Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v8n3p288

A procedure simulating Likert scale item responses

By Boari, Nai Ruscone

Published: 20 December 2015

Electronic Journal of Applied Statistical Analysis
Vol. 08, Issue 03, December 2015, 288-297
DOI: 10.1285/i20705948v8n3p288

A procedure simulating Likert scale item responses

Giuseppe Boari *and Marta Nai Ruscone

Universitá Cattolica del Sacro Cuore and Statistic Department, Largo Gemelli 1, 20123 Milano

Published: 20 December 2015

Data collected during a survey by means of a questionnaire are, in general, expressed with reference to a Likert type scale, giving rise to non-metric data (ordinal categorical). However most of the statistical procedures used to analyze survey data (for example FA or SEM) require at least interval scale measures, that may be obtained for example by using proper scaling procedures. In order to compare, by simulation, scaling techniques most commonly used in the literature, we consider necessary to achieve, in advance, an appropriate algorithm that best reproduces the discretization process followed by the respondents to the Likert questionnaire items. Accordingly, we propose a discretization procedure that, starting from a continuous random variable, describing all possible individual responses to a given stimulus, generates the corresponding categories, choosen among a finite set of integer values.

keywords: Likert scale, continuous variables discretization.

1 Introduction

In the social and behavioral sciences responses to questionnaires items are typically collected by means of the Likert-type scale format: when answering to a question, the participants are asked to choose among a given number, say K, of ordered response categories running, for example, from "never" to "very often", usually coded with the first K integer numbers. However, the statistical techniques commonly used for their analysis (e.g. Structural Equation Models, SEM) assume data being of the metric type. Let us suppose, for example, that we're interested in evaluating the relationship between an output continuous variable Y and two explanatory continuous variables X_1 and X_2, measured by the regression coefficients β_1 and β_2 (see Fig. 1). Let also suppose that

*Corresponding author: giuseppe.boari@unicatt.it

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

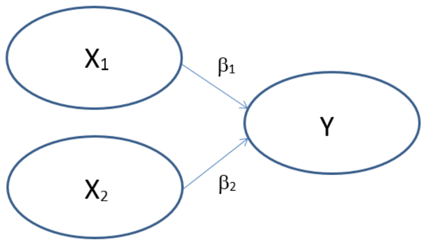

Figure 1: Example of a simple regression model

those variables are not directly observed, but measured by a Likert scale procedure, giving rise to the corresponding discrete variables $\tilde{Y}_1$ $\tilde{X}_1$ and $\tilde{X}_2$, of the ordinal type. Let us remember, with this regards, that psychometric literature on ordinal responses conceptualizes survey observed data as being expression of continuous unobserved underlying variables, whose true values lie in the mind of the specific subject: the Likert scale operates by means of a link function giving a conventional mapping from the stimula space (typically a continuous interval) to the first K integer values (discretization). Therefore, in order to make it possible the estimation of the above mentioned regression coefficients, a new set of variables (Y^*, X_1^* and X_2^*) are produced by means of a typical quantification/scaling procedure. Such scaling/quantification procedures transform the observed K levels, equally spaced, into a corresponding structure possessing at least the interval scale property (see Fig. 2).

The estimation of the regression coefficients β_i may so be realized through the estimates β_i^* obtained with reference to the transformed variables. The whole process previously introduced is summarized in the flow chart of Fig. 3.

Several proposal of quantification/scaling procedures are given in literature: see Thurstone (1927a,b, 1959); Torgerson (1958); Jones (1986); Zanella and Cantaluppi (2004); Zanella and Cerri (2000) for the so-called psychometric approach; Gifi (1990), Michailidis and de Leeuw (1998), Ferrari and Manzi (2010) for linear and non-linear PCA; Manisera et al. (2007); Zani and Berzieri (2008) suggesting a trasformation based on the mid-points of the empirical cumulative function.

In order to evaluate, by simulation, the performance of the most commonly used scaling techniques, it can be useful, for example, to compare the possible bias generated using the previously mentioned estimate β_i^*. With reference to the example previously introduced, the simulation exercise can start generating data from a continuous multidimensional random variable (Y, X_1, X_2) with known structure of covariance corresponding to the β_i. Then, a proper discretization algorithm will be used and subsequently estimates of the β_i^* may be computed on data re-scaled with the different quantification procedures under evaluation.

