Visualising contingency table data

Dongwen Luo, G. R. Wood, G. Jones

Abstract

A geometric object, a simplex, is useful for picturing the joint, conditional and marginal distributions within a contingency table. The joint distribution is represented using weights on all vertices of the simplex, a conditional distribution by weights on vertices of a face of the simplex, and a marginal distribution by weights on the faces containing the conditional distributions. All detailed discussion is based on the simplest case, that of a two-by-two contingency table, for which all distributions are seen in a tetrahedron.

1 Introduction

A contingency table is a cross-tabulation of categorical variables. An example is given in Table 1, using data from an Australian survey of attitudes to genetic engineering of food [4]. The 894 respondents are distributed among four categories defined by income level and attitude to genetic engineering. The question of interest is whether income level and attitude to genetic engineering of food are dependent.

	Attitude	
Income	For	Against
Low	258	222
High	263	151

Table 1. A cross-tabulation of income level against acceptance of genetic engineering of food, with data drawn from a recent Australia-wide survey.

When faced with contingency table data, it is useful for the practitioner to have a quick method for visualising the associated distributions. The primary aim of this article is to bring such a method to a wider audience; the secondary aim is to provide a cameo example of the symbiosis between mathematics and statistics. The article exposits and builds on ideas first introduced by Fienberg [2] and Fienberg and Gilbert [3].

There are three distributional types associated with a contingency table: the joint distribution, conditional distributions and marginal distributions. This article pictures these three types in a simplex. For a given contingency table, the joint distribution can be represented by weights on all vertices of the simplex, a conditional distribution by weights on vertices of a face of the simplex, and a marginal distribution by weights on the faces containing the conditional distributions. All discussion is based on the contents of a two-by-two table, since such a table is complex enough to illustrate all items of interest yet simple enough to be readily pictured.

In the next section we review the three distributions, using notation of Agresti [1]. The three distributional types are described geometrically in Section 3, then the article is completed with a generalisation in Section 4 to tables of arbitrary dimension and a conclusion.

2 Distributions in a two-by-two table

We begin this section by briefly reviewing standard terminology and notation for joint, conditional and marginal distributions in a contingency table. Consider two categorical variables X_1 and X_2, each at two levels. The joint distribution of X_1 and X_2 can be represented in a 2×2 table denoted (π_{ij}), where π_{ij} is the probability of X_1 at the ith level and X_2 at the jth level, for $i = 1, 2$ and $j = 1, 2$.

The marginal distributions of X_1 and X_2 are denoted (π_{1+}, π_{2+}) and (π_{+1}, π_{+2}) respectively. Here the subscript "+" denotes summation over the associated index, so $\pi_{i+} = \sum_j \pi_{ij}$ and $\pi_{+j} = \sum_i \pi_{ij}$. Thus, the marginal distribution of X_1 (X_2) appears as the row (column) totals of the table (π_{ij}).

The distribution of X_2 conditional upon $X_1 = i$ is written as $(\pi_{1|i}, \pi_{2|i})$ so $\pi_{j|i} = \pi_{ij}/\pi_{i+}$ for all j. Symmetrically, we could define the distribution of X_1 for a given level of X_2.

These three distributions associated with a two-by-two table and a numerical example (the frequency table of the Australia survey data) are displayed in Table 2.

	X_2		
X_1	1	2	Total
1	π_{11}	π_{12}	π_{1+}
	$(\pi_{1\|1})$	$(\pi_{2\|1})$	
2	π_{21}	π_{22}	π_{2+}
	$(\pi_{1\|2})$	$(\pi_{2\|2})$	
Total	π_{+1}	π_{+2}	1.00

	Attitude		
Income	For	Against	Total
Low	0.2886	0.2483	0.5369
	(0.5375)	(0.4625)	
High	0.2942	0.1689	0.4631
	(0.6353)	(0.3647)	
Total	0.5828	0.4172	1.00

Table 2. The left panel presents the notation for joint, conditional and marginal distributions of categorical variables X_1 and X_2, each with two levels. The right panel presents the relative frequency table for the Australia survey data. Figures in brackets show the distribution of X_2 for the given level of X_1.

3 Geometry of the three distributions

The joint distribution of categorical variables X_1 and X_2 with two levels each can be represented as

$$(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = \pi_{11}e_1 + \pi_{12}e_2 + \pi_{21}e_3 + \pi_{22}e_4$$

where $e_1 = (1, 0, 0, 0)$, $e_2 = (0, 1, 0, 0)$, $e_3 = (0, 0, 1, 0)$ and $e_4 = (0, 0, 0, 1)$ form the standard basis in $\mathbf{R}^4$ (points A, B, C and D respectively in Figure 1(a)). Thus the joint distribution of X_1 and X_2 can be pictured as weights π_{11}, π_{12}, π_{21} and π_{22} on A, B, C and D respectively.

