Michael Friendly **York University**

Conceptual Models for Visualizing Contingency Table Data **Feb 1996**
Draft 1

Abstract

This paper reviews several graphical methods for visualizing n-way contingency table data: parquet diagrams, mosaic displays and fourfold displays, which all display counts by area or observation density, and asks why visualization methods for categorical data are relatively ill-developed, while analogous methods for quantitative data are both well-developed and highly used.

One answer is suggested by a physical model for categorical data which likens categorical observations to gas molecules in pressure chambers, and provides a conceptual model for the use of area or observation density to display frequency data. The physical model provides concrete interpretations of a surprising number of results in the analysis of categorical data, from degrees of freedom and likelihood ratios, to iterative proportional fitting and Newton-Raphson iteration.

Introduction

For some time I have wondered why graphical methods for categorical data are so poorly developed and little used compared with methods for quantitative data. For quantitative data, graphical methods are commonplace adjuncts to all aspects of statistical analysis, from the basic display of data in a scatterplot, to diagnostic methods for assessing assumptions and finding transformations, to the final presentation of results. In contrast, graphical methods for categorical data are still in infancy. There are not many methods, those that are available in the literature are not accessible in common statistical software, and consequently they are not widely used.

What has made this contrast puzzling is the fact that the statistical methods for categorical data, are in many respects discrete analogs of corresponding methods for quantitative data: log-linear models and logistic regression, for example, are such close parallels of analysis of variance and regression models that they are all special cases of generalized linear models.

Several possible explanations for this apparent puzzle may be suggested. First, it may just be that those who have worked with and developed methods for categorical data are just more comfortable with tabular data, or that frequency tables, representing sums over all cases in a dataset are more easily apprehended in tables than quantitative data. Second, it may be argued that graphical methods for quantitative data are easily generalized so, for example, the scatterplot for two variables provides the basis for visualizing any number of variables in a scatterplot matrix; available graphical methods for categorical data tend to be more specialized. However, a more fundamental reason may be, as I will try to show here, that quantitative data display relies on e well-known natural visual mapping where a magnitude is depicted by length or position along a scale; for categorical data, it will be seen that a count is more naturally displayed by an area or by the visual density of an area.

Some graphical methods for contingency tables

Several schemes for representing contingency tables graphically are based on the fact that when the row and column variables are independent, the estimated expected frequencies, m_{ij}, are products of the row and column totals (divided by the grand total). Then, each cell can be represented by a rectangle whose area shows the cell frequency, n_{ij}, or deviation from independence.

Sieve diagrams

Table 1 shows data on the relation between hair color and eye color among 592 subjects (students in a statistics course) collected by Snee (1974). The Pearson χ^2 for these data is 138.3 with 9 degrees of freedom, indicating substantial departure from independence. The question is how to understand the *nature* of the association between hair and eye color.

For any two-way table, the expected frequencies under independence can be represented by rectangles whose widths are proportional to the total frequency in each column, n_{+j}, and whose heights are proportional to the total frequency in each row, n_{i+}; the area of each rectangle is then proportional to m_{ij}. Figure 1 shows the expected frequencies for the hair and eye color data.

Riedwyl and Schüpbach (1983, 1994) proposed a ***sieve diagram*** (later called a ***parquet diagram***) based on this principle. In this display the area of each rectangle is proportional to expected frequency and observed frequency is shown by the number of squares in each rectangle. Hence, the difference between observed and expected frequency appears as the density of shading, using color to indicate whether the deviation from

Table 1: Hair-color eye-color data

Eye Color	Hair Color BLACK	BROWN	RED	BLOND	Total
Brown	68	119	26	7	220
Blue	20	84	17	94	215
Hazel	15	54	14	10	93
Green	5	29	14	16	64
Total	108	286	71	127	592

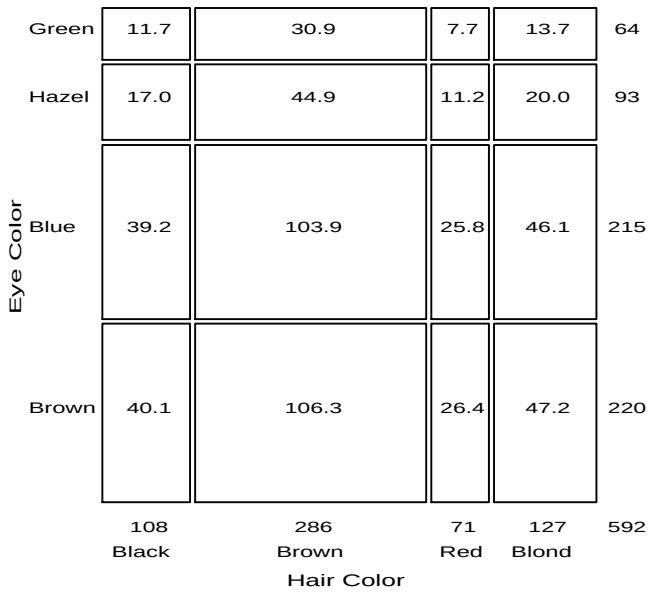

Figure 1: Expected frequencies under independence

independence is positive or negative. (In monochrome versions, positive deviations are shown by solid lines, negative by broken lines.) The sieve diagram for hair color and eye color is shown in Figure 2.

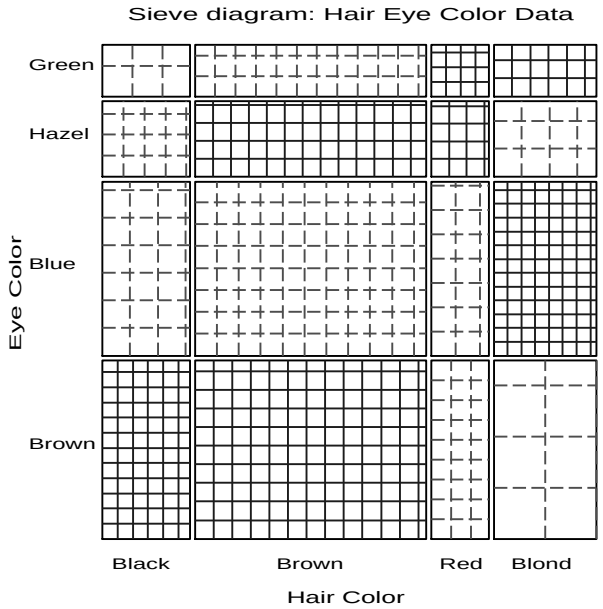

Figure 2: Sieve diagram for hair-eye data

Mosaic displays for n-way tables

The mosaic display, proposed by Hartigan & Kleiner (1981), and extended by Friendly (1994a) represents the counts in a contingency table directly by tiles whose area is proportional to the cell frequency. This display generalizes readily to *n*-way tables and can be used to display the residuals from various log-linear models.

One form of this plot, called the ***condensed mosaic display***, is similar to a divided bar chart. The width of each column of tiles in Figure 3 is proportional to the marginal frequency of hair colors. Again, the area of each box is proportional to the cell frequency, and complete independence is shown when the tiles in each row all have the same height.

Enhanced mosaics: The enhanced mosaic display (Friendly, 1992b, 1994a) achieves greater visual impact by using color and shading to reflect the size of the residual from independence and by reordering rows and columns to make the pattern more coherent. The resulting display shows both the observed frequencies and the pattern of deviations from a specified model.

