Marketing Bulletin, 1997, 8, 59-65, Research Note 2

Scaling Numerical Variables and Information Loss: An Appraisal of Morrison's Work

Pascale Quester and Emanuel Dion

This paper aims to assess the information loss and its effect on r when making a continuous numerical variable discrete by categorising the data. A simulation is used in order to determine the true extent of decrease in r while at the same time correcting for the assumption Morrison made regarding the uniformity of the error distribution. The results are then compared with those obtained using the approach suggested by Morrison.

Keywords: information loss, data categorisation, scaling

Introduction

Marketing research is a discipline which relies heavily on surveys for the purpose of data collection. The results are often acknowledged as influenced by the interviewing techniques and methods but the impact which the actual data manipulation may have on the conclusions is seldom noted. Questionnaires commonly use open-ended questions to gather information on topics such as age, income, and so on. While these open-ended questions may allow for a great and sometimes unlimited number of possible answers, the data often requires subsequent categorisation or scaling before any analysis can take place.

There are two types of questions for which categorisation is required. Quantitative open-ended questions seek numerical answers whereas qualitative open-ended questions call for verbal answers. This paper focuses on the quantitative kind and aims to assess the information loss which may occur when making a continuous numerical variable discrete by categorising the data.

A number of researchers have examined this issue in some detail in the past (Morrison 1972; Martin 1973, 1978; Reynolds & Neter 1980; Lawrence 1981). Reynolds & Neter (1980), guided by Information Theory, concluded that eight categories was the optimal number to use when analysing respondents' age. Lawrence (1981), however, disputed this finding, arguing that Information Theory relies on *proportions* in classes irrespective of absolute numbers in them, thus disguising the fact that analyses conducted with small sample numbers could easily provide differences only arising by chance. According to Lawrence, random variations could often falsely suggest that information had been gained. As a result, Information Theory would appear to be best used in situations when the volume of data available for analysis provides enough statistical validity for meaningful interpretation.

Reynolds & Neter's conclusion that there is an optimal number of categories minimising information loss would, however, remain valid for small numbers of categories, since in this situation there are larger numbers of respondents in each category. In such instances, an increase in information gained is observed when the number of categories is increased to eight, with only marginal additional gains in information beyond that point.

In contrast to the two above approaches, which rely on Information Theory to assess information losses (and gains), Morrison (1972), using a simple algebraic calculation, and subsequently Martin (1978), using a series of simulations, focused on the effects of categorisation or scaling on the correlation coefficient ***r***.

Morrison investigated the information loss in the case of a correlation, initially perfect (***r***=1), between two continuous variables, one independent and the other dependent. He observed the changes that occurred in ***r*** when the dependent variable remained continuous while the independent variable became discrete, as would happen when using an interval scale. He concluded that the loss in correlation resulting from discrete scaling is trivial beyond four or five categories.

However, Morrison assumed that the error distribution is exactly uniform and exhibits a variance equal to 1/12 for all scale intervals, the length of each interval being one (1/12 is the variance of a random variable with uniform distribution between -.5 and +.5). This is not entirely accurate and would only hold true if the error term could be both positive and negative. At the two extremes of the scale, larger values can only be rounded down and smaller values rounded up. More precisely, Morrison appears to calculate a correct error distribution but it is inaccurate since a variance of 1/12 does exist but only for the scaled values and not the initial ones.

Nevertheless, Martin (1978) applied Morrison's approach to a wider range of cases and further examined more practical cases when ***r*** = .1, .2, .5, .7, .8 and .9. His paper, however, fails to detail precisely his methodology and suggests, for instance, that as few as one simulation may have been used in each case. An in-depth examination of his finding can be found elsewhere (Dion & Quester 1996).

This paper relies on a different approach to Morrison's, using simulation in order to determine the true extent of decrease in ***r*** while at the same time correcting for the inaccurate assumption he made regarding the uniformity of the error distribution. The results are then compared with those obtained by Morrison.

Method

The loss of information can be quantified with the use of the correlation coefficient linking the initially continuous variable with its own categorised self. The hypothesis that scaling does not affect the amount of information would imply that this observed correlation equals one. Any loss of information would thus result in the observed correlation differing from 1 and the magnitude of the difference provides a measure for the extent of information loss.

In the first instance, a random uniform linear variable with values between 0 and 1 was generated using the 'ALEAT' function from Lotus 123. Scaling was undertaken, using a variety of scaling point numbers. The linear correlation coefficients between the initial variable and its scaled values was then calculated for a series of observations.

While selecting numbers between 0 and 1 could appear artificial and somewhat unrealistic from a market research perspective, one must note that this is in no way different from working on an interval from 0 to 100 or 0 to 1000, since calculations are conducted on ***r***, which is indifferent to the scale used. Indeed, since all calculations were performed by Lotus

123 to 18 decimal places, results are similar to those which would have been derived from using ’real’ numbers.

