Statistical Methodology 7 (2010) 323–337

Contents lists available at ScienceDirect

Statistical Methodology

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

Informative allocation and consistent treatment selection

Tamás Rudas

Department of Statistics, Faculty of Social Sciences, Eötvös Loránd University, Budapest, Hungary

ARTICLE INFO

Article history:
Received 7 March 2009
Received in revised form
15 September 2009
Accepted 22 September 2009

Keywords:
Cross product ratio
Cross sum ratio
Decision functions
Simpson's paradox

ABSTRACT

When data in the form of a 2×2 treatment by response table are available, the better out of the two treatments is often selected using the odds ratio (cross product ratio). Such decisions do not depend on the allocation of the observations in the different treatment categories and may exhibit a counter-intuitive reversal property, called Simpson's paradox. In cases when those receiving different treatments are potentially different, as in observational studies or in designed experiments with dropout or noncompliance, decisions taking into account the difference in observed allocations may be useful. Using an approach that postulates certain desirable properties of decision functions and derives further characteristics from these, a new decision procedure based on the cross sum ratio is investigated. This procedure is not only sensitive to allocation but also turns out to be the only selection procedure that avoids Simpson's paradox. In addition to these logical advantages, the probability of wrong decision when using the cross sum ratio tends to be smaller than when using the cross product ratio. The application of the new decision procedure is illustrated by the re-analysis of data sets, some of which exhibit Simpson's paradox when analyzed using the cross product ratio. Finally, generalizations of the decision procedures to $2 \times J$ decision tables are considered.

1. Introduction

This paper deals with the comparison of two treatments, $Tr1$ and $Tr2$, based on data in the form of a 2×2 contingency table or decision table T:

E-mail address: rudas@tarki.hu.

1572-3127/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.stamet.2009.09.006

$T =$

Response	Positive	Negative
Tr1	a	b
Tr2	c	d

Such data arise in a variety of contexts as the result of different data collection procedures, including designed experiments (clinical trials), case-control studies [8] and observational studies [19]. The goal of the analysis is to evaluate the treatments relative to each other, either in the context of causal decision theory or that of evidential decision theory (see, e.g., [15,6]). From a practical perspective, several diagnostic [12] and treatment decisions [7] in medicine and also other policy decisions (see, e.g., [24]) are based on the analysis of data as in T.

The comparison is often based on the value of the odds ratio [20] or cross product ratio (cpr)

$$\frac{a/b}{c/d} = \frac{ad}{bc}$$

and on the value of the relative risk (rr), or risk ratio that, depending on the context, is also called success rate (see, e.g., [8]),

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

and deems $Tr1$ better than $Tr2$ if the cpr or the rr is greater than 1.

A formal discussion of treatment comparisons may be given by using decision functions. For all tables T, these functions are defined to yield 1 or -1 or 0, telling which treatment is better (the first one, the second one, none of them). For example, a decision function based on the cpr may defined as

$$CPR = \operatorname{sgn}\left(\log\left(\frac{ad}{bc}\right)\right),$$

where sgn is the sign function. Two such decision functions γ_1 and γ_2 are considered equal, if for all tables T, $\gamma_1(T) = \gamma_2(T)$. Therefore, the decision functions are, in fact, equivalence classes and the same decision functions may be represented by different functional forms. This is similar to considering the functions $f(x) = x$ and $g(x) = (x^3)^{1/3}$ equal. For example, it is easy to see that the CPR is the same decision function as $\operatorname{sgn}(\log(rr))$. In Section 2, six basic properties of such decision functions are postulated and all possible decision functions with these properties are investigated.

The developments in this paper are motivated by two properties of decisions based on the CPR, which may be seen as problematic, at least in certain situations. The first such property is illustrated by the fact that it selects the same treatment (namely $Tr1$) based on the following tables:

$T_1 =$

Response	Positive	Negative
$Tr1$	450	100
$Tr2$	280	70

$T_2 =$

Response	Positive	Negative
$Tr1$	90	20
$Tr2$	600	150

The two sets of data are identical with respect to the conditional distribution of response, given treatment, but they are different with respect to the treatment marginals, that is, with respect to allocation. In other words, cpr and CPR are invariant against changes in allocation, if these changes leave the conditional distribution of response, given treatment, unaffected (see (4)).

Whether arriving at the same conclusion from these two sets of data is appropriate or not, depends on whether the difference in allocation is the result of the action of the experimenter, as in a designed experiment, and in this case reaching the same conclusion seems justified based on an exchangeability argument [14], or rather is the result of some procedure outside of the control of the experimenter, as in an observational study, and in this case it is not obvious whether reaching the same conclusion is justified or not. In the former case, allocation into treatment categories is noninformative and in the latter case it is, at least potentially, informative. This difference needs to be taken into account when choosing the decision function. Note, that even in the case of an originally noninformative allocation, as in a clinical trial, different dropout rates [9] and different compliance rates [9,22] often make the observed treatment marginal informative.

The second characteristic of decisions based on the *CPR* is that the so-called Simpson's paradox [4,14,23,15,2,16] may occur. Simpson's paradox is said to have occurred, when one treatment is better in the data than the other one, but if the data are split according to another variable, the other treatment appears better in both subgroups (briefly referred to as reversal):

$$CPR(T_1) = CPR(T_2) \neq CPR(T_1 + T_2),$$

where the sum of two tables is the table with the sum of the individual frequencies.

It is proved in Section 2 that the two issues related to the use of the *CPR* to select a treatment are, in fact, related. Simpson's paradox may occur with any decision function that is invariant against changes in allocation (including also the risk ratio).