Figure 4 gives the extended the mosaic plot, showing the standardized deviation from independence, $d_{ij} = (n_{ij} - m_{ij)} / \sqrt{m_{ij}}$ by the color and shading of each rectangle: cells with positive deviations are drawn black, outlined with solid lines, with shading slanted from upper left to lower right (NE to SW); negative deviations are drawn red, outlined with broken lines and shaded SE-NW. The absolute value of the deviation is portrayed by shading density: cells with absolute values less than 2 are empty; cells with $|d_{ij}| \geq 2$ are filled; those with $|d_{ij}| \geq 4$ are filled with a darker pattern. Under the assumption of independence, these values roughly correspond to two-tailed probabilities $p < .05$ and $p < .0001$ that a given value of $|d_{ij}|$ exceeds 2 or 4.

When the row or column variables are unordered, we are also free to rearrange the corresponding categories in the plot to help show the nature of association. For example, in Figure 4, the eye color categories have been permuted so that the deviations from independence have an opposite-corner pattern, with positive

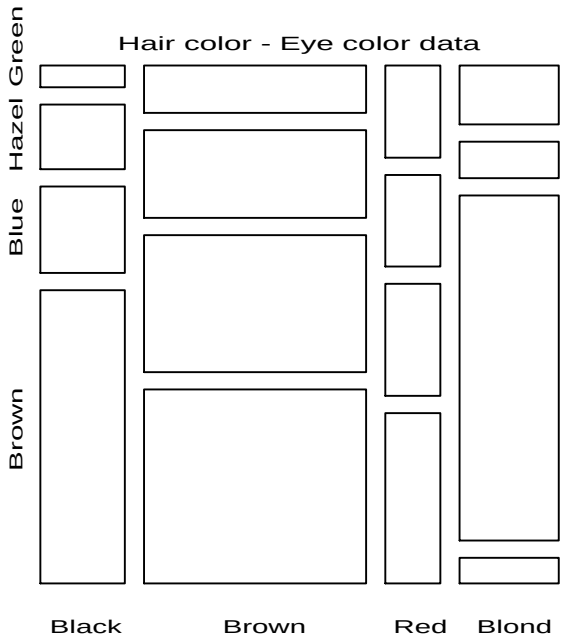

Figure 3: Condensed column proportion mosaic

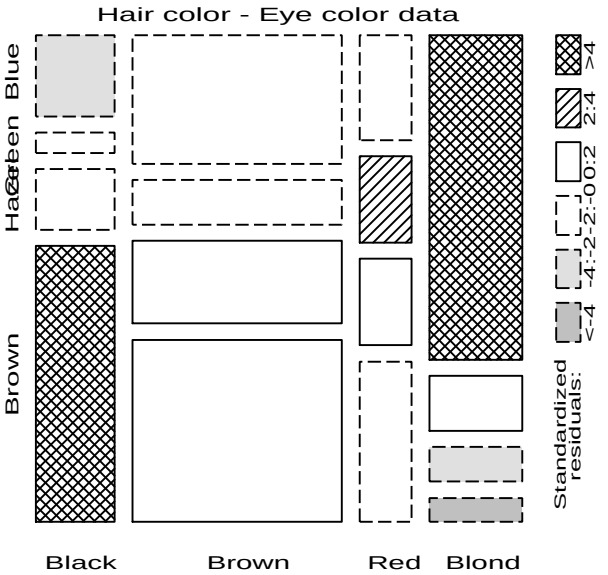

Figure 4: Enhanced mosaic, reordered and shaded

values running from SW to NE corners, negative values along the opposite diagonal. Coupled with size and shading of the tiles, the excess in the black-brown and blond-blue cells, together with the underrepresentation

of brown-haired blonds and people with black hair and blue eyes is now quite apparent. Though the table was reordered based on the d_{ij} values, both dimensions in Figure 4 are ordered from dark to light, suggesting an explanation for the association. (In this example the eye-color categories could be reordered by inspection. A general method (Friendly, 1994a) uses category scores on the largest correspondence analysis dimension.)

Multi-way tables: Like the scatterplot matrix for quantitative data, the mosaic plot generalizes readily to the display of multi-dimensional contingency tables. Imagine that each cell of the two-way table for hair and eye color is further classified by one or more additional variables—sex and level of education, for example. Then each rectangle can be subdivided horizontally to show the proportion of males and females in that cell, and each of those horizontal portions can be subdivided vertically to show the proportions of people at each educational level in the hair-eye-sex group.

Fitting models: When three or more variables are represented in the mosaic, we can fit several different models of independence and display the residuals from that model. We treat these models as null or baseline models, which may not fit the data particularly well. The deviations of observed frequencies from expected, displayed by shading, will often suggest terms to be added to to an explanatory model that achieves a better fit.

- ***Complete independence***: The model of complete independence asserts that all joint probabilities are products of the one-way marginal probabilities:

$$\pi_{ijk} = \pi_{i++}\ \pi_{+j+}\ \pi_{++k} \tag{1}$$

  for all i,j,k in a three-way table. This corresponds to the log-linear model [A] [B] [C]. Fitting this model puts all higher terms, and hence all association among the variables, into the residuals.
- ***Joint independence***: Another possibility is to fit the model in which variable C is jointly independent of variables A and B,

$$\pi_{ijk} = \pi_{ij+}\ \pi_{++k} \quad . \tag{2}$$

  This corresponds to the log-linear model [AB] [C]. Residuals from this model show the extent to which variable C is related to the combinations of variables A and B but they do not show any association between A and B.

For example, with the data from Table 1 broken down by sex, fitting the model [HairEye][Sex] allows us to see the extent to which the joint distribution of hair-color and eye-color is associated with sex. For this model, the likelihood-ratio G^2 is 29.35 on 15 df ($p = .015$), indicating some lack of fit. The three-way mosaic, shown in Figure 5, highlights two cells: males are underrepresented among people with brown hair and brown eyes, and overrepresented among people with brown hair and blue eyes. Females in these cells have the opposite patterns, with residuals just shy of ±2. The d_{ij}^2 for these four cells account for 15.3 of the χ^2 for the model [HairEye] [Sex]. Hence, except for these cells hair color and eye color appear unassociated with sex.

Fourfold Display

A third graphical method based on the use of area as the visual mapping of cell frequency is the "fourfold display" (Friendly, 1994b, 1994c) designed for the display of $2 \times 2 \times k$ tables. In this display the frequency n_{ij} in each cell of a fourfold table is shown by a quarter circle, whose radius is proportional to $\sqrt{n_{ij}}$, so the area is proportional to the cell count.