To come to the point, we consider necessary to have access to a suitable algorithm

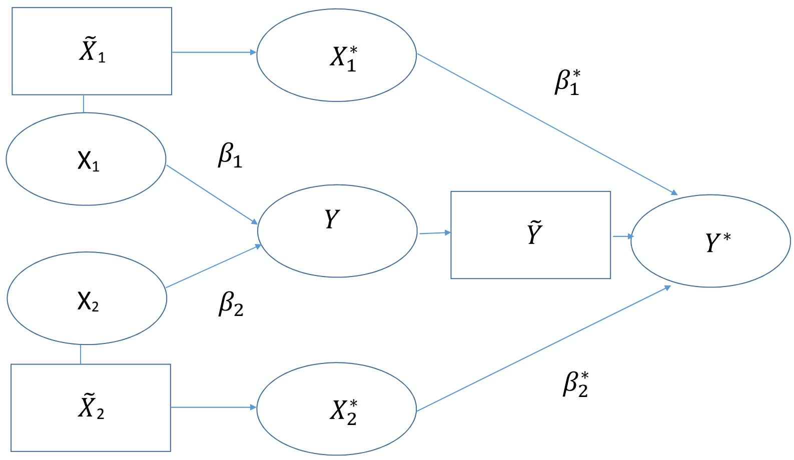

Figure 2: Structure of the relationship between original variables and the corresponding transformed variables

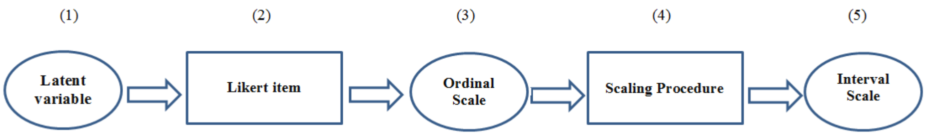

Figure 3: Process flow chart: discretization (1)(2)(3); quantification/scaling (3)(4)(5)

best reproducing the discretization phase of the process followed by the respondents to the Likert questionnaire items (see Fig. 3, steps (1) (2) (3)).

Therefore, section 2 describes the general assumptions concerning the production process of the ordered observed variables; in section 3 the Likert scale characteristics are considered; section 4 gives a short classifications of the discretization methods and in section 5 an innovative algorithm is proposed.

2 Assumptions on the Ordinal Observed Variables

As previously mentioned, the task of rating asks the respondent to a questionnaire item to select the number of the category that best represent his/her opinion on a particular attribute, among a finite ordered set of choices.

This implies, in general, the execution of a process of reduction of the range of a continuous unobserved variable into a finite set of disjoint intervals identified by proper meaningful labels.

We propose, in fact, to adopt the traditional assumptions considering the existence, for each observed ordinal manifest variable, say $\tilde{Y}$, of a corresponding underlying continuous latent variable, say Y Bollen (2014); Bollen and Maydeu-Olivares (2007). Formally, the observed ordinal response $\tilde{Y}$ with K response categories, is defined as:

$$\tilde{Y} = k, \quad if \;\; y_{k-1} < Y \leq y_k \qquad (k = 1, ..., K) \tag{1}$$

where y_k, $k = 0, 1, ..., K$, are thresholds defined in the domain of the underlying latent continuum, which may be spaced at nonequal intervals, satisfying the constraints $-\infty = y_0 < y_1 < ... < y_{K-1} < y_K = \infty$. This approach is often reasonable in the social sciences, in which many variables of interest are conceptually continuous, even though the measurement instruments give data having only ordinal properties. Therefore Likert scales represent a discretization of an underlying latent continuum.

3 Main features of Likert scales

Likert scales are positively characterized by their evident simplicity; nevertheless according to the psychometric literature, their use generally produces the following systematic distortion effects:

- **central tendency bias**, due to propensity of respondents to avoid using extreme response categories;
- **social desirability bias**, deriving from the tendency of raters to portray themselves in the most favorable light;
- **acquiescence bias**, related to the agreement to statements as presented.

In particular, authors commonly share the opinion that central tendency is the effect somewhat more problematic.

Therefore, in simulating the discretization process characterizing the Likert scale behaviour, one should reproduce the effects above illustrated, giving particular attention to the first one; Greenleaf (1992); Smith and Albaum (2005).

