Analyzing Survey Data

On Constructing Interval Scales Using Data Resulting from Categorical Judgements

*This paper is from course material by Professor Glenn Lindsay of the Naval Postgraduate School. It provides a methodology for analyzing survey data, without assuming arbitrary numerical values for the responses to the questions (such as satisfactory = 1, good = 2, excellent = 3, etc.). This method is also valuable for analysis of "**fuzzy**" performance indicators which are based upon worker survey results.*

Categorical Judgments: The Method of Successive Intervals

A frequently used means of obtaining ratings from judges [survey responders] is that of categorical judgment, wherein judges assign instances to ranked categories. For example, corporate bonds may be rated as A, AA, and so on; student opinion forms ask the student to rate an instructor as poor, fair, average, excellent, or outstanding; pollsters often ask people to check one of a set of categories described as strongly agree, agree, no opinion, disagree, and strongly disagree. When an instructor assigns a student's letter grades, he may be viewed as making a categorical judgement in that the possible grades are the categories and the students are the instances. Other examples of ranked categories are found in such diverse applications as restaurant sanitation ratings, personnel appraisals, and motion picture ratings (G, PG, R, NC-17). Usually, there are descriptors associated with each category which serve to help the judge with his rating task.

The method described in this paper is a scaling method which uses categorical ratings provided by judges, and constructs an interval scale which includes not only the instances but also the bounds between the categories. Thus descriptive benchmarks appear on the final scale. Typically, five categories are used. No assumptions are made about the relative interval sizes for the categories. The categories are understood to be a mutually exclusive set of successive intervals which collectively exhaust the property continuum.

Data Assembly

A direct way to aggregate categorical ratings of instances (survey questions) by judges is through a frequency array, with a row for each of the n instances and a column for each of the m categories. Columns in this array should be arranged in ascending order of category value, so that the category representing the least amount of the property is Column 1, and the category representing the greatest amount of the property is Column m. It is not necessary for a judge to rate all instances (questions).

Working with the frequency array, we may cumulate values in each row rightward and divide by the row total to achieve a cumulative frequency array. Since the values in the right hand column of the cumulative frequency array will always be 1.0, this column may be omitted for computational purposes, yielding a cumulative frequency array with n rows, and m minus 1 columns.

In the example given below, 80 judges were asked to rate four political candidates in terms of their "potential effectiveness as President of the USA". The categories were Very Ineffective, Ineffective, Marginal, Effective, and Very Effective. Raw frequency data are:

Candidate	Very Ineffective	Ineffective	Marginal	Effective	Very Effective
A	10	20	27	21	2
B	4	30	35	11	0
C	20	43	15	2	0
D	3	2	34	30	11

From this raw frequency array, the cumulative relative frequency array may be constructed:

Candidate	Very Ineffective	Ineffective	Marginal	Effective
A	0.1250	0.3750	0.7125	0.9750
B	0.0500	0.4250	0.8625	1.0000
C	0.2500	0.7825	0.9750	1.0000
D	0.0375	0.0625	0.4875	0.8625

These results say, for example, that 71.25% of judges found A no better than marginal. Another way of looking at these values would view the columns as upper bounds on adjacent categories, and thus we would say the 71.25% of judges placed Candidate A below the upper bound of the marginal category. That is the interpretation we will use in the work to come. Note that category "Effective" will have an upper bound but the highest category, "Very Effective", will not. Similarly, the lowest category will have no lower bound.

Theory

We assume that a judge's "feelings" about the scale value of an instance (or question) i is a normally distributed random variable with mean S i and standard deviation sigma i. We also assume that judges view the continuum of values for instances as being broken into successive intervals called categories, and that a judge's feelings about a category's upper bound is a normally distributed random variable so that for category j, the upper bound would be normally distributed with mean b j and standard deviation v j.

We want, for each instance i, an estimate of its mean S i. To obtain these estimates, we will also have to obtain estimates of the category upper bounds (b j) since the raw data will be sorted by category.

Since a judge's feelings about instance values and about category upper bounds are normally distributed random variables, the judge's feelings about the distance between an instance value and a category bound will also be a normally distributed random variable

with mean b j minus S i, and variance sigma j squared plus v j squared. It is not unreasonable to assume that the value bound "feelings" are independent random variables, so that the correlation coefficient between all i,j pairs is zero. We also assume that all category bounds have the same standard deviation, so that for all category bounds, the variance equals a constant (c).