For a single 2×2 table the fourfold display described here also shows the frequencies by area, but scaled in a way that depicts the sample odds ratio, $\hat{\theta} = (n_{11} / n_{12}) / (n_{21} / n_{22})$. An association between the variables

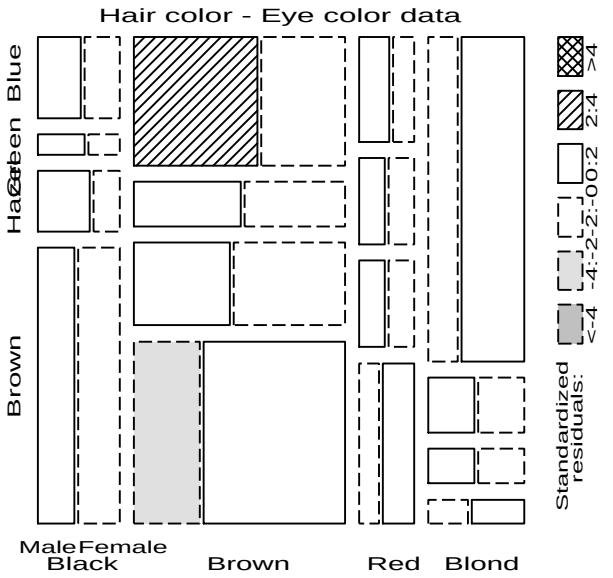

Figure 5: Mosaic display for hair color, eye color, and sex

($\theta \neq 1$) is shown by the tendency of diagonally opposite cells in one direction to differ in size from those in the opposite direction, and the display uses color or shading to show this direction. Confidence rings for the observed θ allow a visual test of the hypothesis H_0: $\theta = 1$. They have the property that the rings for adjacent quadrants overlap *iff* the observed counts are consistent with the null hypothesis.

As an example, Figure 6 shows aggregate data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex. At issue is whether the data show evidence of sex bias in admission practices (Bickel et al., 1975). The figure shows the cell frequencies numerically in the corners of the display. Thus there were 2691 male applicants, of whom 1193 (44.4%) were admitted, compared with 1855 female applicants of whom 557 (30.0%) were admitted. Hence the sample odds ratio, Odds (Admit|Male) / (Admit|Female) is 1.84 indicating that males were almost twice as likely to be admitted.

The frequencies displayed graphically by shaded quadrants in Figure 6 are not the raw frequencies. Instead, the frequencies have been standardized (by iterative proportional fitting) so that all table margins are equal, while preserving the odds ratio. Each quarter circle is then drawn to have an area proportional to this standardized cell frequency. This makes it easier to see the association between admission and sex without being influenced by the overall admission rate or the differential tendency of males and females to apply. With this standardization the four quadrants will align when the odds ratio is 1, regardless of the marginal frequencies.

The shaded quadrants in Figure 6 do not align and the 99% confidence rings around each quadrant do not overlap, indicating that the odds ratio differs significantly from 1. The width of the confidence rings gives a visual indication of the precision of the data.

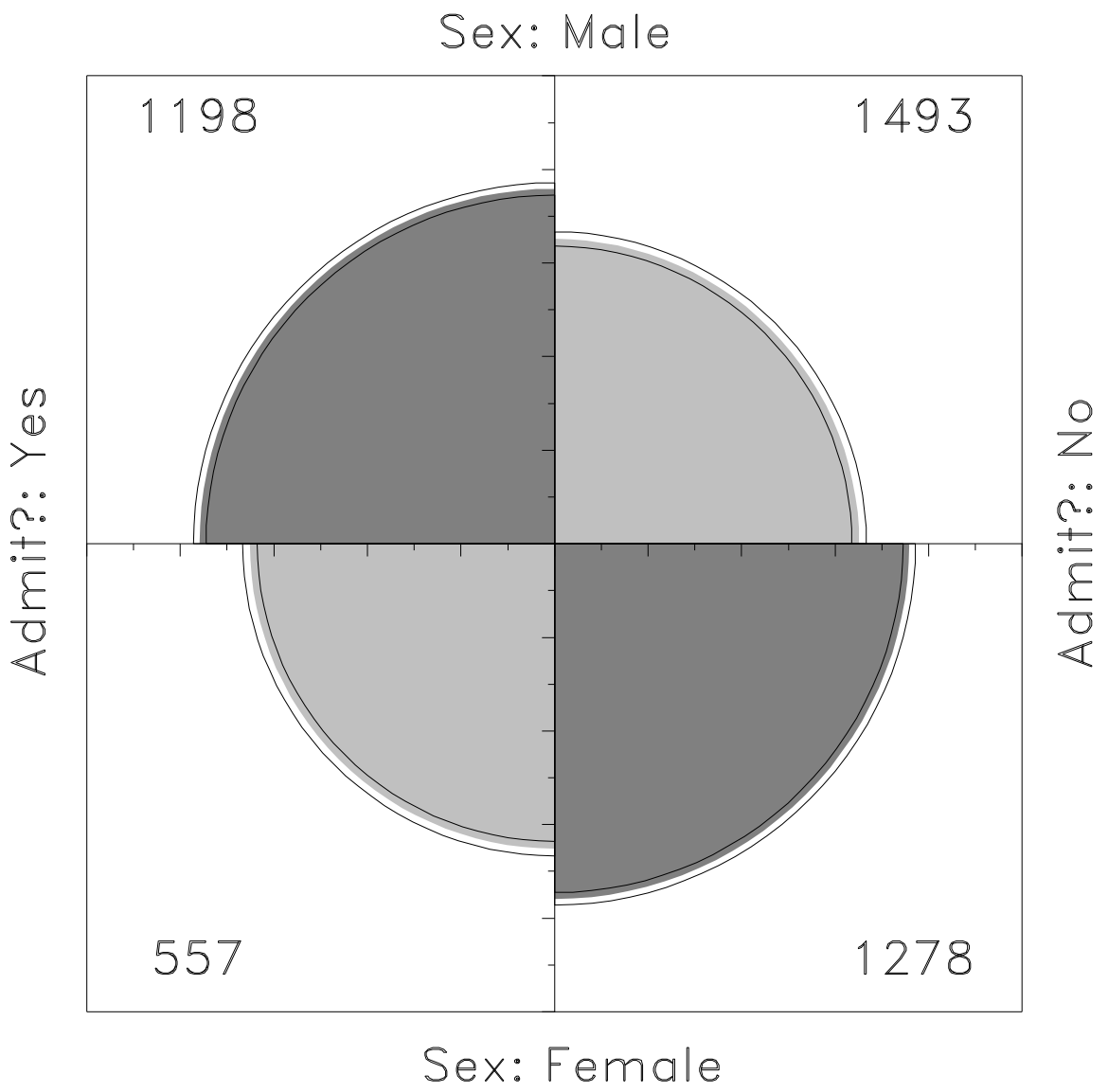

Figure 6: Fourfold display for Berkeley admissions data: Evidence for sex bias?. The area of each shaded quadrant shows the frequency, standardized to equate the margins for sex and admission. Circular arcs show the limits of a 99% confidence interval for the odds ratio.

Multiple strata: In the case of a $2 \times 2 \times k$ table, the last dimension typically corresponds to "strata" or populations, and it is typically of interest to see if the association between the first two variables is homogeneous across strata. The fourfold display allows easy visual comparison of the pattern of association between two dichotomous variables across two or more populations.

The admissions data shown in Figure 6 were obtained from a sample of six departments, so to determine the source of the apparent sex bias in favor of males, we make a new plot, Figure 7, stratified by department. The departments are labelled so that the overall acceptance rate is highest for Department A and decreases steadily to Department F. Again each panel is standardized to equate the marginals for sex and admission. This standardization also equates for the differential total applicants across departments, facilitating visual comparison.

Surprisingly, Figure 7 shows that, for five of the six departments, the odds of admission is approximately the same for both men and women applicants. Department A appears to differs from the others, with women approximately 2.86 (= (313/19) ÷ (512/89)) times as likely to gain admission. This appearance is confirmed by the confidence rings, which in Figure 7 are *joint* 99% intervals for θ_c, $c = 1, \ldots, k$.

