EXHIBIT R

Quality & Quantity (2007) 41:601–626
DOI 10.1007/s11135-007-9089-z

Ordinal Methodology in the Analysis of Likert Scales

RAINER GÖB[1,*], CHRISTOPHER McCOLLIN[2] and MARIA FERNANDA RAMALHOTO[3]
[1]*Institute for Applied Mathematics and Statistics, University of Würzburg, Sanderring 2, D-97070 Würzburg, Germany. E-mail: goeb@mathematik.uni-wuerzburg.de;* [2]*Nottingham Trent University, University Burton Street, Nottingham, NG1 4BU, United Kingdom. E-mail: Chris.McCollin@ntu.ac.uk;* [3]*Instituto Superior Técnico, Maths Dept., Av. Rovisco Pais, 1049-001 Lisbon, Portugal*

Abstract. Likert scales are widely used in survey studies for attitude measuring. In particular, the questionnaires propagated by the SERVQUAL approach are based on Likert scales. Though the problem of attitude suggests an ordinal interpretation of Likert scales, attitude survey data are often evaluated with techniques designed for cardinal measurements. The present paper discusses the interpretation of scales for attitude measuring and gives a survey of data analysis techniques under the proper ordinal understanding.

Key words: attitude measuring, likert scales, ordinal scales, cardinal scales, SERVQUAL, statistical analysis.

1. Introduction

Likert scales are widely used for measuring attitudes, e.g., opinions, psychic and mental dispositions, preferences. Questionnaires and surveys based on Likert scales are used in various areas, e.g., in psychometrics for the analysis of subjective well-being, see Diener et al. (1985) or Watson et al. (1988), in social studies and panels, or for purposes of business administration. The use of Likert scales has increased especially in the service sector with consumer surveys now being commonplace within the hotel, leisure and public utility sectors. In particular, the SERVQUAL approach introduced by Parasuraman et al. (1985, 1988) has received enormous interest. The ways of collecting survey data vary widely from the use of telephone questionnaires to on-line designed web pages for automatic input.

The statistical analysis of survey data can range from simple dot plots to logistic regression and cluster analysis to determine any hidden structure. However, many studies confine themselves to a descriptive analysis.

*Author for correspondence: E-mail: goeb@mathematik.uni-wuerzburg.de

Clason and Dormody (1994) compare 95 articles analyzing Likert scales from the Journal of Agricultural Education. 51 reported only descriptive statistics. In a recent review of some University Business School dissertations, most students opted for questionnaires and/or interviews for their primary research and the main statistical analysis was of an exploratory nature with bar charts, check lists and Pareto plots undertaken. It is interesting to note that in a similar way to Ishikawa's three levels of tools which provide the differentiation between Six Sigma Green and Black Belts, most students will mainly only attempt Ishikawa's level 1 tools (7 basic tools) to carry out their analysis even though they have been taught level 2 and 3 tools such as ANOVA and regression.

Unfortunately, the promotion of ways to analyze data measured in Likert scales is not widely available within textbooks. In fact, there is no common standard accepted by the scientific community for the correct interpretation and analysis of such data. Interpretation and analysis often seem to be in a mismatch. In methodological considerations it is generally acknowledged that attitude measuring scales should be considered as *ordinal*. Nevertheless, many studies use *cardinal* statistics as sample means, sample variances, t-tests to analyze attitude data. Proper ordinal approaches are in the minority. In particular, the SERVQUAL methodology as usually propagated is completely based on cardinal statistics.

The objective of the present paper is to establish a framework for the analysis of survey data under an explicitly ordinal interpretation of the Likert scale. Sections 2 and 3 review the debate on the impact of scale typologies on statistical methodology. Sections 4 and 5 discuss the interpretation of Likert scales. Sections 6 and 7 suggest the multinomial model for modelling data from attitude surveys. Sections 8 through 12 consider the analysis of survey data from a homogeneous sample of respondents. Ways of detecting and analyzing inhomogeneous samples are discussed in Section 13.

2. Ordinal and Cardinal Scales

We consider one-dimensional scales which can be identified with subsets of the real line. Stevens (1946, 1951) characterizes the scale types *nominal, ordinal, interval, ratio* in terms of *permissible transformations*. We use Stevens' (1932) ideas to distinguish between *ordinal* and *cardinal* scales.

Ordinal measure scales consist of categories ordered by a relation of the type "$<$" or "$\leq$", respectively. Any two measure values can be compared in terms of the order relation. The admissibility of strictly increasing scale transformations preserving the order relation is characteristic for ordinal scales. Consequently, differences of scale values are not meaningful. Beyond order, there is no measure for the distance between two scale values. For

instance, the ordinal scales 1, 2, 3, 4, 5 and 1, 3, 9, 27, 81 are equivalent. However, in the first scale the magnitude of differences between successive points is identical, whereas it is increasing in the second scale.

Cardinal measure scales express magnitudes. Differences between scale values are meaningful. In Stevens' terminology, cardinal scales are *interval scales*. Interval scales are characterized by the admissibility of strictly increasing linear transformations. For instance, the cardinal scales 1, 2, 3, 4, 5 and 0, 2, 4, 6, 8 are equivalent.

3. Scale Interpretation and Statistical Methodology

The rationale behind an axiomatic distinction of scales as described in Section 2 is beyond doubt. However, the role of scale identifications in the methodology of statistical data analysis is controversial.

Stevens (1951) and subsequently many other authors, e.g., Luce (1959), Townsend and Ashby (1984) and Luce et al. (1990), postulate the following steps of data analysis:

(S1) Scales for measuring the values of certain attributes are chosen according to criteria provided by measurement theory.

(S2) The measure scale chosen in step (S1) prescribes certain statistics and proscribes others.

In this view, data measured in a specific scale have to be analyzed by statistics which preserve their meaning under the characteristic transformation of the scale. Admissible statistics for ordinal data are frequencies, histograms, order statistics. Methods involving arithmetic or weighted means are appropriate for cardinal data, but they make no sense for the analysis of ordinal data. Andrews et al. (1981) present an elaborate guide to select statistical methods in accordance with measure scales.

The above view of the predominant role of measurement theory in data analysis has been criticized by several authors, see Lord (1953), Savage (1957), Tukey (1961), Adams et al. (1965) and Baker et al. (1986) for instance. More references and a detailed discussion survey are given by Velleman and Wilkinson (1993). Subsequently we consider only one, but substantial critical argument.

The following propositions can be taken for granted: (i) Data analysis is an autonomous discipline. (ii) Among other techniques, data analysis uses formal mathematical methods, without being a part of mathematics. (iii) Any data analysis is motivated by a specific *problem*, i.e., specific interests and objectives of knowledge discovery, occurs in a specific *context*, i.e., a specific scientific or pragmatic environment, and reflects methods with respect to their *solution potential* for the problem in the context. (iv) The criteria of adequacy of methods of data analysis result from the specific problem, the specific context, and the solution potential.

Under propositions (i) through (iv), the description of data analysis by steps (S1), (S2) requires the following further assumption: (v) Measurement theory alone is able to reflect the criteria resulting from problem, context, and solution potential, by determining uniquely an adequate measure scale.

However, considering the actual state of measurement theory as a discipline, see Luce et al. (1990) for instance, it will be difficult to defend Proposition (v). Customary measurement theory deliberately works *without* reflecting the potential of the entire corpus of statistical data analysis regarding problem, context, and solution potential. On the contrary, the succession of steps (S1), (S2) claims that methods of analysis can be selected or excluded without reflecting their potential. Measurement theory claims to be a preliminary fundamental discipline for data analysis. However, it is strongly influenced by formal axiomatic reasoning and fails to provide a conceptual framework to structure data analysis according to the basic matters of problem and context, and solution potential.

The succession of steps (S1), (S2) has to be rejected. Scale type identification is reasonable to avoid conceptual confusion. However, scale type identification by measurement theory is not exclusively decisive for the choice of data analysis methods. The choice of appropriate methods is determined by the three interdependent factors listed above. In this vein, Adams et al. (1965): "Nothing is wrong per se in applying any statistical operation to measurements of given scale, but what may be wrong, depending on what is said about the results of these operations, is that the statement about them will be empirically meaningful or else that it is not scientifically debated."

Examples illustrating the above argument are provided by Lord (1953) and Wright (1997). The subsequent Section 5 discusses the interpretation of Likert scales.

