Chapter 30
Data analysis in flow cytometry

W. A. MOORE & R. A. KAUTZ

Statistics, 30.1 Graphical displays, 30.6

The use of simultaneous four-antibody quantitative immunofluorescence and other multiple dye systems in flow cytometry (see Chapter 29) presents the experimenter with an enormous amount of multidimensional numerical data to process. In order to be useful as an experimental tool, this information must be condensed and displayed to the investigators in a compact and readily interpretable form. Often, little or nothing is known about the distribution of fluorescence beforehand and, therefore, very general methods must be used. In these situations it would be useful to have displays which would guarantee that all significant information is represented. Furthermore, it is very advantageous in terms of computer and investigator efficiency if such a display can be generated entirely automatically, with no previewing by the investigator.

The authors have developed methods which are very satisfactory by these criteria for displaying one and two variable distributions by means of computer generated graphics. Their work has revealed that the most common methods in use heavily emphasize subpopulations with low variability. In addition, the authors have explored the efficiency with respect to sample size (not computer time!) of various procedures for producing such graphical representations. For immunofluorescence these methods have allowed an approximately threefold reduction in the size of sample required (which also means threefold reduction in disk space used).

This article starts with an introduction to the statistical terminology and methods used to analyse multiple immunofluorescence data. Biologists are encouraged to read it, skipping the equations if absolutely necessary. Equations are presented only to restate the argument in more formal terms. Paragraphs in which the argument is essentially mathematical are marked with asterisks and may be omitted by biologically oriented readers.

In the concluding section, the authors will discuss the graphical displays and explain the motivation for and, most important, the interpretation of the two variable displays.

Statistics

In flow cytometry, single cells in suspension are sampled one at a time from a reservoir by passing them in a thin stream past a series of sensors. Since the arrival time of the cells at the sensor stations is not fixed in any way, a value X given by a sensor at some sampling time cannot be exactly predicted ahead of time. It is, in mathematical terms, a 'random variable'. Nevertheless, if the sensors are measuring interesting phenomena, there should be useful information available from these 'random' values. Random variables (such as X) are a distinct type of entity different from the usual 'algebraic' variables, and are traditionally expressed as capital letters to distinguish them from related algebraic variables, which will be written in lower case. A random variable does not have a single 'value' like that assigned to an algebraic variable; it can only be used within some formula which assigns a probability to some set of values it might take. Statistically we can hope to evaluate the probability that a person's height X is less than 72" without knowing that their exact height is, for example, 64.23125" or 74.78543". This probability is usually expressed as

$$\Pr[X < 72]$$

When we want to evaluate such a probability without specifying the exact condition we must combine random and algebraic variables; for example,

$$\Pr[X < x]$$

is 'the probability that the random variable X is less than the algebraic variable x' (whatever value x has and whatever possible values X could take).

The output of each of the sensors is then a random variable which changes with time (a 'stochastic process'). For these purposes the only interesting times are the times (T) when a cell is centred on the sensors (a 'discrete' process). This time, T, is itself a random variable and is called the 'epoch', despite the fact that it is typically measured in microseconds. At such a time, T, we say that the value given by each of the sensors,

$$X^{(1)}, \ldots, X^{(j)}, \ldots, X^{(m)}$$

represents the 'phenotype' of the cell. In this list $X^{(j)}$ represents the value measured by the j^{th} sensor, and is called the j^{th} 'component' of the phenotype. The fact that the phenotype is a random variable does not imply any non-determinacy on the part of the cell (although it may in fact have some) but only that the order and time of arrival cannot be specified ahead of time, i.e. is a corollary to the fact that the epoch T is random.

If we actually record the values for a sequence of cells, this is called a 'realization' of the process. Such recorded values are not random, i.e. they are normal numbers. Random variables are used to describe these values before they are recorded and to derive algebraic formulas which may be used to perform computations on the recorded values. Usually the epochs (T_i for $i = 1, \ldots, n$) and the phenotypes

$$(X_i^{(1)}, \ldots, X_i^{(m)}) \qquad i = 1, \ldots, n$$

are numbered sequentially to keep track of them. It is worth noting that mathematicians are as confused as anybody by formula involving large numbers of super- and subscripts, and frequently combine all the 'components' $X^{(j)}$ into a 'vector' or 'ordered m-tuple' $\mathbf{X}$ for compactness of the formula. This will help when describing a 'sample' of the phenotypes, $\mathbf{X}_1 \ldots \mathbf{X}n$ where

$$\mathbf{X}_i = (X_i^{(1)}, \ldots, X_i^{(m)})$$

In this case the related variable (random or algebraic) is traditionally replaced by a bold (upper/lower) case letter. In any given realization of the process the phenotype is a well defined (vector) value for each observation. The record of a sample of phenotypes is frequently called 'list mode data'.

For most purposes we will assume that the T_i and T_j are independent but that on the average time between events is constant, and that $X_i^{(j)}$ and $X_k^{(j)}$ are also independent, i.e. the cells do not influence each other in their transit through the system. Such a process is said to be a 'Poisson point process'. The behaviour of such processes has been extensively studied, and it is theoretically possible to test the independence assumption (i.e. do the cells influence each other?) on the basis of measured T_i. This theory is also relevant to the design and evaluation of cell sorters. However, we will not consider any of these problems. We will also assume that the $X_i^{(j)}$ are also independent of time. This sort of process is called 'stationary'. Many interesting processes are not stationary, and indeed it is possible to study kinetics in flow. However, we will not do so other than noting that by including T_i into $\mathbf{X}_i$ (say as $X_i^{(0)}$) all of the following methods may be used.

