A Scaled Sample Space Cube Used to Illustrate Attributable Fractions
A three-dimensional graphic method is proposed for displaying the association structure between multiple explanatory variables and their relation to a categorical response. The method combines the techniques of mosaic displays and scaled Venn diagrams, and is especially useful for illustrating attributable fractions in epidemiology. The primary purpose is to show the reduction of disease risk in a population if the joint exposure distribution or the conditional risk function is modified, and the method can be extended to illustrate the potential effects of successive removal of exposures on the overall risk of disease. The scaled sample space cube may be used for communicating the difficult concept of attributable fraction to statisticians, the medical community and the general public in an easily understandable way. Demonstrations of the method use theoretical models as well as data from the Hordaland study on the effect of smoking and occupational exposure on obstructive lung disease. Also, the general principle of adding a third dimension to a mosaic display, instead of using shading or coloring, to show an attribute of a cell in a multiway contingency table, can be helpful for other purposes, as the residual analysis of a loglinear model fitting.

Key Words: Logistic regression; Mosaic Display; Scaled Venn Diagram; Generalized Attributable Fraction; Ordered Preventive Strategies; Obstructive Lung Disease
1 INTRODUCTION

In a recent paper Eide and Heuch (2001) proposed a particular kind of scaled Venn diagram as a tool for illustrating the excess risk of a disease attributable to raised levels of a specified set of exposures. Such scaled Venn diagrams clearly illustrate relations with risk, but do not represent the association structure between the explanatory variables very well. Hartigan and Kleiner (1984) developed the mosaic display for visualizing the structure of a multi-dimensional contingency table, and the scaled Venn diagram proposed by Eide and Heuch (2001) may be thought of as such a mosaic display (without spacings) for a two-dimensional table where the cross-classification of all the explanatory variables constitutes the first dimension and the conditional distribution of the response variable constitutes the second. Friendly (1994, 1995) further developed the mosaic display of Hartigan and Kleiner (1984) as a visual aid in fitting loglinear models to a multi-way contingency table. Also, for describing clinical data relationships, Marshall (2001) introduced an adaptation of a scaled Venn diagram with frequencies represented by partly overlapping rectangles.

In the present paper a new method is proposed for illustrating excess risk that uses the mosaic display in combination with the scaled Venn diagram of Eide and Heuch (2001). The procedure proposed describes both the multi-dimensional structure within a set of categorical explanatory variables and their influence on the risk of a disease or more generally a categorical response. This will be particularly useful in interpreting various aspects of attributable fractions.
These measures are used in epidemiology to quantify the relative reduction in morbidity or mortality that would take place if the exposure to some specified risk factors were reduced to a counterfactual distribution while the distribution of all other risk factors remained the same.

2 THE SCALED SAMPLE SPACE CUBE

Assume that $Y$ is an indicator variable for a binary response of particular interest. Typically, $Y = 1$ indicates the event $D$, occurrence of a particular disease, while $Y = 0$ represents the complement, $\overline{D}$. Let $X_1, X_2, \ldots, X_L$ be a set of categorical explanatory variables generating an $L$-dimensional contingency table.

For the moment, assume that the joint probability distribution of all the variables is known. Now, in principle, the mosaic display (Hartigan and Kleiner 1984) can be drawn in $L!$ different ways, each representing a unique ordering of the $L$ explanatory variables. For a selected ordering, say $(X_{(1)}, X_{(2)}, \ldots, X_{(L)})$, this is done by starting with a unit square. Then, the square is partitioned in vertical strips with widths (and thus areas) equal to the marginal point probabilities of $X_{(1)}$. Next, for each strip a horizontal partition is made according to the distribution of $X_{(2)}$ conditional on the value of $X_{(1)}$ for this strip. The procedure continues by doing likewise for $X_{(3)}$ conditional on the joint values of $X_{(1)}$ and $X_{(2)}$ giving new vertical partitions, and proceeds in this way until all $L$ explanatory variables have been run through. The mosaic display thus appearing is a representation of the joint distribution of the $L$ explanatory variables.
given the selected ordering. Every point probability of the \( L \)-dimensional distribution is now represented by a rectangle with area equal to the corresponding probability. Of course, all rectangles considered collectively also partition the whole unit square.

