eRum 2016 – european R users meeting

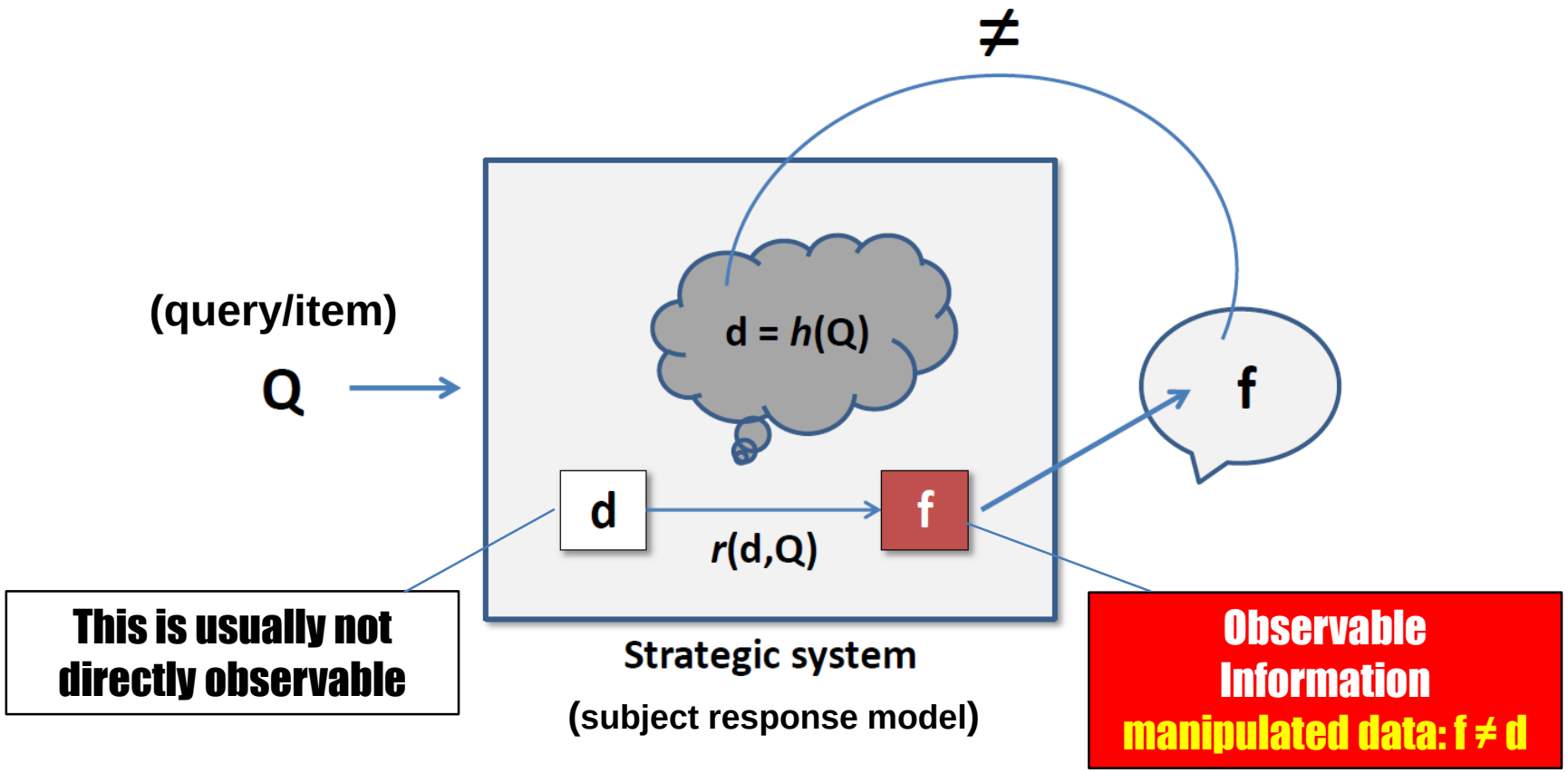
Analyzing the statistical effects of manipulated data

Luigi Lombardi and Marco Bressan
Dept. of Psychology and Cognitive Science, University of Trento

UNIVERSITÀ DEGLI STUDI
DI TRENTO

1 The Sample Generation by Replacement (SGR) approach

The SGR logic in self-report measures



Examples of manipulated data

- Faking good and faking bad responses
- Voluntarily random responses
- Cheating responses bias