This result, which contradicts the display for the aggregate data in Figure 6, is a classic example of Simpson's paradox. The resolution of this contradiction can be found in the large differences in admission rates among departments. Men and women apply to different departments differentially, and in these data women apply in larger numbers to departments that have a low acceptance rate. The aggregate results are misleading because they falsely assume men and women are equally likely to apply in each field. (This explanation ignores the possibility of structural bias against women, e.g., lack of resources allocated to departments that attract women applicants.)

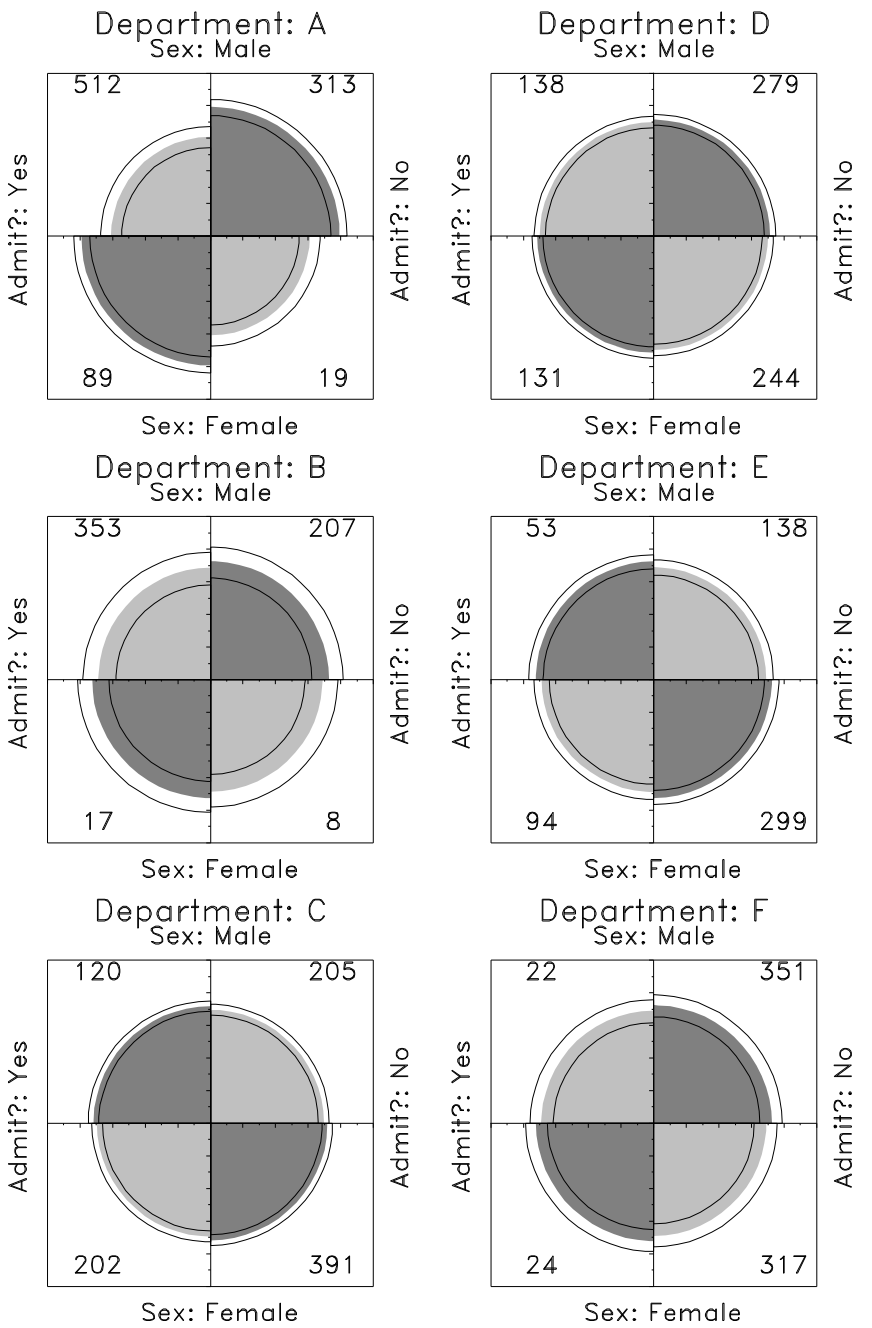

Figure 7: Fourfold display of Berkeley admissions, by department. In each panel the confidence rings for adjacent quadrants overlap if the odds ratio for admission and sex does not differ significantly from 1. The data in each panel have been standardized as in Figure 6.

Conceptual Models for Visual Displays

Visual representation of data depends fundamentally on an appropriate visual scheme for mapping numbers into graphic patterns (Bertin 1983). The widespread use of graphical methods for quantitative data relies on the availability of a natural visual mapping: magnitude can be represented by length, as in a bar chart, or by position along a scale, as in dot charts and scatterplots. One reason for the relative paucity of graphical methods for categorical data may be that a natural visual mapping for frequency data is not so apparent. And, as I have just shown, the mapping of frequency to area appears to work well for categorical data.

Closely associated with the idea of a visual metaphor is a conceptual model that helps you interpret what is shown in a graph. A good conceptual model for a graphical display will have deeper connections with underlying statistical ideas as well. In this section I will describe conceptual models for both quantitative and frequency data which have these properties and help to elucidate the differences between their graphical displays. The discussion borrows from Sall (1991a, 1991b), Farebrother (1987) and Friendly (1995).

Quantitative Data

The simplest conceptual model for quantitative data is the balance beam (Figure 8), often used in introductory statistics texts to illustrate the sample mean as the point along an axis where the positive and negative deviations balance.

Springs: A more powerful model (Sall, 1991a) likens observations to fixed points connected to a moveable junction by springs of equal spring constant, $k \sim 1 / \sigma$, as shown in Figure 9. By this model, each observation exerts a force proportional to $(X - \mu) / \sigma$ on the junction, and the sample mean is seen as the point which not only balances the forces, but also minimizes the total potential energy in the system.