Alternatively, since $\pi_{ij} \geq 0$ for all i, j and $\sum_{ij} \pi_{ij} = 1$, the joint distribution of X_1 and X_2 can be represented by the centre of mass J (more formally known as the "resultant" or "barycentre") of these weights on A, B, C and D in the three dimensional simplex given by

$$S_3 = \{(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) : \sum_{ij} \pi_{ij} = 1 \text{ and } \pi_{ij} \geq 0 \text{ for all } i, j\}$$

as illustrated in Figure 1(a).

The distribution of X_2 conditional on $X_1 = 1$ can be represented as $(\pi_{1|1}, \pi_{2|1}, 0, 0)$, an ordered 4-tuple in $\mathbf{R}^4$, and since we have the representation

$$C_1 = \pi_{1|1}e_1 + \pi_{2|1}e_2$$

evidently this distribution can be represented by weights $\pi_{1|1}$ and $\pi_{2|1}$ on A and B alone.

Alternatively, since $\pi_{j|1} \geq 0$ for all j with $\sum_j \pi_{j|1} = 1$, the distribution of X_2 conditional on $X_1 = 1$ is the resultant of these weights on A and B, so is a point C_1 in line segment AB. Similarly, the distribution of X_2 conditional on $X_1 = 2$ can be represented as $(0, 0, \pi_{1|2}, \pi_{2|2})$, so as a point C_2, the resultant of weights $\pi_{1|2}$ and $\pi_{2|2}$ on C and D respectively (illustrated in Figure 1(b)).

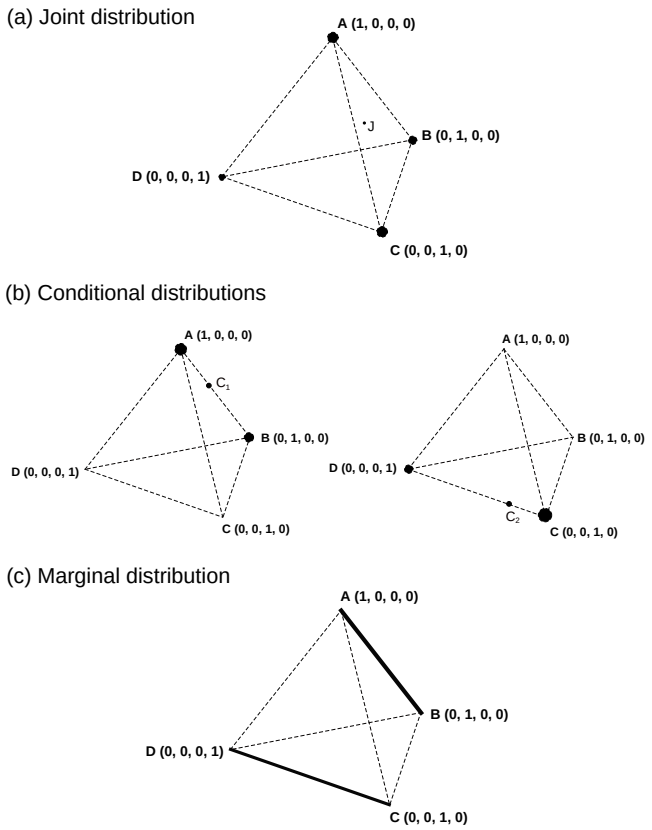

Figure 1. The three distributions of categorical variables X_1 and X_2, each with two levels. In (a) the joint distribution of X_1 and X_2 is seen as weights π_{11}, π_{12}, π_{21} and π_{22} on A, B, C and D, with resultant J. In (b) the conditional distribution of X_2 when $X_1 = 1$ is seen as weights $\pi_{1|1}$ and $\pi_{2|1}$ on A and B, having resultant C_1, while the the conditional distribution of X_2 when $X_1 = 2$ is weights $\pi_{1|2}$ and $\pi_{2|2}$ on C and D, having resultant C_2. In (c) the marginal distribution of X_1 is seen as weights π_{1+} and π_{2+} on edges AB and CD.

Joint distributions lying on AB oblige X_1 to equal one, so arguably line segment AB corresponds to $X_1 = 1$. Similarly, line segment CD corresponds to $X_1 = 2$. For this reason the marginal distribution of X_1, (π_{1+}, π_{2+}), can be represented as these weights on edges AB and CD, pictured by weighting these edges in Figure 1(c).

From the definition of conditional probability we have that

$$(\pi_{11},\ \pi_{12},\ \pi_{21},\ \pi_{22}) = \pi_{1+}(\pi_{1|1},\ \pi_{2|1},\ 0,\ 0) + \pi_{2+}(0,\ 0,\ \pi_{1|2},\ \pi_{2|2})$$

or

$$J = \pi_{1+}C_1 + \pi_{2+}C_2$$

In this special case where the joint distribution J and the conditional distributions C_1 and C_2 are known, the marginal distribution of X_1 can be represented as the weights π_{1+} and π_{2+} on C_1 and C_2 (still on AB and CD respectively) having resultant J.