Let this mosaic display be the base of a unit cube. Above each rectangle in the base, plot a congruent rectangle at a height equal to the probability of \( Y = 1 \) conditional on the corresponding values of the explanatory variables. Thus a \textit{scaled sample space cube} appears: each pair of rectangles constitutes the base and top of a rectangular box with volume equal to the joint probability of \( Y = 1 \) and the combination of values for the explanatory variables. Also, the corresponding probability of \( Y = 0 \) is represented by the volume of the box on top of the box for \( Y = 1 \). The total volume of the boxes for \( Y = 1 \) equals the total probability of \( Y = 1 \).

Figure 1 gives an example of the scaled sample space cube for a model with two dichotomous explanatory variables corresponding to events \( C \) and \( E \), in this order. In medicine, \( C \) and \( E \) may be the events of being exposed to two harmful agents increasing the risk of a certain disease (\( D \)). The interest then lies in the reduction in risk if the exposures to \( C \) or \( E \) were reduced in the population. The joint distribution of the explanatory variables is chosen as \( P(C \cap E) = 3/14 \), \( P(C \cap \bar{E}) = 5/14 \), \( P(\bar{C} \cap E) = 4/14 \), and \( P(C \cap \bar{E}) = 2/14 \), and the corresponding conditional probability distribution of disease, \( P(D|..) \), as \( 2/20 \), \( 4/20 \), \( 3/20 \) and \( 11/20 \). For comparison the scaled Venn diagram for this example is shown in Figure 2 (as also given in Figure 12 of Eide and Heuch (2001)).
3 EXCESS RISK AND THE ATTRIBUTABLE FRACTION

The attributable fraction is defined to be the fraction by which the probability of disease would be reduced if the levels of exposure were to be changed to the lowest level (not exposed), i.e.

$$\lambda = \frac{P(D) - P(D|E)}{P(D)}$$

where $E$ is the lowest exposure level. This epidemiologic concept was first introduced with a different formula by Levin (1953) and described as 'the maximum proportion of' the disease (i.e. cancer) 'attributable to' the exposure (i.e. smoking) (page 436). However, the more intuitive expression above was formulated by MacMahon and Pugh (1970) and termed 'attributable risk', and shown by Leviton (1973) to be algebraically equivalent to Levin’s formulation.

Other terms like population attributable risk (PAR), etiologic fraction (Miettinen 1974) and excess fraction (Greenland and Robins 1988) have also been used for $\lambda$; for an overview see Benichou (2000). With multiple exposures, $E$ may represent the combination where all exposures are at their lowest level. In epidemiology, the attributable fraction is an increasingly popular measure for quantifying the potential impact on the occurrence of a disease of hypothetically eliminating excess exposure in a population (Uter and Pfahlberg 2001), and recently available software (Brady, 1998) will further facilitate its use.

The numerator of $\lambda$ is the excess risk which is the reduction in risk if the exposure is completely eliminated. The scaled sample space cube is now convenient
for displaying this excess risk by plotting the conditional probabilities of disease both before reduction of the risks, \( p(x) = P(D|x) \), and after, \( p_0(x) = P(D|x_0) \). The volume between the two surfaces in the cube is then equal to the excess risk, and the ratio of this volume to the total volume under \( p(x) \) equals the attributable fraction. Notice that \( x \) does not refer to the ordinary Cartesian coordinates in the base of the cube but rather to the combination of exposure levels whose joint probability a rectangle in the base mosaic represents.