The main formal *result* of the paper is that there exists only one consistent (see (3)) decision function, that is such that Simpson's paradox cannot occur with it, and this is

$$CSR = \operatorname{sgn}\left(\log\left(\frac{a+d}{b+c}\right)\right).$$

This decision function is based on the cross sum ratio

$$csr = \frac{a+d}{b+c}.$$

The main *message* of the paper is that when allocation is potentially informative, a decision function, like the *CSR* that is sensitive to allocation may be preferable to the *CPR*. Obviously, there are data sets where the *CPR* and the *CSR* lead to different conclusions. For example, with the above data, $CPR(T_2) = 1$ and $CSR(T_2) = -1$, a conclusion that may appear surprising to one's intuition—an intuition that has been educated using the *cpr*. The decision based on the *CSR* reflects the fact that in T_2, $Tr2$ was beneficial to 450 subjects in excess of those to whom it was detrimental, while the same quantity is only 70 for $Tr1$, because the *CSR* depends on how $a - b$ compares to $c - d$. In fact, for $g_i = (g_{i1}, g_{i2})$, and $\delta(g_i) = g_{i1} - g_{i2}$,

$$\operatorname{sgn}(\delta(g_1) - \delta(g_2)) = CSR \quad \text{if } g_1 = (a, b) \quad \text{and} \quad g_2 = (c, d) \tag{1}$$

and

$$\operatorname{sgn}(\delta(g_1) - \delta(g_2)) = CPR \quad \text{if } g_1 = \left(\frac{a}{a+b}, \frac{b}{a+b}\right) \quad \text{and} \quad g_2 = \left(\frac{c}{c+d}, \frac{d}{c+d}\right). \tag{2}$$

That is, the *CSR* and the *CPR* are the same function applied to the observed conditional frequency distributions or to the observed conditional probability distributions in the decision table. The difference between the two decisions is exactly whether or not the $a + b$ and $c + d$ marginal distributions, that is the allocations into treatment categories, are taken into account.

It has to be emphasized again, that the *CSR* is investigated here as a decision function and it is a simple representative of an equivalence class of decision functions. Therefore, the *csr* is not considered a measure of effect strength and issues related to the *cpr* in that context (see, e.g., [5,21]) need not be considered.

Section 3 deals with sampling properties of the *CSR*, in particular the probability of wrong decision based on a sample. In this comparison, the *CSR* presents itself as a more reliable decision function than the *CPR*. Further, confidence intervals for the true value of the *CSR* are proposed.

Section 4 discusses real examples when $CPR \neq CSR$, including the Berkeley admissions data [3], data concerning medical school applications [24] and data regarding administering penicillin to children with meningococcal disease [13]. The analysis based on the *CSR* removes the paradoxical elements present in the published analyses of these data sets and gives new insight into the relationships among the variables involved. The practice of using the net approval rate in opinion polls is also discussed and is shown to be equivalent to using the *CSR*.

Section 5 points to possible extensions of the *CPR* and of the *CSR* beyond 2×2 decision tables. When comparing several treatments, both the *CPR* and the *CSR* are transitive. Extensions of the *CSR* and of the *CPR* based on (1) and on (2) to situations with more than two different responses show that Simpson's paradox may occur with the generalized *CPR* but the generalized *CSR* remains free from it.

Section 6 concludes the paper with a brief discussion.

2. Decision functions

This section gives an axiomatic treatment of decision functions. Obvious properties that decision functions have to possess are postulated. Based on these properties, the relationship between invariance against changes in allocation and the possibility of Simpson's paradox to occur, and the roles played by the *CPR* and the *CSR* among decision functions will be clarified. To simplify presentation, all cell entires will be assumed to be positive.

The first two properties describe noninformative sets of data, that is data that imply that no treatment is preferable to the other. A data set is not informative for treatment selection if the number of positive and negative responses is the same within both treatments.

Property 1.

$$\gamma\left(\begin{array}{|c|c|}\hline a & a\\ \hline b & b\\ \hline\end{array}\right)=0.$$

A data set is also not informative if the same frequencies are observed in the response categories of both treatments.

Property 2.

$$\gamma\left(\begin{array}{|c|c|}\hline a & b\\ \hline a & b\\ \hline\end{array}\right)=0.$$

The next two properties postulate simple antisymmetries: if the rows are interchanged, then the decision changes to the opposite,

Property 3.

$$\gamma\left(\begin{array}{|c|c|}\hline a & b\\ \hline c & d\\ \hline\end{array}\right)=-\gamma\left(\begin{array}{|c|c|}\hline c & d\\ \hline a & b\\ \hline\end{array}\right),$$

and if the positive and negative responses are interchanged, then the decision changes to its opposite.

Property 4.

$$\gamma\left(\begin{array}{|c|c|}\hline a & b\\ \hline c & d\\ \hline\end{array}\right)=-\gamma\left(\begin{array}{|c|c|}\hline b & a\\ \hline d & c\\ \hline\end{array}\right)$$

The last two properties describe data sets that imply that the first treatment is better than the second one. This is the case if $Tr1$ has more positive than negative responses and $Tr2$ does not have more positive than negative responses,

Property 5.

$$a>b,\ c\le d\Rightarrow\gamma\left(\begin{array}{|c|c|}\hline a & b\\ \hline c & d\\ \hline\end{array}\right)=1,$$

and this is also the case when the two treatments have the same number of negative responses but $Tr1$ has more positive responses than $Tr2$:

Property 6.

$$a>c,\ d=b\Rightarrow\gamma\left(\begin{array}{|c|c|}\hline a & b\\ \hline c & d\\ \hline\end{array}\right)=1.$$

In what follows, only decision functions with the above properties will be considered. Three further properties, not necessarily possessed by a decision function, will play a central role in the analysis of decision functions. The first one is related to combining data sets and means that Simpson's paradox cannot occur.

Consistency

$$\text{If } \gamma(T_1) = \gamma(T_2), \text{ then } \gamma(T_1 + T_2) = \gamma(T_1). \tag{3}$$

The second one addresses a related issue, namely the combination of decisive data (that is, data where one treatment appears to be better than the other one) with noninformative data in Properties 1 and 2.

Indifference

$$\text{If } \gamma(T) = 0 \quad \text{by Property 1 or Property 2, then} \gamma(T_1 + T) = \gamma(T_1), \text{ for all } T_1.$$

Note that indifference does not assume that adding any T with $\gamma(T) = 0$ leaves the decision unchanged.

An important aspect of the argument in this paper is the distinction between decision functions that are and that are not sensitive to changes in allocation into treatment categories.

Invariance against changes in allocation

$$\gamma\left(\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline\end{array}\right) = \gamma\left(\begin{array}{|c|c|}\hline ta & tb \\ \hline uc & ud \\ \hline\end{array}\right) \tag{4}$$

for every table and all positive t and u.

The following result may be verified directly.

Proposition 1. *Properties 1–6 hold true for both CPR and CSR. Further, CPR is not consistent and not indifferent but is invariant against changes in allocation, CSR is consistent and indifferent but is not invariant against changes in allocation.* □

Proposition 2. *If a decision function is invariant against changes in allocation, then it is equal to the CPR.* □

The proofs of the results of the paper are given in the Appendix.