Voluntarily manipulations

- Desirability responses bias
- Extreme responses bias
- Mid-value responses bias

Involuntarily manipulations

The sgr package (2014; *The R Journal*, 6(1), 164-177)

sgr: A Package for Simulating Conditional Fake Ordinal Data

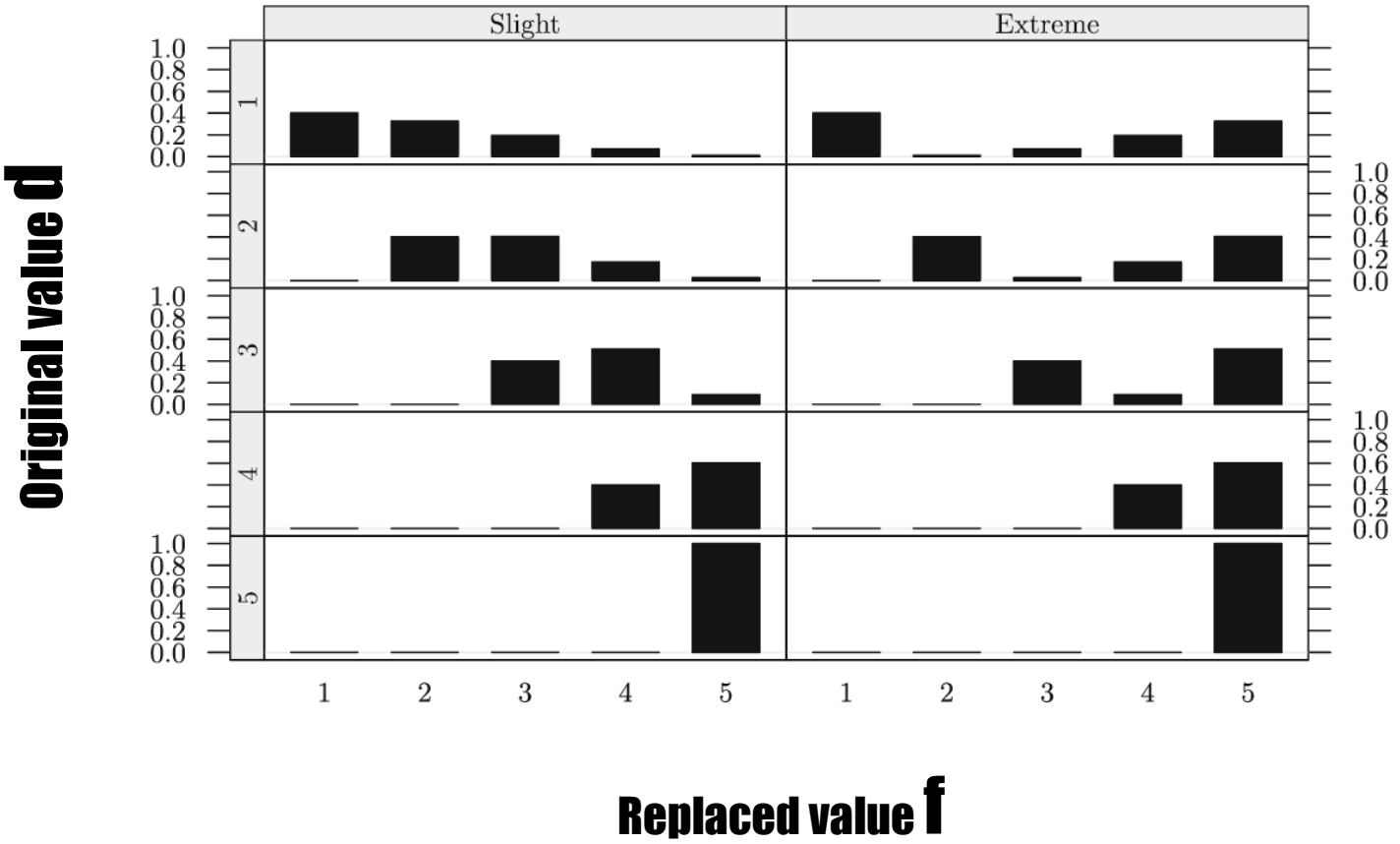
by Luigi Lombardi and Massimiliano Pastore