4 Simulating the discretization process

The aim of the discretization of a continuous variable is to find a set of cut points defining a partition of its domain into a small number, say K, of contiguous intervals, corresponding to K categories. The discretization algorithms may be distinguished between supervised and unsupervised procedures; the latter, most commonly used, divide the continuous range into K subranges, whose width or whose frequency (number of instances in each interval) are specified by the user (the so-called equal-width intervals or equal-frequency intervals). However, the unsupervised procedures may not give good results when the continuous distribution is not uniform or in the presence of outliers Catlett (1991). The supervised discretization methods are typically characterized by the following iterative process, defining the optimal solution to the problem.

1. Fix the number K of intervals
2. Choose two finite extreme values $[y_0, y_K]$ defining the finite range of the observations
3. Define a set of cut-points $y_1, y_2, ..., y_{K-1}$
4. Evaluate a proper target function (divergence measure)
5. Iterate steps 3-5 until a stopping rule will be satisfied
6. Define the K representative values $\overline{y}_1, \overline{y}_2, ..., \overline{y}_K$ summarizing each interval (tipically the conditional mean or median or simply the first K integers numbers)

Likert scoring applies the intervals into the first K integer numbers. The typical discretization procedure is outlined in the following figure 4: the continuous distribution $f(y)$, represented in panel (a), will be transformed into a discrete probability distribution p_k, see panel (b), by selecting the optimal partition of the range $[y_0, y_K]$ and assigning the probabilities

$$p_k = \int_{y_{k-1}}^{y_k} f(y)dy \tag{2}$$

to the conventional values $1, 2, ..., K$ defining the generic item X_i. In order to find the

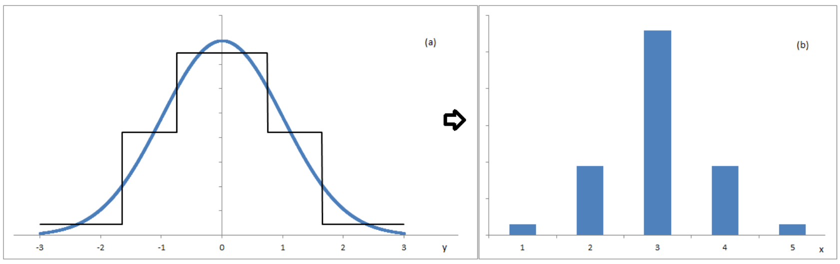

Figure 4: Discretization, panel (b), of a continuous pdf, panel (a)

optimal set of cut-points $y_1, y_2, ..., y_{K-1}$ the following type of target function is in general employed:

$$\min_{y_1,...,y_{K-1}} \sum_{k=1}^{K} d[(f(y), p_k(y)]w(y) \tag{3}$$

where $w(y)$ is a weight factor, $d[\cdot, \cdot]$ is a proper measure of divergence between the continuous function $f(y)$ and the step function $p_k(y)$ assuming the constant values

$p_k/(y_k - y_{k-1})$, $k = 1, 2, ..., K$ in each interval, see panel (a). Observe that some quantification procedures, dealing with the particular problem of the discretization of a continuous wave signal (performing, for example, an analog to digital transformation), define the levels of the step function together with the corresponding cut-points, so not imposing constraint (2), see e.g. Max (1960); Gifi (1990).
Concerning the divergence measures $d[\cdot,\cdot]$ most frequently employed in the literature, we can make the following, non exhaustive, list.

- Max quantizer:
$$d[\cdot,\cdot] = [f(y) - p_k(y)]^2 \qquad w(y) = f(y) \tag{4}$$
- Kullback Leibler (K-L):
$$d[\cdot,\cdot] = \ln[f(y)/p_k(y)] \qquad w(y) = f(y) \tag{5}$$
- χ^2:
$$d[\cdot,\cdot] = [f(y) - p_k(y)]^2/f(y) \qquad w(y) = 1 \tag{6}$$
- weighted χ^2:
$$d[\cdot,\cdot] = [f(y) - p_k(y)]^2/f(y) \qquad w(y) = f(y) \tag{7}$$
- weighted absolute value (A-V):
$$d[\cdot,\cdot] = |f(y) - p_k(y)| \qquad w(y) = f(y) \tag{8}$$