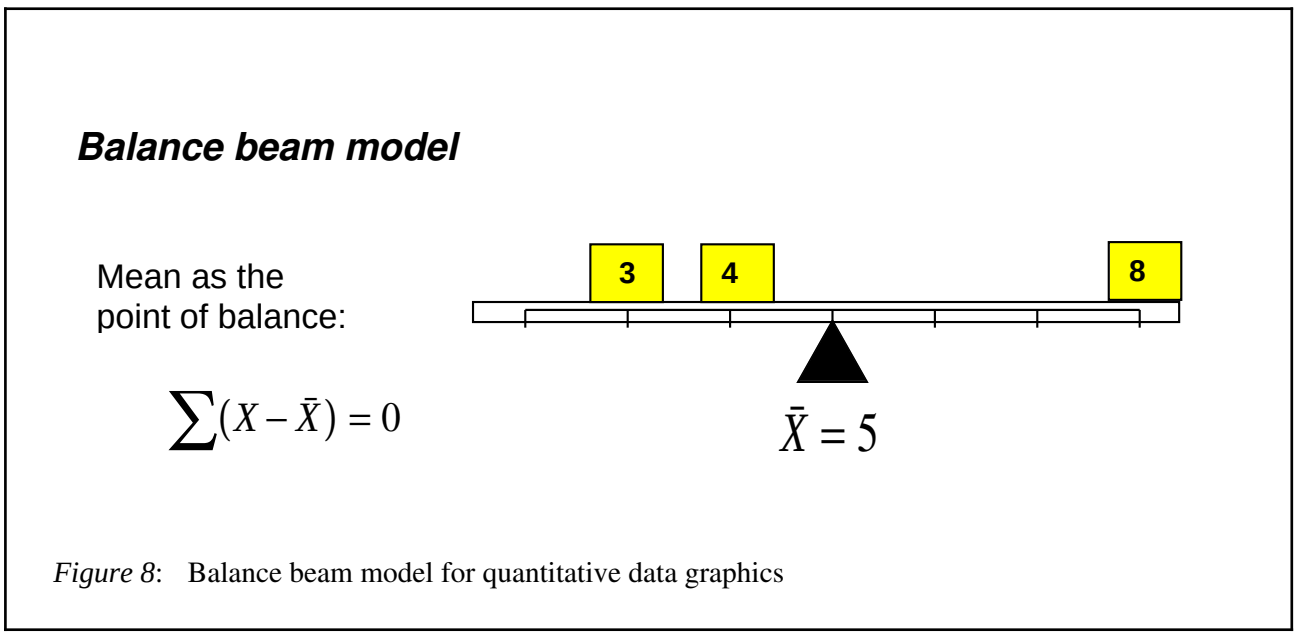

Figure 8: Balance beam model for quantitative data graphics

The spring model is more powerful because it provides a basis for understanding a wide class of both graphical displays and statistical principles for quantitative data. For example, least square regression can be

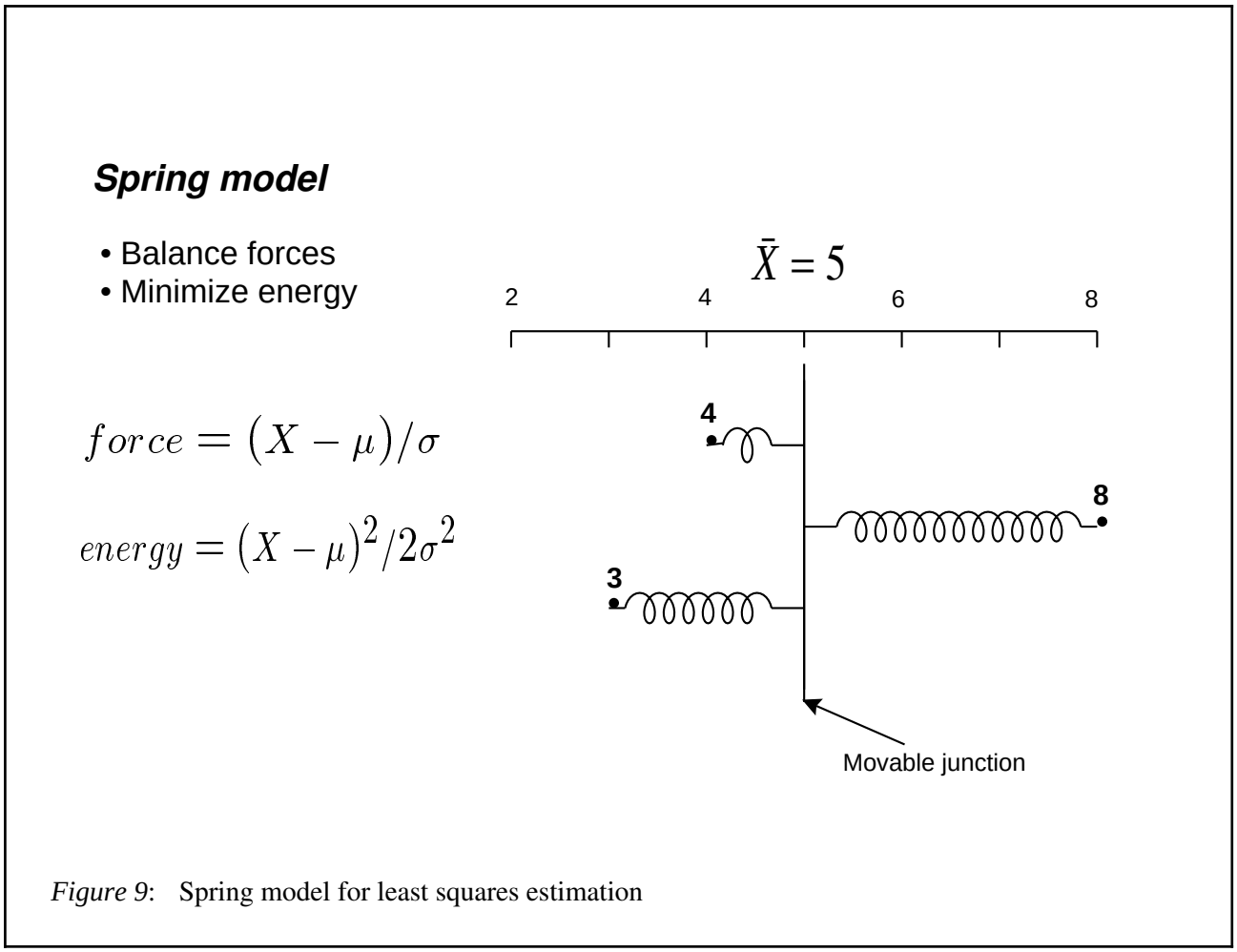

Figure 9: Spring model for least squares estimation

represented as shown in Figure 10, where the points are again fixed and attached to a moveable rod by unit length, equally stiff springs. If the springs are constrained to be kept vertical, the rod, when released, moves to the position of balance and minimum potential energy, the least square solution. The normal equations,

$$\Sigma e_i = \sum_{i=1}^{n} (y_i - a - bx_i) = 0 \tag{3}$$

$$\Sigma x_i\, e_i = \sum_{i=1}^{n} (y_i - a - bx_i)\, x_i = 0 \tag{4}$$

are seen as conditions that (Eq. (3)) the vertical forces balance, and (Eq. (4)) the rotational moments about the intercept $(0, a)$ balance. Letting $\mathbf{X} = \{\mathbf{1}, \mathbf{x}\}$, the normal equations provide the derivation $\mathbf{X}'\,\boldsymbol{e} = \mathbf{X}'(\boldsymbol{y} - \mathbf{X}\boldsymbol{b}) = 0 \Rightarrow \mathbf{X}'\boldsymbol{y} = \mathbf{X}'\mathbf{X}\boldsymbol{b} \Rightarrow \boldsymbol{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$, but springs get it right without inverting a matrix.

Neat explanations by springs: The appeal of the spring model lies in the intuitive explanations it provides for many statistical phenomena and the understanding it can bring to our perception of graphical displays. Without much explanation here (but see Sall, 1991a and Farebrother, 1987) I will simply list some of these.