Thus a judge's feelings about the distance from bound j to instance i can be viewed as a normally distributed random variable with mean b j minus S i and variance sigma i squared plus c. It follows that the probability that an instance i is rated below bound j is equal to the probability that z is less than (b j minus S i) divided by the square root of (sigma i squared plus c), where z is normally distributed with mean 0 and variance 1. From the frequency data from judges we obtain estimates of all of these probabilities for each i,j pair, and then retrieve the z values from a normal distribution table. We now have n times (m-1) equations to solve.

Editor's note: The original paper derives the solution to these simultaneous equations.

For this method to work, we must have a complete array of values (every question must have at least one response in every category). The best method to deal with this is to consolidate two columns together (as is done in the example at the end of this paper).

Step by Step Procedure for Obtaining Scale Values with a Complete Array

An EXCEL spreadsheet which accomplishes the operations below is available. It uses the data in the example in this paper. Please contact Steve Prevette or call at 509-373-9371.

1. Arrange the raw frequency data in a table where the rows are instances (questions) and the columns the categories. Columns should be in rank order, with Column 1 representing the least favorable category, etc.
2. Compute relative cumulative frequencies for each row, and record these in a new table. The last column of this new table will consist of 1's and may be omitted.
3. Treating these values as leftward areas under a Normal (0,1) curve, go to a table of the normal distribution and find the z values for these areas. Record these in a new n by (m-1) table. This is the z ij array for the computations which follow.
4. For each row i in the z ij array, compute the row average, z bar i.
5. For each column j in the z ij array, compute the column average. Call these column averages b j, and note that b j is the value of the upper bound of category j on our scale.
6. Compute a grand average of all the values in the z ij array. This is readily done by simply averaging the column averages. Call the grand average B bar.
7. Compute B = sum from j = 1 to m-1 of (b j minus B bar) squared.
8. For each row, compute Ai = sum from j = 1 to m-1 of (z ij minus z bar i) squared.
9. For each row, compute the square root of (B divided by Ai). This is an estimate of the standard deviation of the response for the question, the square root of sigma i squared plus c.

10. Finally, for each row (question) compute S i = b bar minus z bar i times the square root of (B divided by Ai).

The values of S i are the scale values of the instances (questions), and they are on the same interval scale as the category bounds b j. We now have the desired scale, and may perform any linear transformation to move the scale where we want it. Remember to use the same transformation to move both instance values and the category bounds.

Example

We will continue the example started at the beginning of this paper. Because candidates B and C had no responses in the Very Effective column, we must pool the Effective and Very Effective categories together. Steps 1 and 2 have already been accomplished. Steps 3, 4, 5, and 6 are reflected in the table below:

Candidate	**Potential Effectiveness**				
z ij	**Very Ineffective**	**Ineffective**	**Marginal**	**Row Total**	**Row Average z bar i**
A 1	-1.15	-0.32	0.56	-0.91	- 0.303
B 2	-1.64	-0.19	1.09	-0.74	- 0.247
C 3	- 0.67	0.78	1.96	2.07	0.690
D 4	-1.78	-1.53	-0.03	- 3.34	-1.113
Column Totals	-5.24	-1.26	3.58	- 2.92	= Grand Total
Column Averages: b j	-1.310	- 0.315	0.895	-0.243	= Grand Average: b bar

*Note: In the formulae below, the notation **2 represents "Squared", and SQRT represents "Square Root".*

Step 7. B = (-1.310 - (-0.243)) **2 + (-0.315 - (-0.243)) **2 + (0.895 - (- 0.243)) **2 = 2.439

Step 8. A1 = (-1.15 - (-0.303)) **2 + (-0.32 - (-0.303)) **2 + (0.56 - (- 0.303)) **2 = 1.462
Similarly, A2 = 3.731, A3 = 3.471, and A4 = 1.792