Furthermore, an alternative approach using a limited range of whole numbers instead of those generated by ‘ALEAT’ would present a number of shortcomings. Firstly, it would multiply the number of analyses to conduct (for each type of non-uniform distribution selected) and thus would make a synthesis more difficult. It would also involve reducing the range further whereas the most common practice is to scale truly continuous or semi-continuous variables into discrete ones, whether the available secondary data consist of continuous variables, such as age, wage or cost or whether respondents are asked to indicate their answers using pre-formed categories. Finally, using a uniform distribution of numbers allows for a direct comparison with Morrison’s results.

The process was computerised with an appropriate program so as to repeat the calculation a great number of times in order to eliminate the risk of random variations affecting the results. In particular, the number of observations was varied so as to enable the assessment of this factor on the standard deviation of the results obtained. A schematic representation of the process is shown in Figure 1.

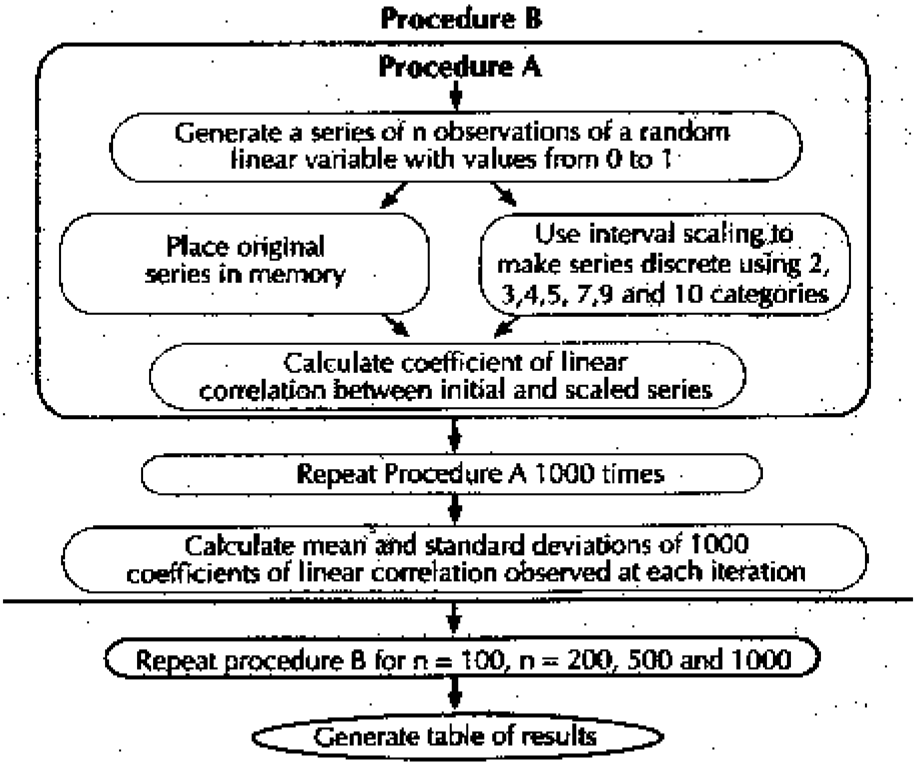
Procedure B
Procedure A
Generate a series of n observations of a random linear variable with values from 0 to 1
Place original series in memory
Use interval scaling to make series discrete using 2, 3,4,5, 7,9 and 10 categories
Calculate coefficient of linear correlation between initial and scaled series
Repeat Procedure A 1000 times
Calculate mean and standard deviations of 1000 coefficients of linear correlation observed at each iteration
Repeat procedure B for n = 100, n = 200, 500 and 1000
Generate table of results

Figure 1. Schematic representation of the simulation process

Results

The results of the simulations are shown in Table 1.

The case of n=1000 provides, as expected, the smallest standard deviations. It is possible to develop confidence intervals around the value of interest as demonstrated in Table 2.

As Table 2 shows, the results using Morrison's method appear more consistent with ours for the larger categories. When nine categories are used, the figure calculated using his method differs from ours by a mere 0.5%. For smaller categories, however, his findings are not confirmed by our simulation. Indeed, when only two categories were used, our calculations differ from those produced by his method by 5%. This suggests a more substantial loss of information than previously thought. However, when three or more categories are used, the loss of information was relatively slight.

These results suggest that the theoretical calculations using Morrison's method provide a satisfactory solution for large numbers of intervals but fail to do so for smaller numbers of intervals; the smaller the number of intervals, the greater the difference found between our calculations and Morrison's.

Table 1. Results of simulations

		Number of categories							
n*	**Indicator**	**2**	**3**	**4**	**5**	**7**	**9**	**12**	**20**
1000	**Mean**	.8662	.9429	.9682	.9799	.9897	.9938	.9965	.9987
	SD	.0049	.0023	.0012	.0008	.0004	.0002	.0001	.0001
500	**Mean**	.8663	.9435	.9682	.9799	.9898	.9938	.9965	.9987
	SD	.0069	.0028	.0021	.0012	.0005	.0003	.0002	.0001
200	**Mean**	.8661	.9429	.9686	.9796	.9898	.9938	.9966	.9988
	SD	.0105	.0054	.0027	.0018	.0009	.0006	.0003	.0001
100	**Mean**	.8668	.9429	.9688	.9799	.9897	.9938	.9966	.9988
	SD	.0153	.0074	.0038	.0024	.0013	.0008	.0004	.9988