4. The Likert Scale

Rensis Likert (1932) introduced a scale and technique for attitude measurement. An individual is confronted with statements which are essentially value judgements. The value judgements may concern the individual's reflections of reality or the individual's psychic dispositions as feelings, wants, desires, conative dispositions. The individual is invited to define his attitude towards each statement by choosing among a number of r grades (*scores, degrees*) on the *r-grade Likert scale*. Most popular are five-grade and seven-grade Likert scales. The grades (scores, degrees) $1, \ldots, r$ are ordered in ascending order of agreement or approval of the individual with respect to the value statement. In case of $r = 5$, the grades are usually interpreted by *strongly disagree, disagree, neutral (undecided), agree, strongly agree*.

Likert scales are widely used in different areas for attitude measurement by surveys, e.g., in psychology, sociology, health care, marketing, quality control. Popular applications are in the assessment of customers' quality perceptions or expectations, and of subjective well-being. Subjective well-being has become an important topic in research and practical fields like health care, see Diener (1984) or Diener et al. (1999).

Lots of differently structured attitude measuring techniques based on Likert scales are used. We describe some standard schemes which have been widely used for many years: SERVQUAL, PANAS, SWLS, GSOEP.

4.1. SERVQUAL

Attitude surveys structured according to the SERVQUAL approach introduced by Parasuraman et al. (1985, 1988) are currently among the most popular applications of sample surveys in industry. SERVQUAL surveys intend to inquire customers' attitudes towards service quality. Service quality is considered with respect to ν dimensions which are addressed by a questionnaire regarding M performance items. Parasuraman et al. (1985, 1988) suggest $M = 22$ items grouped into $\nu = 5$ dimensions of service quality: tangibles (environmental factors), reliability, responsiveness, assurance, empathy. This setting has widely been accepted in applications. More subtle investigations use statistical instruments like principal components analysis to confirm or modify the setting, see Asubonteng et al. (1996) for a literature survey.

The questionnaire contains a statement on each of the M performance items. The respondent is invited to qualify his attitude towards each statement in a response scale of Likert type with grades or scores ranging from 1 ("strongly disagree") to r ("strongly agree"). Most popular in SERVQUAL manuals and case studies are scales with $r = 7$ or $r = 5$ grades, see Parasuraman et al. (1985, 1988). Occasionally, other values of r, e.g., $r = 10$ are also used, see Asubonteng et al. (1996).

SERVQUAL distinguishes between two attitudes: *expectations* on quality, i.e., what a customer expects from the service, and *perceptions* of quality, i.e., the customer's view of what actually happened. SERVQUAL intends to measure the *gap* between expectations and perceptions. To this end, SERVQUAL questionnaires are doubled: The respondent is invited to qualify his attitude towards each of the M statements once in the sense of expectation, irrespective of what actually happened, once in the sense of perception of what actually happened.

The SERVQUAL community has adopted some standard quantitative methodology for the evaluation of SERVQUAL surveys which was essentially coined by Parasuraman et al. (1985, 1988, 1991). These methods are based on an implicit cardinal interpretation of the Likert scale. For

each respondent, dimension scores and a total SERVQUAL score are calculated as arithmetic or appropriately weighted averages. Survey scores are calculated as arithmetic averages of the respondent scores. Gap scores are calculated as differences of perception score minus expectation score.

4.2. PANAS

The *positive and negative affect scale* (PANAS) introduced by Watson et al. (1988) is concerned with measuring subjective dispositions in the sense of moods, momentary, mid-term, or long-term. PANAS refers to 20 feelings or emotions in two dimensions, positive and negative. The respondent is asked to notify the degree of realizing the feeling or emotion in a five-grade Likert scale with values *very slightly or not at all, a little, moderately, quite a bit, extremely*.

The quantitative methodology suggested by Watson et al. (1988) uses an implicit cardinal interpretation of Likert scores.

4.3. SWLS

To measure global and rather persistent judgements on individual life, Diener et al. (1985) suggest the *satisfaction with life scale* (SWLS). SWLS considers only five statements: "In most ways my life is close to my ideal", "The conditions of my life are excellent", "I am satisfied with my life", "So far I have gotten the important things I want in life", "If I could live my life over, I would change almost nothing". The respondent is asked to notify the degree of approval with each statement in a seven-grade Likert scale ranging from *strongly disagree* to *strongly agree*.

The quantitative methodology used by Diener et al. (1985) to evaluate the SWLS technique is based on an implicit cardinal interpretation of Likert scores.

4.4. GSOEP

The *German Socio-Economic Panel* (GSOEP) has been conducted as a longitudinal panel in Germany since 1984. It includes 11 questions concerning satisfaction with work, income, health, housing, leisure, consumption. The answers are notified in an eleven-grade Likert scale ranging from 0 (*totally dissatisfied*) to 10 (*totally satisfied*). Further information about GSOEP can be obtained on the WWW at http://www.diw.de/. GSOEP contains no advice for scale interpretation or methods of analysis.

5. The Character of Likert Scales

Following the conclusion of Section 3, the specific *problem*, the *context* of data analysis, and the *problem solving potential* of methods are crucial for

the deciding about the scale type and appropriate methods of analysis. For a Likert scale, the alternative is between an ordinal and a cardinal scale type.

5.1. THE PROBLEM BEHIND LIKERT SCALES

The *problem* behind the use of Likert scales is measuring attitudes. Accordingly, the interpretation of such scales is discussed in psychology, sociology and economics. Attitude measuring has to satisfy the following criteria:

- Longitudinal consistency or retesting reliability: At repeated measuring times under invariant relevant side conditions respondents exhibit the same rating.
- Longitudinal comparibility: Responses given by an individual at different times with respect to the same item can be compared on the scale.
- Internal consistency.
- Interpersonal comparibility: Responses from different inviduals can be compared on the scale.
- Plausibility: The measuring method has to conform to naive assessments of attitudes.

By definition in terms of admissible transformations, an ordinal scale is less restrictive in interpretation than a cardinal scale. Hence it is easier to satisfy the above criteria under an ordinal interpretation than under a cardinal interpretation of the Likert scale.

In particular, ordinal scales facilitate achieving comparibility. For instance, consider a seven-grade Likert scale to measure satisfaction and let an individual report grade 2 at time t_1 and grade 4 at time t_2. Under a cardinal interpretation this amounts to the controversial conclusion that the individual at time t_2 experiences twice the satisfaction experienced at time t_1. Under an ordinal interpretation it only means that the indvidual's satisfaction increased from the second to the fourth position on the scale. Comparability has been discussed extensively in the theory of utility and of social choice, see Georgescu-Roegen (1968), van Praag (1991) or Sen (1999). Many authors agree that under ordinal scaling interpersonal comparibility is a justified working hypothesis, see Ferrer-I-Carbonell and van Praag (2003). Cardinal interpretations, however, involve considerable difficulties in guaranteeing comparability.

Naive cardinal interpretations of ordinal scales may violate internal consistency and interpersonal comparibility. Hart (1996) reports the results of an experiment suggested by Lodge (1981) for quantifying the grades in a Likert scale by magnitudes. A sample of respondents is invited to assign magnitudes to the grades of a 7-grade Likert scale with the interpretations *atrocious, very bad, bad, so-so, good, very good, excellent*. The result shows

considerable differences in the weights assigned to distances between the grades on the Likert scale. For instance, the step from *atrocious* to *very bad* is quantified by 0.6, whereas the step from *so-so* to *good* is quantified by 1.9.

5.2. THE CONTEXT OF THE USE OF LIKERT SCALES

Consider the scientific and pragmatic *context* of the use of Likert scales in survey techniques. In view of the context, methodology is rated by the following criteria:

- Acceptance by communities in practice or research.
- Standardization.
- Comparibility of results.

Section 4 lists four popular standardized attitude measuring techniques based on Likert scales: SERVQUAL, SWLS, PANAS, GSOEP. All are widely accepted, definite, standardized. They differ in advice for scale interpretation and for methods of data analysis.

GSOEP contains no advice for scale interpretation or methods of analysis. Recent studies in GSOEP attitude data include explicitly ordinal scale interpretations, see Ferrer-I-Carbonell and van Praag (2003) or Nolte and McKee (2004), and implicitly cardinal ones, see Lucas et al. (2003) or Ronellenfitsch and Razum (2004).