Figure 3 shows the excess risk in a scaled sample space cube for the example with the two dichotomous explanatory variables if both exposures were eliminated. The angle of view is slightly changed from Figure 1 to better display \( p_0(x) \) (the green box). The cube illustrates the fact that the persons unexposed to the first factor (\( C \)) constitute slightly more than half of the population (4/7 actually), and that within this group, less than one half are unexposed to \( E \) (3/8 actually) while among those exposed to \( C \) there is a larger proportion not exposed to \( E \) (2/3 actually). Furthermore, Figure 3 shows that the conditional risk of disease is the highest for those exposed to both exposures (11/20). The unconditional excess risks are 0, 2/20, 1/20, 9/20, respectively, and are represented as the corresponding excess volume between the nonexposed level (green) and the three exposed levels (yellow), i.e. the total volume of the yellow boxes. They give rise to the components of the attributable fraction (Eide and Gefeller 1995) of 0, 0.167, 0.067 and 0.300, which sum to the combined attributable fraction of 0.533. Such a component is not an attributable fraction in the classical sense, but is indeed a generalized attributable fraction (Eide and Heuch 2001).
showing the preventive gain in the whole population from eliminating excess risk in the specific cross-classified subpopulation only. An equivalent term is level-specific attributable fraction (Coughlin, Benichou and Weed 1994; Benichou, Chow, McLaughlin, Mandel and Fraumeni Jr 1998; Benichou 2001). The largest potential gain is in the subpopulation exposed to both $E$ and $C$, as clearly shown by the larger volume in Figure 3 (equal to 0.300) between $p(x)$ and $p_0(x)$ for this combination.

One may also illustrate the effect of modifying the joint exposure distribution (Figure 4), or modifying the conditional risk function (Figure 5a), in both cases without removing the excess exposure completely. In Figure 4 the joint distribution of the two dichotomous explanatory variables has been modified to $P^*(C \cap E) = 6/14$, $P^*(C \cap \bar{E}) = 4/14$, $P^*(\bar{C} \cap E) = 3/14$, and $P^*(\bar{C} \cap \bar{E}) = 1/14$, and in Figure 5 the conditional probabilities of disease have been modified to $P^*(D|\bar{C} \cap E) = 2/20$ (unchanged), $P^*(D|C \cap E) = 9/50$, $P^*(D|C \cap \bar{E}) = 11/80$, and $P^*(D|\bar{C} \cap \bar{E}) = 13/40$, both modifications reducing the overall probability of disease by 20% (i.e. the generalized attributable fraction is $\lambda = 0.20$).

4 EXTENDED APPLICATIONS

The method outlined above can be used in any situation with multiple categorical explanatory variables and a dichotomous response variable. It also applies to adjusted attributable fractions (Whittemore 1982), for which the modified risk of disease, $p^*(x)$, reflects changes in exposure for a particular set of explanatory
variables while adjusting for another set. In that case, the surface representing the probability of disease after exposure modification will be constant within categories defined by the latter set, but differ between such categories.

Furthermore, the scaled sample space cubes can be used for displaying excess risk in the context of generalized attributable fractions, and one may animate successive changes of risk levels according to a prespecified plan. Using modern graphical software for three-dimensional plotting, the cubes may be rotated (interactively if one wishes) to possibly give a better impression of various features of the risk and exposure modification. As an example, Figures 5b-d show three rotations of the scaled sample space cube of Figure 5a. For instance, the levels of risk after modification are easier to spot with rotations a and c than with rotation b and maybe d. All three-dimensional figures in this paper were created with Maple 8.00 (Monagan, Geddes, Heal, Labahn, Vorkoetter, McCarron, DeMarco 2002) and can be interactively rotated and manipulated on the screen.

5 **EMPIRICAL EXAMPLE**

For didactical reasons the method above was described for theoretical probability models, but its utility is probably greater for displaying observed data or models estimated from observations. An example of the latter is based on random sample data of the Hordaland study of obstructive lung disease (Bakke, Eide, Hanoa and Gulsvik 1991). In this cross-sectional study of 4270 responders to a mail questionnaire, 396 subjects reported to have chronic cough. In
a stepwise logistic regression analysis of chronic cough the categorical exposure variables residence (urban, rural), occupational dust exposure (no, yes) and smoking habits (never, ex, 1-9, 10-19, 20+ cigarettes/day) were retained in the final model (Eide and Gefeller 1995).