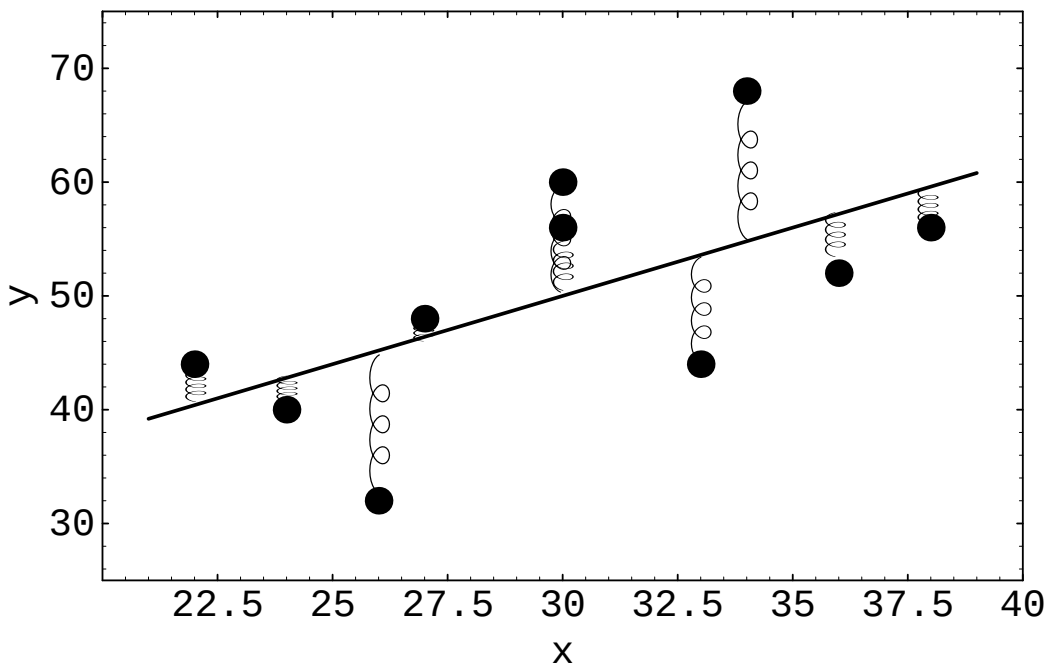

Figure 10: Spring model for least squares regression. Fixed data points are connected to a moveable rod by springs constrained to remain vertical. The least squares line is the position of balanced forces.

- ***least squares*** $\longleftrightarrow$ minimum energy
- ***orthogonality*** $\longleftrightarrow$ balancing of forces
- ***hypothesis testing***: To test the hypothesis H_0: $\beta = 0$, simply force the rod to the horizontal position, measuring the additional energy required to make it obey the hypothesis (Figure 11). This is the regression sum of squares, $SSR(X)$.
- ***sample size and power***: Adding more points (more springs) causes the position of the rod (and thus the coefficient estimates) to be held more tightly. It is therefore easier to reject the null hypothesis.
- ***error variance***: The spring constant is inversely proportional to σ, so smaller error variance means stiffer springs, which require more force to impose the null hypothesis.
- ***outliers***: Points far from the regression line pull the line towards themselves in proportion to the square of their vertical distance from the line.
- ***leverage***: Observations exert a force on the regression line proportional to the square of their horizontal distance from $\bar{x}$. Influential observations are those with
- ***multiple regression***: With two (or more) predictors, fit a plane (or hyperplane) to points with vertical springs.
- ***partial F tests***: With two predictors, fit a plane to (X_1,X_2). How much more energy is required to make the plane flat in the x_2 direction? This is the extra sum of squares, $SSR(X_2 \mid X_1)$
- ***robust estimation***: Use springs which deform (so they exert little or no force) when stretched past some limit.

- ***smoothing splines***: Make the "regression line" from a material which is flexible but stiff.
- ***principal components***: If we relax the restriction that the springs are constrained to remain vertical, the rod moves to the position of the first principal component (Figure 12).

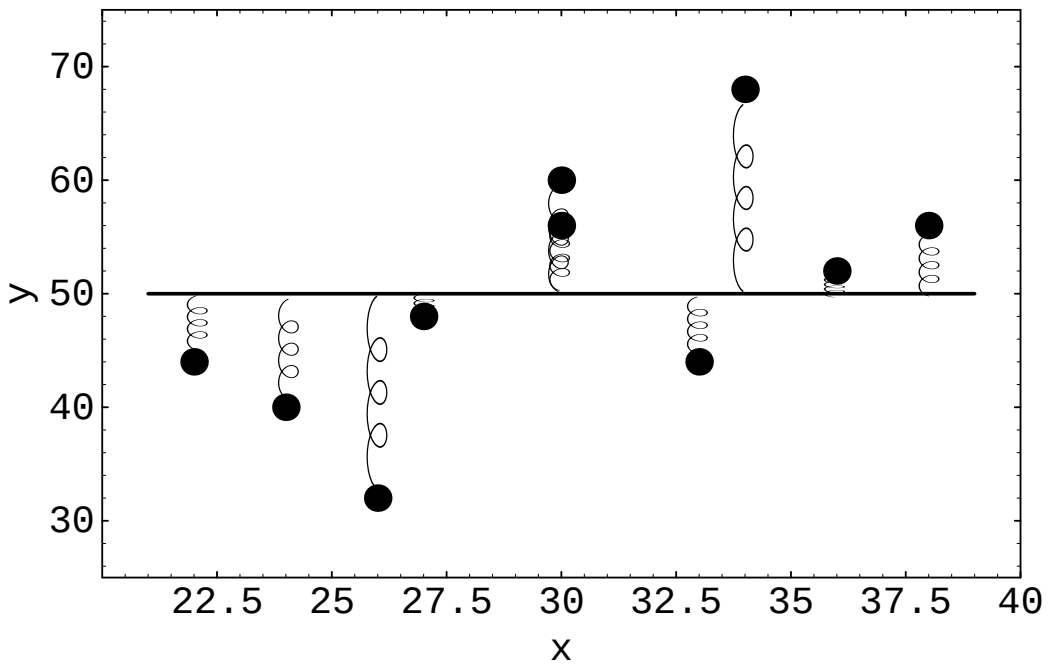

Figure 11: Testing a hypothesis by springs. The regression sum of squares for testing H_0: $\beta = 0$ measures the additional energy required to force the rod to be horizontal.

Categorical Data

For categorical data, we need a visual analog for the sample frequency in *k* mutually exclusive and exhaustive categories. Consider first the one-way marginal frequencies of hair color from Table 2.

Urn model: The simplest physical model represents the hair color categories by urns containing marbles representing the observations (Figure 13). This model is sometimes used in texts to describe multinomial sampling, and provides a visual representation that equates the count n_i with the area filled in each urn, as in the familiar bar chart. (When the urns are of equal width, count is also reflected by height, but in the general case, count is proportional to area.) However, the urn model is a static one and provides no further insights. It does not relate to the concept of likelihood or to the constraint that the probabilities sum to 1.

