's paradox may potentially affect grouping of measurements whatever about the metrological area of application. This may happen if data are acquired in experiments where not all the involved variables are under the control of the experimenter: in fact, difficulties related to the presence of hidden confounding interactions, spurious relationships, lurking variables at work, may arise.

Moreover, the metrologist may be challenged by analyses of results of comparative experiments performed by third-parties, in a variety of sensible sectors, including, e.g., health, food and nutrition, and ecology. Further research of metrological interest is envisaged to address the topic of bias caused by reversal phenomena and related effects on uncertainty evaluation and conformity assessment.

In the light of the analysis presented in this paper, the following conclusions are pointed-out:

- paradoxes may arise due to the fact that global data can be diversely partitioned according to given subtotals: statistical dependence is introduced when partitioned data are unbalanced over sub-partitions;
- design of experiments is crucial in view of aggregation of partition-wide comparisons without high risk of (paradoxical) reversal of interpretation of relevant results;
- a probabilistic interpretation of global results, expressed in terms of relative frequency, should be based on the rules of conditional probabilities for correct weighting of mean values over the partitions level.

References

[1] C.R. Blyth, On Simpson's paradox and the sure-thing principle, J. Am. Stat. Assoc. 67 (338) (1972) 364–366.

[2] E.H. Simpson, The interpretation of interaction in contingency tables, J. R. Stat. Soc. Ser. B 13 (2) (1951) 238–241.

[3] G.U. Yule, Notes on the theory of association of attributes in statistics, Biometrika 2 (2) (1903) 16–134.

[4] I.J. Good, Y. Mittal, The amalgamation and geometry of two-by-two contingency tables, Ann. Stat. 15 (2) (1987) 694–711. + Addendum, Ann. Statist. 17 (2) (1989), p. 947.

[5] J.L. Petit, Généralisation du paradoxe de Simpson, Rev. Statistique Appliquée 40 (3) (1992) 47–61.

[6] J.P. Vandenbroucke, The history of confounding, Soz.-Präventimed. 47 (2002) 216–224.

[7] Y.-K. Tu, D. Gunnell, M.S. Gilthorpe, Simpson's Paradox, Lord's paradox and suppression effects are the same phenomenon – the reversal paradox, Emerging Themes Epidemiol. 5 (2) (2008), http://dx.doi.org/10.1186/1742-7622-5-2.

[8] G. Rücher, M. Schumacher, Simpson's paradox visualized: the example of the rosiglitazone meta-analysis, BMC Med. Res. Meth. 8 (34) (2008), http://dx.doi.org/10.1186/1471-2288-8-34.

[9] R.A. Kievit, W.E. Frankenhuis, L.J. Waldorp, D. Borsboom, Simpson's paradox in psychological science: a practical guide, Front. Psychol. 4 (2013), http://dx.doi.org/10.3389/fpsyg.2013.00513 (art. 513).

[10] J.A. Stenmark, C.-S.P. Wu, Simpson's paradox, confounding variables and insurance ratemaking, Proc. Casualty Actuarial Soc. 91 (174) (2004) 133–198.

[11] S. Ramanana-Rahary, M. Zitt, R. Rousseau, Aggregation property of relative impact and other classical indicators: convexity issues and the Yule–Simpson paradox, Scientometrics 79 (2) (2009) 311–327.

[12] H.H. Goltz, M.L. Smith, Yule–Simpson's paradox in research, Pract. Assess. Res. Eval. 15 (15) (2010). <http://pareonline.net/getvn.asp?v=15&n=15>.

[13] V.G. Hardcastle, Partitions, probabilistic casual laws and simpson's paradox, Synthese 86 (1991) 209–228.

[14] P.S. Bandyoapdhyay, D. Nelson, M. Greenwood, G. Brittan, J. Berwald, The logic of Simpson's paradox, Synthese 181 (2) (2011) 185–208, http://dx.doi.org/10.1007/s11229-010-9797-0.

[15] D.G. Saari, K.K. Sieberg, Are partwise comparisons reliable?, Res Eng. Des. 15 (2004) 62–71.

[16] G.E. D'Errico, Paradigms for uncertainty treatments: a comparative analysis with application to measurement, Measurement 42 (2009) 494–500.

[17] G.E. D'Errico, Issues in significance testing, Measurement 42 (2009) 1478–1481.

[18] H.R. van der Vaart, Some extensions of the idea of bias, Ann. Stat. 32 (2) (1961) 436–447.

[19] D.V. Lindley, M.R. Novick, The role of exchangeability in inference, Ann. Stat. 9 (1) (1981) 45–58.

[20] G.J Szekely, Paradoxes in Probability Theory and Mathematical Statistics, Reidel, Dordrect (Holland), 1986, p. 58.

[21] J. Kocik, Proof without words: Simpson's paradox, Math. Mag. 74 (5) (2001) 399.

[22] A.P. Dawid, Conditional independence in statistical theory (with comments), J. R. Stat. Soc. Ser. B 41 (1) (1979) 1–31.

[23] S. Greenland, J.M. Robins, J. Pearl, Confounding and collapsibility in causal inference, Stat. Sci. 14 (1) (1999) 29–46.

[24] O.E. Percus, J.K. Percus, How to win without overtly cheating: the inverse Simpson paradox, Math. Intell. 32 (4) (2010) 49–52.

[25] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, Evaluation of measurement data - the role of measurement uncertainty in conformity assessment, Joint Committee for Guides in Metrology, JCGM 106:2012, 2012.