SERVQUAL, SWLS, and PANAS, in the manner originally conceived by their authors, see Parasuraman et al. (1988), Diener et al. (1985) and Watson et al. (1988), contain advices on data analysis. These advices imply a cardinal interpretation of the Likert scale: empirical sums, means, variances and correlation coefficients of scores are calculated. Such approaches are mainly motivated by pragmatic reasoning since cardinal statistics are widely available in textbooks and software. However, they contradict principles of attitude measuring which suggest ordinal scaling, see Section 5.1, above. Essentially, two types of misleading conclusions may follow from the conflict of intrinsic ordinality and imposed cardinality. (1) Complete distortion of results by applying strictly monotonous transformations to a scale which bears a cardinal interpretation. Fortunately, this type of misinterpretation is prevented by the pragmatic context, where SERVQUAL, SWLS, and PANAS are strictly linked to unambiguous Likert scales with grades $1,\ldots,7$. The idea of subjecting the scale to transformations is purely academic. (2) Interpretation of attitude grades in terms of magnitudes. This is a serious misinterpretation supported by approaches like SERVQUAL, SWLS, or PANAS. Often enough, practitioners report results of surveys by statements like "We've increased customer satisfaction by 150% in one year."

5.3. THE ANALYSIS OF ATTITUDE SURVEYS

Consider the *problem solving potential* of methods for the analysis of attitude surveys. The major criteria are:

- Clarity.
- Exactness.
- Informational value.
- Simplicity.
- Availability.

The cardinal scale approach suggested by standard descriptions of SERVQUAL, PANAS, SWLS excels by *simplicity* and *availability*. Methods like principal components analysis, factor analysis, correlation analysis, t-testing or ANOVA are from the conventional statistical toolbox, readily available in textbooks or software packages.

Deficiencies of the cardinal scale approach are in *clarity* and *exactness*. The basic problem of scale interpretation generally remains unmentioned in the SERVQUAL environment and is discussed by few authors only, see Hart (1996) or Hart (1999). Many of the methods usually recommended are based on normality assumptions. These assumptions mostly remain undiscussed. Attempts of substantiating by asymptotics are not made.

The *informational value* of methods recommended in SERVQUAL, PANAS or SWLS schemes is undoubted. Summed or averaged scores convey information about respondents' attitudes. However, cardinal statistics also may hide or distort information. For instance, strong agreements and strong disagreements may be averaged, providing a misleading impression of average agreement.

5.4. CONCLUSION ON THE INTERPRETATION OF LIKERT SCALES

The problem of attitude measuring clearly suggests an ordinal interpretation of Likert scales. The context of use has established some implicit cardinal interpretation. To some extent, cardinal statistics have successfully been applied in the analysis of attitude surveys.

In summary, ordinal methodology for Likert scale analysis conforms to the problem of attitude measuring, but it differs from widely used and sufficiently successful practice. To be acceptable for practitioners, a proper ordinal approach to Likert scale analysis has to substantiate its problem solving potential according to the criteria listed in Section 5.3, in particular with respect to simplicity and availability. The subsequent sections give an overview of ordinal methodology which is competitive in this sense.

6. Formal Description of Attitude Questionnaires

The discussion of quantitative analysis of attitude surveys requires a formal description of attitude questionnaires.

Consider a questionnaire with M statements expressing ν dimensions. In SERVQUAL usually $\nu = 5, M = 22$. Let $1 = m_1 < \cdots < m_{\nu+1} = M + 1$ and let the statements (items) $m_\rho, \ldots, m_{\rho+1} - 1$ be associated with dimension ρ, $\rho = 1, \ldots, \nu$. Responses are notified in a Likert scale with r grades represented by the numbers $1, \ldots, r$. The survey is conducted with n respondents $i = 1, \ldots, n$. The response of respondent i with respect to statement j is denoted by an r-tuple $\boldsymbol{X}_{ij} = (X_{ij1}, \ldots, X_{ijr})$ with entries from $\{0, 1\}$, $X_{ij1} + \cdots + X_{ijr} = 1$. $X_{ijl} = 1$ means: with respect to statement j, respondent i exhibits agreement grade $1 \le l \le r$ on the Likert scale. Then

$$\boldsymbol{X}_i^{(\rho)} = \sum_{j=m_\rho}^{m_{\rho+1}-1} \boldsymbol{X}_{ij}, \quad \rho = 1, \ldots, \nu, \qquad \boldsymbol{X}_i = \sum_{\rho=1}^{\nu} \boldsymbol{X}_i^{(\rho)} = \sum_{j=1}^{M} \boldsymbol{X}_{ij} \tag{1}$$

is the response vector of respondent i in dimension ρ, respectively the total response vector in all M items.

The above scheme can be used to describe the response on perception as well as on expectation. The response of respondent i with respect to statement j in terms of the gap between perception and expectation is denoted by a $(2r-1)$-tuple

$$\boldsymbol{Z}_{ij} = (Z_{i,j,-(r-1)}, \ldots, Z_{i,j,0}, \ldots, Z_{i,j,r-1}) \tag{2}$$

with entries from $\{0, 1\}$, $Z_{i,j,-(r-1)} + \cdots + Z_{i,j,r-1} = 1$. $Z_{ijl} = 1$, $l < 0$ means: with respect to item j, the perception of respondent i is $|l|$ degrees below the expectation. $Z_{ijl} = 1$, $l > 0$ means: with respect to item j, the perception of respondent i exceeds the expectation by l degrees. $Z_{ij0} = 1$ means: with respect to item j, the perception of respondent i equals the expectation.

The above interpretation of gaps is consistent with an ordinal interpretation of the Likert scale. The cyphers $-(r-1), \ldots, 0, \ldots, r-1$ indicate distances of expectation and perception in terms of degrees, not of magnitudes. These distances remain invariant under strictly monotonous scale transformations.

The vectors

$$\boldsymbol{Z}_i^{(\rho)} = \sum_{j=m_\rho}^{m_{\rho+1}-1} \boldsymbol{Z}_{ij}, \quad \rho = 1, \ldots, \nu, \qquad \boldsymbol{Z}_i = \sum_{\rho=1}^{\nu} \boldsymbol{Z}_i^{(\rho)} = \sum_{j=1}^{M} \boldsymbol{Z}_{ij} \tag{3}$$

are the gap vectors of respondent i in dimensions $\rho = 1, \ldots, \nu$, respectively the total gap vector of respondent i in all M items on the questionnaire.

7. The Multinomial Model for Attitude Responses

The notation of Section 6 can be used to express survey evaluations based on a cardinal view of the Likert scale by forming weighted sums and averages of scores. Under an ordinal interpretation, quantitative analysis is primarily interested in the *proportions* of respondents choosing a certain grade on the attitude scale. In view of this interest, the *multinomial distribution* is a natural stochastic model of response behaviour. The multinomial framework in terms of the multinomial logit is often used in explanatory modelling of choices or preferences, see Powers and Xie (1999). However, in the analysis of sample surveys based on Likert scales the multinomial model is not very popular, see Maravelakis et al. (2003) for instance.

An s-dimensional random vector $\boldsymbol{Y}$ has multinomial distribution $MULT(k, p_1, \ldots, p_s)$, briefly $\boldsymbol{Y} \sim MULT(k, p_1, \ldots, p_s)$, if the probability density function is given by

$$P(\boldsymbol{Y}=\boldsymbol{y}) = \frac{k!}{y_1! \ldots y_s!} P_1^{y_1} \cdots P_s^{y_s} \quad \text{for } y_1, \ldots, y_s \in \mathbb{N}_0, \quad y_1 + \cdots + y_s = k, \tag{4}$$

with parameters $k \in \mathbb{N}$ and $p_1, \ldots, p_s \geq 0$, $p_1 + \cdots + p_s = 1$. Formula (4) gives the probability of choosing yl times the category l in k experiments where the *choice probability* of choosing category l in one experiment is p_l.

The above interpretation suggests to assume

$$\begin{aligned} \boldsymbol{X}_{ij} &\sim MULT(1, p_{ij1}, \ldots, p_{ijr}), \\ \boldsymbol{Z}_{ij} &\sim MULT(1, q_{i,j,-(r-1)}, \ldots, q_{i,j,r-1}) \end{aligned} \tag{5}$$

for the response vector $\boldsymbol{X}_{ij}$ of respondent i on item j and for the gap vector $\boldsymbol{Z}_{ij}$ between perception and expectation of respondent i on item j. The choice probability p_{ijl} respectively q_{ijl} quantifies respondent i's average inclination to exhibit attitude l towards statement j.