Figure 13 about here

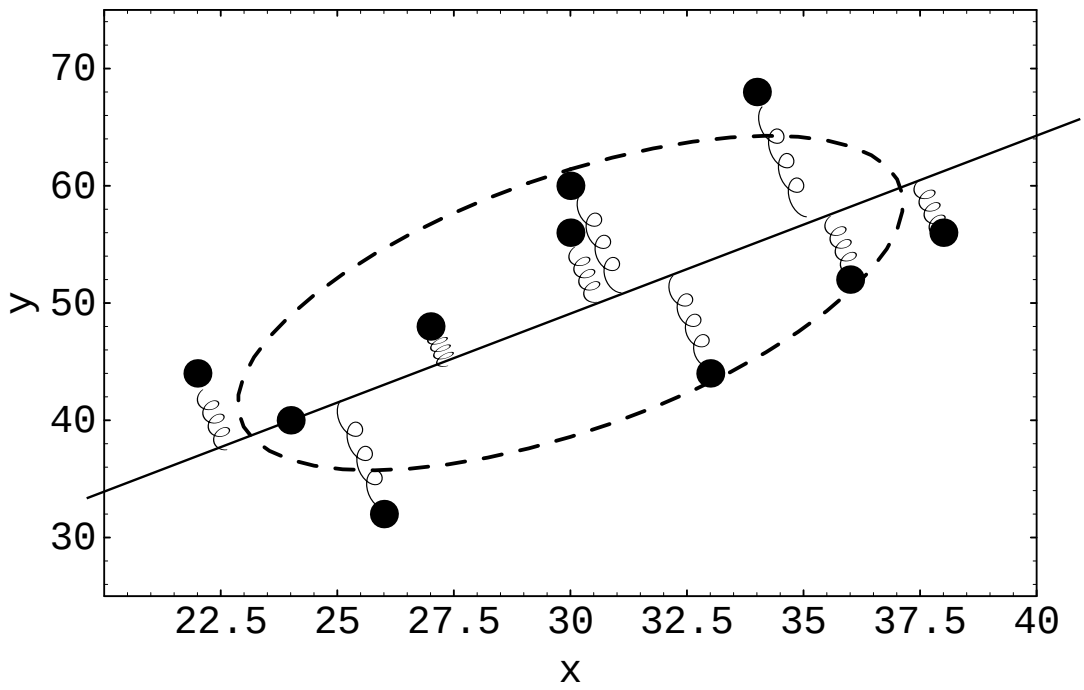

Figure 12: Principal components by springs

Pressure and Energy: A dynamic model gives each observation a force (Figure 14). Consider the observations in a given category (red hair, say) as molecules of an ideal gas confined to a cylinder whose volume can be varied with a movable piston (Sall, 1991b), set up so that a probability of 1.0 corresponds to ambient pressure, with no force exerted on the piston. An actual probability of red hair equal to p means that the same number of observations are squeezed down to a chamber of height p. By Boyle's law (that pressure × volume is constant) the pressure is proportional to $1 / p$. In the figure, pressure is shown by *observation density*, the number of observations per unit area. Hence, the graphical metaphor is that a count can be represented visually by observation density when the count is fixed and area is varied (or by area when the observation density is fixed as in Figure 13.)

The work done on the gas (or potential energy imparted to it) by compressing a small distance δy is the force on the piston times δy, which equals the pressure times the change in volume. Hence, the potential energy of a gas at height=p is $\int_p^1 (1 / y)\, dy$, which is $-\log(p)$, so the energy in this model corresponds to negative log likelihood.

Figure 14 about here

Fitting probabilities: Minimum energy, balanced forces: Maximum likelihood estimation means literally finding the values, $\hat{\pi}_i$, of the parameters under which the observed data would have the highest probability of occurrence. We take derivatives of the (log-) likelihood function with respect to the parameters, set these to zero, and solve:

$$\frac{\partial \log L}{\partial \pi_i} = 0 \quad \Rightarrow \quad \frac{n_1}{\pi_1} = \frac{n_2}{\pi_2} = \ ... \ = \frac{n_c}{\pi_c} \quad \Rightarrow \quad \hat{\pi}_i = \frac{n_i}{n} = p_i$$

As with the spring model, setting derivatives to zero means minimizing the potential energy; the maximum likelihood solution simply sets parameter values equal to corresponding sample quantities, where the forces are also balanced.

In the mechanical model (Figure 15) this corresponds to stacking the gas containers with movable partitions between them, with one end of the bottom and top containers fixed at 0 and 1. The observations exert pressure on the partitions, the likelihood equations are precisely the conditions for the forces to balance, and the partitions move so that each chamber is of size $p_i = n_i / n$. Each chamber has potential energy of $-\log p_i$, and the total energy, $-\sum_i^c n_i \log p_i$ is minimized. The constrained top and bottom force the probability estimates to sum to 1, and the number of movable partitions is literally and statistically the degrees of freedom of the system.

Figure 15 about here

Testing a hypothesis: This mechanical model also explains how we test hypotheses about the true probabilities (Figure 16). To test the hypothesis that the four hair color categories are equally probable, $H_0: \ \pi_1 = \pi_2 = \pi_3 = \pi_4 = {}^1/_4$ simply force the partitions to move to the hypothesized values and measure how much energy is required to force the constraint. Some of the chambers will then exert more pressure, some less than when the forces are allowed to balance without these additional restraints. The change in energy in each compartment is then $-(\log p_i - \log \pi_i) = -\log(p_i / \pi_i)$, the change in negative log-likelihood. Sum these up and multiply by 2 to get the likelihood ratio G^2.

Figure 16 about here

The pressure model also provides simple explanations of other results. For example, increased sample size increases power, because more observations means more pressure in each compartment, so it takes more energy to move the partitions and the test is sensitive to smaller differences between observed and hypothesized probabilities.

Multi-way Tables: The dynamic pressure model extends readily to multi-way tables. For a two-way table of hair color and eye color, partition the sample space according to the marginal proportions of eye color, and then partition the observations for each eye color according to hair color as before (Figure 17). Within each column the forces balance as before, so that the height of each chamber is n_{ij} / n_{i+}. Then the area of each cell is proportional to the MLE of the cell probabilities, $(n_{i+} / n) \ \ (n_{ij} / n_{i+}) = n_{ij} / n = p_{ij}$,, which again is the sample cell proportion.

Figure 17 about here

For a three-way table, the physical model is a cube with its third dimension partitioned according to conditional frequencies of the third variable, given the first two. If the third dimension is represented instead by partitioning a two-dimensional graph, the result is the mosaic display.

Testing Independence: For a two-way table of size $I \times J$ independence is formally the same as the hypothesis that conditional probabilities (of hair color) are the same in all strata (eye colors). To test this hypothesis, force the partitions to align and measure the total additional energy required to effect the change (Figure 18). The degrees of freedom for the test is again the number of movable partitions, (I-1)(J-1).

Figure 18 about here

Each log-linear model for three-way tables can be interpreted analogously. For example, the log-linear model [A] [B] [C] (complete independence), corresponds to the cube in which all chambers are forced to conform to the one way marginals, $\pi_{ijk} = \pi_{i++}\ \pi_{+j+}\ \pi_{++k}$ for all i,j,k. G^2 is again the total additional energy required to move the partitions from their positions in the saturated model in which the volume of each cell is $p_{ijk} = n_{ijk} / n$ (so the pressures balance), to the positions where each cell is a cube of size $\pi_{i++} \times \pi_{+j+} \times \pi_{++k}$. Other models have a similar representation in the pressure model.

Iterative Proportional Fitting: For three-way (and higher) tables some log-linear models have direct estimates of expected cell frequencies. The cases where direct estimates exist are analogous to the two-way case, where the estimates under the hypothesized model are products of the sufficient marginals. Here we see that the partitions in the observation space can be moved directly in planar slices to their positions under the hypothesis, so that iteration is unnecessary.