Propositions 1 and 2 readily imply

Proposition 3. *If a decision function is consistent, then it cannot be invariant against changes in allocation.* □

The variation independence of the odds ratio from the marginals of the two-way contingency table, which is a very desirable property of it when used as a measure of association (see, e. g., [20]), turns out to be the 'reason' that Simpson's paradox may occur, when used as a decision function.

Proposition 4. *Any indifferent decision function is equal to the CSR.* □

Propositions 1 and 4 imply

Proposition 5. *If γ is indifferent then it is also consistent.* □

The next important result is

Proposition 6. *Any consistent decision function is equal to the CSR.* □

The foregoing results are summarized in

Proposition 7. *The following three statements are equivalent*

(a) *γ is consistent*
(b) *γ is indifferent*
(c) *γ = CSR.* □

Indifference and consistency are equivalent and every decision function with these properties is equal to the *CSR*, if Properties 1–6 are assumed to hold. Invariance against changes in allocation, on the other hand, implies that the reversal may occur and every such decision function is equal to the *CPR*. However, if there are equal allocations into the two treatment categories of T, then the *CSR* and the *CPR* reach the same conclusion.

Proposition 8. *If $a + b = c + d$, then $CSR(T) = CPR(T)$.* □

To complete this section, we note that the indifference property used so far implies a more general indifference.

Proposition 9. *If γ is indifferent, then $\gamma(T_1) = 0$ implies that*

$$\gamma(T_1 + T_2) = \gamma(T_2)$$

for all T_2. □

3. Statistical properties of the *CSR* and the *CPR*

Because the *CSR* is proposed here as a decision function, out of its statistical properties, the probability of wrong decision is of central interest. A wrong decision is to choose $Tr1$ ($Tr2$) based on the observed value of the *CSR* when its population value is equal to -1 or 0 ($+1$ or 0). This probability is easily evaluated under multinomial sampling. Suppose that the treatment by response table in the population has the following distribution

p_a	p_b
p_c	p_d

and the relevant random variables behind observations a, b, c, d are A, B, C, D, respectively. When (A, B, C, D) has a multinomial joint distribution with parameters $n(= a+b+c+d)$ and (p_a, p_b, p_c, p_d), then $A + D \sim B(n, p_a + p_d)$.

Proposition 10. *The probability of wrong decision under multinomial sampling with the CSR if $p_a + p_d > p_b + p_c$ is*

$$\Phi\left(\frac{1}{2}\sqrt{n}\frac{1 - csr}{\sqrt{csr}}\right) + o(1),$$

where Φ is the standard normal distribution function. □

Exact probabilities of wrong decision are given for selected values of n and of $p_a + p_d$ in Table 1. As one would expect, the probability of choosing $Tr2$, when $Tr1$ is better, reduces with increasing values of the *csr* and with increasing sample sizes. For values of $p_a + p_d$, or of the *csr*, larger than those reported in Table 3, the probability of wrong decision is zero for the sample sizes investigated. The same table may be used to read off the probability of wrong decision when $CSR = -1$: one simply has to use the entry $1/csr$ in these cases.

Threshold values of *csr*, for different sample sizes, such that if the true *csr* is greater than the respective threshold value, the probability of wrong decision (concluding from the sample that $CSR = -1$ or $CSR = 0$) is less than 0.05, are given in Table 2. The reciprocals of these values are similar thresholds in cases when $csr < 1$.

Comparable results are difficult to obtain for the *CPR*, mainly because its variance, even asymptotically that could be used to calculate an approximation, does not only depend on the value of the *cpr*, but rather on the individual cell probabilities.

For purposes of comparison, for each sample size investigated, 1000 2×2 tables were generated that had the *csr* values given in Table 2. These tables had uniformly distributed p_a and p_b values and the p_c and p_d values were implied. Then the probability of wrong decision was approximated using the asymptotic distribution of the logarithm of the *cpr* (see e.g., [1] for a review of the large sample properties of the *cpr*). In each case, the probability of wrong decision was determined with reference to the correct decision according to the *CPR* which, in some cases, was the opposite of the decision suggested by the *CSR*. Table 3 reports the average approximate error probabilities. Note that the average is taken here as a simple descriptive statistic and no assumption concerning a superpopulation of 2×2 tables is made. The average error probabilities for the *CPR* were considerably higher than those for the *CSR* but the advantage of the *CSR* seemed to decrease with increasing sample size. These results

Table 1
Exact probabilities of wrong decision ($CSR = -1$ or $CSR = 0$) for various sample sizes and population values of the *csr*, when $CSR = 1$.

$p_a + p_d$	$p_b + p_c$	*csr*	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
0.51	0.49	1.04	0.4599	0.4158	0.3436	0.2739
0.52	0.48	1.08	0.3816	0.3100	0.1975	0.1086
0.53	0.47	1.13	0.3078	0.2178	0.0970	0.0309
0.54	0.46	1.17	0.2409	0.1437	0.0402	0.0062
0.55	0.45	1.22	0.1827	0.0887	0.0140	0.0008
0.56	0.44	1.27	0.1341	0.0511	0.0040	0.0001
0.57	0.43	1.33	0.0950	0.0273	0.0010	0.0000
0.58	0.42	1.38	0.0650	0.0136	0.0002	0.0000
0.59	0.41	1.44	0.0428	0.0062	0.0000	0.0000
0.60	0.40	1.50	0.0271	0.0026	0.0000	0.0000
0.61	0.39	1.56	0.0165	0.0010	0.0000	0.0000
0.62	0.38	1.63	0.0096	0.0004	0.0000	0.0000
0.63	0.37	1.70	0.0054	0.0001	0.0000	0.0000
0.64	0.36	1.78	0.0029	0.0000	0.0000	0.0000
0.65	0.35	1.86	0.0015	0.0000	0.0000	0.0000
0.66	0.34	1.94	0.0007	0.0000	0.0000	0.0000
0.67	0.33	2.03	0.0003	0.0000	0.0000	0.0000
0.68	0.32	2.13	0.0001	0.0000	0.0000	0.0000
0.69	0.31	2.23	0.0001	0.0000	0.0000	0.0000
0.70	0.30	2.33	0.0000	0.0000	0.0000	0.0000

Table 2
Minimum required population values of the *csr* to have wrong decision ($CSR = -1$ or $CSR = 0$) probabilities not exceeding 0.05 for various samples sizes.