Analogously, we assume multinomial distributions for dimension responses or dimension gaps and for total responses and for total gaps, i.e.,

$$\boldsymbol{X}_i^{(\rho)} \sim MULT(m_{\rho+1} - m_\rho, p_{i1}^{(\rho)}, \ldots, p_{ir}^{(\rho)}), \quad \rho = 1, \ldots, \nu, \tag{6}$$

$$\boldsymbol{X}_i \sim MULT(M, p_{i1}, \ldots, p_{ir}), \tag{7}$$

$$\boldsymbol{Z}_i^{(\rho)} \sim MULT(m_{\rho+1} - m_\rho, q_{i,-(r-1)}^{(\rho)}, \ldots, q_{i,r-1}^{(\rho)}), \quad \rho = 1, \ldots, \nu, \tag{8}$$

$$\boldsymbol{Z}_i \sim MULT(M, q_{i,-(r-1)}, \ldots, q_{i,r-1}). \tag{9}$$

Above, choice probabilities are indexed in the respondent i. In most cases, choice probabilities are identical for groups of individuals or for the entire population. We distinguish two assumptions:

(A1) *Homogeneous population and sample*: The choice probabilities are invariant for all respondents from a given population, and in partic-

ular for all respondents $1, \ldots, n$. The respondent index i in formulae (5) through (9) can be omitted.

(A2) *Clustered population and sample*: The choice probabilities are invariant in mutually exclusive subgroups (clusters) of the population, and in particular in subgroups $C_1, \ldots, C_Q, C_1 \cup \cdots \cup C_Q = \{1, \ldots, n\}$ of respondents in the sample. Choice probabilities corresponding to different clusters are different.

8. Estimation and Confidence Intervals for Choice Probabilities in a Homogeneous Population

We consider a homogeneous sample according to assumption (A1) where the choice probabilities are identical for all respondents. Responses of different individuals can be assumed to be independent. The respective vectors of choice probabilities, see formulae (5) through (9), are estimated by the vectors

$$\overline{\boldsymbol{X}}_j = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_{ij}, \quad \overline{\boldsymbol{X}}^{(\rho)} = \frac{1}{n(m_{\rho+1}-m_\rho)}\sum_{i=1}^{n} \boldsymbol{X}_i^{(\rho)}, \qquad \overline{\boldsymbol{X}} = \frac{1}{nM}\sum_{i=1}^{n} \boldsymbol{X}_i, \tag{10}$$

$$\overline{\boldsymbol{Z}}_j = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Z}_{ij}, \qquad \overline{\boldsymbol{Z}}^{(\rho)} = \frac{1}{n(m_{\rho+1}-m_\rho)}\sum_{i=1}^{n} \boldsymbol{Z}_i^{(\rho)}, \qquad \overline{\boldsymbol{Z}} = \frac{1}{nM}\sum_{i=1}^{n} \boldsymbol{Z}_i, \tag{11}$$

of empirical survey averages (empirical proportions). The components of the survey averages are uniformly minimum variance unbiased (UMVU) estimators for the corresponding choice probabilities, for instance Lehmann (1983) for the background in estimation theory.

The accuracy of a parameter estimate is best supported by providing a *confidence region* at a sufficiently high confidence level γ. In case of a vector parameter, *simultaneous confidence intervals* are particularly useful since the accuracy of each component estimate can be evaluated separately. For the vector parameter $\boldsymbol{p} = (p_1, \ldots, p_s)$ in an s-dimensional multinomial distribution, the s simultaneous confidence intervals $I_1 = (LCL_1; UCL_1), \ldots, I_s = (LCL_s; UCL_s)$ at a nominal confidence level γ have to satisfy

$$\mathrm{P}_{\boldsymbol{p}}(p_1 \in I_1, \ldots, p_s \in I_s) \overset{!}{\geq} \gamma \quad \text{for all values of } \boldsymbol{p} = (p_1, \ldots, p_s). \tag{12}$$

We consider an i.i.d. sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ of size N from a multinomial distribution $MULT\ (1, p_1, \ldots, p_s)$. The confidence intervals for p_l are centered around the UMVU estimator for p_l, i.e., the sample average (empirical proportion) $\overline{Y}_{\cdot l} = \frac{1}{N}\sum_{d=1}^{N} Y_{dl}$ with respect to component l.

Asymptotic simultaneous confidence intervals for the choice probabilities have been constructed by Quesenberry and Hurst (1964) and Goodman

(1965). Both Quesenberry and Hurst (1964) and Goodman (1965) suggest for p_l an interval with lower/upper limit

$$LCL_l/UCL_l = \frac{z + 2N\overline{Y}\cdot_l - / + \sqrt{z\left(z + 4N\overline{Y}_{\cdot l}(1-\overline{Y}_{\cdot l})\right)}}{2(N+z)}, \tag{13}$$

where z is a suitable $100\beta\%$ quantile of χ^2-distribution. Quesenberry and Hurst (1964) show that the requirement (12) is satisfied asymptotically for large N, i.e., the nominal confidence level is guaranteed asymptotically, by choosing the $100\gamma\%$ quantile $z = z_{s-1}(\gamma)$ of the χ^2- distribution. $\chi^2(s-1)$ with degree of freedom $s-1$. By a theorem of Wilks (1962) based on the Bonferroni inequality, Goodman (1965) shows that (12) can be satisfied asymptotically with narrower intervals by choosing the $100(1-(1-\gamma)/s)\%$ quantile $z = z_1(1-(1-\gamma)/s)$ of the χ^2-distribution $\chi^2(1)$ with degree of freedom 1.

Further simultaneous confidence intervals for the choice probabilities are discussed in literature. Bailey's (1980) approach based on a normalizing transformation of the estimators produces shorter intervals than Goodman's for small values of the estimators. The method of Sison and Glaz (1995) is quite involved and cannot be used without software support. Fitzpatrick and Scott (1987) suggest the simple intervals

$$I_l = \left(\overline{Y}_{\cdot l} - \frac{c(\gamma)}{\sqrt{N}}; \overline{Y}_{\cdot l} + \frac{c(\gamma)}{\sqrt{N}}\right), \tag{14}$$

where $c(0.90) = 1.00, c(0.95) = 1.13, c(0.99) = 1.40$.

May and Johnson (1997) compare the approaches of Quesenberry and Hurst (1964) Goodman (1965), Fitzpatrick and Scott (1987), Sison and Glaz (1995) and some more in a simulation study. Fitzpatrick and Scott (1987) intervals are recommended for quick and rough calculations. Quesenberry and Hurst (1964) intervals are wide and conservative, agreeing with formula (12) generally also for relatively small sample sizes. The narrower Goodman (1965) intervals should be used only if the dimension s of the multinomial distribution is small or if the expected occupancy in each degree $l = 1, \ldots, s$ is at least 10.

To apply formulae (13) and (14) for constructing simultaneous confidence intervals for choice probabilities in the setting introduced by Section 7, the estimators $\overline{Y}_{\cdot l}$ and the sample size N have to be identified appropriately with parameters from formulae (10) and (11). The identifications can be found in Table 1.

Table I. Interpretation of choice probabilities in formulae (5) through (9) for a homogeneous sample and corresponding estimators and sample sizes N to be used in confidence interval formulae (13) and (14)

Number s of Likert Degrees	Parameter	Estimator	Sample Size
$s=r$	p_{jl}, probability that a respondent chooses degree l with respect to statement j	$\overline{Y}_{\cdot l} = \overline{X}_{\cdot jl} = \frac{1}{n}\sum_{i=1}^{n} X_{ijl}$, sample average number of respondents choosing degree l with respect to statement j	$N=n$
$s=r$	$p_l^{(\rho)}$, average probability that a respondent chooses degree l with respect to statements in dimension ρ	$\overline{Y}_{\cdot l} = \overline{X}_{\cdot l}^{(\rho)} = \sum_{i=1}^{n} X_{il}^{(\rho)} / n(m_{\rho+1}-m_{\rho})$, sample average number of respondents choosing degree l with respect to statements in dimension ρ	$N=n(m_{\rho+1}-m_{\rho})$
$s=r$	pl, average probability that a respondent chooses degree l with respect to statements on the questionnaire	$\overline{Y}_{\cdot l} = \frac{1}{n}\sum_{i=1}^{n} X_{ijl} = \overline{X}_{\cdot jl}$, sample average number of respondents choosing degree l with respect to statements on the questionnaire	$N=nM$
$s=2r-1$	q_{jl}, probability that a respondent chooses the gap degree l with respect to statement j	$\overline{Y}_{\cdot l} = \overline{Z}_{\cdot jl} = \frac{1}{n}\sum_{i=l}^{n} Z_{ijl}$, sample average number of respondents choosing gap degree l with respect to statement j	$N=n$
$s=2r-1$	$q_l^{(\rho)}$, average probability that a respondent chooses the gap degree l with respect to statements in dimension ρ	$\overline{Y}_{\cdot l} = \overline{Z}_{\cdot l}^{(\rho)} = \sum_{i=1}^{n} Z_{il}^{(\rho)} / n(m_{\rho+1}-m_{\rho})$, sample average number of respondents choosing gap degree l with respect to statements in dimension ρ	$N=n(m_{\rho+1}-m_{\rho})$
$s=2r-1$	ql, average probability that a respondent chooses gap degree l with respect to statements on the questionnaire	$\overline{Y}_{\cdot l} = \overline{Z}_{\cdot l} = \frac{1}{n}\sum_{i=1}^{n} Z_{il}$, sample average number of respondents choosing the gap degree l with respect to statements on the questionnaire	$N=nM$

9. Choice of Sample Size

A reasonable criterion for sample size selection is imposing an upper limit on the width of simultaneous confidence intervals, see Tortora (1978).

