Analyzing faking-good response data: Combination of a Replacement and a Binomial (CRB) distribution approach

Analisi di dati soggetti a processi tipo faking-good mediante approcci di mistura

Luigi Lombardi, Antonio Calcagnì

Abstract In this short paper, we describe a novel approach to model and analyse ordinal data in the presence of faking behavior, namely the tendency of survey's participants to falsify their responses in order to achieve a particular purpose. The proposal relies on the use of two statistical approaches commonly used to analyse faking and preference data: the Sampling Generation by Replacement (SGR) and Combination of Uniform and Binomial distributions (CUBE). By combining both SGR and CUBE, we propose CRB (Combination of Replacement and Binomial distributions), where the response ordinal measure is modeled as a convex combination of the shifted-Binomial distribution and the Replacement distribution. Thus, the first component aims to represent the response measure unaffected by faking behavior whereas the second element of the linear model represents the result of a faking strategy. As for the CUBE models, CRB parameters are estimated via Maximum likelihood by means of the EM algorithm. Finally, an application to ordinal data is proposed to show how the CRB model can be used to analyse self-reported data potentially affected by faking behavior.

Abstract *Abstract in Italian*. Questo lavoro presenta alcuni risultati preliminari per la definizione di un modello di analisi dei dati in presenza di *faking* o *malingering*. Dato un campione di misure ordinali - come quelle ottenute nel contesto delle *surveys* o questionari *self-report* - si definisce *faking* quel processo per il quale parte (o la totalità) dei rispondenti modifica la propria risposta in modo deliberato con l'obiettivo di ottenerne un vantaggio. Per l'analisi di tale tipologia di dati, si propone un nuovo approccio, denominato CRB (*Combination of a Replacement and Binomial distributions*), derivante dall'integrazione di due approcci statistici indipendenti, ossia SGR (*Sample Generation by Replacement*) and CUBE (Combination of Uniform and Binomial distributions). CRB modella la risposta ordinale mediante una combinazione lineare convessa di due componenti, una distribuzione Binomiale traslata per la componente ordinale della risposta ed una distribuzione di *Replace-*

Luigi Lombardi, University of Trento, e-mail: luigi.lombardi@unitn.it · Antonio Calcagnì, University of Padova, e-mail: antonio.calcagni@unipd.it

ment per la componente faking. Come per i modelli CUBE, la stima dei parametri del modello è effettuata per massima verosimiglianza. Infine, un breve caso studio è utilizzato per mostrare il funzionamento del modello CRB per l'analisi e valutazione di dati soggetti potenzialmente a faking.

Key words: Fake-good data, CRB approach, Ordinal data, Generalized Mixture distribution, CUBE approach

1 Introduction

Faking behavior in self-report measures, a type of response set, is a tendency to falsify item responses in order to meet strategic goals (e.g., avoiding being charged with a crime, see [1]). This behavior may be observed in some sensitive contexts such as, for example, risky sexual behaviors and drug addictions (e.g., [2, 3]) where individuals may react by hiding their real opinions or honest responses.

SGR (Sample Generation by Replacement) is a probabilistic resampling procedure [4, 5] that can be used to study and evaluate uncertainty in inferences based on possible fake responses as well as to study the implications of fake data for empirical results. In general, a SGR analysis takes an interpretation perspective which incorporates in a global model all the available information (empirical or hypothetical) about the process of faking and the underlying true model representation. In particular, SGR has a statistical descriptive nature which tries to capture the phenomenological effect of faking according to an informational, data-oriented perspective based on a data replacement (information replacement) paradigm. SGR has been normally used as a methodology to study, using Monte Carlo simulation designs, the impact of fake data on parameter estimations and model fit evaluations.

Unlike SGR, CUBE models (Combination of a Uniform and a Binomial distribution) is a class of statistical models that is grounded on the data generating process of the discrete response choice [6, 7] which allows the modeling of rating data expressing preferences and evaluations. The CUBE approach considers the final discrete response as the combination of two components: *feeling* and *uncertainty*. The shifted Binomial component regards the expression of feeling and takes into account for the fraction of responses associated with a precise opinion on the rating. By contrast, the uniform component concerns to uncertainty in rating and mimics aspects not directly associated to the content of the item. This representation allows finer model specifications which include refuge options, response styles and possible overdispersion. Moreover, unlike SGR models, CUBE models are also supported by Maximum likelihood (ML) estimation procedures based on EM algorithms.