Note that the χ^2 divergence is naturally weighted by the factor $1/f(y)$; moreover the Max quantizer criterion may be expressed as

$$\min_{y_1,\ldots,y_{K-1},p_1,\ldots,p_K} \sum_k \int_{y_{k-1}}^{y_k} [f(y) - p_k(y)]^2 f(y)dy \tag{9}$$

and the weighted χ^2 criterion assumes the following simple expression

$$\min_{y_1,\ldots,y_{K-1}} \sum_k \int_{y_{k-1}}^{y_k} [f(y) - p_k(y)]^2 dy \tag{10}$$

where the weight factor $f(y)$ reduces the presence of the implicit weight $1/f(y)$. Observe that the criteria based on Max quantizer and K-L, weighted with $f(y)$, tend to assign higher divergences to the intervals corresponding to the higher values of $f(y)$, possibly giving them a narrower width. On the other hand, the χ^2 criteria, using the natural weight $1/f(y)$, tends to be more accurate in the tails of the distribution. Moreover, it can be noted that A-V divergence corresponds to the square root of the weighted χ^2 criterion, further weighted by $w(y) = f(y)$. In Gifi (1990); Manisera et al. (2007); Carpita and Manisera (2012) the Max quantizer is used, for simulating the discretization phases (1) (2) (3) see fig 3.

5 The proposed discretization algorithm

Recall that, as previously stated, a discretization algorithm simulating the likert scale process should reproduce the main peculiarity of those scales: the central tendency bias. Then, in our opinion, a discretization algorithm should generally assign higher probability to the intervals (classes) closer to the mode of the continuous distribution.

For this reason we decided to start the algorithm with testing the presence of the asymmetry; then, once the total number of intervals has been fixed, choose the number of intervals to be located to the right and to the left of the modal value. For this reason, process the iterative phase begins from the modal class. In order to discretize a continuous variable Y, such that $P(y_{min} \leq Y \leq y_{max}) \simeq 1$ the following procedure is suggested:

- choose the number K of classes
- evaluate the range $R = y_{max} - y_{min}$, the mode $\widetilde{y}$ and the median of Y
- compute $h = int(\frac{|mode-median|}{R/K})$; when $h > 0$ then asymmetry exists
  - if K odd then assign
    $(K-1)/2 + h$ classes to the right of the mode
    $(K-1)/2 - h$ to the left of the mode
  - if K even then assign
    $K/2 + h$ classes to the right of the mode
    $K/2 + h$ to the left of the mode
- compute $f(\widetilde{y})$ and fix $\epsilon > 0$, an appropriate small positive tolerance value
  - if K odd
    - optimization of the modal class
    - for $j = 1, 2, ...$
    - calculate y_{k-1} and y_k corresponding to $f^{-1}(\widetilde{y} - j\epsilon)$
    - evaluate the probability p_k pertaining to the class $(y_{k-1}, y_k]$
    - compute the height $p_k(y) = p_k/(y_k - y_{k-1})$ of the k-th interval of the step function
    - set the remaining right and left intervals to be equally spaced
    - evaluate the divergence measure for the currently defined step function
    - stop the j-iteration whenever the minimum divergence is reached
    - for the optimization of the remaining intervals refer to the K even case
  - if K even
    - optimization of the right tail intervals
    - for each interval increasingly far from the modal value

- ∗ for $j = 1, 2, ...$
- ∗ calculate y_{k-1} and y_k corresponding to $f^{-1}(y_k - j\epsilon)$
- ∗ evaluate the probability p_k pertaining to the class $(y_{k-1}, y_k]$
- ∗ compute the height $p_k(y)$ of the k-th interval of the step function
- ∗ set the remaining right intervals to be equally spaced
- ∗ evaluate the divergence measure for currently defined step function relatively to the right tail of the distribution
- ∗ stop the j-iteration whenever the minimum divergence is reached
- ∗ for the optimization of the lower (left) intervals repeat the previous iteration substituting right with left

- the threshold values $y_1, y_2, ..., y_{K-1}$ and the corresponding p_k define the optimal discretization setting

6 A brief comparison of six discretization procedures

In order to evaluate the performance of the proposed algorithm we give, in the following Table 1, some essential results referring to the discretization of a continuous standard normal distribution into a discrete one, with $K = 5$ classes, having fixed $y_{min} = -3$ and $y_{max} = 3$. For sake of simplicity we restricted the comparison among four supervised procedures (Weighted χ^2, A-V, Max pseudo-optimal, Max optimal) and two unsupervised (Equally spaced intervals, Equal probability intervals), often mentioned in the literature as a reference point. We may compare the different p_k distributions, $k = 1, ..., 5$, with reference to the Equally spaced one, which gives the probabilities of the std normal corresponding to the five intervals of size 1.20 partitioning the range $[-3, 3]$.