Abstract Many self-report measures of attitudes, beliefs, personality, and pathology include items that can be easily manipulated by respondents. For example, an individual may deliberately attempt to manipulate or distort responses to simulate grossly exaggerated physical or psychological symptoms in order to reach specific goals such as, for example, obtaining financial compensation, avoiding being charged with a crime, avoiding military duty, or obtaining drugs. This article introduces the package **sgr** that can be used to perform fake data analysis according to the sample generation by replacement approach. The package includes functions for making simple inferences about discrete/ordinal fake data. The package allows to quantify uncertainty in inferences based on possible fake data as well as to study the implications of fake data for empirical results.

Introduction

How can we evaluate the impact of fake information in real life contexts? In nature, some individuals tend to distort their behaviors or actions in order to reach specific goals. In some species, for example, wimpy animals may not signal their real social value by faking a higher status to deceive other

The replacement distribution



SGR data analysis: data generation scheme

Observed data

Observed sample
(subject by item response matrix)

$$\mathbf{X} \quad (n \times m)$$

$$T(\mathbf{X})$$

Observed statistic

3 SGR data analysis: data generation scheme

Observed data

Observed sample
(subject by item response matrix)

$\mathbf{X} \quad (n \times m)$

$T(\mathbf{X})$
Observed statistic

SGR simulated data space

Generative model

$g(\mathbf{D}|\theta)$

$\mathbf{D}_1 \quad \mathbf{D}_2 \quad \cdots \quad \mathbf{D}_H \quad (n \times m)$

3 SGR data analysis: data generation scheme

Observed data

Observed sample
(subject by item response matrix)

$\mathbf{X} \quad (n \times m)$

$T(\mathbf{X})$

Observed statistic

SGR simulated data space

Generative model

$g(\mathbf{D}|\theta)$

$\mathbf{D}_1 \quad \mathbf{D}_2 \quad \cdots \quad \mathbf{D}_H \quad (n \times m)$

Replacement model

$R(\mathbf{F}|\mathbf{D}, \theta_F)$

$\mathbf{F}_1 \quad \mathbf{F}_2 \quad \cdots \quad \mathbf{F}_H \quad (n \times m)$

3 SGR data analysis: data generation scheme

Observed data

Observed sample
(subject by item response matrix)

$\mathbf{X} \quad (n \times m)$

$T(\mathbf{X})$
Observed statistic

SGR simulated data space

Generative model

$g(\mathbf{D}|\theta)$

$\mathbf{D}_1 \quad \mathbf{D}_2 \quad \cdots \quad \mathbf{D}_H \quad (n \times m)$

Replacement model

$R(\mathbf{F}|\mathbf{D}, \theta_F)$

$\mathbf{F}_1 \quad \mathbf{F}_2 \quad \cdots \quad \mathbf{F}_H \quad (n \times m)$

