Effects of malingering in self-report measures: A scenario analysis approach.

Massimiliano Pastore[1], Luigi Lombardi[2], and Francesca Mereu[3]

[1] Dipartimento di Psicologia dello Sviluppo e della Socializzazione
Università di Padova
Via Venezia, 8 I-35131, Padova Italy
(e-mail: massimiliano.pastore@unipd.it)

[2] Dipartimento di Scienze della Cognizione e della Formazione
Università di Trento
Via Matteo del Ben, 5 I-38068 Rovereto (TN), Italy
(e-mail: luigi.lombardi@unitn.it)

[3] Dipartimento di Psicologia
Università di Cagliari
Via Is Mirrionis, 1 I-09123 Cagliari, Italy
(e-mail: franceschina.mereu@tiscali.it)

Abstract. In many psychological questionnaires (i.e., personnel selection surveys and diagnostic tests) the collected samples often include fraudulent records. This confronts the researcher with the crucial problem of biases yielded by the usage of standard statistical models. In this paper we generalize a recent combinatorial perturbation procedure, called SGR (Sample Generation by Replacements; [Lombardi *et al.*, 2004]), to the analysis of structured malingering scenarios for dichotomous data. Combinatorial aspects of the approach are discussed and an application to a simple data set on the drug addiction domain is presented. Finally, the close relationships with Monte Carlo simulation studies are explored.

Keywords: Sample Generation by Replacements, fraudulent data, scenario analysis.

1 Introduction

In some circumstances social desirability biases may drastically limit the validity of self-report measures. In general, faking and demand characteristics represent serious threats to the psychometric validity of both social competence tests and self-report measures of socially undesirable behaviors. In particular, possible fake data confront the researcher with the problem of evaluating the effect of malingering responses to final statistical results. It is worth mentioning that even in the presence of simple undirected (uniform) malinger data the answer to this problem is not necessarily obvious, as even the random perturbation of data constitutes a biased information which decreases the efficiency of parameter estimates and weakens the accuracy of statistical results.

A case of particular empirical interest is the situation in which a researcher wants to evaluate the impact of structured malinger data in testing a given target model. For example, within a simple dichotomous scenario, we might be interested in studying how the result of an exact Fisher's test applied to a 2×2 contingency data-table varies as a function of different malingering scenarios. A fake-scenario analysis is a methodology for analyzing observed data by considering hypothetical malingering processes and may be considered as an additional analysis a researcher can run to broaden the sources of information she/he is interested in. Therefore, fake-scenario analysis is supposed to allow improved decision-making by allowing more complete consideration of outcomes and their eventual implications.

In this paper we propose a simple combinatorial procedure for treating structured fake data in a dichotomous setting. The new procedure extends a recent data generating procedure called SGR (Sample Generation by Replacements, [Lombardi *et al.*, 2004]) developed to provide a perturbation model and a sampling procedure to generate structured collections of perturbations.

Section 2 of this paper will first outline the basic principles of the new replacement approach. Section 3 will then present an illustrative application of the SGR approach to the analysis of a small data set on the use of ecstasy in an adolescent population. Finally, Section 4 will discuss the relation of the SGR method with Monte Carlo simulation studies. At the end of the section some possible extensions of the SGR approach are also outlined.

2 The method of replacements

Our procedure implements a combinatorial method that can be applied to discrete data with a restricted number of values (e.g., dichotomous or Likert-type scale) and consists of two different components:

1. a perturbation model,
2. a sampling procedure to generate perturbed samples from a given real data set.