Thus, in general, our data consist of a finite 'sample' $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of the (vector) values which the stochastic process took, or in other terms a list of the phenotypes which were observed in the sample of n cells. From this data we would like to describe the phenotypes present in some more useful and compact form than $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Since the $\mathbf{X}_i$ are independent of the T_i and of each other, the order in which we consider the $\mathbf{X}_i$ is irrelevant. Therefore we are free to impose our own notions of order on the data. In particular, since the phenotypes are quantified, we can use the usual ordering of real numbers. For example, given some range of phenotypes (an empirically described subpopulation) we would like to know the number of cells which displayed this phenotype. We will consider first a simple type of subpopulation, namely

$$\{\mathbf{X}_i \mid X_i^{(j)} < x\}$$

those cells that measured less than x by the j^{th} sensor. (Using < rather than > may seem awkward at this point but it gives a positive derivative which will be more convenient later on.) We can summarize the number of cells in the sample which had such a phenotype for every value of x with the function

$$F_n^{(j)}(x) = 1/n \sum_{i=1}^{n} I_{[0,x)}(X_i^{(j)})$$

the 'sample marginal cumulative distribution function', where I is the indicator function

$$I_{(x,y)}(z) = \begin{cases} 1, \text{ if } x \le z < y \\ 0, \text{ otherwise} \end{cases}$$

Since potentially we want to compare samples of different sizes, we have normalized this by dividing by the number of cells in the sample. Restating this simply, $F_n^{(j)}(x)$ is the frequency of cells which have phenotype whose j^{th} component is less than x. An easily appreciated example of such a function would be the proportion of individuals who were less than a given height, tabulated for all reasonable heights.

We can compute the marginal distributions for each sensor, so it is natural to wonder if they can completely describe the sample. Unfortunately, they cannot because they do not include any information about interactions between $X^{(j)}$ and $X^{(k)}$, i.e. correlations between various components of the phenotype of one cell. Recall that we said that there were no correlations between the phenotypes of successive observations, but nothing was said about those within each observation. For example, we can make separate tables for height and weight, but if we know someone's height we can make a better guess as to their weight than that given by the general weights table.

In order to completely describe the sample, the

'sample joint cumulative distribution function' must be used.

$$F_n(x_1, \ldots, x_m) = 1/n \sum_{i=1}^{n} \left\{ \prod_{j=1}^{m} I_{[0,x_j)}(X_i^{(j)}) \right\}$$

This corresponds to the frequency of the phenotype

$$X^{(j)} < x_j \text{ for } j = 1, \ldots, m$$

in the sample. Notice that the function (and the phenotype) are described in terms of all the components simultaneously. This corresponds to combining height and weight into one (much larger) table in the analogy used above. This is the main disadvantage to this representation, since computationally the amount of work involved in dealing with such a function goes up exponentially with m. Also if $m > 1$, the function cannot be drawn on paper, and if $m > 3$ it has no physical counterparts at all. In our practice, m ranges from 3 to 6 although values of 32 and higher have been used. From this function we can obviously derive each of the previous functions.

$$F_n^{(j)}(x_j) = F_n(\infty, \ldots \infty, x_j, \infty, \ldots, \infty)$$

Although the phenotypes present in the sample represent only a finite set of values, we still feel that the underlying phenomoma are continuous in nature, i.e. may take on infinitely many possible values. This means that we do not usually expect another sample to take on exactly the same values. In particular, if we have two samples even from the same source, we will not expect that F_n will be identical to F'_n, only that they will be more similar to each other than to other unrelated samples. This similarity would be expected to become greater as the number of cells sampled is increased. In the limit as $n \to \infty$ we will assume that frequency (as defined a rational number) also becomes continuous (a real number) or a probability. Of course no real organism has an infinite number of cells to sample, but we gloss over this quibble. Assuming then that such a continuous function exists, we will call it $F(x_1, \ldots, x_m)$, (or $F(x)$ in vector notation) the '(joint) cumulative distribution function' (CDF). In an ideal world, $F(x)$ will correspond to $X(t)$, the stochastic process, in a similar way to that which $F_n(x_1, \ldots, x_n)$ corresponds to $X_1, \ldots, X_n$, a realization of that process, or a sample.

The function $F(x_1, \ldots, x_n)$ describes the probability of a particular subpopulation of phenotypes, namely those that are small in all their components, small being defined in each case by the arguments x_i. However, this is not a particularly useful one for many purposes. If we want to calculate the probability of more complex subpopulations, we will notice that for one dimension

$$\Pr[X < x] = F(x)$$

$$\Pr[x < X < y] = F(y) - F(x)$$

then if $F(x)$ is a smooth (differentiable) function

$$\begin{aligned} f(x) &= \lim_{h \to 0} \Pr[x < X < x + h]/h \\ &= \lim_{h \to 0} \{F(x+h) - F(x)\}/h \\ &= dF(x)/dx \end{aligned}$$

exists and is called the 'probability' density function' (PDF), because it represents (in the limit) the probability corresponding to a small subpopulation of phenotypes $x < X < x + h$. In order to compute the probability of an arbitrary set of phenotypes, we can reverse this process by dividing the specified phenotypes up into a bunch of small disjoint phenotype groups and adding together their probabilities. Taken to the limit this process is integration of the PDF over the phenotypes in the subpopulation. When $f(x)$ is itself a smooth function, it is usually easier to visualize $f(x)$ rather than $F(x)$. However, as long as they both exist they are in a sense two aspects of the same object. The above argument can be extended into m dimensions, to produce a unique joint PDF corresponding to the joint CDF. We will not go into the details, but simply note that from now on x may be vector valued.