Sample size	100	200	500	1000
Minimum value of the csr	1.4178	1.2743	1.1631	1.1119

Table 3
Average approximate error probabilities for 1000 tables each having the *csr* values given in Table 2, when the *CPR* is used for decision. In all tables, $CSR = 1$ and the probability of wrong decision is 0.05 when the *CSR* is used. In some tables, $CPR = -1$, in others $CPR = 1$ and wrong decision is always the opposite or $CPR = 0$.

Sample size	100	200	500	1000
Average error probability of the *CPR*	0.12	0.10	0.08	0.07

indicate that decisions based on the *CSR* are more stable in the statistical sense than those based on the *CPR*.

Because the *CSR* takes on three values only, the construction of confidence intervals simplifies to the question whether or not a confidence interval with a given level of confidence contains any value of the *CSR*, other than the observed one.

Proposition 11. *If $CSR = 1$ is observed under multinomial sampling, an asymptotic $1-\alpha$ level confidence interval for the true value of CSR does not contain 0 or -1 if*

$$1 - \Phi\left(\frac{2(a+b) - n}{\sqrt{n}}\right) < \alpha. \quad \square \tag{5}$$

When the value of $CSR = 1$ is observed in the data, one can say that a 95% confidence interval does not contain 0 or -1, if the *csr* is such that the left hand side of (5) is less than 0.05. Threshold values of the *csr*, such that the left hand side of (5) is equal to 0.05, for various values of n, are given in Table 4.

When the sampling scheme specifies the row marginals of the decision table, allocation in the treatment categories is not informative and the use of the *CPR* is not suggested. When the column

Table 4
One-sided 95% confidence bounds for various sample sizes for the true value of the *CSR*. When the observed value of *csr* exceeds (is less than) the value given, a one-sided 95% confidence interval does not contain $CSR = 0$ and $CSR = -1$ ($CSR = 0$ and $CSR = 1$).

Sample size	50	100	200	500	1000	1500	2000	3000	5000
Observed $csr > 1$	1.61	1.39	1.26	1.16	1.11	1.09	1.08	1.06	1.05
Observed $csr < 1$	0.62	0.71	0.79	0.86	0.90	0.92	0.93	0.94	0.95

marginals are fixed, say N_+ and N_-, as in a case-control study, and sampling is product multinomial, the probability of wrong decision with the *CSR* if $p_a + p_d > p_b + p_c$ is

$$\sum_{0\le A\le N_+}\left(\binom{N_+}{A}p_a^A p_c^{N_+-A}\sum_{A+(N_--N_+)/2\le B\le N_-}\binom{N_-}{B}p_b^B p_d^{N_--B}\right),$$

where [.] means integer part. Using $X = A - C$ and $Y = B - D$, similarly to Proposition 10, a normal approximation of this probability may be obtained.

Proposition 12. *Under product multinomial sampling, with fixed N_+ and N_-, the probability of wrong decision with the CSR, when $p_a + p_d > p_b + p_c$, is*

$$\Phi\left(\frac{E_Y - E_X}{\sqrt{V_X + V_Y}}\right) + o(1),$$

where

$$E_X = N_+(p_a - p_c), \quad E_Y = N_-(p_b - p_d)$$
$$V_X = 4N_+\frac{p_a}{p_a + p_c}\frac{p_c}{p_a + p_c}$$
$$V_Y = 4N_-\frac{p_b}{p_b + p_d}\frac{p_d}{p_b + p_d}. \quad \square$$

4. Applications of the *CSR*

In this section we revisit three data sets where inconsistent findings using the *CPR* have been discussed in the literature and compare the standard analysis to that of obtained by using the *CSR*, and also discuss the practice of using the net approval rates in opinion polls that is closely related to the *CSR*.

4.1. Berkeley admissions

The first example we consider is the famous Berkeley admissions data [3]. Here, the gender by admission status table shows that male applicants had a higher chance of being admitted to the graduate school than female applicants had, and this observation raised the question of gender discrimination. When the data are studied at the departmental level, it turns out that female applicants had an advantage in all departments where there was a considerable difference between the admission rates. The marginal admission rate was favorable to male applicants because they applied in larger fractions to departments where the overall admission rate was high, while female applicants tended to apply to departments with lower overall admission rates. Because admission decisions were made at the departmental level, the suspicion of gender discrimination was refuted. This analysis is widely accepted today and is also reported in textbooks (see, e.g., [10]).

The foregoing analysis used, essentially, the *CPR* and gave an explanation of the paradox that relies on different allocations. In other words, an analysis that is not sensitive to allocation, suggests that allocation should be taken into account. If so, the analysis based on the *CPR* is not appropriate, rather,

Table 5
Berkeley admissions: *cpr* and *csr* values for the six largest departments.

	Dept A	Dept B	Dept C	Dept D	Dept E	Dept F	Total Depts A–F	Total grad school
cpr	0.36	0.80	1.14	0.91	1.23	0.85	1.84	1.46
csr	1.32	1.61	1.26	0.93	1.52	0.91	1.21	1.05

Table 6
Medical school applications: *cpr* and *csr* values for different MCAT-BS scores.

Score	3 or less	4	5	6	7	8	9	10 or more	Total
cpr	1.75	1.81	2.00	2.42	2.40	1.86	1.55	0.93	0.93
csr	1.17	1.57	1.72	1.62	1.20	0.87	0.61	0.48	0.77

an analysis using the *CSR*, which is sensitive to allocation, is needed. Table 5 contains the values of these decision functions for the six largest departments based on the data as reported in [10].

The analysis based on the *CSR* shows that in four out of the six departments and in both totals, men have an advantage, including Department A, where the *CPR* found the largest female advantage. In the remaining two departments (D and F) women have an advantage according to the *CSR*, although a 95% level confidence interval for the true value of the *CSR* contains 1 (that is, male advantage). The *CSR* shows an overwhelming male advantage in the entire admission procedure, including application and selection. Although there was no male advantage when selection was looked at only using the *CPR*, the fact that men tended to apply to programs with high admission rates (or programs where men applied tended to have high admission rates), points to a male advantage indicated by the *CSR* that is sensitive to allocation and, therefore, can incorporate the effects of both application and selection in the admission process. Of course, whether or not this advantage has a discriminatory nature, cannot be decided based on the data only.