As in Section 8, above, we consider an i.i.d. sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ of size N from a multinomial distribution $MULT(1, p_1, \ldots, p_s)$. The length of the intervals by Quesenberry and Hurst (1964) and Goodman (1965), see formula (13), is

$$UCL_l - LCL_l = \frac{\sqrt{z\left(z + 4N\overline{Y}_{\cdot l}(1-\overline{Y}_{\cdot l})\right)}}{N+z}, \tag{15}$$

where z is a suitable quantile of a χ^2-distribution as discussed in Section 8. The interval length depends on the estimator $\overline{Y}_{\cdot l}$ for p_l and attains a maximum at $\overline{Y}_{\cdot l} = 0.5$ with asymptotic length $\sqrt{z/N}$.

Sample size is determined according to the following criterion: The confidence limits should not differ from the estimate by more than a prescribed amount ε, i.e., the length of the confidence interval for each $p_l, l = 1, \ldots, s$ should not exceed 2ε. Hence we obtain

$$N = \left\lceil \frac{z}{4\varepsilon^2} \right\rceil. \tag{16}$$

To identify the parameters $s, p_1, \ldots, p_s$ and N in the setting of Section 7, consider Table 1.The following example 9.1 shows that sample size calculations by formula (16) are very sensitive with respect to the type of confidence intervals. For a conservative assessment of confidence, the technique of Quesenberry and Hurst (1964) should be used.

9.1. EXAMPLE

Consider a SERVQUAL survey based on a 7-grade Likert scale. Interest is in estimating the choice probabilities $q_{j,-6}, \ldots, q_{j,6}$ for gap degrees in each item. Hence $s = 13, N = n$. Estimates should be accurate up to $\mp 10\%$ at 90% confidence.

Under the more conservative confidence intervals of Quesenberry and Hurst (1964) we have to use the quantile $z = z_{12}(0.90) = 18.55$ of the χ^2-distribution $\chi^2(12)$ with degree of freedom 12. Hence from formula (16) we obtain sample size $N = 464$.

Under the looser confidence intervals of Goodman (1965) we have to use the quantile $z = z_1(0.9923) = 7.10$ of the χ^2-distribution $\chi^2(1)$ with degree of freedom 1. Hence from formula (16) we obtain sample size $N = 177$.

10. Testing the Equality of Choice Probabilities in a Homogeneous Population

As in Section 8 we consider the survey as an independent sample from a homogeneous population of respondents. The estimates together with simultaneous confidence intervals for the choice probabilities appearing in formulae (10) and (11) provide a good insight into the attitudes of individuals in the population. Descriptive statistics like histograms, bar charts, pie charts, Pareto charts should be used for presentation. In the present and the subsequent Section 11 we present methods of statistical inference for the comparison of choice probabilities: Tests of significance for the equality of choice probabilities, see below, and tests of significance for rank orders of choice probabilities, see Section 11.

The tests are based on an i.i.d. sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ of size n from a multinomial distribution $MULT(M, p_1, \ldots, p_s)$. The sum $\boldsymbol{Y} = \boldsymbol{Y}_1 + \cdots + \boldsymbol{Y}_n$ is a sufficient statistic for the probabilities $p_1, \ldots, p_s$, see Lehmann (1983). Hence testing can be based on $\boldsymbol{Y}$ which has multinomial distribution $MULT(nM, p_1, \ldots, p_s)$, see the reproduction theorem for multinomial distribution in Appendix A.1. The interpretation of the quantities $s, p_1, \ldots, p_s, \boldsymbol{Y}$, of the general scheme with the appropriate quantities in the special settings of Section 7 is obvious from Table 2.

We want to find out whether respondents prefer certain attitudes or whether all among a given number of pairwise different attitudes $i_1, \ldots, i_t$ have the same probability to be chosen. To this end we consider the equality hypothesis

$$H : p_{i_1} = \cdots = p_{i_t}. \tag{17}$$

Similar to Fisher's well-known test for comparing binomial probabilities, a reasonable test of (17) compares the results $Y_{i_1}, \ldots, Y_{i_t}$ relative to the total number $y = Y_{i_1} + \cdots + Y_{i_t}$ of observed choices in degrees $i_1, \ldots, i_t$. According to assertion (c) of Theorem A.2.1 in Appendix A.2, the conditional distribution of $Y_{i_1}, \ldots, Y_{i_t}$ under the condition $y = Y_{i_1} + \cdots + Y_{i_t}$ is the multinomial distribution $MULT(y, \pi_1, \ldots, \pi_t)$ where $\pi_l = p_{i_l}/(p_{i_1} + \cdots + p_{i_t})$. The equality hypothesis H is equivalent to $H' : \pi_1 = \cdots = \pi_t = 1/t$.

A reasonable test statistic should be a measure of variation of the responses $Y_{i_1}, \ldots, Y_{i_t}$. Light and Margolin (1971) suggest the variation measure

$$V = V(Y_{i_1}, \ldots, Y_{i_t}) = \frac{y}{2} - \frac{1}{2y} \sum_{l=1}^{t} Y_{i_l} \tag{18}$$

which is derived from Gini's (1955) well-known variation measure for categorical data. The equality hypothesis $H' : \pi_1 = \cdots = \pi_t = 1/t$ is rejected if the

Table II. Assignment of choice probabilities and statistics from formulae (5) through (9) to the general testing schemes of Sections 10 and 11

Number s of Likert Degrees	Compared Probabilities	Test Statistic
$s=r$	$p_{j_1},\dots,p_{jr}$, probabilities that a respondent chooses degree $1,\dots,r$ with respect to statement j	$\boldsymbol{Y}=\sum_{i=1}^{n}\boldsymbol{X}_{ij}$, vector of total numbers of respondents choosing degrees $1,\dots,r$ with respect to statement j
$s=r$	$p_1^{(\rho)},\dots,p_r^{(\rho)}$, average probabilities that a respondent chooses degrees $1,\dots,r$ with respect to statements in dimension ρ	$\boldsymbol{Y}=\sum_{i=1}^{n}\boldsymbol{X}_i^{(\rho)}$, vector of total numbers of respondents choosing degree $1,\dots,r$ with respect to statements in dimension ρ
$s=r$	$p_1,\dots,p_r$, average probabilities that a respondent chooses degrees $1,\dots,r$ with respect to statements on the questionnaire	$\boldsymbol{Y}=\sum_{i=1}^{n}\boldsymbol{X}_i$, vector of total numbers of respondents choosing degrees $1,\dots,r$ with respect to statements on the questionnaire
$s=2r-1$	$q_{j_1},\dots,q_{jr}$, probabilities that a respondent chooses gap degree $1,\dots,r$ with respect to statement j	$\boldsymbol{Y}=\sum_{i=1}^{n}Z_{i_j}$, vector of total numbers of respondents choosing gap degrees $-(r-1),\dots,r-1$ with respect to statement j
$s=2r-1$	$q_1^{(\rho)},\dots,q_r^{(\rho)}$, average probabilities that a respondent chooses gap degrees $-(r-1),\dots,r-1$ with respect to statements in dimension ρ	$\boldsymbol{Y}=\sum_{i=1}^{n}\boldsymbol{Z}_i^{(\rho)}$, vector of total numbers of respondents choosing gap degree $-(r-1),\dots,r-1$ with respect to statements in dimension ρ
$s=2r-1$	$q_1,\dots,q_r$, average probabilities that a respondent chooses gap degrees $-(r-1),\dots,r-1$ with respect to statements on the questionnaire	$\boldsymbol{Y}=\sum_{i=1}^{n}\boldsymbol{Z}_i$, vector of total numbers of respondents choosing gap degrees $-(r-1),\dots,r-1$ with respect to statements on the questionnaire

sample variation is too large, i.e., if $V \geq c$. The p-value of this test under sample realizations $Y_{i_1}=y_1,\dots,Y_{i_t}=y_t$, $y_1+\cdots+y_t=y$, is given by

$$\frac{1}{t^y} \sum_{\substack{x_1,\ldots,x_t \geq 0 \\ x_1+\cdots+x_t=y \\ V(x_1,\ldots,x_t) \geq V(y_1,\ldots,y_t)}} \frac{y!}{x_1! \cdots .x_t!}. \tag{19}$$

Further research on simplifying approximations of expression (19) is necessary.