Assuming that smooth functions exist corresponding to $F(x)$ and $f(x)$, how can we find them or estimate their values? An estimator of an unknown value or function is usually written in the same form as the unknown but with a 'hat'. Since $F_n(x)$ is approximately equal to $F(x)$ the simplest choice of $\hat{F}(x)$ (estimate of $F(x)$) would seem to be $F_n(x)$. However, this implies that $\hat{f}_n(x) = dF_n(x)/dx$, i.e. we must take the derivative of the sample CDF. Unfortunately, this derivative is not useful because it is infinite at the sample points and zero elsewhere (i.e. is composed of 'Dirac delta' functions). One of the main advantages of using $F(x)$ (which we do not know exactly) in our formulae rather than $F_n(x)$ (which we do) is that the derivative of $F(x)$ is well behaved. Since the most 'natural' estimator of $f(x)$ is not useful, any use we make of $\hat{f}(x)$ (and it is the preferred form for display) will be influenced by the arbitrary selection of an estimator. This is also the reason for the use of the sample CDF, since it is free of any model we put on the data.

The preferred approach to finding estimators, due to R.A. Fisher [1], is called maximum likelihood. If we assume for a moment that a unique $f(x)$ does exist and if the samples are independent, then the likelihood function L_f,

$$L_f(X_1, \ldots, X_n) = \prod_{i=1}^{n} f(X_i)$$

is, roughly speaking, the probability that the sample would have been drawn. (Strictly speaking, this probability is zero; hence the use of likelihood.) In general, we would like to select some estimator $\hat{f}$ which would give the greatest possible likelihood to the observed sample. Unfortunately, if f is allowed to be an arbitrary function, such an estimator always exists and in fact corresponds to the derivative of the sample CDF. It is clear that we shall have to put additional conditions on $\hat{f}$ if we want it to behave neatly.

Ideas for doing so date back to 1661, when London haberdasher John Graunt constructed a pseudo-histogram and a version of the CDF in order to summarize the 'bills of mortality' for the London area. He attempted to calculate the probability of dying between various ages, and the probability of living at least to a given age (i.e. dying later). To construct a true histogram we must first divide the observed phenotypes into a finite number of classes. Let Φ be the sample universe or the set of all possible phenotypes; then we can cover Φ by a finite set of classes ϕ_j such that every possible phenotype is in some class

$$\Phi = \bigcup_{i=1}^{k} \phi_i$$

and no phenotype is present in two classes

$$\text{if} \quad \phi_i \bigcap_{i \neq j} \phi_j = 0 \quad \forall\, i \neq j$$

We will look for a function which is constant on each ϕ_j (called a set function) and is also a valid PDF or

$$\hat{f}(x;c_1, \ldots, c_k) = \sum_{j=1}^{k} c_j I_{\phi_j}(x)$$

and

$$\int_{-\infty}^{\infty} \hat{f}(x;c_1, \ldots, c_k)\, dx = 1$$

or

$$\sum_{j=1}^{k} c_j \mu(\phi_j) = 1$$

The function μ is the Lebesque measure of ϕ_j, loosely equivalent to the number of phenotypes possible in class ϕ_j as opposed the number q_j observed.

The sample CDF is piecewise constant and, therefore, discontinuous and fails to have a well behaved derivative. The simplest form of continuous function that we can approximate F with will be piecewise linear, i.e. composed of linear segments joined together. The first derivative of such a function is piecewise constant, i.e. a histogram estimator. If we let q_j be the number of $\mathbf{X}_i$ which fall into class ϕ_j

$$q_j = \sum_{i=1}^{n} I_{\phi_j}(\mathbf{X}_i)$$

then

$$c_j = \frac{q_j}{n\mu(\phi_j)}$$

gives a maximum likelihood estimator for f(x). The estimated probability density for each class is simply the number of samples which fell into the set divided by the size of the set and again normalized for the total number of cells, n. We can see that the estimated CDF (in one dimension) is simply a piecewise linear function which agrees exactly with the sample CDF at the boundaries between classes.

The most common way to form such a histogram is to divide up the sample universe (set of all possible phenotypes) so that each of the classes has an equal size ($\mu(\phi_i) = \mu(\phi_j)$). The main advantage of this approach is that it is easy to implement, but it has two disadvantages. It sets an arbitrary lower bound on the size of a feature which can be resolved, so that extra classes may be needed in order to increase resolution. Furthermore, in real distributions a great many of the classes will be empty or nearly so, i.e. many of our estimators c_j are approximations of 0—noisy and not very useful. Another approach is to fix the number of events in each class q_j and then seek a set of classes ϕ_j which would yield a maximum likelihood estimator.

In one dimension, this approach has a particularly neat solution, which has been explored by Wegman [2,3]. If the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is sorted into ascending order, the result is a set of (non-independent) random variables called the order statistics

$$Y_1 < \ldots < Y_i < Y_{i+1} < \ldots < Y_n$$

If we ask that every class have an equal number of phenotypes, say $q_j = q$, and seek a set of ϕ_j which give maximum likelihood estimators of the form above, we find that the boundaries of the histogram classes are simply the corresponding order statistics $Y_{(jq)}$.