4.2. Medical school applications

Wainer and Brown [24] reported medical school application percentages for MCAT Biological Sciences test takers with various scores. In all score groups, black test takers applied to medical schools in larger fractions than white test takers did. In spite of this, among all test takers, whites applied to medical school in larger fraction than blacks did. Again, the paradox goes away when the *CSR* is used to decide whether blacks or whites tend to apply more typically to medical school. For test results up to 7, blacks tend to apply to medical school and for scores higher than 7, white test takers apply more typically to medical school, and the same is true when the test scores are disregarded. The *cpr* and *csr* values are given in Table 6.

4.3. Administering penicillin to meningococcal disease patients

The effect of administering parenteral penicillin to children with meningococcal disease (MC) before admission to hospital has been the subject of considerable debate. Harnden et al. [13] reported a *cpr* of 5.96 indicating a sizable association between administering penicillin and death for children with MC diagnosed prior to hospital admission. Although they warned that it might very well be the case that severity of the disease increased the chances of administering penicillin and the increased odds of death among those who received penicillin might be a consequence of this selection, a *British Medical Journal* editorial, based on the *CPR*, went as far as saying that 'Prehospital penicillin for MC may be harmful. . .'. Other contributions supported the practice of administering penicillin [25]. In a commentary [17], the contributing statistician described that the odds ratio obtained depended strongly on whether all children with MC or only those for whom the GP diagnosed the disease were taken into account. For the former group the odds ratio of death as opposed to survival, for those who did versus those who did not receive penicillin, was 0.86. Thus, overall, there appeared to be a small protective effect but for those with a diagnosed MC, there was harm associated with administering

Table 7
Patients with meningococcal disease diagnosed before hospital admission [18]. $CPR = 1$ ($cpr = 5.96$) and $CSR = -1$ ($csr = 0.79$).

Response	Died	Survived
Penicillin	22	83
No penicillin	2	45

Table 8
All patients with meningococcal disease [18]. $CPR = -1$ ($cpr = 0.86$) and $CSR = 1$ ($csr = 1.73$).

Response	Died	Survived
Penicillin	22	83
No penicillin	81	262

penicillin, a conclusion that seemed somewhat paradoxical. Given the relatively limited information about covariates, in conclusion, [17] warned that 'strong associations are not necessarily causal'.

The two data sets [18] are reported here as Tables 7 and 8. They are examples of the situation when $CSR = -CPR$. The data came essentially from an observational study. Allocation is informative, not only concerning the numbers treated but also because of the best judgment applied by the GP's. Consequently, the application of the *CPR* may not be appropriate. For children with MC diagnosed before hospital admission, $csr = 0.79$ ($CSR = -1$, not significantly different from $CSR = 0$ or 1). For all children with MC, $csr = 1.73$ ($CSR = 1$, significantly different form $CSR = 0$ or -1).

For children with MC diagnosed before hospital admission, the *CSR* suggests a beneficial effect of penicillin, in conformity with expectation, however this effect is weak. For all children, the *CSR* suggests harm associated with penicillin. The conclusion based on the *CSR* confirms the intuition of [13], who assumed that those who did receive penicillin could be more seriously ill than those who did not receive penicillin. However, as suggested by the *CSR*, this applies not within the group of those with an early MC diagnosis but rather to all children who later turned out to have MC. Our results, based on the *CSR*, support the practice of administering penicillin to children who are diagnosed with MC.

4.4. Net ratings in opinion polls

It is very common in the polling industry to use net ratings (or net approval rates) to compare the public's view on two politicians (e.g., two candidates for an office). The net rating is the difference between the percentage of those who say a certain characteristic applies and the percentage of those who say it does not apply to the politician (see, e. g., [11]).

If the following data are observed for a given characteristic

Response	Applies	Does Not Apply	No Opinion
Politician A	a	b	e
Politician B	c	d	f

then the net approval rates of the two politicians are $a - b$ and $c - d$, respectively. A comparison of the two politicians' approvals based on the net approval rates is exactly the same as their comparison based on the *CSR* in the following table:

Response	Applies	Does Not Apply
Politician A	a	b
Politician B	c	d

Note, that in the first table $a+b+e = b+c+f$ (the same sample is interviewed about the politicians) but in the second table $a + b \neq c + d$ (because of differences in the numbers of people who know them). The use of the net ratings is suggested to reduce or eliminate the effect of different levels of name recognition that is, different allocations of those who have an opinion. On the other hand, there seems to be no trace in the published literature of the ratio of positive to negative ratings being used

as a means of comparing the overall approval rates, which would be parallel to a decision based on the *CPR*.

5. More general decision tables

A reviewer asked whether any of the above results can be extended beyond 2×2 tables. This section gives a brief discussion of such possibilities.

First note that Simpson's paradox is a lack of collapsibility property in $2 \times 2 \times 2$, or more generally, $K \times 2 \times 2$ decision tables, where the two treatments are observed under 2 (or K) different scenarios. Here, decisions based on the *CSR* are collapsible in the sense that if the same treatment is better in each of the 2 (or K) conditions, then the same treatment is also better overall. This is not always true for decisions based on the *CPR*.

When several treatments are compared in an $I \times 2$ decision table, both the *CSR* and the *CPR* lead to decisions that are transitive. More precisely, it is easy to see that

Proposition 13. *Consider three treatments Tr*1*, Tr*2*, Tr*3 *and decision tables* T_{ij} *comparing Tri to Trj. For* $\gamma = CSR$ *and* $\gamma = CPR$*, if* $\gamma(T_{12}) = 1$ *and* $\gamma(T_{23}) = 1$*, then* $\gamma(T_{13}) = 1$*, that is if Tr*1 *is better than Tr*2 *and Tr*2 *is better than Tr*3*, then Tr*1 *is also better than Tr*3*.* □

When two treatments are compared using polytomous outcomes, like categories very good, good, bad, very bad, an extension of the decision procedure from 2×2 tables to $2 \times J$ tables is needed.