Step 9. SQRT(B / A1) = 1.292, SQRT(B / A2) = 0.809, SQRT(B / A3) = 0.838, SQRT(B / A4) = 1.167

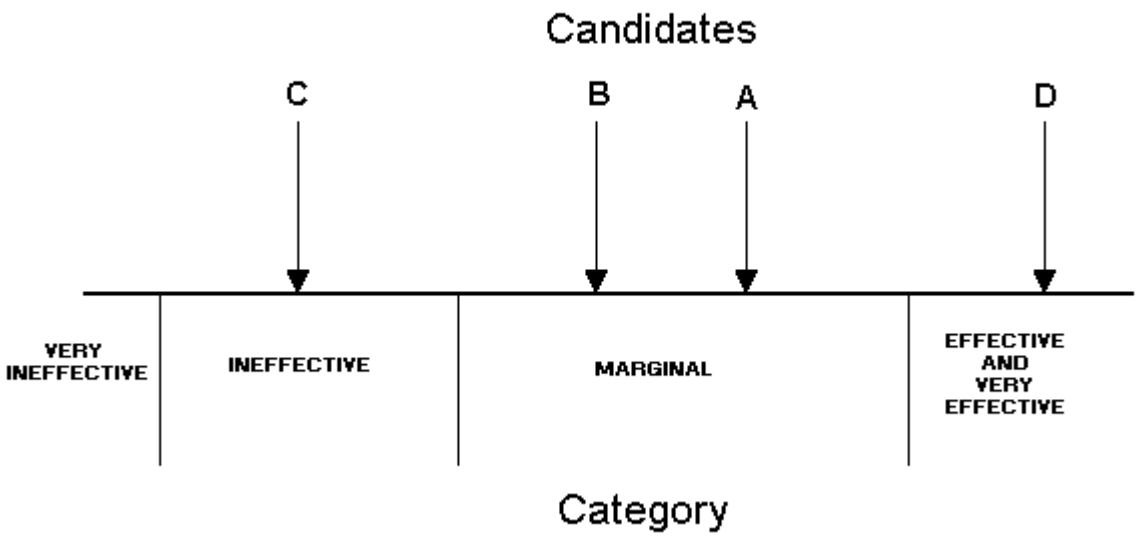
Step 10. S1 = -0.243 - (-0.303) (1.292) = 0.148
S2 = -0.243 - (-0.247) (0.809) = -0.043
S3 = -0.243 - (-0.690) (0.838) = -0.821
S4 = -0.243 - (-1.113) (1.167) = 1.055

Also, Upper Bound on the Very Ineffective Category is -1.310
Upper Bound on the Ineffective Category is -0.315
Upper Bound on the Marginal Category is 0.895

Final Results of the Example

Category	Candidate	Score
Very Ineffective		Less than -1.31
Ineffective		-1.31 to -0.315
	Candidate C	**-0.821**
Marginal		-0.315 to 0.895
	Candidate B	**-0.043**
	Candidate A	**0.148**
Effective and Very Effective		greater than 0.895
	Candidate D	**1.055**

Graphical Representation of Example Results:



Incomplete Zij Arrays

Zij array entries corresponding to Pij > 0.98 and Pij < 0.02 should be omitted to avoid undue influence by a small number of judges. A problem also occurs when the response array is incomplete. There must be at least one response by at least one judge in each of the categories for all of the candidates for the response array to be complete.

Because of the variety of situations that can occur, it is probably best not to attempt to provide here specific instructions on how to cope with an incomplete Zij array. Three tactics are suggested below.

1. One may delete those rows with missing Zij values to obtain a smaller but complete Zij array, and apply the method given in this paper. This means, of course, that instances represented by those deleted rows will not be scaled directly. One either discards these instances, or "pieces" them onto the scale in some way that will hopefully be defendable (but will seldom be altogether satisfactory).
2. One may pool extreme categories to obtain a Zij array void of missing values. For example, if Column 1 has missing Zij values and Column 2 is complete, we combine Categories 1 and 2 into a single category, and use the Zij values of Column 2. As another example, if the last column (column m-1) has missing Zij values and the next to last column is complete, we combine these last two categories together. This method was used in the example above where Categories Effective and Very Effective were combined together.
3. A third approach is to break the Zij array down into smaller arrays, applying the previously described tactics so that one has several complete, but smaller Zij arrays. These arrays are scaled separately. If one has been clever in dividing the original array, the resulting set of scales will have two overlapping points in common so that linear transformations will place all instances and bounds on the same scale.

Conclusion

This methodology allows the transformation of categorical judgements in a survey to an interval scale. This method is useful for the "fuzzy" performance measure methods using survey results being applied at certain Department of Energy sites. This is a robust statistical method that does not rely upon arbitrarily assigning numerical values to the judgement categories.

Credits:

Author: Glenn F. Lindsay, Naval Postgraduate School, Monterey CA
Date: September, 1981.
Reference: Torgenson, W. S., Theory and Methods of Scaling, Wiley, 1958.