$$\phi_j = [Y_{((j-1)q)}, Y_{(jq)})$$

*Recall that we said that the sample CDF $F_n(x)$ became closer to $F(x)$ as more cells were sampled. It is to be expected that $\hat{f}_n$ will do the same thing. What we do not know is how fast we can expect it to do so, i.e. how big a sample size, n, we need before it converges to a useful extent. Convergence is usually measured by the mean squared error (MSE) at each point x:

$$\text{MSE}[\hat{f}_n(x)] = \text{E}[\{\hat{f}_n(x) - f(x)\}^2]$$

This of course depends both on $f(x)$ and on the ϕ_j chosen, so we will consider a specific example, namely a single variable histogram with equal width classes. For a particular n, let the width of the interval be h_n. Then

$$\phi_j = [(j-1)h_n, jh_n)$$
$$\mu(\phi_j) = h_n$$

i.e. equal half-open intervals on the real line (we assume them starting at 0). The mean squared error is [4]

$$\mathrm{MSE}[\hat{f}(x)] \leq \frac{2f(x')}{nh_n} + \frac{1}{2}|f'(x')|^2 h_n^2 + 0\left(\frac{1}{n} + h_n^3\right)$$

where x' is the midpoint of the interval containing x. Integrating the MSE over all possible x gives the integrated mean square error or IMSE. Now it can be shown that

$$h_n = \left[\frac{2}{\int (f'(x))^2\,dx}\right]^{1/3} n^{-1/3}$$

gives a globally optimum interval in the sense of minimizing

$$\mathrm{IMSE}(\hat{f}_n) < 3[\tfrac{1}{2}\int (f'(x))^2\,dx]^{1/3}\, n^{-2/3} + 0\left(\frac{1}{n} + h_n^3\right)$$

Unfortunately, in order to compute this optimum, we must know the first derivative of the function we are seeking. The formula does, however, give us a means of evaluating the rate of convergence in a qualitative manner.

We have seen that the error in our histogram decreases proportionally to $n^{-2/3}$ as the sample size, n, goes up. However, MSE decrease of $0(n^{-1})$ is possible theoretically (the Cramer–Rao lower bound), so we see that the histogram is not an efficient estimator, i.e. it requires more cells than are theoretically required for a given degree of accuracy. Although the estimated CDF is continuous, the histogram itself is not and, therefore, is not continuously differentiable, which makes displaying it inconvenient. It is also not 'robust' with respect to h_n, the class interval. If h_n is too large, useful information will be lost. If it is too small, the function $\hat{f}$ can fluctuate unacceptably, and approaches the derivative of the sample CDF (as do all maximum likelihood estimators with large numbers of parameters). Also, in order to find the optimal h_n, we must know the first derivative of the true density $f'(x)$. This quantity will be large (which means h_n small and IMSE large) when $f'(x)$ is large, i.e. when the distribution has very tight peaks in it, like those given by beads or chromosomes.

We have seen then that histograms in general are not ideal estimators. The search for a better estimator was taken up by Rosenblatt in 1956 [5], and later this method was generalized and investigated by Parzan [6]. Rosenblatt estimators are functions of the form

$$\hat{f}_n(x) = \frac{1}{2nh_n} \sum_{i=1}^{n} I_{[x-h_n,\, x+h_n)}(X_i)$$

which we can also calculated directly from the sample CDF

$$\hat{f}_n(x) = \frac{F_n(x+h_n) - F_n(x-h_n)}{2h_n}$$

which can be thought of as using a classification which shifts along the axis staying centred on x rather than being a fixed division of the line, i.e. at each possible phenotype we count the number of cells with phenotypes within $\pm h_n$ of the chosen one.

If we select

$$h_n = \left[\frac{9}{2\int (f''(x))^2\,dx}\right]^{1/5} n^{-1/5}$$

then approximately

$$\mathrm{IMSE} = 5/4(9^{-1/5}2^{-4/5})\,[\int (f''(x))^2\,dx]^{1/5}\, n^{-4/5}$$

[4] which is substantially closer to the theoretical lower bound. If we do not know $f''(x)$, which is usually the case, we can take an iterative approach to finding h_n by estimating f'' from the data. It is in fact possible to achieve MSE error decrease of $0(n^{-1} \ln n)$, but apparently only by relaxing the constraint that the estimator $\hat{f}$ be itself a probability density and allowing it to take negative values in some cases [4]. Since this would cause difficulties for many applications, we have to accept $0(n^{-4/5})$ as the best we are likely to get.

In the above formulation, the indicator function, I, which is flat in the middle and discontinuous at the ends, may be replaced by a more suitable smooth function (Parzen kernel), $K(x)$, in which case the estimator takes the form

$$\hat{f}_n(x) = \int \frac{1}{h_n} K\left(\frac{x-y}{h_n}\right) dF_n(y)p$$

or

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)$$

This corresponds to covering over the infinities in the maximum likelihood estimator with smooth functions (the kernel function $K(y)$). Devotees of Fourier transform theory may recognize this as a version of the convolution integral, and indeed we can regard this process as filtering of the sample data by a filter with band width controlled by h_n and a non-physical impulse response $K(y/h_n)$. A physical analogy would be looking at a dot-plot with diffraction limited dots.

*A variety of kernel functions may be used. The Gaussian distribution is an obvious candidate. The Epanechnikov kernel [7] may be optimal for some purposes

$$K(y) = \frac{3}{4}(1-y^2) \text{ if } |y| \leq 1$$

but when $f''(x)$ is not known, a kernel which is smooth

at the endpoints will probably be better for estimating it so

$$K(y) = \frac{15}{16}(1-y^2)^2 \text{ for } |y| \leq 1$$

may then be a better choice. Both of these are preferable to the Gaussian kernel for computation because they are zero outside of a finite interval. All these kernel estimators have $0(n^{-4/5})$ error decrease and in fact nearly identical convergence properties (at their respective optimal h_n).

The 'proportional area approach' may be extended to kernel estimation, by replacing h_n with the distance to the q^{th} nearest point. This becomes difficult when more than one variable is involved, but a suitable local h_n can be computed from a preliminary approximation of $f(x)$ locally, i.e. using a large h_n when the preliminary estimate of the density is low and vice versa. This operation is no longer linear with respect to Fourier transform theory, but does offer the advantages of not putting a global limit on the resolution and of depending on a (crude) preliminary estimate of the value of the function rather than its second derivative which is much more attractive numerically. This is the method we rely on for all our standard displays.