If in the $2 \times J$ decision table the observed response frequencies of *Tri* are $\mathbf{f}_i = (f_{i,1}, \ldots, f_{i,J})$, in a descending order of desirability, define δ_j, $j = 1, \ldots, J-1$, for an arbitrary nonnegative vector $\mathbf{v} = (v_1, \ldots, v_J)$ of dimension J, as

$$\delta_j(\mathbf{v}) = (v_1 + \cdots + v_j) - (v_{j+1} + \cdots + v_J),$$

so $\delta_j(\mathbf{f}_i)$ is the difference of the number of responses in the j best categories and of the number of responses in the $J-j$ worst response categories for treatment i. A generalization of the *CSR* (see (1)) is obtained by considering *Tr*1 better than *Tr*2 if

$$\operatorname{sgn}(\delta_j(\mathbf{f}_1) - \delta_j(\mathbf{f}_2)) = 1 \quad \text{for all } j = 1, \ldots, J-1,$$

and a generalization of the *CPR* (see (2)) is obtained by considering *Tr*1 better than *Tr*2 if

$$\operatorname{sgn}\left(\delta_j\left(\frac{\mathbf{f}_1}{f_{1+}}\right) - \delta_j\left(\frac{\mathbf{f}_2}{f_{2+}}\right)\right) = 1 \quad \text{for all } i = 1, \ldots, J-1,$$

where f_{i+} is the sum of all observations for *Tri*. Again, the generalizations of the *CSR* and of the *CPR* are based on the same function but the *CSR* uses the observed conditional frequencies and the *CPR* uses the observed conditional probabilities, just like in (1) and (2). Thus, the *CSR* remains sensitive to allocation and the *CPR* remains not sensitive to allocation in these generalized forms. Also, Simpson's paradox may occur with the generalized *CPR* and does not occur with the generalized *CSR*.

The generalizations suggested above are related to the concept of stochastic dominance (of responses to *Tr*1 versus those to *Tr*2) but this aspect is not going to be pursued in this paper. For *Tr*1 to be better than *Tr*2, these generalizations require the same relation to hold in all 2×2 decision tables that are obtained from the $2 \times J$ decision table by a monotone combination of responses into two groups. Whether or not this is an appropriate definition, depends on the actual decision problem, of course.

6. Discussion

This paper considered the problem of selecting the better one out of two treatments. Such decisions are mostly obtained by using functions of the data closely related to the odds ratio. The possibility or the actual occurrence of Simpson's paradox in such decisions has been the subject of much scholarly work, some explaining under what conditions the paradox may or may not occur, some making suggestions for choosing the better treatment if it did occur. This paper takes the position that a

real evaluation of various decision procedures is not possible unless at least minimal requirements for the behavior of decision procedures are postulated. Based on six such 'axioms', various decision procedures were investigated. It was argued that when allocation into the treatment categories is informative (observational studies or designed experiments with dropout or noncompliance), it is desirable to use decision functions which are sensitive to allocation. In addition, it turned out that all decision procedures which are not sensitive to allocation may exhibit Simpson's paradox, so a good way to avoid it is to use decision procedures which are sensitive to allocation, at least when the allocation is informative. The main results of the paper included that all decision procedures which are not sensitive to allocation always reach the same decision as the *CPR* and all decision procedures which never exhibit Simpson's paradox always reach the same conclusion as the *CSR* does. The paper also discussed statistical properties of the *CSR* and suggested possible generalizations to more complex decision problems.

The illustrative applications of the *CSR* to data that had been analyzed by the *CSR* showed that the conclusions are sometimes different. Our statistical education relies strongly on the *CPR* and we all may feel the conclusions reached by the *CSR* counter-intuitive in some cases. The frequent controversies around decisions reached by the *CPR* should suggest to consider using, in the case of informative allocation, the *CSR* that is a logically sounder decision procedure. However, the *CSR* and the *CPR* are not that different: they are based on the same function applied to conditional frequency or to conditional probability distributions of the responses in the two treatment categories.

Acknowledgments

The author is indebted to Antonio Forcina, Adrian Raftery and Joe Shafer for helpful discussions and to Rafael Perera for kindly allowing to use the data in Tables 7 and 8. Part of this research was done while the author was a visitor at the Center for Statistics and the Social Sciences and the Department of Statistics, University of Washington, where he is now an Affiliate Professor. The author is also a Recurrent Visiting Professor at the Central European University and the moral support received is acknowledged.

Appendix

Proof of Proposition 2. Invariance against changes in allocation implies that

$$\gamma\left(\begin{array}{|c|c|}\hline a & b\\ \hline c & d\\ \hline\end{array}\right)=\gamma\left(\begin{array}{|c|c|}\hline a & b\\ \hline \frac{b}{d}c & \frac{b}{d}d\\ \hline\end{array}\right)=\gamma\left(\begin{array}{|c|c|}\hline a & b\\ \hline \frac{b}{d}c & b\\ \hline\end{array}\right)$$

and by Property 6, $\gamma(T)=1$ if $CPR(T)=1$, and by Property 3, $\gamma(T)=-1$ if $CPR(T)=-1$. Finally, by Property 2, $\gamma(T)=0$ if $CPR(T)=0$. □

Proof of Proposition 4. One has to see that for indifferent decision functions γ

(a) if $a+d>b+c$ then $\gamma(T)=1$
(b) if $a+d<b+c$ then $\gamma(T)=-1$
(c) if $a+d=b+c$ then $\gamma(T)=0$

Part (a): If $a=b$ then $d>c$ and because of Property 5

$$\gamma\left(\begin{array}{|c|c|}\hline d & c\\ \hline b & a\\ \hline\end{array}\right)=1$$

and Properties 3 and 4 imply that

$$\gamma\left(\begin{array}{|c|c|}\hline a & b\\ \hline c & d\\ \hline\end{array}\right)=1.$$

If $a > b$ then if $d \geq c$, Property 5 implies the required result. If $c > d$, then $a-b > 0$ and $c-d > 0$, therefore $T = T_1 + T_2$ with

$$T_1 = \begin{array}{|c|c|}\hline a-b+\frac{1}{2}min(b,d) & \frac{1}{2}min(b,d) \\ \hline c-d+\frac{1}{2}min(b,d) & \frac{1}{2}min(b,d) \\ \hline \end{array} \quad T_2 = \begin{array}{|c|c|}\hline b-\frac{1}{2}min(b,d) & b-\frac{1}{2}min(b,d) \\ \hline d-\frac{1}{2}min(b,d) & d-\frac{1}{2}min(b,d) \\ \hline \end{array}$$

and $\gamma(T_1) = 1$ because of Property 6, for the condition in (a) may be written as $a - b > c - d$, and $\gamma(T_2) = 0$ because of Property 1. Then indifference implies that $\gamma(T) = 1$.