Figure 1 in fact illustrates these ideas using the frequency table of the Australia survey data shown in the right panel of Table 2. Here we can represent the joint distribution of Income and Attitude as

$$(0.2886,\ 0.2483,\ 0.2942,\ 0.1689) \in \mathbf{R}^4$$

which corresponds to point J in the tetrahedron. The distributions of Attitude conditional on Income Low and Income High can be represented by $C_1 = (0.5375, 0.4625, 0, 0)$ and $C_2 = (0, 0, 0.6353, 0.3647)$ respectively. Since $J = 0.5369C_1 + 0.4631C_2$, the marginal distribution of Income, $(0.5369,\ 0.4631)$, can be specialized now as weights 0.5369 and 0.4631 on C_1 and C_2 having resultant J.

Fienberg and Gilbert [3] showed that the loci of all points corresponding to independence of rows and columns in a 2×2 table is a portion of a hyperbolic paraboloid in the tetrahedron, illustrated in Figure 2. In the figure, the point J (the joint distribution of Income and Attitude) is seen to be a small distance away from the independence surface; further analysis would confirm that, with a sample size as large as 894, this indicates dependence between Income and Attitude. Loosely speaking, for a given sample size the further J is from the independence surface, the greater the dependence between X_1 and X_2.

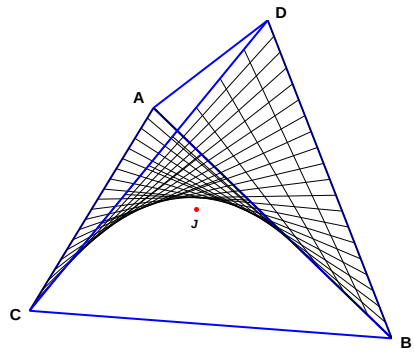

Figure 2. A graphic illustrating the locus of all points corresponding to independent 2×2 tables (a portion of a hyperbolic paraboloid) and the joint distribution J of Income and Attitude in the tetrahedron $ABCD$.

4 Tables of higher dimension

For a general contingency table, the three distributional types can be pictured in a higher dimensional simplex, having as many vertices as cells of the table. The joint distribution appears as weights on all vertices of the simplex. Conditioning on the levels of a subset of the variables partitions all vertices of the simplex; the convex hull of each partition set forms a face of the simplex. A distribution conditional on levels of the chosen variables appears as weights on vertices of the associated face. The marginal distribution of the random variables used for conditioning appears as weights on the simplicial faces determined by the partition sets. For example, for a 4×4 table with variables X_1 and X_2, the joint distribution is the weights on the sixteen vertices of the simplex S_{15}. To picture the distribution of X_2

conditional upon X_1, the vertices of S_{15} are partitioned into four sets of four using the levels of X_1. Four faces of S_{15} are then constructed as convex hulls of each set of vertices; the distribution of X_2 conditional upon a given level of X_1 is weights on the vertices of the associated face. The marginal distribution of X_1 is weights on the four faces. These ideas are illustrated in Figure 3.

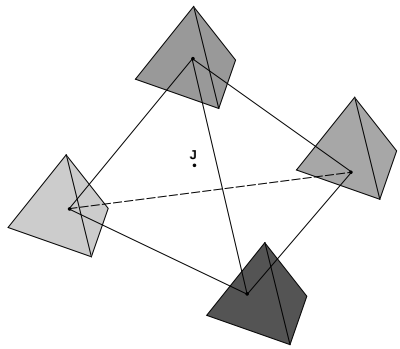

Figure 3. A schematic illustration showing that for a multi-way table the joint distribution J appears as weights on all vertices of a higher dimensional simplex; the resultant is a point in the simplex. Conditioning on values of a subset of all variables leads to a partitioning of the vertex set. Such a partition is shown as the four shaded simplexes. A conditional distribution is a weighting of the vertices of a partition set, for example, a weighting on the vertices of the upper shaded simplex. The associated marginal distribution of the subset of variables is the weighting of the facial simplexes formed by the partition, shown here using shading. The diagram presented here is strictly appropriate for a 4×4 table.

5 Conclusion

The three distributional types associated with a 2×2 table have been pictured in a tetrahedron. The joint distribution appears as weights on all vertices of the tetrahedron with resultant a point in the tetrahedron. A conditional distribution can be viewed as weights on vertices of an edge of the tetrahedron with resultant a point in the edge. A marginal distribution can be viewed as weights on the edges containing the conditional distributions. These ideas directly generalize to multi-way tables.

References

[1] A. Agresti, *Categorical Data Analysis* (Wiley New York 1990).

[2] S.E. Fienberg, *The geometry of an $r \times c$ contingency table*, The Annals of Mathematical Statistics **39** (1968), 1186–1190.

[3] S.E. Fienberg and J.P. Gilbert, *The geometry of a two by two contingency table*, Journal of the American Statistical Association **65** (1970), 694–701.

[4] J. Norton, G. Lawrence, and G.R. Wood, *The Australian public's perception of genetically-engineered foods*, Australasian Biotechnology **8** (1998), 172–181.

Department of Statistics, Macquarie University, NSW 2109
E-mail: gwood@efs.mq.edu.au

Institute of Information Sciences and Technology, College of Sciences, Massey University, Palmerston North, New Zealand

Received 26 May 2004, accepted 8 July 2004.