In this contribution, we introduce a novel model representation, called CRB (Combination of a Replacement and a Binomial distribution), which combines the two approaches by integrating into a common framework some nice features of the two perspectives to provide an effective data analysis strategy for faking behavior

in self-report measures. In particular, the new representation substitutes the second component (*uncertainty*) of CUBE with a replacement distribution mimicking the faking process in self-report measures.

2 Model

In this section, we will first highlight some connections between the two approaches according to a general probabilistic representation. Next, we will formally describe our proposal.

2.1 CUBE and SGR: similarities and differences

Let Y be a discrete (observed) random variable with a finite support $\{1,2,\ldots,m\}$ (e.g., a rating-type variable). In its general terms, the CUBE representation can be defined as follows:

$$P(Y=y) = \pi P(Z=y) + (1-\pi)P(V=y) \tag{1}$$

here Z and V are two hidden variables with the same support of Y. Note that, Eq. 1 constitutes a mixture representation for the observed variable Y. In particular, let $C \in \{0,1\}$ be a Bernoulli variable with parameter $\pi \in]0,1]$, then Eq. 1 can be rewritten as follows:

$$P(Y=y|C=1) = P(Z=y) \quad \text{and} \quad P(Y=y|C=0) = P(V=y) \tag{2}$$

Therefore, the mixture distribution reduces to a two step process where we first draw a coin C (with probability π of observing the target event), and next we sample the value of Y according to the previous dichotomous result observed on C. Note that, Eq. 1 implies a hidden joint distribution $P(C,Z,V)$ which in its general form *does not require* to satisfy the independence condition for the pair (Z,V). Therefore, $P(Y)$ represents the probability distribution of the transformed random variable

$$Y = CZ + (1-C)V. \tag{3}$$

Unlike CUBE, the SGR representation is defined as follows:

$$P(W=w) = \sum_x P(W=w|X=x)P(X=x) \tag{4}$$

where W and X are hidden variables with the same support of Y. In this context, the conditional distribution $P(W|X)$ is called the *replacement distribution*, whereas $P(X)$ is named the *prior distribution* for the true variable. In the SGR perspective, the random variable W is called the fake response, whereas X represents the true

hidden (and unknown) response. Note that $P(X)$ identifies the prior distribution of the true value X before any direct inspection of the observed data.

Now, we are in the position to link CUBE and SGR by setting the new transformed variable:

$$Y = CZ + (1-C)W \tag{5}$$

where, in this context, W denotes the fake random variable defined in Eq. 4. Note that a similar representation has been adopted in a recent SGR contribution called *mixture* SGR (see Eq. 11 in [8]). Recollecting all the terms we finally have:

$$P(Y=y) = \pi P(Z=y) + (1-\pi)P(W=y) \tag{6}$$

which, at a general level, directly connects SGR with the CUBE representation.

2.2 The CRB model

We now define the model instances for the CRB distribution. The first component is

$$b_y(\xi) = P(Z=y) = \binom{m-1}{y-1}(1-\xi)^{y-1}\xi^{m-y}, \qquad y = 1,2,\ldots,m, \tag{7}$$

the so-called shifted Binomial distribution with parameter $\xi \in [0,1]$. The second component of the CRB distribution is

$$p_y = P(W=y) = \frac{1}{m}\sum_{x=1}^{m} p_{y|x}, \qquad y = 1,2,\ldots,m, \tag{8}$$

with replacement distribution

$$p_{y|x} = \begin{cases} 1, & x = y = m \\ \frac{1}{m-x}, & 1 \le x < y \le m \\ 0, & 1 \le y \le \min\{x, m-1\} \end{cases} \tag{9}$$

The latter component corresponds to a fake good distribution (e.g., see Table 1 and Figure 1). Here we assume an uninformative prior for $P(X)$, that is to say, $P(X = x) = \frac{1}{m}$ for all $x = 1,2,\ldots,m$. Therefore, the CRB distribution takes the following form:

$$g_y \equiv P(Y=y) = \pi b_y(\xi) + (1-\pi)p_y. \tag{10}$$

Clearly, the distribution in Eq. 10 is a discrete one with a well defined two-dimensional parameter space $\Omega = \{(\pi,\xi) : 0 < \pi \le 1, 0 \le \xi \le 1\}$. Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be a random sample of n i.i.d. rating responses on the finite support $\{1,2,\ldots,m\}$. Then, the CRB log-likelihood function is expressed by:

$$L(\theta) = \sum_{i=1}^{n} \log\{\pi b_{y_i}(\xi) + (1-\pi) p_{y_i}\}. \tag{11}$$

Note that Maximum likelihood (ML) estimates can be obtained by EM algorithm according to the procedure outlined, for instance, by [9] for standard CUBE models.