2.1 Basic elements

In many social and psychological surveys the resulted dataset often includes incomplete records (missing data) and/or fake records (fake data). In particular, as regards the dichotomous fake-data problem, we think of the dataset as being represented by a collection of pairs $\mathbf{d} = \{(g_i, y_i) : i = 1, \ldots, I\}$ where g_i is a group variable with $g_i = k$ denoting that individual i belongs to group k $(k = 1, \ldots, K)$; y_i is a Boolean response variable where $y_i = 1$ means that individual i gives an affirmative answer to a possibly sensitive target question Q. We may assume that a certain portion of the response vector $\mathbf{y}$ is actually fake-data. The fake-portion $\mathbf{y}^f$ of $\mathbf{y}$ together with the uncorrupted portion $\mathbf{y}^u$ of $\mathbf{y}$, constitutes the full data set, that is to say $\mathbf{y} = \mathbf{y}^f \cup \mathbf{y}^u$. The exact fake-portion $\mathbf{y}^f$ of $\mathbf{y}$ is assumed to be an unknown parameter and only the number $0 < N \leq I$ of fake data points in $\mathbf{y}$ is supposed to be known. The general idea is the following: in order to analyze the data and provide an uncertainty analysis of some statistic of interest we replace some portions $\mathbf{y}_1, \ldots, \mathbf{y}_H$ of $\mathbf{y}$, each of which contains exactly N elements, with new components $\mathbf{y}_1^r, \ldots, \mathbf{y}_H^r$ such that $\mathbf{y}_h^r = \mathbf{1}_h - \mathbf{y}_h$ for all $h = 1, \ldots, H$. In the SGR approach these new components are generated from an appropriate population, and, therefore, the complete datasets $\mathbf{y}_1^*, \ldots, \mathbf{y}_H^*$ (with $\mathbf{y}_h^* = \mathbf{y}_h^r \cup \mathbf{y}_h^u$; $h = 1, \ldots, H$), are analyzed. We call the data array $\mathbf{y}_h^*$ and $\mathbf{y}_h^r$ the h^{th}-perturbed array of $\mathbf{y}$ and the h^{th}-replaced portion of $\mathbf{y}$, respectively. Finally, it is worth mentioning that within a dichotomous scenario each perturbed array $\mathbf{y}_h^*$ represents a node of the I dimensional Boolean hypercube $\{0,1\}^I$ having $\mathbf{y}$ as its origin. In the Boolean hypercube each perturbed array $\mathbf{y}_h^*$ has the same Hamming distance

$$d(\mathbf{y}_h^*, \mathbf{y}) = \sum_i |y_{h,i}^* - y_i| = N$$

from the original data array $\mathbf{y}$.

2.2 Malingering scenarios

Several malingering scenarios can be proposed according to both the typology of the investigated groups and the sensitivity of the self-report measure. The most elementary malingering scenario can be described by means of the *principle of indifference*. This principle reflects the fact that in the absence of further knowledge all entries in $\mathbf{y}$ are assumed to be equally likely in the process of faking (across the K different groups). In other words, it assumes a random malingering model compatible with uniform randomly fake-data. In contrast, the availability of external knowledge about process faking may suggest the modeling of more complex scenarios. For example, in personnel selection some subjects are likely to fake a personality questionnaire to match the ideal candidate's profile (positive impression management or fake-good process, [Ballenger *et al.*, 2001]). Similarly, in the administration of diagnostic tests individuals often attempt to malinger posttraumatic stress disorder (PTSD) in order to secure financial gain and/or treatment, or to avoid being charged with a crime (fake-bad process; [Elhai *et al.*, 2001]). In these latter scenarios it could be reasonable to consider a conditional replacement model, where the conditioning is a function of response polarity (e.g., negative for fake-good and positive for fake-bad).

In general we may define a $K \times 2$ probability matrix $\mathbf{P} = [p_{kj}]$, where p_{k1} (resp. p_{k2}) denotes the probability that each of the N fake data points in $\mathbf{y}$ is associated to an affirmative (resp. negative) response of an individual belonging to group k ($k = 1, \ldots, K$). We impose that $\sum_k \sum_j p_{kj} = 1$. So, for example, to simulate a uniform random malingering model we set $p_{kj} = \frac{1}{2K}$ $(\forall k, \forall j)$. The system

$$\mathcal{M} = \langle \mathbf{d}, \mathbf{P}, N, (\mathbf{y}_h^*, h = 1, \ldots, H) \rangle$$