Graphical displays

Once an estimator $\hat{f}^{(J)}(x)$ for a marginal (or conditional) probability density function has been found, it is natural to inspect it by graphical methods. Indeed this is usually one of the first techniques taught in algebra. For one variable, a line drawing of the estimated probability density as a function of x (the measured parameter) is adequate. This display should be normalized, so that a fixed area on the graph is under the curve, because in probability analysis it is this area which indicates the relative number of cells. (This is because we are working with the density function (derivative) rather than the distribution function directly.) Normalizing the area allows the eye to compare curves from different samples in terms of probability.

If the joint density of two of the variables (usually called X and Y rather than $X^{(J)}$ and $X^{(J')}$) $\hat{f}(x,y)$ is to be displayed, more complex methods are necessary. The most common approach is to note that the density function can be used to define a two-dimensional surface in R^3, namely

$$S_f = \{(x,y,z) \mid z = f(x,y)\}$$

Such a surface may be physically modelled by a solid block with its top sculpted into relief, as specified by $f(x,y)$, i.e. like a plaster 'mountains and valleys' model. This model is commonly simulated with computer graphics by drawing perspective or orthometric projections of this surface (Fig. 30.1). When implemented using current graphics techniques, such algorithms take up a lot of computer time due to the need for hidden line removal to make the surface appear opaque and the desirability of making images of the surface from multiple view points. This model of the surface has several free parameters, including the scaling of z (probability density), and the position of the 'virtual camera' in spherical coordinates, which makes it difficult for most users to specify the view they want drawn.

A more subtle point that must be considered is that the eye can be misled by the non-physical nature of the surface being drawn. In the projected displays we are exploiting depth cues used by the human visual system. By simulating hidden lines and perspective we cue the eye to the relative depth of the facets of the surface. (When real time hardware rotation is available the kinetic depth effect may be used for the same purpose.) When analysing such scenes the eye seems to decompose them by edges and by boundary concavities. This means that the eye tends to evaluate objects based on their shape and perimeter, whereas in analysis of probability density functions it is the volume of the block underneath the feature which is important (because we are using the derivative of probability). In more familiar terms, the problem is that the eye cannot readily compare the volume of objects of radically different shapes, and thus cannot compare the significance of a broad (high variance) distribution, compared to a narrow (low variance) one. For example, most people could not compare the amount of stone in the Washington monument with the amount of water in the reflecting pool with any sort of accuracy. This type of situation is unfortunately all too common when a highly variable stained subpopulation of cells is present together with an unstained one. (Fig. 30.1 is of chromosome data, for which this is not usually a problem.)

Another common technique for displaying bivariate distributions is to draw a selection of the level curves of the density function, or a contour map M_f, as follows:

$$C_f(z) = \{(x,y) \mid z = f(x,y)\}$$

where

$$z_i < z_{i+1} \text{ and } i = 1, \ldots, n-1$$

are the 'contour levels'. The most natural way to assign z_i, is equally spaced, i.e. a topographic or 'geological survey' map (Figs. 30.2 and 30.3, panel A).

$$z_i = iz_1$$

This approach uses one free parameter (namely z_1, the height of the first contour) which must be assigned by

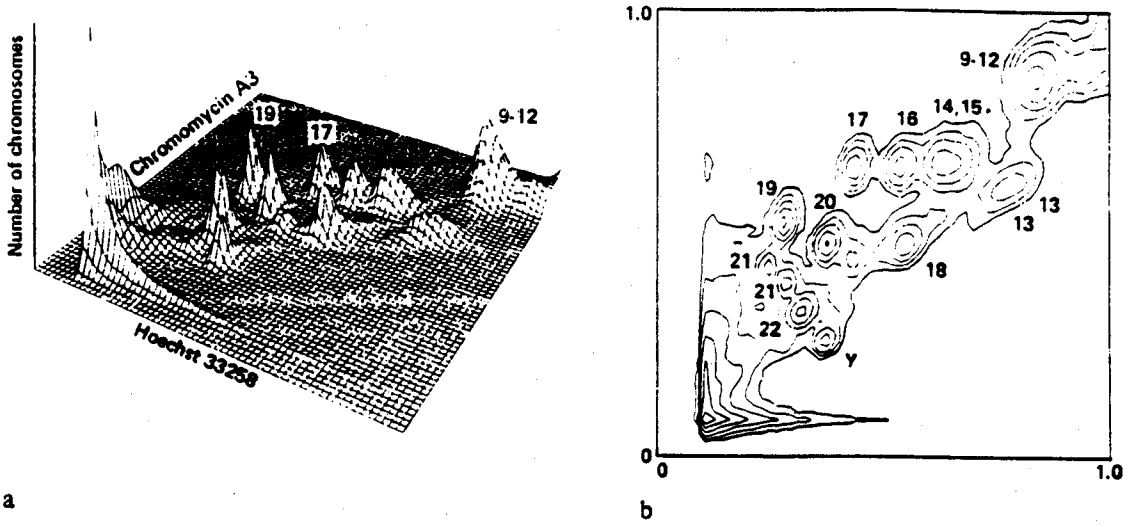

Fig. 30.1. (a) A perspective view of a dual variable histogram. This presentation shows the similarity to a mountain range. The mathematical function which maps the x,y plane to the surface of this 'mountain range' is the probability density function in the text. (b) A contour map of the same data. The data shown is a flow karyotype of human chromosomes. (Kindly supplied by Dr Joe Gray of the Lawrence Livermore Laboratory.)

the user. In practice, the display is very dependent on this parameter. Therefore the user must learn to manipulate a single relatively arbitrary parameter whose significance is even less intuitive than spherical coordinates were. Indeed, the user probably still needs to make multiple maps in order to validate a choice of z_1. When the choice of z_1 is inadvertently made too small, inordinate amounts of computer time can be wasted, blacking large regions of the paper. Alternatively, if it is chosen too high, significant features may not show up at all. On the positive side, these maps are generally easier and quicker to generate than projected views, especially when floating point hardware and large amounts of memory are not available. In the authors' experience (with the limitations cited above) they are easier to interpret than the projective displays but the authors have never generated a fool-proof heuristic for selecting z_1 automatically.