11. Ranking of Choice Probabilities in a Homogeneous Population

If choice probabilities are apparently not identical, major interest is in a hypothesis on the *rank order* of the choice probabilities. Such a hypothesis may be formulated by the rank order of the empirical proportions in the sample as expressed by a Pareto chart. Methods for confirming this hypothesis are required. Simultaneous confidence regions are no help for this purpose. In the sequel, we develop a method of testing the ranking hypothesis by multiple comparisons as used in comparative treatment analysis, see Hsu (1996) for instance.

Consider the comparison of choice probabilities $p_1, \ldots, p_s, \sum p_l = 1$. We wish to confirm the composite rank order hypothesis

$$K : p_{i_1} > p_{j_1}, \ldots, p_{i_t} > p_{j_t} \tag{20}$$

where $i_m \neq j_m$. To this end, we consider the negation $H = \neg K$ as the null hypothesis. K is confirmed by a significance test if H can be rejected. H is the disjunction $H = H_1 \cup \cdots \cup H_t$ of the null hypotheses in the t partial testing problems $H_1 : p_{i_1} \leq p_{j_1}$ against $K_1 : p_{i_1} > p_{j_1}$, $H_2 : p_{i_2} \leq p_{j_2}$ against $K_2 : p_{i_2} > p_{j_2}$, and so on until $H_t : p_{i_t} \leq p_{j_t}$ against $K_t : p_{i_t} > p_{j_t}$. We test H against K by successively testing H_m against K_m. H is rejected in favour of K if each H_m is rejected in favour of K_m.

In the following Section 11.1 we describe the design of partial tests under a prescribed level of significance. The subsequent Section 11.2 considers the test of the composite hypothesis H against K.

11.1. DESIGN OF PARTIAL TESTS FOR RANK ORDER

Consider the partial problem

$$H_m : p_{i_m} \leq p_{j_m} \quad \text{against} \quad K_m : p_{i_m} > p_{j_m}, \tag{21}$$

$m \in \{1, \ldots, t\}$, where $p_{i_m} + p_{j_m} > 0$. An equivalent formulation of problem (21) is

$$H'_m : \pi_m \leq 0.5 \quad \text{against} \quad K'_m : \pi_m > 0.5, \text{ where } \pi_m = \frac{p_{i_m}}{p_{i_m} + p_{j_m}}. \tag{22}$$

Similar to Fisher's well-known test for comparing binomial probabilities, a reasonable test of (21) compares the results Y_{i_m}, Y_{j_m} relative to the total number $y = Y_{i_m} + Y_{j_m}$ of observed choices in degrees i_m and j_m. H'_m is rejected in favour of K'_m if $Y_{i_m} > c$, i.e., if Y_{i_m} is a too large amount of y.

We have to determine the critical value $c = c_{\alpha 0} \in \{0, \ldots, y\}$ under a prescribed level of significance $0 < \alpha_0 < 1$. Clearly in case of $0 = y = Y_{i_m} + Y_{j_m} H'_m$ cannot be rejected under any level α_0, so formally $c = c_{\alpha 0} = 0 = y$ in case of $y = 0$. Consider the case $0 < y = Y_{i_m} + Y_{j_m}$. According to assertion (d) of Theorem A.2.1 in Appendix A.2, the conditional distribution of Y_{i_m} under the condition $Y_{i_m} + Y_{j_m} = y$ is the binomial distribution $Bi(y, \pi_1)$. Hence (22) can be tested by the well-known test for binomial probabilities. The critical value $c = c_{\alpha_0}$ is determined as the minimum integer $c \in \{0, \ldots, y\}$ satisfying the inequalities

$$1 - L_{y,c}(0.5) = 0.5^y \sum_{l=c+1}^{y} \binom{y}{l} \overset{!}{=} \alpha_0 \overset{!}{<} 0.5^y \sum_{l=c}^{y} \binom{y}{l} = 1 - L_{y,c-1}(0.5), \quad (23)$$

where $L_{y,c}(0.5)$ is the distribution function of the binomial distribution $Bi(y, 0.5)$. These values are available in tables and are provided by any modern statistical software package.

11.2. DESIGN OF THE TEST FOR THE COMPOSITE RANK ORDER HYPOTHESIS

Let the significance level $0 < \alpha < 1$ be prescribed for a test of the composite hypothesis H against K. This level can be guaranteed by prescribing $\alpha_0 = \frac{\alpha}{t}$ for each of the t partial problems H_m against K_m. Let R_m denote the event that the m-th partial test, $m \in \{1, \ldots, t\}$, rejects H_m in favour of K_m, and let R denote the event that H is rejected in favour of K. Then $\mathrm{P}(R_m | H_m) \leq \frac{\alpha}{t}$ by the design of the partial test and hence by the well-known Bonferroni inequality

$$\mathrm{P}(R|H) = \mathrm{P}(R_1 \cup \cdots \cup R_t | H) \leq \sum_{m=1}^{t} \mathrm{P}(R_m|H) = \sum_{m=1}^{t} \mathrm{P}(R_m|H_m) \leq \alpha. \quad (24)$$

11.3. THE p-VALUE OF THE TESTS FOR RANK ORDER

Under small $y = Y_{i_m} + Y_{j_m}$ it may be impossible to satisfy (23) with $\alpha_0 = \frac{\alpha}{t}$, i.e., to guarantee a partial test of prescribed significance level $\alpha_0 = \frac{\alpha}{t}$. Hence it may be more adequate to consider the p-value for rejecting $H = H_1 \cup \cdots \cup H_t$ under sample realizations $Y_1 = y_1, \ldots, Y_s = y_s$. The p-value of the partial test of H_m against K_m with the rejection region of type $Y_{i_m} > c$ as described in Section 11.1 is

$$1 - L_{y_{i_m}+y_{j_m},\,y_{i_m}-1}(0.5) = 0.5^{y_{i_m}+y_{j_m}} \sum_{l=y_{i_m}}^{y_{i_m}+y_{j_m}} \binom{y_{i_m}+y_{j_m}}{l}. \tag{25}$$

By the Bonferroni inequality, an upper bound for the p-value of the test of the composite hypothesis H is

$$\sum_{m=1}^{t} \left(1 - L_{y_{i_m}+y_{j_m},\,y_{i_m}-1}(0.5)\right) = \sum_{m=1}^{t} 0.5^{y_{i_m}+y_{j_m}} \sum_{l=y_{i_m}}^{y_{i_m}+y_{j_m}} \binom{y_{i_m}+y_{j_m}}{l}. \tag{26}$$

11.4. COMPARISON OF CHOICE PROBABILITIES BY SIMULTANEOUS CONFIDENCE INTERVALS

Choice probabilities p_{i_m}, p_{j_m} may also be compared by simultaneous confidence intervals for the difference $p_{i_m} - p_{j_m}$ or the ratio p_{i_m}/p_{j_m}. Such simultaneous confidence intervals are provided by Goodman (1965).

12. In-Questionnaire Association

An important topic in survey data analysis is in-questionnaire *association* or *dependence*, i.e., association or dependence between responses on certain items or dimensions or the entire questionnaire. Are responses tending to be similar or do they diverge? Particularly important is item-to-total association, i.e., the relationship between responses on an individual item and responses on the entire questionnaire. A detailed analysis goes beyond the scope of the present paper. Important ordinal measures of association are Spearman's well-known rank correlation coefficient, the gamma statistic by Goodman and Kruskal (1954), Kendall's (1945) tau-b, Somer's (1962) d statistic. These measures should be investigated under the multinomial response model to develop efficient estimation and testing procedures.

13. Comparison of Vectors of Choice Probabilities

Sections 10 and 11 compare the choice probabilities in a given multinomial parameter vector. A further important topic is the comparison of entire parameter vectors. Two topics are interesting:

- Comparison of parameter vectors corresponding to questionnaire items or questionnaire dimensions or the entire questionnaire. This topic is related to in-questionnaire association, see Section 12, above.