If, finally, $a < b$ then $d > c$ and for the table

$$\begin{array}{|c|c|}\hline d & c \\ \hline b & a \\ \hline \end{array}$$

one has the same situation as for the

$$\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline \end{array}$$

table in the last case and, therefore

$$\gamma\left(\begin{array}{|c|c|}\hline d & c \\ \hline b & a \\ \hline \end{array}\right) = 1$$

and Properties 3 and 4 imply that $\gamma(T) = 1$ also in this case.

Part (b): Apply part (a) to

$$\begin{array}{|c|c|}\hline c & d \\ \hline a & b \\ \hline \end{array}$$

to obtain that if $a + d < b + c$ then for any indifferent γ,

$$\gamma\left(\begin{array}{|c|c|}\hline c & d \\ \hline a & b \\ \hline \end{array}\right) = 1,$$

and then Property 3 implies the required result.

Part (c): In this case $a - b = c - d$. If $a \geq b$ then $T = T_1 + T_2$ with

$$T_1 = \begin{array}{|c|c|}\hline a-b+\frac{1}{2}min(b,d) & \frac{1}{2}min(b,d) \\ \hline c-d+\frac{1}{2}min(b,d) & \frac{1}{2}min(b,d) \\ \hline \end{array} \quad T_2 = \begin{array}{|c|c|}\hline b-\frac{1}{2}min(b,d) & b-\frac{1}{2}min(b,d) \\ \hline d-\frac{1}{2}min(b,d) & d-\frac{1}{2}min(b,d) \\ \hline \end{array}$$

and $\gamma(T_1) = 0$ because of Property 2 and $\gamma(T_2) = 0$ because of Property 1. Then indifference implies the required result.

If $a < b$, then $T = T_1 + T_2$ with

$$T_1 = \begin{array}{|c|c|}\hline \frac{1}{2}min(a,c) & b-a+\frac{1}{2}min(a,c) \\ \hline \frac{1}{2}min(a,c) & d-c+\frac{1}{2}min(a,c) \\ \hline \end{array} \quad T_2 = \begin{array}{|c|c|}\hline a-\frac{1}{2}min(a,c) & a-\frac{1}{2}min(a,c) \\ \hline c-\frac{1}{2}min(a,c) & c-\frac{1}{2}min(a,c) \\ \hline \end{array}$$

and $\gamma(T_1) = 0$ because of Property 2 and $\gamma(T_2) = 0$ because of Property 1. Then indifference implies the required result. □

Proof of Proposition 6. We have to prove that for consistent decision functions γ

(a) if $a + d > b + c$ then $\gamma(T) = 1$,
(b) if $a + d < b + c$ then $\gamma(T) = -1$,
(c) if $a + d = b + c$ then $\gamma(T) = 0$.

Part(a): Assume first that $a > b$ and $c < d$. In this case Property 5 implies that $\gamma(T) = 1$. If $a \leq b$ and $c < d$ then let x be any number such that $d - c > x > b - a$. Then T splits into the following tables

$$T_1 = \begin{array}{|c|c|}\hline a-\frac{1}{2}min(a,c) & a-\frac{1}{2}min(a,c) \\ \hline c-\frac{1}{2}min(a,c) & d-x-\frac{1}{2}min(a,c) \\ \hline \end{array} \quad T_2 = \begin{array}{|c|c|}\hline \frac{1}{2}min(a,c) & b-a+\frac{1}{2}min(a,c) \\ \hline \frac{1}{2}min(a,c) & x+\frac{1}{2}min(a,c) \\ \hline \end{array}$$

Then $\gamma(T_1) = 1$ because of Properties 3–5 and $\gamma(T_2) = 1$ because of Properties 3, 4 and 6. Then by consistency, $\gamma(T) = 1$. If $c \geq d$ then let x be any number such that $a - b > x > c - d$. Then T splits into the following tables

$$T_1 = \begin{array}{|c|c|}\hline a - x - \frac{1}{2}min(b,d) & b - \frac{1}{2}min(b,d) \\ \hline d - \frac{1}{2}min(b,d) & d - \frac{1}{2}min(b,d) \\ \hline \end{array} \quad T_2 = \begin{array}{|c|c|}\hline x + \frac{1}{2}min(b,d) & \frac{1}{2}min(b,d) \\ \hline c - d + \frac{1}{2}min(b,d) & \frac{1}{2}min(b,d) \\ \hline \end{array}$$

Here $\gamma(T_1) = 1$ because of Property 5 and $\gamma(T_2) = 1$ because of Property 6. Consistency implies that $\gamma(T) = 1$.

Part(b): Swapping the two columns of T and applying Property 4 shows that $\gamma(T) = -1 = CSR(T)$.

Part(c): If $a < c$, choose a positive number x such that $x < min(a, b, c - a)$. Then the split T into

$$T_1 = \begin{array}{|c|c|}\hline x & x \\ \hline c - a + x & c - a + x \\ \hline \end{array} \quad T_2 = \begin{array}{|c|c|}\hline a - x & b - x \\ \hline a - x & b - x \\ \hline \end{array}$$

With reference to Properties 1 and 2, consistency implies that $\gamma(T) = 0$. If $a > c$, apply Property 3 and the previous result to obtain that $\gamma(T) = 0$. If $a = c$, apply Property 1. □

Proof of Proposition 7. Proposition 1 is that (c) implies both (a) and (b). Proposition 4 is that (b) implies (c). Proposition 5 is that (b) implies (a). Proposition 6 is that (a) implies (c). □

Proof of Proposition 8. Let $e = a + b = c + d$ so that $b = e - a$ and $d = e - c$. Then

$$ad = a(e - c) > / = / < (e - a)c = bc$$

if and only if

$$a > / = / < c,$$

which happens if and only if

$$a + d = a + e - c > / = / < e - a + c = b + c. \quad \square$$

Proof of Proposition 9. By Proposition 4, $\gamma = CSR$ and the statement is true for the CSR. □

Proof of Proposition 10. When $p_a + p_d > p_b + p_c$, a wrong decision occurs if $A + D \leq B + C$ or, equivalently, if $A + D \leq n/2$.

Using the normal approximation to the binomial distribution, the probability of this event is

$$P\left(X \leq \frac{n/2 - n(p_a + p_d)}{\sqrt{n(p_a + p_d)(1 - p_a - p_d)}}\right) + o(1)$$
$$= P\left(X \leq \sqrt{n}\frac{0.5 - p_a - p_d}{\sqrt{(p_a + p_d)(1 - p_a - p_d)}}\right) + o(1), \tag{6}$$

where X is a standard normal random variable.