When direct estimates do not exist, the MLEs can be estimated by iterative proportional fitting (IPF). This process simply matches the partitions corresponding to each of the sufficient marginals of the fitted frequencies to the same marginals of the data. For example, for the log-linear model [AB] [BC] [AC], the sufficient statistics are $\{n_{ij+}\}$, $\{n_{i+k}\}$, and $\{n_{+jk}\}$. The conditions that the fitted margins must equal these observed margins from are

$$\frac{n_{ij+}}{\hat{m}_{ij+}} = \frac{n_{i+k}}{\hat{m}_{i+k}} = \frac{n_{+jk}}{\hat{m}_{+jk}} = 1 \quad , \tag{5}$$

which is equivalent to balancing the forces in each fitted marginal. The steps in IPF follow directly from Equation (5). For example, the first step in cycle t+1 of IPF matches the frequencies in the [AB] marginal table,

$$\hat{m}_{ijk}^{(t+1)} = \hat{m}_{ijk}^{(t)} \left(\frac{n_{ij+}}{\hat{m}_{ij+}^{(t)}} \right) \quad , \tag{6}$$

which makes the forces balance when Equation (6) is summed over variable C: $\hat{m}_{ij+}^{(t+1)} = \hat{n}_{ij+}$. The other steps in each cycle make the forces balance in the [BC], and [AC] margins.

The iterative process can be shown visually (Friendly, 1995), in a way that is graphically exact, by drawing chambers whose area is proportional to the fitted frequencies, $\hat{m}_{ijk}$, and which are filled with a number of points equal to the observed n_{ijk} Such a figure will then show equal density of points in cells which are fit well, but relatively high or low density where $n_{ijk} > \hat{m}_{ijk}$ or $n_{ijk} < \hat{m}_{ijk}$, respectively. The IPF algorithm can in fact be animated, by drawing one such frame for each step in the iterative process. When this is done, it is remarkable how quickly IPF converges, at least for small tables.

Likewise, numerical methods for minimizing the negative log likelihood directly can also be interpreted in terms of the dynamic model (Farebrother, 1988; Friendly, 1995). For example, in steepest descent and Newton-Raphson iteration, the update step changes the estimated model parameters $\beta^{(t+1)}$ in proportion to the score vector $\boldsymbol{f}^{(t)}$ of derivatives of the likelihood function, $\boldsymbol{f}^{(t)} = \partial \log L / \partial \beta = \mathbf{X}' (\boldsymbol{n} - \boldsymbol{m}^{(t)})$ to give $\beta^{(t+1)} = \beta^{(t)} + \lambda f^{(t)}$. But $\boldsymbol{f}^{(t)}$ is just the vector of forces in the mechanical model attributed to the differences between $\boldsymbol{n}$ and $\boldsymbol{m}^{(t)}$as a function of the model parameters.

Conclusion

I began this paper with the puzzling contrast in use and generality between graphical methods for quantitative data and those for categorical data, despite strong formal similarities in their underlying methods. The explanation I believe I have demonstrated is that categorical data require a different graphic metaphor, and hence a different visual representation (count ~ area) from that which has been useful for quantitative data (magnitude ~ position along a scale). The sieve diagram, mosaic, and the fourfold display all show frequencies in this way, and are valuable tools for both the analysis and presentation of categorical data.

In the second part of this paper I have outlined concrete, physical models for both quantitative and categorical data and their graphic representation and have shown these to yield a wide range of interpretation for statistical principles and phenomena. Although the spring and pressure models differ fundamentally in their mechanics, both can be understood in terms of balancing of forces and the minimization of energy. The recognition of these conceptual models can make a graphical display a tool for thinking, as well as a tool for data summarization and exposure.

Finally, as I look to the future development of graphical methods for categorical data, I see two areas where our report card, perhaps reflected in this volume, may be marked “needs improvement”: First, much of the power of graphical methods for quantitative data stems from the availability of tools which generalize readily to multivariable data, and which can make important contributions to model building, model criticism, and model interpretation. The mosaic display possesses some of these properties, and other papers here attest to the widespread utility of biplots and correspondence analysis. However, I believe there is need for further development of such methods, particularly as tools for constructing models and communicating their import.

Second, I am reminded of the statement (Cochran ???) that the ***practical power*** of any statistical tool is the product of its statistical power times its probability of use. It follows that statistical and graphical methods are of practical value to the extent that they are implemented in standard software, available, and easy to use. Statistical methods for categorical data analysis have nearly reached that point. Graphical methods still have some way to go.

References

Bertin, J. (1983), *Semiology of Graphics* (trans. W. Berg). Madison, WI: University of Wisconsin Press.

Bickel, P. J., Hammel, J. W. & O'Connell, J. W. (1975). Sex bias in graduate admissions: data from Berkeley. *Science*, *187*, 398-403.

Farebrother, R. W. (1987), "Mechanical representations of the L1 and L2 estimation problems", In Y. Dodge (ed.) *Statistical data analysis based on the L1 norm and related methods*, Amsterdam: North-Holland., 455-464.

Farebrother, R. W. (1988), "On an analogy between classical mechanics and maximum likelihood estimation", *Osterrische Zeitschrift fur Statistik und Informatik*, *18*, 303-305.

Friendly, M. (1992a), "Graphical methods for categorical data". *Proceedings of the SAS User's Group International Conference*, *17*, 1367-1373.

Friendly, M. (1992b), "Mosaic Displays for Loglinear Models". American Statistical Association, *Proceedings of the Statistical Graphics Section*, 61-68.

Friendly, M. (1994a), Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, *84*, 190-200.

Friendly, M. (1994b), *A fourfold display for 2 by 2 by k tables*. Department of Psychology Reports, 1994, No. 217.

Friendly, M. (1994c). SAS/IML graphics for fourfold displays. *Observations*, *3(4)*, 47-56.

Friendly, M. (1995). Conceptual and visual models for categorical data. *American Statistician*, 1995, *49*, 153-160.

Hartigan, J. A., and Kleiner, B. (1981). Mosaics for contingency tables. In W. F. Eddy (Ed.), *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*. New York: Springer-Verlag.

Riedwyl, H., & Schüpbach, M. (1983). Siebdiagramme: Graphische Darstellung von Kontingenztafeln. Technical Report No. 12, Institute for Mathematical Statistics, University of Bern, Bern, Switzerland.

Riedwyl, H., and Schüpbach, M. (1994). Parquet diagram to plot contingency tables. In *Softstat '93: Advances in Statistical Software*, F. Faulbaum (Ed.). New York: Gustav Fischer, 293-299.

Sall, J. (1991a), "The conceptual model behind the picture", *ASA Statistical Computing and Statistical Graphics Newsletter*, *2* (April), 5-8.

Sall, J. (1991b), "The conceptual model for categorical responses", *ASA Statistical Computing and Statistical Graphics Newsletter*, *3* (November), 33-36.

Snee, R. D. (1974), "Graphical display of two-way contingency tables", *The American Statistician*, 28, 9-12.

Figure Captions

Figure 13: Urn model for multinomial sampling. Each observation in Table 1 is represented by a token classified by hair color into the appropriate urn. This model provides a basis for the bar chart, but does not yield any further insights.

Figure 14: Pressure model for categorical data. Frequency of observations corresponds to pressure of gas in a chamber, shown visually as observation density; negative log likelihood corresponds to the energy required to compress the gas to a height p.

Figure 15: Fitting probabilities for a one-way table. The movable partitions naturally adjust to positions of balanced forces, which is the minimum energy configuration.

Figure 16: Testing a hypothesis. The likelihood ratio G^2 measures how much energy is required to move the partitions to constrain the data to the hypothesized probabilities. The components of G^2 indicate the degree to which each chamber has low or high pressure, relative to the balanced state.

Figure 17: Two-way tables. For multiple samples, the model represents each sample by a stack of pressure chambers whose width is proportional to the marginal frequencies of one variable.

Figure 18: Testing independence. The chambers are forced to align with both sets of marginal frequencies, and the likelihood ratio G^2 again measures the additional energy required.