

Letter

Quantifying Information Content in Survey Data by Entropy

Fredrik A. Dahl ^{1,*} and Nina Østerås ²

¹ Helse Sør-Øst Health Services Research Centre, Akershus University Hospital, Mail drawer 95, NO-1474 Lørenskog, Norway

² The National Resource Centre in Rheumatology, Diakonhjemmet hospital, P.O. Box 23 Vinderen, NO-0319 Oslo, Norway; E-Mail: nina.osteras@medisin.uio.no

* Author to whom correspondence should be addressed; E-Mail: Fredrik.dahl@ahus.no.

Received: 4 November 2009; in revised form: 6 January 2010 / Accepted: 26 January 2010 /

Published: 28 January 2010

Abstract: We apply Shannon entropy as a measure of information content in survey data, and define information efficiency as the empirical entropy divided by the maximum attainable entropy. In a case study of the Norwegian Function Assessment Scale, entropy calculations show that the 5-point response version has higher information efficiency than the 4-point version.

Keywords: entropy; information; survey; data; questionnaire

1. Introduction

When we invest the time and effort of researchers and participants in a population survey, we naturally want the collected information to be as valuable as possible, and informally we may express the value as a product of information quality and quantity. Researchers routinely evaluate the quality of the information in survey data using advanced concepts of reliability (absence of random noise) and validity (whether or not we are measuring the right thing). To this end it is standard practice to apply advanced statistics like Cronbach's alpha, Cohen's Kappa, item-to-item correlations, *etc.* [1].

Information *quantity*, however, is often evaluated much more crudely. Usually, it is measured as the number of respondents, the number of questions and the number of response options, at best. We see this as an imbalance, and claim that information quantity should be evaluated according to information theory. Hence, we argue that the Shannon entropy [2] of the response distribution is the natural scale for quantifying information content for the responses to a survey question:

$$-n \sum_{i=1}^k p_i \log_2(p_i)$$

where n is the number of respondents, p_i are the probabilities of the k different response values. The maximum entropy is of course attained when $p_1 = p_2 = \dots = p_k = 1/k$, giving $n \log_2(k)$. We define the *information efficiency* of a question in a questionnaire as the empirical entropy divided by the maximum entropy obtainable for the given question. It measures to what extent the responders use the available options—on a scale from 0 to 1—and is likely to be more intuitively appealing than the entropy number itself.

Based on our literature search, our approach appears to be novel. Wu and Zhang [3] apply an information-theoretic approach to the use of auxiliary information from survey data, including entropy evaluations. However, their purpose is to create statistical estimators with low variance, rather than quantifying information. In formal diagnostic reasoning, the utility of performing a test is sometimes evaluated in terms of reduction of the entropy of the distribution of alternative diagnoses [4]. Along the same lines, Tu *et al.* [5] use entropy to evaluate the informativeness of an HIV screening program. Cox [6] applies entropy computations to questionnaire design, with focus on signal-to-noise relations, which aim at evaluating information quality, rather than quantity.

2. Application to the Norwegian Function Assessment Scale

The Norwegian Function Assessment Scale is a self-administered instrument containing 39 items, which exists in a 4- and a 5-point response version. A randomized comparison of these was performed by Østerås *et al.* [7], with a total 3,325 respondents. We refer to [7] for further information on the surveys. For the separate questions, the entropies divided by the number of respondents varied from 0.232 to 1.186 for the 4-points version, and from 0.406 to 1.580 for the one with 5 options. This gives a range of information efficiencies for the 4-point scale of (0.116, 0.593), and (0.175, 0.680) for the 5-point version. The average information efficiencies were 0.345 and 0.401, respectively. The interpretation of these numbers is that the 4-point version collected 34.5% of the information possible, while 5-point version collected 40.1%.

3. Discussion

To increase information efficiency, one should seek to define response alternatives that will be chosen with approximately equal frequency. Beyond this general advice, information efficiency provides a theoretically sound measure of response spread. A special case of low information efficiency occurs when a large portion of the respondents choose the highest (or lowest), from a list of ordered categories. We recognize this as a so-called ceiling (or floor) effect. A possible countermeasure is to refine the scale near the end where most subjects respond, which was actually done in the 4- to 5-point scale example above. The response entropy can also be increased through selective sampling, where respondents that are more likely to give unusual responses are oversampled, which is common in epidemiology.

Our motivation for introducing an entropy-based measure was a perceived imbalance between researchers' focus on quality and quantity of information in survey data. There is often a trade-off

between these, and although we believe that quantity has in general received too little attention, one should not go overboard by focusing on entropy alone. In particular, one should be aware that random noise in the responses will normally increase the information quantity at the expense of the quality. In this case, quality must of course be given priority.

4. Conclusions

Our conclusion supports that of Østerås *et al.* [7], in that the 5-point version should be preferred. Not only did it collect more information in absolute terms, it did so also according to our information efficiency criterion, which controls for the number of response categories. In this study, entropy-based information efficiency appears to be a useful concept, and we believe this will be the case for most questionnaire surveys.

References

1. DeVellis, R.F. Scale development: theory and applications. *Appl. Soc. Res. Method. Ser.* **2003**, *26*, 27–60.
2. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656.
3. Wu, C.C.; Zhang, R.C. An information-theoretic approach to the effective usage of auxiliary information from survey data. *Ann. Inst. Statist. Math.* **2006**, *58*, 499–509.
4. Luciani, D.; Marchesi, M.; Bertolini, G. The role of Bayesian networks in the diagnosis of pulmonary embolism. *J. Thromb. Haemostasis* **2003**, *1*, 698–707.
5. Tu, X.M.; Litvak, E.; Pagano, M. Issues in Human Immunodeficiency Virus (HIV) screening programs. *Am. J. Epidemiol.* **1992**, *136*, 244–255.
6. Cox, E.P., III The optimal number of response alternatives for a scale: A review. *J. Market Res.* **1980**, *17*, 407–422.
7. Østerås, N.; Gulbrandsen, P.; Garratt, A.; Saltyte Benth, J.; Dahl, F.A.; Natvig, B.; Brage, S. A randomised comparison of a four- and a five-point scale version of the Norwegian Function Assessment Scale. *Health Qual. Life Outcomes* **2008**, *6*, 14.

© 2010 by the authors; licensee Molecular Diversity Preservation International, Basel, Switzerland. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).