- Comparison of parameter vectors corresponding to different respondents. Here, we question the assumption of a homogeneous population of respondents, made throughout in Sections 8–11, see assumption (A1) of Section 7.

Methods for comparing multinomial parameter vectors are provided by literature. A simple approach is to use Pearson's χ^2-test, see Clason and Dormody (1994). Light and Margolin (1971) and Margolin and Light (1974) present an ANOVA scheme which tests the hypothesis of equality of m multinomial probability vectors. A survey of measures of agreement between respondents is provided by Adejumo et al. (2004).

In practice, populations of respondents are often inhomogeneous, i.e., the hypothesis of equality of the n multinomial probability vectors of n respondents will often be rejected. Groups (clusters) of customers, consumers, patients, social classes, age brackets, genders, may differ substantially in their attitudes. Some distinguishing factors may be quite obvious and known beforehand so that stratified surveys may be conducted. In other cases, however, stratified sampling is impossible. Major obstacles are: (1) Unknown factors. (2) Known factors, but unknown distribution of factors in the population. (3) Practical unfeasibility of stratification, e.g., due to economic restrictions. In such cases, groups of substantially different respondents have to be identified from survey data.

Standard model free clustering algorithms based on distance measures can contribute to clustering in attitude surveys. However, model based clustering is generally more efficient. By means of assumption (A2), the multinomial model of Section 7 can describe clusters as groups sharing the same vector of choice probabilities. Recently, advances in clustering multinomial samples have been made in genetics, see Medvedovic et al. (2000). This type of probabilistic clustering has the potential to be more efficient than model free techniques. The description of multinomial clustering goes beyond the scope of the present paper. However, in view of the potential of such methods business statistics should reflect and adopt such approaches.

14. Conclusion

The problem of attitude measuring suggests an ordinal interpretation of the Likert scale. The above Sections 7–13 show that plenty of proper ordinal methods exist for the analyis of data measured in Likert scales. However, such methods are not as easily available in textbooks and statistical packages as cardinal statistics. Some new methods were introduced in Sections 8, 10 and 11. Further work should concentrate on developing con-

venient and customized versions of ordinal statistics and on propagating these among researchers and practitioners.

Appendix

A. Properties of Multinomial Random Vectors

A.1. REPRODUCTION THEOREM

Sums of independent multinomial random vectors with identical vectors of choice probabilities follow a multinomial distribution:

THEOREM A.1.1. Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ be s-dimensional independent random vectors, each with multinomial distribution $MULT(k_i, p_1, \ldots, p_s)$.

Then the sum $\sum_{i=1}^{n} \boldsymbol{Y}_i$ has multinomial distribution $MULT(\sum k_i, p_1, \ldots, p_s)$. For a proof of Theorem A.1.1 see Wilks (1962).

A.2. MARGINAL AND CONDITIONAL DISTRIBUTIONS

The following Theorem A.2.1 gives marginal and conditional distributions in a multinomial vector.

THEOREM A.2.1. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_r)$ be an r-dimensional random vector with multinomial distribution $MULT(N, p_1, \ldots, p_r)$. Let $1 \le i_1 < \cdots < i_m \le r$. Then we have the following results:

(a) For $y_{i_1}, \ldots, y_{i_m} \ge 0$, $y_{i_1} + \cdots + y_{i_m} = y \le N$ we have

$$\begin{aligned} &\mathrm{P}(Y_{i_1} = y_{i_1}, \ldots, Y_{i_m} = y_{i_m}) \\ &= \frac{N!}{y_{i_1}! \cdots y_{i_m}! (N-y)!} p_{i_1}^{y_{i_1}} \cdots p_{i_m}^{y_{i_1}} (1 - p_{i_1} - \cdots - p_{i_m})^{N-y}. \end{aligned} \tag{A.1}$$

(b) The sum $Y_{i_1} + \cdots + Y_{i_m}$ follows the binomial distribution $Bi(N, p_{i_1} + \cdots + p_{i_m})$.

(c) Let $1 \le y \le N$. Then the conditional distribution of the m-dimensional random vector $(Y_{i_1}, \ldots, Y_{i_m})$ under the condition $Y_{i_1} + \cdots + Y_{i_m} = y$ is the multinomial distribution $MULT(y, \pi_1, \ldots, \pi_m)$ where

$$\pi_l = \frac{p_{i_l}}{p_{i_1} + \cdots + p_{i_m}} \quad \text{for } l = 1, \ldots, m. \tag{A.2}$$

(d) Let $1 \le y \le N$, $m = 2$. Then the conditional distribution of the univariate random variable Y_{i_1} under the condition $Y_{i_1} + Y_{i_2} = y$ is the binomial distribution $Bi(y, p_{i_1}/p_{i_1} + p_{i_2})$.

The proof of Theorem A.2.1 makes use of the well-known theorem on multinomial expansions:

$$(p_{j_1}+\cdots+p_{j_k})^z=\sum_{\substack{y_{j_1},\ldots,y_{j_k}\geq 0\\ y_{j_1},\ldots,y_{j_k}=z}}\frac{z!}{y_{j_1}!\cdots y_{j_k}!}p_{j_1}^{y_{j_1}}\cdots p_{j_k}^{y_{j_k}}. \tag{A.3}$$

Hence formula (A.1) in assertion (a) of theorem A.2.1 is obtained by calculating

$$\mathrm{P}(Y_{i_1}=y_{i_1},\ldots,Y_{i_m}=y_{i_m})=\frac{N!}{y_{i_1}!\cdots y_{i_m}!(N-y)!}p_{i_1}^{y_{i_1}}\cdots p_{i_m}^{y_{i_m}}$$
$$\sum_{\substack{y_{j_1},\ldots,y_{j_{N-m}}\geq 0\\ y_{j_1},\ldots,y_{j_{N-m}}=N-y}}\frac{(N-y)!}{y_{j_1}!\cdots y_{j_{N-m}}!}p_{j_1}^{y_{j_1}}\cdots p_{j_{N-m}}^{y_{j_{N-m}}}.$$

For the proof of assertion (b) let $0\leq y\leq N$. Using formula (A.3) on multinomial expansion we obtain

$$\begin{aligned}\mathrm{P}(Y_{i_1}+\cdots+Y_{i_m}=y)&=\sum_{\substack{y_{i_1},\ldots,y_{i_m}\geq 0\\ y_{i_1},\ldots,y_{i_m}=y}}\frac{N!}{y_{i_1}!\cdots y_{i_m}!(N-y)!}\\ &\quad p_{i_1}^{y_{i_1}}\cdots p_{i_m}^{y_{i_m}}(1-p_{i_1}-\cdots-p_{i_m})^{N-y}\\ &=\binom{N}{y}(1-p_{i_1}-\cdots-p_{i_m})^{N-y}\\ &\quad\sum_{\substack{y_{i_1},\ldots,y_{i_m}\geq 0\\ y_{i_1},\ldots,y_{i_m}=y}}\frac{y!}{y_{i_1}!\cdots y_{i_m}!}p_{i_1}^{y_{i_1}}\cdots p_{i_m}^{y_{i_m}}\\ &=\binom{N}{y}(p_{i_1}+\cdots+p_{i_m})^y(1-p_{i_1}-\cdots-p_{i_m})^{N-y}.\end{aligned}$$

This proves assertion (b).

Investigating the conditional distribution considered in assertion (c) we obtain for $y_{i_1}, \ldots, y_{i_m} \geq 0$, $y_{i_1} + \cdots + y_{i_m} = y$:

$$\begin{aligned}
&\mathrm{P}(Y_{i_1} = y_{i_1}, \ldots, Y_{i_m} = y_{i_m} | Y_{i_1} + \cdots + Y_{i_m} = y) \\
&\quad = \frac{\mathrm{P}(Y_{i_1} = y_{i_1}, \ldots, Y_{i_m} = y_{i_m})}{\mathrm{P}(Y_{i_1} + \cdots + Y_{i_m} = y)} \\
&\quad = \frac{\frac{N!}{y_{i_1}! \cdots y_{i_m}! (N-y)!} p_{i_1}^{y_{i_1}} \cdots p_{i_m}^{y_{i_1}} (1 - p_{i_1} - \cdots - p_{i_m})^{N-y}}{\binom{N}{y} (p_{i_1} + \cdots + p_{i_m})^{y} (1 - p_{i_1} - \cdots - p_{i_m})^{N-y}} \\
&\quad = \frac{y!}{y_{i_1}! \cdots y_{i_m}!} \left(\frac{p_{i_1}}{p_{i_1} + \cdots + p_{i_m}} \right)^{y_{i_1}} \cdots \left(\frac{p_{i_m}}{p_{i_1} + \cdots + p_{i_m}} \right)^{y_{i_m}}.
\end{aligned}$$

Acknowledgements

This paper is supported by funding from the "Growth" program of the European Community and was prepared in collaboration by member organizations of the Thematic Network-Pro- ENBIS-EC contract number G6RT-CT-2001-05059.