In the case of contour maps, a learned (for those of us who were boy scouts) ability to correlate topographical maps with landmarks is exploited. In such maps, the eye responds to the contours as the edges of figures; in particular, those in places where many such curves are close together are emphasized. This means that the eye's response is roughly proportional to the grade of the surface or the magnitude of the gradient of the density

$$\| \nabla f(x,y) \| = \{(df/dx)^2+(df/dy)^2\}^{1/2}$$

This is the same way that you would measure the steepness of a road, e.g. 6 ft climb per 100 ft travelled. Now, unfortunately, the non-physical nature of the situation comes in to play again. In maps drawn of the real world, this quantity is both important, because a large gradient indicates the presence of cliffs which the average pedestrian will avoid, and is also limited by the strength of the materials involved. The relief of features on the surface of the earth is quite small compared to the diameter of the earth, and is relatively small even compared to the area covered in most topographic maps. If this were not true, such maps would be nearly useless. As any rock climber could certainly tell you, such a map gives no information as to how or even whether a given cliff can be climbed. The earth is smooth at such large scales mainly because rocks crumble and weather. However, probability densities do neither, and, therefore, the grade can assume really large values over small areas (the Washington monument analogy is severely understated). The result is that the eye is again fooled into overvaluing tight distributions with respect to broad ones.

In order to overcome these problems and select a more useful set of contour levels, z_i, we will steal a page from image analysis and use the $f(x,y)$ to represent intensity over the plane, i.e. a TV image, rather than a surface or solid in R^3. Again, the non-physical nature of the function defeats the naive approach since no usable display device has anything like the dynamic range of a PDF. However, the eye can readily interpret

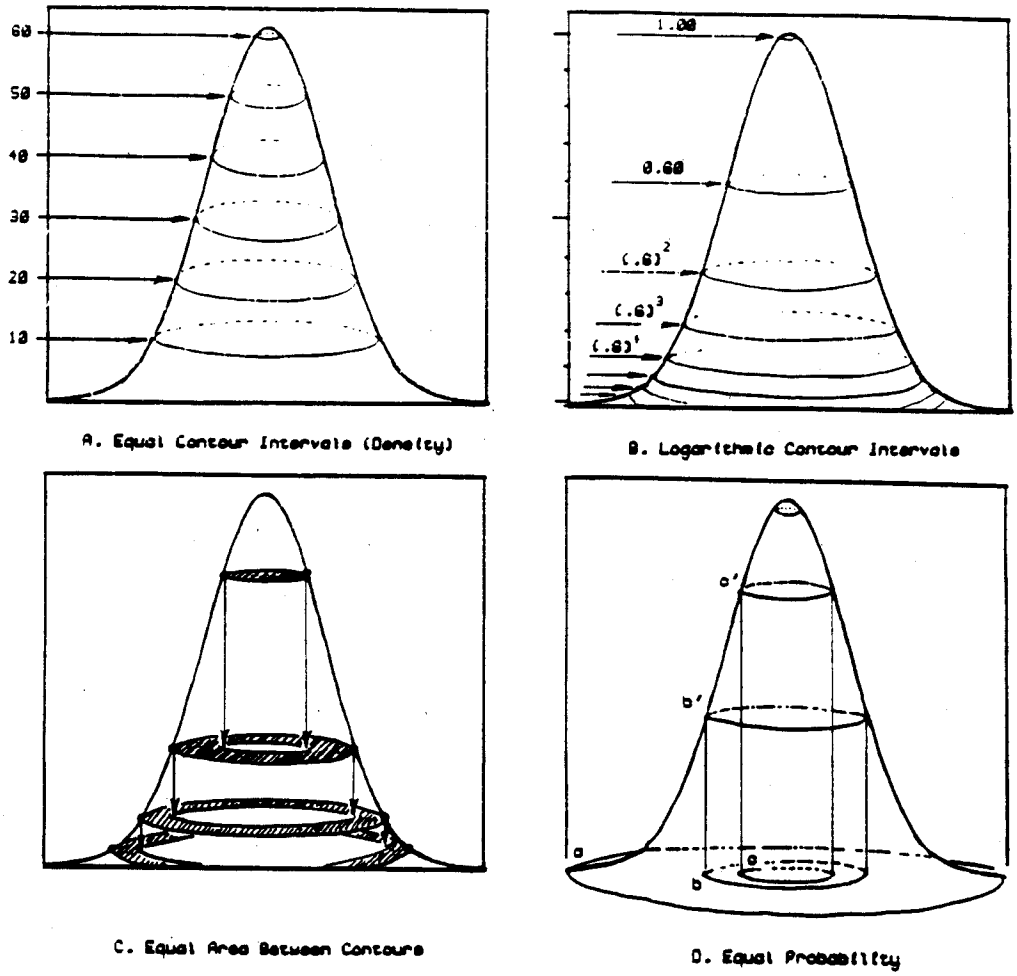

Fig. 30.2. Four different contouring methods. In these figures the vertical axis corresponds to the z axis of Fig. 30.1, the heights of the mountains.