defines the formal representation of the malingering scenario. $\mathcal{M}$ is said to be consistent if the replaced portion $\mathbf{y}_h^r$ of $\mathbf{y}$ is stochastically consistent with both $\mathbf{d}$ and $\mathbf{P}$ for all $h = 1, \ldots, H$, otherwise $\mathcal{M}$ is inconsistent. Notice that $\mathbf{P}$ induces a constrained random path $\mathbf{y} = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N = \mathbf{y}_h^*$ of length N on the Boolean hypercube $\{0, 1\}^I$. The random path starts from node $\mathbf{y}$ and continues through the nodes of the Boolean hypercube by steps satisfying the following constraint

$$d(\mathbf{x}_{n+1}, \mathbf{y}) = d(\mathbf{x}_n, \mathbf{y}) + 1, \quad n = 1, \ldots, N - 1 \tag{1}$$

where node $\mathbf{x}_n$ represents a transition node in the path which is governed by the probability matrix $\mathbf{P}$. A malingering scenario $\mathcal{M}$ is inconsistent with respect to a final node $\mathbf{x}_N$ whenever it does not exist a random path linking $\mathbf{y}$ and $\mathbf{x}_N$ according to the probability matrix $\mathbf{P}$ and the full data set $\mathbf{d}$. For example, suppose that $K = 2$, $N = 10$, and $p_{11} = 1.0$ (that is to say, each of the 10 fake data points in $\mathbf{y}$ is associated with probability 1.0 to an affirmative response of an individual belonging to the first of the two groups, $k = 1$). Moreover, suppose also that in the original data $\mathbf{d}$, if $g_i = 1$ then $y_i = 0$ which means that all subjects in the first group gave a negative response. It is straightforward to verify that the above malingering scenario is stochastically inconsistent.

Finally, let T be a statistical test and let $t = T(\mathbf{d})$ be its value when the statistic is computed using the original data set $\mathbf{d}$. The main goal of a replacement analysis is the evaluation of some properties of T under the perturbed sample space

$$\{\mathbf{d}_h^* = (\mathbf{g}, \mathbf{y}_h^*), h = 1, \ldots, H\}$$

generated according to a consistent malingering scenario $\mathcal{M}$. Alternatively, we may also consider the evaluation of T across the nodes of the random walk $\mathbf{y} = \mathbf{x}_1, \ldots, \mathbf{x}_N = \mathbf{y}_h^*$ $(\forall h = 1, \ldots, H)$.

3 Empirical data example

In this exploratory study we tested the new procedure on a small data-set from a study in the substance abuse domain. The current section is divided into three subsections: the first introduces the empirical data set and the statistical test; the second discusses the use of malingering scenarios to generate a family of perturbed datasets; and the third evaluates the statistical test with respect to the malingering scenarios.

3.1 Original dataset and test statistic

We illustrate the entire procedure using data on the interrelation between gender and ecstasy use. Participants were 22 undergraduate students from an high school in the Sardinia district (Italy). Ages ranged from 18 to 26, with a mean of 22.09 and a standard deviation of 2.15. Data was collected using a single item selected from a survey regarding the use of alcohol and other drugs in adolescents. In particular, the item consisted in a self-report measure of annual ecstasy use. The item was represented by the following question: 'Have you used ecstasy in the last 12 months?' For purposes of this analysis, annual ecstasy use was considered a dichotomous outcome (at least once = 1/none = 0). A contingency table summarizing the data is reported below (Table 1). The resulting (22×2) data matrix $\mathbf{d}$ was subjected to an exact Fisher's test to evaluate the association between gender and ecstasy use. The test statistic was not significant ($p = 0.476$).

	no	yes
m	9	2
f	11	0

Table 1. Contingency table for data **d**.