3 Application

In this application we analyzed an hypothetical set of ordinal data about illicit drug use (cannabis consumption) among young people (see Table 1). In particular, the response variable uses a four-point ordinal scale ranging from 1 = *never* to 4 = *often*, with intermediate levels being 2 = *once* and 3 = *sometimes*. The observed frequencies for the four values were 27, 8, 5, and 3, respectively. To apply the faking-good model, as defined in Eq. 10, we reversed the rating scale. We finally ran the CRB model to the data sample with $n = 43$ independent rating observations. The estimated parameters were as follows:

$$\widehat{\pi} = 0.12090; \qquad \widehat{\xi} = 0.83336;$$

with a log-likelihood function $L(\widehat{\theta}) = -44.76983$ and a very low dissimilarity index $\delta = 0.01139$, which was calculated as the normalized difference between observed relative frequencies, $\widehat{p}_h$, and fitted probabilities, $\widehat{g}_h$:

$$\delta = \sum_{h=1}^{m} |\widehat{p}_h - \widehat{g}_h|.$$

The results suggested the prominent role of the replacement distribution component p_{y_i}, which modeled the faking-good style of response, in modulating the response process as reflected by the mixture parameter $(1-\pi)$.

Table 1 Replacement distribution $p_{y|x}$ and its marginal representation p_y. This corresponds to a fake good scenario such that $Y > X$ with a discrete uniform kernel.

	$X = x$					
$y\|x$	1	2	3	4	5	p_y
1	0.0	0.0	0.0	0.0	0.0	0.0
2	1/4	0.0	0.0	0.0	0.0	0.05
3	1/4	1/3	0.0	0.0	0.0	0.11667
4	1/4	1/3	1/2	0.0	0.0	0.21667
5	1/4	1/3	1/2	1.0	1.0	0.61667

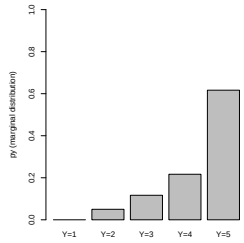

Fig. 1 Marginal distribution p_y.

4 Results and conclusion

In this short contribution we described a new model to analyse self-reported measures which could potentially be affected by faking behavior. Indeed, as many research have previously shown (e.g., see [2]), the latter plays an important role in social research and surveys based on self-reported questionnaires. We proposed a *combination of Replacement and Binomial* (CRB) distributions approach which takes the advantages of two statistical methodologies, namely SGR [4] and CUBE [7] that were independently proposed to analyse faking and preference data respectively. The new CRB approach uses a statistical rationale based on a mixture distributions approach where ordinal measures are represented as convex linear combination of a shifted-Binomial distribution, modeling the component unaffected by faking, and a Replacement distribution, which models instead the faking component. Model parameters were estimated via Maximum likelihood as offered by the EM algorithm in the general CUBE framework [9]. We showed the novel CRB approach on a simple application involving ordinal data from a hypothetical case study. Results suggested how faking response styles should deserve more attention, especially in those research involving analyses based on self-reported surveys and questionnaires.

References

1. M. Ziegler, C. MacCann, R. Roberts, *New perspectives on faking in personality assessment* (Oxford University Press, 2011)
2. M.J. Zickar, C. Robie, Journal of Applied Psychology **84**(4), 551 (1999)
3. L.A. McFarland, A.M. Ryan, Journal of Applied Psychology **85**(5), 812 (2000)
4. L. Lombardi, M. Pastore, Multivariate behavioral research **47**(4), 519 (2012)
5. M. Pastore, L. Lombardi, Quality & Quantity **48**(3), 1191 (2014)
6. D. Piccolo, Quaderni di Statistica **5**(1), 85 (2003)
7. A. D'Elia, D. Piccolo, Computational Statistics & Data Analysis **49**(3), 917 (2005)
8. M. Bressan, Y. Rosseel, L. Lombardi, Frontiers in psychology **9**, 1876 (2018)
9. M. Iannario, Metron **68**(1), 87 (2010)