Panel A shows equal contour spacing. The contour elevations are chosen such that the intervals between them (along the z axis) are the same. This is the form of contouring found on geographic topo maps, and is what is offered on most simple computer systems. Equal spacing tends to put all of the contour lines on the highest peaks, and often does not show enough detail of lower features. Also, adjacent contour lines at different elevations may contain grossly different numbers of cells between them (see panel D). This method is referred to in our literature as equal 'density'.

Panel B shows logarithmic contour spacing. The contour elevations are chosen such that their elevations have fixed ratios, in this case 50%. This method puts more contours in the lower regions, while still showing the heights of the peaks.

Panel C shows equal area contouring. The contour elevations are chosen such that the resulting contour lines have equal areas of paper between them. The coloured annular sections in the figure all have the same area. Because the width of the annulus must decrease as the radius increases, this method tends to put more contour lines in areas furthest from peaks. Thus this method is useful in defining the regions between cell populations (i.e. for setting sort windows).

Panel D shows equal probability contours. The contour elevations are chosen such that there are equal numbers of cells between the resulting contour lines. The number of cells is proportional to the volume of mountain. The volume defined by the region abb' (volume of rotation about the axis of the peak) is the same as the volume defined by the region bb'c'c.

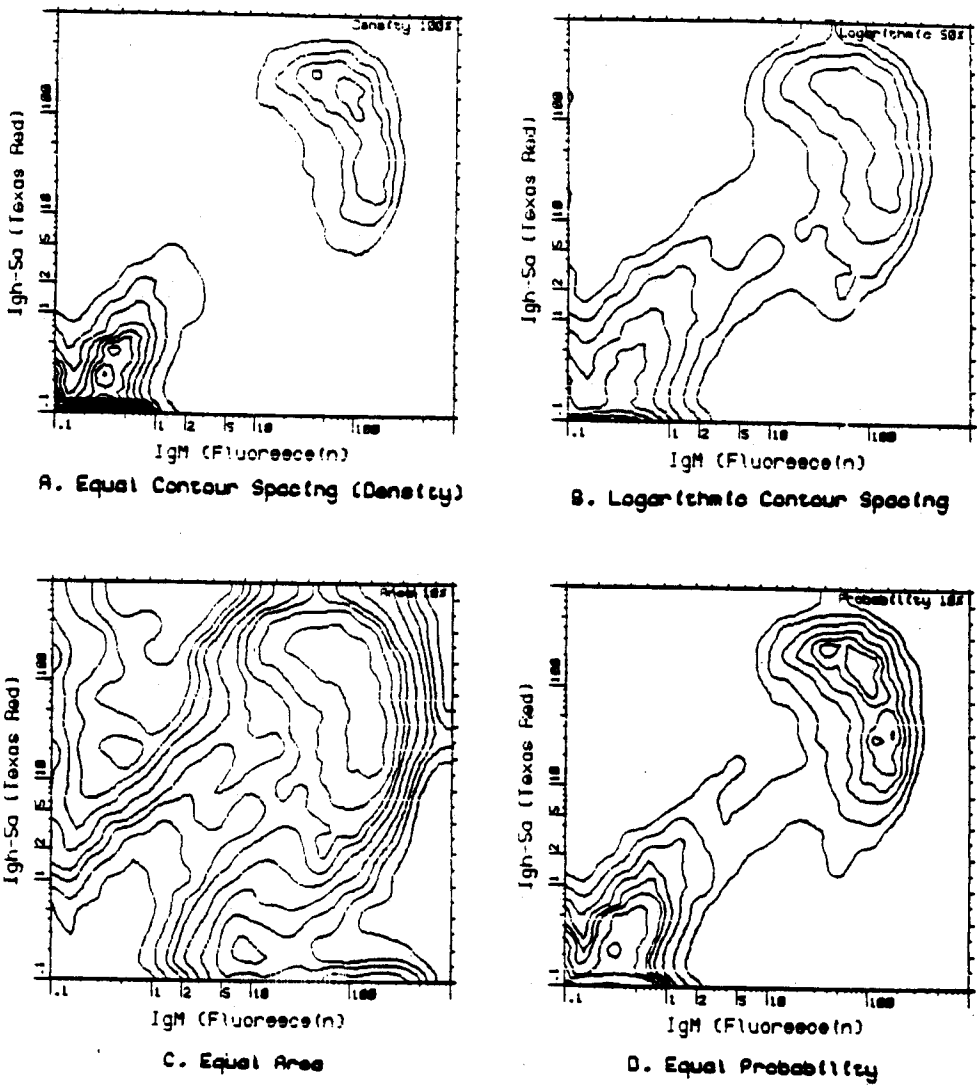

Fig. 30.3. Contour maps produced by each of the four contouring methods explained in Fig. 30.2. The data shown is IgM vs. Igh-5a for whole spleen cells from a CBA/N (immunodeficient) mouse.

Panel A: note that almost half of the contour lines are crowded into the peak of negatives near the origin, and that there is no detail of lower slopes.

Panel B: there is no crowding at the peak of negatives, and some detail of lower features can be seen.

Panel C: the valley between the two main populations, the optimum point to separate the two populations, is made clear.

Panel D: both high and low features can be seen: 10% of the cells are in the region between the first contour and the border; 10% are in the region between the first and second contours, etc. Note that the region between the second and third contours is divided between the two peaks. The number of cells in the union of these two regions add up to 10% of the total.

an image whose dynamic range is compressed as long as relative intensities are preserved (or sometimes even inverted) and sufficient edge contrast remains. (Witness the success of TV.) Therefore we will turn to the technique of monotonic point transforms, i.e. for a strictly monotonic function on R

$$t: R \rightarrow R \text{ such that } x > y \Rightarrow t(x) > t(y)$$

the function f may be transformed 'point-wise' by it as

$$T: f \rightarrow t \circ f$$
$$T[f]: (x,y) \rightarrow t[f(x,y)]$$

One obvious function, which preserves this property, is the logarithm, and indeed graphs and contours of ln[f] are frequently used (Figs. 30.2 and 30.3, panel B). This model has two free parameters, namely the ratio $r = z_i/z_{i+1} < 1$ of successive contour levels and implicitly the number of contours (or equivalently the minimum contour level) since the range of ln[f] extends to $-\infty$. It is not too difficult to select these parameters heuristically so that they usually produce a reasonable display; however, there is no obvious relation between visual features and volume (probability) and it is difficult to compare maps drawn with different values of r.