3.2 Modeling malingering scenarios

In the following analyses we supposed that there were no more than a total of nine fake responses in the observed sample (approximately 50% of the sample) and according to this hypothesis we defined three different malingering scenarios:

1. ($\mathcal{M}_1$) An undirected uniform malingering model: $p_{kj} = \frac{1}{4}$ ($\forall k = 1, 2, \forall j = 1, 2$).
2. ($\mathcal{M}_2$) An oriented and gender symmetric malingering model assigning positive probabilities of faking to negative answers only: $p_{k2} = \frac{1}{2}$ ($\forall k = 1, 2$).
3. ($\mathcal{M}_3$) An oriented, but gender asymmetric, malingering model such that $p_{12} = .60$ (males) and $p_{22} = .40$ (females).

In order to evaluate the uncertainty of the statistical test we resort to generating a family of $H = 3000$ different perturbed matrices with exactly N replacements in accordance to the procedure described in Section 2.2.

The three scenarios are based on three different probability matrices each of which represents a different malingering process. According to $\mathcal{M}_1$ a simple uniform random malingering model is implemented. It reflects the absence of further knowledge about the process of faking governing the transaction between the original data set and the final perturbed array. Unlike, $\mathcal{M}_1$, the oriented malingering scenario $\mathcal{M}_2$ subsumes a different psychological process. In particular, $\mathcal{M}_2$ models the generation of fake good responses. Finally, also $\mathcal{M}_3$ models a fake good-type process, but unlike $\mathcal{M}_2$ it assumes that the probability of faking is asymmetric in the two groups with more fraudulent responses in the male group as compared to the female one.

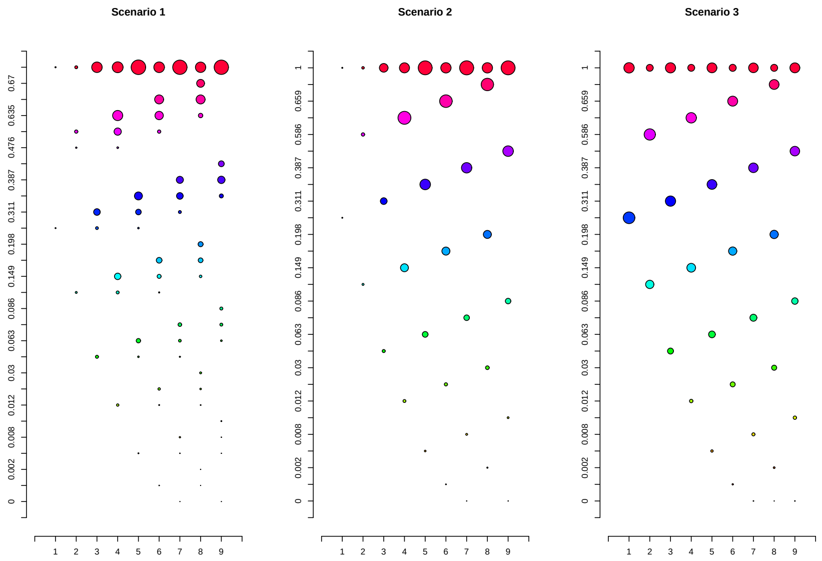

Fig. 1. Exact Fisher's test probabilities as a function of replacement.

3.3 Results

Figure 1 shows the Fisher's Exact test probabilities as a function of N (number of assumed fraudulent points in the original sample $\mathbf{y}$) for the three malingering scenarios. In its basic form, a large value of the test probability is evidence of a null hypothesis of independence between sex and ecstasy assumption. A Larger circle in the bubble plot indicates a larger size of the equivalence class of the perturbed arrays associated to the same contingency table. Figure 2 shows the proportion of significant Exact Fisher's tests as a function of N for the three malingering scenarios. The pattern associated to $\mathcal{M}_1$ showed that the uniform random malingering scenario was in general less sensitive to replacements than both the oriented malingering scenario ($\mathcal{M}_2$) and the asymmetric malingering scenario ($\mathcal{M}_3$), the latter being clearly the most sensitive to number of replacements.