The true (population) value of the csr is $(p_a + p_d)/(p_b + p_c)$ and $p_a + p_d = csr/(1 + csr)$, thus (6) may be rewritten as

$$P\left(X \leq \frac{1}{2}\sqrt{n}\frac{1 - csr}{\sqrt{csr}}\right) + o(1). \quad \square$$

Proof of Proposition 11. A $1 - \alpha$ level confidence interval for the true value of CSR does not contain 0 or -1, when $CSR = 1$ was observed, if

$$t\left(\frac{a + d}{b + c}, u\right) = P\left(\frac{A + D}{B + C} > \frac{a + d}{b + c}\middle|\operatorname{sgn}\left(\log\frac{p_a + p_d}{p_b + p_c}\right) = u\right) < \alpha,$$

for $u = 0$ and $u = -1$. It is easily seen that

$$t\left(\frac{a + d}{b + c}, u\right) = P\left(A + D > a + d|\operatorname{sgn}\left(\log\frac{p_a + p_d}{p_b + p_c}\right) = u\right).$$

The hypothesis $CSR = 0$ is simple but the hypothesis $CSR = -1$ is composite and asymptotically

$$\sup_{csr<1} \; t\left(\frac{a+d}{b+c}, -1\right) = t\left(\frac{a+d}{b+c}, 0\right)$$

and it is sufficient to determine the value of t for $u = 0$. Under the hypothesis of $CSR = 0$, that is $p_a + p_d = 0.5$, the distribution of $A + D$ is $B(n, 0.5)$ and asymptotically

$$t\left(\frac{a+d}{b+c}, 0\right) = 1 - \Phi\left(\frac{2(a+b)-n}{\sqrt{n}}\right). \quad \square$$

Proof of Proposition 12. X and Y are independent and asymptotically normal. Therefore, asymptotically, $X = 2A - N_+$ is normal with $E_X = N_+(p_a - p_c)$ and variance $V_X = 4N_+ \frac{p_a}{p_a+p_c} \frac{p_c}{p_a+p_c}$ and $Y = 2B - N_-$ is normal with expectation $E_Y = N_-(p_b - p_d)$ and variance $V_Y = 4N_- \frac{p_b}{p_b+p_d} \frac{p_d}{p_b+p_d}$. Using independence, $X - Y$ is asymptotically normal, with expectation $E_X - E_Y$ and variance $\sqrt{V_X + V_Y}$. The probability of wrong decision is

$$\begin{aligned} P(B + C > A + D) &= P(B - D > A - C) = P(Y > X) \\ &= P(X - Y < 0) \end{aligned}$$

and $P(X - Y < 0)$ is the same, asymptotically, as the probability that a standard normal variable is less than

$$\frac{E_Y - E_X}{\sqrt{V_X + V_Y}}. \quad \square$$

References

[1] A. Agresti, Categorical Data Analysis, 2nd edition, Wiley, New York, 2002.
[2] J. Aldrich, Correlations genuine and spurious in Pearson and Yule, Statistical Science 10 (1995) 364–376.
[3] P.J. Bickel, E.A. Hammel, J.W. O'Connell, Sex bias in graduate admissions: Data from Berkeley, Science 187 (1975) 398–404.
[4] C.R. Blyth, On Simpson's paradox and the sure-thing principle, Journal of the American Statistician Association 67 (1972) 364–366.
[5] H.C. Craemer, Reconsidering the odds ratio as a measure of 2 × 2 association in a population, Statistics in Medicine 23 (2004) 257–270.
[6] S.P. Curley, G.J. Browne, Normative and descriptive analyses of Simpson's paradox in decision making, Organizational Behavior and Human Decision Processes 84 (2001) 308–333.
[7] H.T.O. Davies, I.K. Crombie, M. Tavakoli, When can odds ratios mislead?, British Medical Journal 316 (1998) 989–991.
[8] B. Dawson, R.G. Trapp, Basic and Clinical Biostatistics, McGraw-Hill Professional, 2004.
[9] P. Diggle, P. Heagerty, K.Y. Liang, S.L. Zeger, Analysis of Longitudinal Data, Oxford University Press, 2002.
[10] D. Freedman, R. Pisani, R. Purves, Statistics, Norton & Company, New York, 1998.
[11] A.M. Gallup, F. Newport, The Gallup Poll: Public Opinion 2004, Rowman & Littlefield, Lanham, MD, 2006.
[12] L.B. Goldstein, D.L. Simel, Is this patient having a stroke?, JAMA 293 (2005) 2391–2402.
[13] A. Harnden, N. Ninis, M. Thompson, R. Perera, M. Levin, D. Mant, R. Mayon-White, Parenteral penicillin for children with meningococcal disease before hospital admission: Case-control study, British Medical Journal 332 (2006) 1295–1297.
[14] D.V. Lindley, M.R. Novick, The role of exchangeability in inference, The Annals of Statistics 9 (1981) 45–58.
[15] C. Meek, C. Glymour, Conditioning and intervening, The British Journal for the Philosophy of Science 45 (1994) 1001–1021.
[16] J. Pearl, Causality, Cambridge University Press, 2000.
[17] R. Perera, Commentary: Statistics and death from meningococcal disease in children, British Medical Journal 332 (2006) 1297–1298.
[18] R. Perera, Personal communication, 2008.
[19] P. Rosenbaum, Observational Studies, Springer, New York, 1995.
[20] T. Rudas, Odds Ratios in the Analysis of Contingency Tables, Sage, Newbury Park, 1998.
[21] T. Rudas, W. Bergsma, Letter to the editor, Statistics in Medicine 23 (2004) 3545–3547.
[22] K.B. Schechtman, Patient compliance, in: S.C. Chow (Ed.), Encyclopedia of Biopharmaceutical Statistics, Marcel Dekker, New York, 2000, pp. 712–717.
[23] C.H. Wagner, Simpson's paradox in real life, The American Statistician 36 (1982) 46–48.
[24] H. Wainer, L.M. Brown, Two statistical paradoxes in the interpretation of group differences illustrated with medical school admission and licensing data, The American Statistician 58 (2004) 117–123.
[25] J. York, British GPs practice excellent medicine, British Medical Journal (2006) Rapid Responses http://www.bmj.com/cgi/eletters/332/7553/1295.