Figure 6 shows a mosaic display for the joint observed distribution of the three explanatory variables. It is seen that the urban and rural subpopulations are approximately of equal size, and that never smoking is slightly more prevalent in the rural areas, but that the distributions of smoking habits are otherwise fairly equal in the two subpopulations. Furthermore, dust exposure follows approximately the same pattern within the smoking categories for rural compared to urban residence, but seems to increase in frequency with increasing smoking levels, although ex-smokers have a level of dust exposure equal to the medium smokers. Figure 6 suggests that dust exposure is associated with smoking and maybe with residence. A backwards loglinear model fitting confirmed this with significant second order terms between dust and smoking (likelihood ratio test: \( p < 0.0001 \)), dust and residence \( (p = 0.0024) \), and smoking and residence \( (p = 0.0001) \), but no three-way interaction \( (p = 0.1429) \). A logistic regression model, as used here to study the prevalence of chronic cough, does, however, implicitly assume the presence of all such interactions (including the three-way interaction) between the explanatory variables (Fienberg 1980).

All attributable fraction estimates based on the logistic model, as computed by Eide and Gefeller (1995), are shown in Figure 7. Figure 8a shows the scaled sample space cube for this model, with the green surface representing the preva-
lence as estimated from the data by the logistic model. It may be noticed that the ex-smokers who are not exposed to dust or gas have a lower prevalence than the never-smokers who are not so exposed. A scaled Venn diagram for the same example was shown by Eide and Heuch (2001).

Figures 8b-d show the excess risk that can be eliminated by removing exposures in the sequence of occupational dust or gas, all smoking, and urban residence, as indicated by the path of whole line arrows in Figure 7. To get a closer view of the excess risk, Figures 8a-d have been zoomed in and the top of the unit cube removed. Figure 8c may be the most interesting as it shows the effect of removing the two exposures judged to be modifiable, which corresponds to reducing the prevalence of chronic cough by 51.2 %. It appears, for instance, that the subpopulation of heavy smokers in urban areas who are occupationally exposed is relatively small (2.3 % of the population), but contributes to the adjusted attributable fraction with 8.2 % due to a very high prevalence (37.1 %) of chronic cough. In contrast, non-occupationally exposed, moderate smokers in urban areas contribute about the same amount (8.1 %), but this is because the subpopulation comprises a substantial proportion of the population (7.6 %), and not because it has a high prevalence (14.7 %). The essential feature of each box in the plot illustrating the potential for disease prevention is the volume, which is clearly shown by the figures. Figure 9 shows a rotated version of Figure 8c with a different light setting to improve on the visual impression. The mosaic display for the joint distribution of the three-exposure variables in Figure 6 was obtained by a rotation looking along the z-axis.
A two-dimensional alternative to the three-dimensional scaled sample space cube is the mosaic display based on all the explanatory variables as well as the response variable, which is introduced as the last variable in the sequence. Figure 10 shows such a display for the chronic cough example of the Hordaland study as produced by the Mondrian data visualization system (freeware from http://www1.math.uni-augsburg.de/Mondrian/index.html). The total area of all blue and green rectangles equals the observed prevalence of chronic cough. Eliminating occupational dust or gas exposure would reduce this prevalence by an amount equal to the green areas, and the fraction of prevalence of chronic cough attributable to occupational dust and gas adjusting for residence (0.168) equals the ratio of the total green areas to the total blue and green areas. One can proceed by modifying the prevalence in all categories of smoking to the level of the never-smokers for each residence stratum. Finally, one can reduce the prevalence in the urban areas to the level of the rural areas. Alternatively, one can draw new mosaic plots for each step based on the modified structure of the exposure variables when one exposure category is eliminated.

The complete two-dimensional mosaic display has the advantage that it gives a more accurate impression of population frequencies and interrelations between the explanatory variables. It also lends itself to a natural representation of adjusted attributable fractions. With an increasing subdivision, however, the rectangles of excess risk become small, scattered and difficult to interpret. The scaled sample space cube is more easily adapted to such situations by removing and zooming in at parts of the display, and it will often show more clearly which
population subgroups can gain most by risk modification.