So far we have looked at contour intervals which were completely general, i.e. independent of $f(x,y)$. Now we will look at two 'data driven' approaches, in which the monotonic transform used is itself derived from the distribution. Consider the set of points where the density is less than a specified level, i.e. the 'background' of the image

$$B_f(z) = \{(x,y) \mid f(x,y) < z\}$$

This can be thought of loosely as a background for biological data as well since it corresponds to the set of observed phenotypes whose estimated frequency is less than z, i.e. the rare phenotypes. If f is a well behaved function (C^k) then this set is also well behaved (Borel, or 'an event'). There are two obvious means of measuring this set. We will consider first the area

$$a_f(z) = \mu\{(x,y) \mid f(x,y) < z\} = \mu\{B_f(z)\}$$

where μ is Lebesque measure, i.e. the area over which the density is less than the specific z. It is well defined and monotonic if $f(x,y)$ has finite support. (The area of the complement of B_f is clearly always finite but defining A[f] in terms of it would introduce a plethora of $-$ signs and physical observations that are by nature finite.) A few moments' reflection should reveal that this function is a sort of 'invariant' operator for our set of monotonic point transforms, because for a density f and any two other monotonic point transforms, G and H,

$$A \circ H[f] = A \circ G[f] = A[f]$$

and it thus might be a good candidate for display. Equally spaced contours drawn on A[f] separate regions of equal area (Figs. 30.2 and 30.3, panel C).

*This technique is known as 'histogramming' in image analysis, because the usual way to compute a_f is to histogram the number of times each grey level occurs in the image. This usage conflicts with ours where the histogram refers to a density estimator. Its use in image analysis is related to its invariance property, since such an image is invariant when transmitted through a non-linear transmission system as long as the transfer function remains monotonic. For our purposes (since the range of f is not as precisely known as the number of grey levels usually is) it will be more expedient simply to sort a finite set of values of $f(x,y)$, say $f_{ij} = f(ix_1, jy_1)$ (i.e. our version of the histogram), into ascending or descending order. Then, knowing that the measure of each class in our histogram is x_1y_1, we can interpolate the sorted list to approximate each of the equal area contour levels. Note that rather than actually computing the transformed value of each point of the surface, we usually prefer to compute the z_i corresponding to equally spaced contours on A[f].

* Because A[f] preserves only the relative order of the values of f when equally spaced contours are drawn on A[f], they are as likely to occur where f is small as where it is large. This means that in some sense this transform preserves the least possible structural information (i.e. only order) and in fact, when averaged over a suitable area (the contour interval), such a contour map contains no information. This means that if a slide made by this method is sufficiently out of focus, the contour map will disappear into a uniform grey. The useful aspect of this property is that it does visualize the distribution in regions where the probability density is low. These regions are of interest mainly in elucidating the exact boundaries or boundary overlaps (say for sorting) of a previously identified subpopulation.

The second measure of the background $B_f(z)$ and the most interesting transform with regard to automatic first pass analysis, is based on the function

$$p_f(z) = \Pr\{f(X,Y) < z\} = \Pr\{B_f(z)\}$$

Where X and Y are random variables distributed as $f(x,y)$ and therefore $f(X,Y)$ is also a random variable. This function corresponds to the probability of the background event $B_f(z)$ discussed above, i.e. the probability that a cell will have a phenotype whose density is less than z or, in other words, the total frequency of background (rare) phenotypes.

Equally spaced contours on the surface P[f] will separate regions (events) of equal probability (Figs. 30.2 and 30.3, panel D). Finally, we have a represen-

tation in which the visual cues (the edges) have equal significance in terms of the probability distribution. Also, the spacing between contours in such a map is a true probability, and indeed any subpopulation (simply connected Borel set) which has a probability greater than the interval must be crossed by at least one contour line. Thus the 'equi-probability' contour model has a natural significance value which the user can associate with the data, namely the size of the smallest subpopulation of interest, which gives very satisfactory results entirely automatically from that point on.

In order to compute the function p_f, we first sort the histogram class values f_{ij} as above, and note that this function is the first difference of the function we need. Then performing a finite summation on the sorted data allows us to interpolate a suitable set of contour levels z_i corresponding approximately to equally spaced contours on P[f].

References

1 FISHER R.A. (1922) On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. Lond.* (Series A), **222**, 309.

2 WEGMAN EDWARD J. (1970) Maximum likelihood estimation of a unimodal density function. *Ann. math. Statist.* **41**, 457.

3 WEGMAN EDWARD J. (1970) Maximum likelihood estimation of a unimodal density, II. *Ann. math. Statist.* **41**, 2169.

4 TAPIA RICHARD A. & TOMPSON JAMES R. (1978) *Nonparametric Probability Density Estimation*. The Johns Hopkins University Press, Baltimore.

5 ROSENBLATT M. (1956) Remarks on some nonparametric estimates of a density function. *Ann. math. Statist.* **27**, 832.

6 PARZEN E. (1962) On estimation of a probability density function and mode. *Ann. math. Statist.* **33**, 1065.

7 EPANECHNIKOV V.A. (1969) Nonparametric estimates of a multivariate probability density. *Theory Probab. Applic.* **14**, 153.