*n= number of repetitions of procedure B (refer Fig. 1)

Table 2. Confidence intervals of the linear correlation coefficients between the random continuous linear variable and its discrete form

	Number of categories							
	2	**3**	**4**	**5**	**7**	**9**	**12**	**20**
	95% confidence intervals							
Upper Bound	.8665	.9430	.9683	.9799	.9898	.9938	.9965	.9988
Lower Bound	.8659	.9428	.9681	.9798	.9897	.9938	.9965	.9987
Morrison's method	.8165	.9354	.9661	.9789	.9856	.9895	.9920	.9937

If one aimed to assess accurately the loss in shared variance occurring between the two variables, the correlation coefficients must be squared. This is shown in Table 3.

Once again, it is clear that the main difference between Morrison's conclusions and ours concerns the magnitude, rather than the existence, of the information loss that occurs when categorising data. Both methods converge as the number of categories approached six, after which the observed losses become unsubstantial.

Table 3. Percentage loss in explained variance according to the number of categories of the independent variable

Number of categories	**r -square values**		**% of explained variance lost as a result of categorisation**		
	Morrison's	Simulation*	Morrison's	Simulation*	Difference
2	66.7	75.0	33.3	25.0	8.3
3	87.5	88.7	12.5	11.3	1.2
4	93.3	93.7	6.7	6.3	0.4
5	95.8	96.0	4.2	4.0	0.2
6	97.1	97.2	2.9	2.8	0.1
7	97.9	97.9	2.1	2.1	0
8	98.4	98.4	1.6	1.6	0
9	98.8	98.8	1.2	1.2	0

Conclusions

Despite Morrison's mistaken assumption regarding the uniformity of error distribution, his results are generally upheld. Overall, our results suggest that reducing a continuous variable to six to eight categories results in minimal information losses; Morrison suggested that such losses become trivial beyond five categories. A market researcher relying on this rule, however, should remain mindful that it is always ill-informed to write a close-ended question of this type with no prior knowledge of the response distribution. Exploratory research should always provide the justification for such categorisation decisions. The decision to use more or fewer categories would also be subject to other statistical circumstances, including the level of correlation between the variables under study.

When trading off market research realities, such as the need to simplify questionnaires for the purpose of reducing interviewing time, against market research's goals, such as the need to collect as much reliable information as possible, the researcher has to take into account the potential danger that exists in over-reducing the number of proposed categories. While increasing the number of categories never results in poorer information transfer, reducing it requires caution in order to minimise information losses. Increasing category numbers will, however, only contribute a decreasing amount of additional information (law of diminishing returns).

There does not appear to be a unique solution or universal rule regarding the optimal number of categories to use. Thus, Lawrence's recommendation to gather and enter as much data as possible with as many categories as possible, seems sensible. Then, initial calculations, such as correlation, need to be undertaken to ascertain which type of simplifying categorisation may be suitable, if any.

Market researchers often must adopt categorisations which are both more relevant and practical for their ultimate purpose, but it rarely makes sense to divide the data in equal size categories. For example, dividing some population on the basis of age in order to obtain an equal number of observations in each category may result in such odd age classes as 15-24, 33-38, 52-54 and so forth whereas more natural classes such as 15-29, 30-39, 40-49 and the like would make more sense for the purpose of further analysis, even though numbers of respondents in each class may vary. The interpretability of the data, therefore, should also contribute to such decisions in categorisation. Furthermore, the number of categories may need to be determined in relation to the type of statistical analysis to be carried out. For example, a researcher intending to analyse the data using a Chi-Square analysis would probably wish to avoid cells with low expected values.

What is useful and more realistic than any attempt to eradicate information loss when categorising a continuous variable is the ability to quantify such losses. A knowledge of such losses allows a more accurate assessment of the trade-offs which must be made with other market research imperatives, such as duration of questionnaire administration and confidentiality issues.

References

Dion E & Quester PG (1996). Impact of categorisation on information retention: An appraisal of Martin's work. *The Australian Journal of Market Research,* July, *4* (2), 21-26.

Lawrence R (1981). How many categories for respondent classification? *Journal of the Market Research Society*, *23* (4), 220-238.

Martin W (1973). The effects of scaling on the correlation coefficient: A test of validity. *Journal of Marketing Research*, *10*, 316-318.

Martin W (1978). The effects of scaling on the correlation coefficient: Additional considerations. *Journal of Marketing Research*, *15*, 304-308.

Morrison D (1972). Regression with discrete dependant variables: The effect on R ^{2}. *Journal of Marketing Research*, *9*, 338-340.

Reynolds F & Neter J (1980). Toward a method for assessing adequacy of classifications of consumer characteristics. *Journal of the Market Research Society*, *22* (1), 3-27.

Pascale Quester is a Senior Lecturer at the University of Adelaide, and Emanuel Dion was a doctoral candidate at ESC Nantes Atlantique

Acknowledgment: This paper is based on Emanuel Dion's Doctorate thesis and originated during Dr Quester's study leave in 1995 at ESC Nantes Atlantique.