References

Adams, E., Fagot, R. F. & Robinson, R. E. (1965). A theory of appropriate statistics. *Pschometrika* 30 (2): 99–127.

Adejumo, A. O., Heumann, C. & Toutenburg, H. (2004). A review of agreement measure as a subset of association measure between raters. SFB386-Discussion Paper 385, Ludwig-Maximilians-Universität, München, Germany.

Andrews, F. M., Klem, L., Davidson, T. N., O'Malley, P. M. & Rodgers, W. L. (1981). *A Guide for Selecting Statistical Techniques for Analyzing Social Science Data*. Ann Arbor: Institute for Social Research, University of Michigan.

Asubonteng, A., McClearly, K. J. & Swan, J. E. (1996). SERVQUAL revisited: A critical review of service quality. *J. Service. Market*. 10 (6): 62–81.

Bailey, B. J. R. (1980). Large sample simultaneous confidence intervals for the multinomial probabilities based on transformations of the cell frequencies. *Technometrics* 22 (4): 583–589.

Baker, B. O., Hardyck, C. D. & Petrinovich, L. F. (1986). Weak measurements versus strong statistics: An empirical critique of S. S. Stevens' proscriptions on statistics. *Education. Psychol. Measure.* 26: 291–309.

Clason, D. L. & Dormody, T. J. (1994). Analyzing data measured by individual likert-type items. J. Agric. Education 35 (4): 31–35.

Diener, E.(1984). Subjective well-being. *Psychol. Bull.* 95: 542–575.

Diener, E., Emmons, R. A., Larsen, R. J. & Griffin, S. (1985). The satisfaction with life scale. *J. Personal. Assess*. 49 (1): 71–75.

Diener, E., Suh, E. M., Lucas, R. E. & Smith, H. L. (1999). Subjective well-being: Three decades of progress. *Psychol. Bull.* 125 (2): 276–302.

Ferrer-I-Carbonell, A. & van Praag, B. M. S. (2003). Income satisfaction inequality and its Causes. *J. Econ. Inequality* 1: 107–127.

Fitzpatrick, S. & Scott, A. (1987). Quick simultaneous confidence intervals for multinomial proportions. *J. Am. Stat. Assoc*. 82: 399.

Gini, C. W. (1955). *Variabilità e Concentrazione*. Vol. 1: *Memorie di metodologia statistica*.

Georgescu-Roegen, N. (1968). Utility. in *International Encyclopedia of Social Sciences*, Vol. 16, New York: MacMillan Co. & The Free Press, pp. 236–267.

Goodman, L. A. (1965). On simultaneous confidence intervals for multinomial proportions. *Technometrics* 7 (2): 247–254.

Goodman, L. A. & Kruskal, W. H. (1954). Measuring of association for cross classifications. *J. Am. Stat. Assoc.* 49: 732–768.

Hart, M. C. (1996). Improving the discrimination of SERVQUAL by using magnitude scaling. In London: G. K. Kanji (ed.) Chapman & Hall *Total Quality Management in Action.*

Hart, M. C. (1999). The quantification of patient satisfaction. In *Managing Quality: Strategic Issues in Health Care Management*. H. T. O. Davies, M. Tavakoli, M. Malek, and A. Neilson Ashgate (eds.), Aldershot.

Hsu, J. C. (1996). *Mutiple Comparisons. Theory and Methods*. Boca Raton, London: Chapman & Hall

Kendall, M. G. (1945). The treatment of ties in ranK problems. *Biometrika*, 33: 239–251.

Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. New York: John Wiley & Sons.

Lehmann, E. L. (1983). *Theory of Point Estimation*. New York: John Wiley & Sons.

Light, J. & Margolin, B. H. (1971). An analysis of variance for categorical data. *J. Am. Stat. Assoc*. 66 (335): 534–544.

Likert, R. (1932). A technique for the measurement of attitudes. *J. Social. Psychol.* 5: 228–238.

Lodge, M. (1981). Magnitude scaling. *Sage University Paper Series on Quantitative Applications in the Social Sciences*, 07-025, Beverly Hills and London: Sage Publications

Lord, F. M. (1953). On the statistical treatment of football numbers. *Am. Psychol*. 8: 750–751.

Lucas, R. E., Clark, A. E., Georgellis, Y. & Diener, E. (2003). Reexamining adaptation and the set point model of happiness: Reactions to changes in marital status. *J. Personal. Social Psychol*. 84 (3): 527–539.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychol. Rev.* 66: 81–95.

Luce, R. D., Krantz, D. H., Suppes, P. & Tversky, A. (1990). *Foundations of Measurement*. Vol. III. New York: Academic Press.

Maravelakis, P. E., Perakis, M., Psarakis, S. & Panaretos, J. (2003). The use of indices in surveys. *Qual. Quant*. 37: 1–19.

Margolin, B. H. & Light, J. (1974). An analysis of variance for categorical data, II. *J. Am. Stat. Assoc*. 69 (347): 755–764.

May, W. L. & Johnson, W. D. (1997). Properties of simultaneous confidence intervals for multinomial proportions. *Commun. Stat. Simulat. Comput.* 26 (2): 495–518.

Medvedovic, M., Succop, P., Shukla, R. & Dixon, K. (2000). Clustering mutational spectra via classification Likelihood and Markov chain Monte Carlo algorithms. *J. Agric. Biol. Environ. Stat*. 6 (1): 19–37.

Nolte, E. & McKee, M. (2004). Changing health inequalities in east and west Germany since unification. *Social Sci. Med*. 58 (1): 119–136.

Parasuraman, A., Berry, L. L. & Zeithaml, V. A. (1991). Refinement and assessment of the SERVQUAL. *J. Retail.* 67(4): 420–449.

Parasuraman, A., Zeithaml, V. A. & Berry, L. L. (1985). A conceptual model for service quality and its implication for future research. *J. Market*. 41–50.

Parasuraman, A., Zeithaml, V. A. & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *J. Retail*. 64 (1): 12–40.

Powers, D. & Xie, Y. (1999). *Statistical Methods for Categorical Data Analysis*. Academic Press, Inc.

Quesenberry, C. P. & Hurst, D. C. (1964). Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics* 6 (2): 191–195.

Ronellenfitsch, U. & Razum, R. (2004). Deteriorating health satisfaction among immigrants from eastern Europe to Germany. *Int. J. Equity Health* 3 (1):4

Savage, I. R. (1957). Nonparametric Statistics. *J. Am. Stat. Assoc*. 52: 331–334.

Sen, A. (1999). The possibility of social choice. *Am. Econ. Rev*. 89: 349–378.

Sison, C. P. & Glaz, J. (1995). Simultaneous confidence intervals and sample size determination for multinomial proportions. *J. Am. Stat. Assoc*. 90 (429): 366–369.

Somer, R. H. (1962). A new asymmetric measure of association of ordinal variables. *Am. Sociol. Rev*. 27: 799–811.

Stevens, S. S. (1946). On the theory of scales of measurements. *Science* 103: 677–680.

Stevens, S. S. (1951). Mathematics, measurement and psychphysics. In *Handbook of Experimental Psychology*. S. S. Stevens (ed.), New York: John Wiley & Sons pp. 1–49.

Tortora, R. D. (1978). A note on sample size estimation for multinomial populations. *Am. Stat.* 32 (3): 100–102.

Townsend, J. T. & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychol. Bull*. 96 (2): 394–401.

Tukey, J.W. (1961). Data analysis and behavioral science or learning to bear the quantitative burden by Shunning Badmandments. In *The Collected Works of John W. Tukey*, Vol. III, L. V. Jones (ed.), Belmont: Wadsworth. pp. 391–484.

van Praag, B. M. S. (1999). Ordinal and cardinal utility. *J. Econom*et. 50: 69–89.

Velleman, P. F. & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *Am. Stat*. 47 (1): 65–72.

Watson, D., Clark, L. & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The panas scales. *J. Personal. Social Psychol*. 54 (6): 1063–1070.

Wilks, S. S. (1962). *Mathematical Statistics*. New York: John Wiley & Sons

Wright, D. B. (1997). Football standings and measurement levels. *Statistician* 46 (1): 105–110.