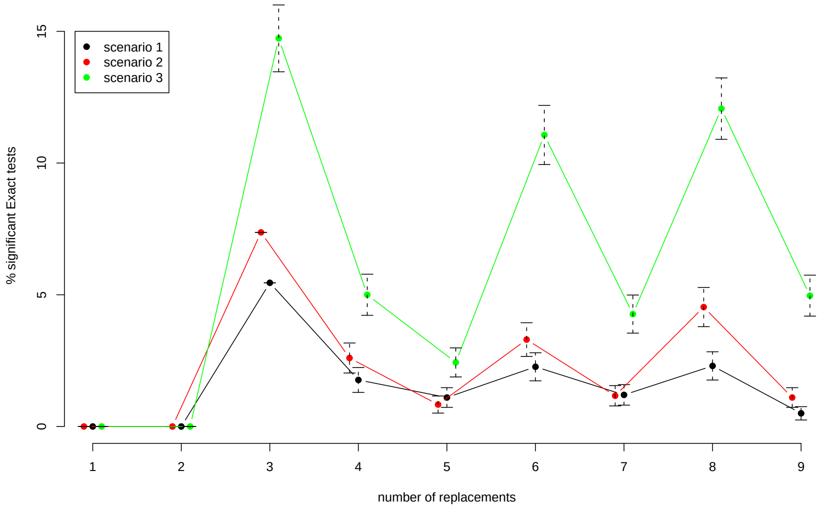

Fig. 2. Proportion of significant Exact Fisher's tests as a function of replacement. Vertical segments represent 95% confidence intervals

4 Concluding remarks

The reader may have already noticed some similarities between the approach proposed here and standard Monte Carlo experiments. For example, the idea of generating new data sets. However, the two approaches are substantially different. Usually a Monte Carlo experiment uses a hypothesized model to generate new data under various conditions (e.g. [Robert and Casella, 2004]). Therefore the simulated data are used to evaluate some characteristics of the model. This, of course, implies that the distribution of the random component in the assumed model must be known, and it must be possible to generate pseudorandom samples from that distribution under the desired conditions planned by the researcher.

Instead of using the hypothesized model structure to generate simulated data sets, our approach uses the original data sample in order to generate a new family of data sets. In particular, these new data sets are obtained by adding structured perturbations in the original data set. The availability of external knowledge about process faking may suggest the modeling of highly structured malingering scenarios. In these more complex scenarios it could be reasonable to consider conditional replacement models, where the conditioning is a function of some response polarity (e.g., fake-good or for fake-bad). In the latter case, each new sample represents an alternative malingering scenario which is directly derived from both the original sample and the assumed malingering model. Next, the result of a target criterion can be compared with the ones obtained from the perturbed samples.

Several possible extensions of our approach may be considered. In the present paper, under the assumption of different malingering scenarios, a very simple SGR model has been proposed as a model for Boolean data. However, the current approach can be straightforwardly extended to categorical data as well as to continuous data. In particular, a SGR model for continuous data would imply a different kind of metric, for example either the city-block distance (L_1) or the standard Euclidean distance (L_2). These new extensions would enlarge the general replacement schema by adding more complex constraints with which we could provide more structured perturbed scenarios.

References

[Ballenger *et al.*, 2001]J.F. Ballenger, A. Caldwell-Andrews, and R.A. Baer. Effects of positive impression management on the neo personality inventory-revised in a clinical population. *Psychological assessment*, pages 254–260, 2001.

[Elhai *et al.*, 2001]J.D. Elhai, S.N. Gold, A.H. Sellers, and W.I. Dorfman. The detection of malingered posttraumatic stress disorder with mmpi-2 fake bad indices. *Assesment*, pages 221–236, 2001.

[Lombardi *et al.*, 2004]L. Lombardi, M. Pastore, and M. Nucci. Evaluating uncertainty of model acceptability in empirical applications: A replacement approach. In K. van Montfort, J. Oud, and Dordrecht (NE) Satorra, A. Kluwer, editors, *Recent Development on Structural Equation Models*, pages 69–82, 2004.

[Robert and Casella, 2004]C.P. Robert and G. Casella. *Monte Carlo Statistical Methods (second edition).* Springer-Verlag, New York, 2004.