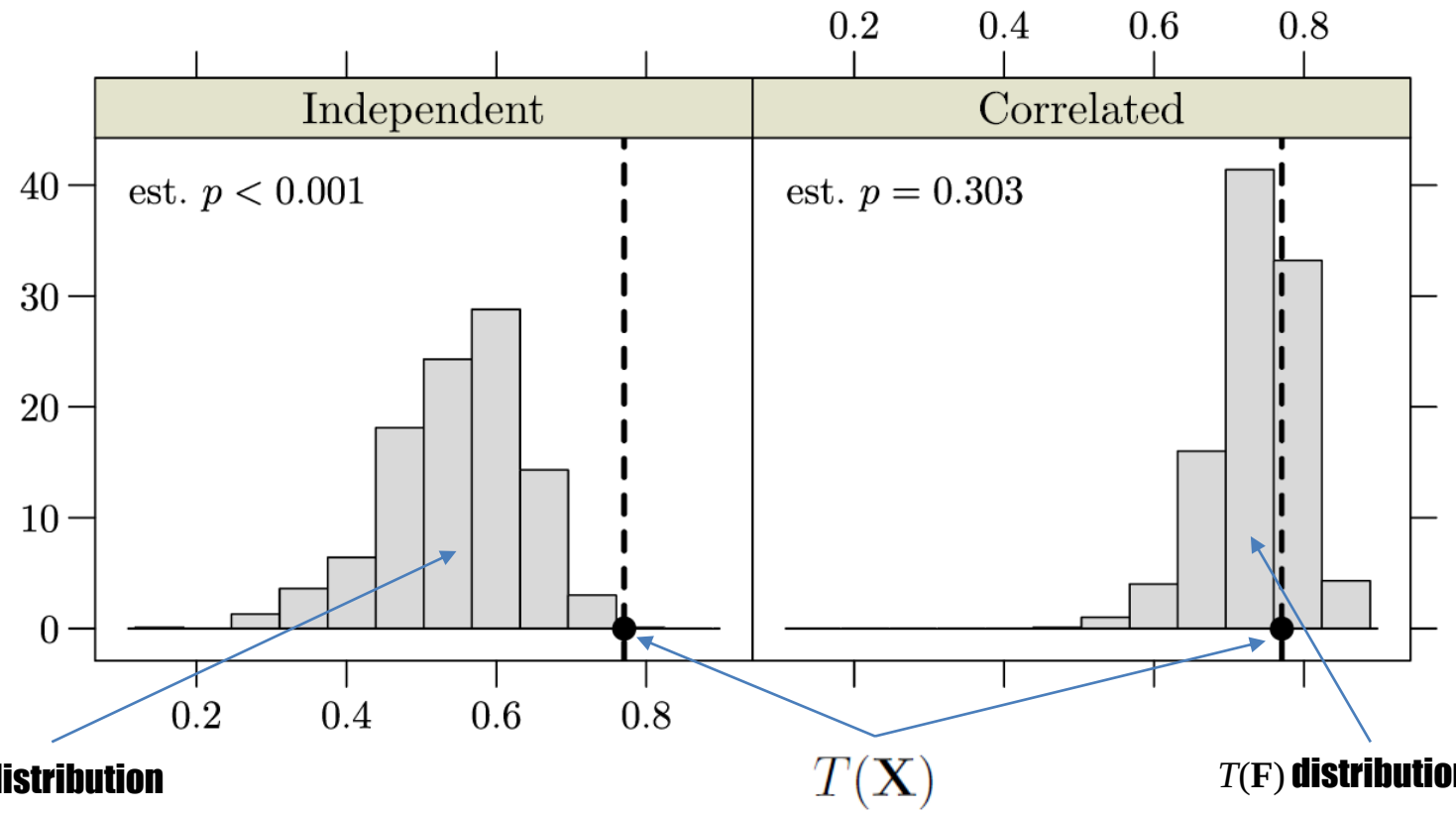
$T(\mathbf{F}_1) \quad T(\mathbf{F}_2) \quad \cdots \quad T(\mathbf{F}_H)$

Recomputed observed statistic results

3 SGR data analysis

SGR allows to test and compare different manipulated data models

example in Lombardi, Pastore, Nucci, & Bobbio (2015)



$T(\mathbf{F})$ **distribution**

$T(\mathbf{F})$ **distribution**

SGR data analysis

SGR works with the following types of data

- Binary data
- Ordinal data
- Categorical data (New!)

...and the following model representations

- General unparameterized probability models
  (categorical models, simple discrete probability models)
- UVA (Underlying-Variables Approach)
  (factorial models, SEM models, path analysis models)
- IRT (Item Response Theory) New!

In sum:

SGR can be used to **quantify uncertainty in inferences based on possible manipulated data** as well as to **evaluate the implications of manipulated data for statistical results**.

collaborators from the University of Padua (Massimiliano Pastore, Massimo Nucci, Andrea Bobbio) working on the SGR project

Main SGR references

- Lombardi L. & Pastore M. (2012). Sensitivity of fit indices to fake perturbation of ordinal data: A sample by replacement approach. *Multivariate Behavioral Research, 47,* 519-546.
- Pastore M. & Lombardi L. (2014). The impact of faking on Cronbach's Alpha for dichotomous and ordered rating scores. *Quality & Quantity, 48*, 1191-1211,
- Lombardi L. & Pastore M. (2014). sgr: A package for simulating conditional fake ordinal data. *The R Journal*, 6(1), 164-177.
- Lombardi L., Pastore M., Nucci M., & Bobbio A. (2015). SGR modeling of correlational effects in fake good self-report measures. *Methodology and Computing in Applied Probability, 17*, 1037-1055.

Other examples of replacement distribution