Table 1: Comparison of some discr. procedures (std normal distr. and $K = 5$ classes)

procedure	1	2	3	4	5	mid class width
Weighted χ^2	.063	.103	.669	.103	.063	1.93
A-V	.083	.220	.560	.220	.083	1.57
Equally spaced intervals	.036	.262	.452	.262	.036	1.20
Max pseudo-optimal	.100	.200	.400	.200	.100	1.04
Max optimal	.107	.244	.300	.244	.107	0.76
Equal prob. intervals	.200	.200	.200	.200	.200	0.50

7 Concluding remarks

Remind, as initially stated, that a good discretization algorithm is expected to best reproduce the Likert scale distortion effect of central tendency. For this reason we suggest to take into account the probabilities assigned to the central classes or, equivalently, the mid class widths. Observe that the Equally spaced interval method gives to the middle interval a probability of 45.2%; both of Max Procedures and the Equal Probability intervals are positioned below; on the contrary A-V and Weighted χ^2 do emphasize that level (56.0% and 66.9% respectively). In conclusion, is our opinion that the A-V criterion looks to reproduce the central tendency bias in a satisfactory manner, avoiding, at the same time, excessive effects.

References

Bollen, K. A. (2014). *Structural equations with latent variables.* John Wiley, New York.

Bollen, K. A. and Maydeu-Olivares, A. (2007). A polychoric instrumental variable (piv) estimator for structural equation models with categorical variables. *Psychometrika*, 72(3):309–326.

Carpita, M. and Manisera, M. (2012). Constructing indicators of unobservable variables from parallel measurements. *Electronic Journal of Applied Statistical Analysis*, 5(3):320–326.

Catlett, J. (1991). On changing continuous attributes into ordered discrete attributes. In *Machine learning EWSL-91*, pages 164–178. Springer.

Ferrari, P. and Manzi, G. (2010). Nonlinear principal component analysis as a tool for the evaluation of customer satisfaction. *Qualitative Technology and Quantitative Management*, 7(2):117–132.

Gifi, A. (1990). *Nonlinear multivariate analysis.*

Greenleaf, E. A. (1992). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research*, 29(2):176–188.

Jones, L. V. (1986). Psychological scaling. *Encyclopedia of Statistical Sciences*, 7:340–343.

Manisera, M. et al. (2007). Scoring ordinal variables for constructing composite indicators. *Statistica*, 67(3):309–324.

Max, J. (1960). Quantizing for minimum distortion. *Information Theory, IRE Transactions on*, 6(1):7–12.

Michailidis, G. and de Leeuw, J. (1998). The gifi system of descriptive multivariate analysis. *Statistical Science*, 13(4):307–336.

Smith, S. M. and Albaum, G. S. (2005). *Fundamentals of marketing research.* Sage, Thousand Oaks.

Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological review*, 34(4):273–286.

Thurstone, L. L. (1927b). The method of paired comparisons for social values. *The Journal of Abnormal and Social Psychology*, 21(4):384–400.

Thurstone, L. L. (1959). *The measurement of values.* Univer. Chicago Press, Chicago.

Torgerson, W. S. (1958). Theory and methods of scaling.

Zanella, A. and Cantaluppi, G. (2004). Simultaneous transformation into interval scales for a set of categorical variables. *Statistica*, 64(2):401–426.

Zanella, A. and Cerri, M. (2000). La misura di customer satisfaction: qualche riflessione sulla scelta delle scale di punteggio. In *Valutazione della qualitá e Customer Satisfaction: il ruolo della statistica*, pages 217–231. Vita e pensiero, Milano.

Zani, S. and Berzieri, L. (2008). Measuring customer satisfaction using ordinal variables: an application in a survey on a contact center. *Statistica applicata*, 20:331–351.