6 DISCUSSION

Attributable fractions in various forms are being applied to an increasing extent in medical research, especially in epidemiology (Benichou et al. 1998; Eide and Heuch 2001; Uter and Pfahlberg 2001; Gefeller 2001; Land, Vogel and Gefeller 2001). Magnus and Beaglehole (2001) recently showed how proper use of these concepts should change the longstanding professional opinion about one of the world’s great public health challenges, coronary heart disease. Ezzati, Lopez, Rodgers, Vander Hoorn, Murray and the Comparative Risk Assessment Collaborating Group (2002) reported from the Global Burden of Disease 2000 study (http://www.WHO.int/whr) attributable fraction estimates to quantify the contributions of 26 major risk factors to global and regional burden of disease. The distribution of each risk factor was reduced to a counterfactual 'low-risk' distribution, with all the other factors’ distributions remaining the same. In general, the applicability of the attributable fraction should easily be extended to a wider spectrum of scientific fields like environmental research, tort-law and general causal modelling. To take full advantage of new research findings described with these measures it is important to convey their meaning not only to the medical profession and health authorities, but also to statisticians and researchers in other fields, political bodies, media and the general public. Graphic displays which are intuitively easy to understand may be of great help, for in-
stance to illustrate the potential effect of suggested intervention programmes on
a population.

The method of scaled sample space cubes as proposed in this article, makes
possible a dynamic three-dimensional description of the impact of reducing risk
or changing exposure distribution on the prevalence or incidence of disease in
a population. It applies to statistical models for risk with theoretical as well
as empirical foundation, and with single or multiple exposures to be modified.
Although exposure variables are assumed to be categorical in the scaled sam-
ple space cube, continuous variables can easily be accommodated by reasonable
categorization. Indeed, in empirical studies variables are always recorded on
a discrete scale according to the level of measurement accuracy. As for the
scaled Venn diagram (Eide and Heuch 2001), the cube can easily be adapted to
the situation with data from a case-control study, by plotting the odds ratios
along the z-axis, provided that an approximate estimate of the joint exposure
distribution is available.

The construction of a scaled sample space cube depends on the ordering
of the explanatory variables through their specified ordering in the mosaic dis-
play in the base of the cube. This ordering may be chosen quite arbitrarily
and the cube will have a different appearance for each ordering. However, al-
though the rectangles in the mosaic will occupy different positions, they will still
have areas equal to the probabilities or relative frequencies of the various cross-
classifications and, thus, this applies to the volumes of the rectangular boxes
as well. The overall impression of the excess risk given by the cube will be the
same, regardless of the order chosen for the variables in the base mosaic. When a certain stepwise strategy of exposure removal is illustrated it does, however, seem natural to choose the order of variables in the base mosaic as the reverse of the order of exposure removal, as has been the convention applied in this paper.

For the figures in this article special programming in Maple (Monagan et al 2002) was necessary. The technique of drawing polygons in three-dimensional space was applied. In principle, a program is needed that can draw general two-dimensional histograms where the base of the three-dimensional bars can be user specified rectangles and not only regular squares in a grid. Such options would be welcomed in any general statistics package also for other purposes, but are lacking. Interactive rotation of the figures should be provided, for example in the way figures can be rotated in Maple (Monagan et al 2002). Animation facilities further enhance the applicability of such software. Another solution could be to extend existing software for the drawing of mosaic displays of which a brief survey was recently given by Friendly (2002).

Although the response variable in the scaled sample space cube has been assumed to be dichotomous in the presentation above, the cube may also be generalized to the situation with a polytomous response. For instance, one may classify diseased persons with regard to severity of illness and plot the conditional probability of the different classes (severely ill, mildly ill, not ill) on top of each other along the $z$-axis. This may be done for the situation before and after intervention to illustrate the impact on the probabilities of illness from
changing the joint exposure distribution or reducing the harmfulness of various exposures. An example is given in Figure 11 where the diseased set \( D \) in the theoretical example in Figure 1 has been partitioned into two subsets \( D_1 \) (‘mild disease’) and \( D_2 \) (‘severe disease’) represented by yellow and red boxes, respectively.

The scaled sample space cube proposed in this article is a three-dimensional derivative of the scaled Venn diagram of Eide and Heuch (2001) and the mosaic display of Hartigan and Kleiner (1984). Marshall (2001) used the term scaled Venn diagram in a different sense than Eide and Heuch (2001) in his development of the method of scaled rectangle diagrams. One could think of using a scaled rectangle diagram instead of a mosaic display in the base of the scaled sample space cube. The scaled rectangle diagram is restricted to binary variables, however, and it is not always possible to construct a scaled rectangle diagram in which area and probability are exactly congruent (Marshall 2001). The mosaic display without spacings was termed a Mondrian diagram by Theus (1997), and in a similar fashion one could imagine using different colors (or shading) of the rectangles in the base mosaic display instead of a third dimension to represent the values of the conditional disease risk function. The direct interpretation of the box volumes as probabilities or relative frequencies disappears, however, and the exact size of the excess risk is harder to see. Thus, the scaled sample space cube seems to be more attractive for use in the context of attributable fractions. Conversely, the use of coloring or highlighting as used in the mosaic display to show various attributes of the cells in a multiway contingency table (Theus and
Lauer 1999) could be replaced by plotting the attribute in the third dimension. For instance the fit of a loglinear model could be displayed by plotting the negative and positive residuals for each cell orthogonally to the plane of the mosaic display in both directions. Such extensions could add a new chapter to the long history of development of the mosaic display (Friendly 2002).

REFERENCES


Epidemiology, 5, 645-655.


**Figure Captions.**

Figure 1. Scaled sample space cube for example in text with exposures C and E.

Figure 2. Scaled Venn diagram for same example as shown in Figure 1.

Figure 3. Scaled sample space cube showing original risk of disease, $p(x)$, and risk of disease after eliminating all excess risk, $p_0(x)$.

Figure 4. Excess risk from modifying the exposures’ joint distribution, i.e. the volume of the yellow boxes.

Figure 5. Same amount of excess risk (yellow boxes) as in Figure 4, but from reducing the conditional risk $P(D|\cdot)$ from $p(x)$ (yellow top) to $p^*(x)$ (green top).

Various viewing angles.

Figure 6. Mosaic display for the explanatory variables $X_1$: Residence, $X_2$: Smoking, and $X_3$: Occupational dust or gas exposure.

Figure 7. Logistic model adjusted attributable fraction percents for the prevalence of chronic cough in ordered stepwise strategies.

Figure 8. Scaled sample space cubes for marked preventive strategy of Figure 7. Categories for residence, smoking and dust or gas exposure are defined as in Figure 6.

Figure 9. Excess prevalence due to occupational dust or gas exposure and smoking adjusted for residence. Rotated version of Figure 8c with extra light setting.

Figure 10. Mosaic display of the four-dimensional contingency table given the specified ordering. Green areas equal excess prevalence of chronic cough.
due to occupational dust or gas exposure. Red lines are empty boxes (no excess prevalence). Screen output as produced by the Mondrian data visualization system with text an blue color added.

Figure 11. Scaled sample space cube for the theoretical example where disease \( (D) \) has been further divided into \( D_1 \): mild disease (yellow boxes) and \( D_2 \): severe disease (red boxes), and no disease \( \overline{D} = D_0 \) (invisible).
Probability of disease (D) dependent on two exposures
Excess risk for D due to C or E
Effect of modifying jointly C and E distribution
Rotation c.
a. Prevalence of chronic cough by logistic regression model
b. Excess prevalence of chronic cough due to dust only
c. Excess prevalence of chronic cough due to smoking or dust...
Excess prev. of chronic cough due to smoking, dust or residence.
Excess prevalence of chronic cough due to smoking or dust