1

Introduction

Flow cytometry is now a well established technique in cell biology and is gaining increasing use in clinical medicine. The major applications to date in the latter have been in DNA histogram analysis to determine "ploidy" (DNA index, Hidderman et al. 1984) and S-phase fractions for prognostic purposes in cancer patients and in immunophenotyping (Parker 1988). However, more recent applications in cancer work include determination of tumour cell production rate using bromodeoxyuridine (Begg et al. 1985) and estimations which relate to therapy resistance including glutathione, drug efflux mechanisms and membrane transport (Watson 1991). The power of the technology relates to its capacity to make very rapid multiple simultaneous measurements of fluorescence and light scatter at the individual cell level and hence to analyse heterogeneity in mixed populations.

The early commercial instruments were somewhat fearsome beasts with vast arrays of knobs, switches, dials, oscilloscopes and wires hanging out all over the place. At best, they tended to be regarded as "user non-friendly" and at worst as "non–user friendly". However, the recent generation of machines have been simplified considerably, with the in-house computer taking over many of the tasks which the operator previously had to perform manually. The undoubted "user-friendliness" of these modern instruments, together with the relative reduction in initial capital outlay, is a considerable advantage as it makes the technology available to many more users. In turn, this makes it possible for relatively untrained persons, who may not be fully aware of potential problems and pitfalls, to stain samples and operate the instruments to produce "numbers". There appears to be a prevalent philsophy amongst the instrument manufacturers to produce bench-top devices that require a minimum of operator interaction, so that all that needs to be done is stain up the cells, shove them in the instrument, and out come the numbers.

I'm sure this philosophy is fine from the manufacturers' standpoint, as this approach helps to sell more machines because you purchase an instrument specifically designed to do a particular task. Under test conditions the instrument will perform very well the particular task for which it was designed. However, there are a number of disadvantages to this philosophy. First, a particular instrument designed specifically for a particular task may not do so well with an apparently similar task using different combinations of fluorochromes for different purposes. Second, the "new" generation of flow cytometry users and operators may not even be aware that such problems could exist. Third, the operator is usually insufficiently aware of deficiencies or potential deficiencies in a particular instrument, as no manufacturer will ever say it's not very good at doing this or that. Finally, many operators are completely at the mercy of the software data-handling package supplied, which may contain deficiencies that the manufacturers do not appreciate.

Flow cytometers produce a vast amount of data, which is one of their many attractions, but this can be a two-edged sword. Data, which are just a series of numbers, must be converted to information. Moreover, the information produced from those numbers not only must have meaning, but also must be shown to have meaning. This is the most important single aspect of flow cytometry, but it has received relatively little attention.

One of the frequently voiced advantages of the technology is that it produces "good statistics" because large numbers of cells have been analysed. However, confidence limits are seldom placed on results, and hence the reader has little or no feel for the inherent variability in the information produced. This variability is important and has three major components. The first is due to the measuring system, and applies not just to flow cytometry but to every measurement system. Manufacturers will tend to downplay or ignore this component. The second component is due to variability in the processes involved in making the measurement possible, and in flow studies this includes variability in fluorescence staining procedures (including the various reagents) as well as the technical competence with which the procedures are carried out. The last, and most important, source of variability is within the biology being studied, and it is from this that we might gain some extra information.

This short monograph was compiled from a series of notes originally intended for users of the custom-built instrument in the MRC Clinical Oncology Unit at Cambridge. All of the procedures described in this book are contained within our computer analysis package, which has been updated continuously over the past decade, and most of the examples are drawn from our data base. The statistics sections are limited to those we have found most useful, but I hope this will provide newcomers to flow cytometry with insight into some of the potential problems to be faced in data handling, data analysis and interpretation. The statistics begin with some very basic concepts including measurement of central tendency, distances between points and hence assessments of distributions, and what these various parameters mean. These basic concepts, of which everyone is well aware, are then developed to show how they can be used to analyse data and help convert them into information. A distinction is made here between data handling – for example, gating and counting the numbers of cells within that gate (a process commonly regarded as data analysis but which, in reality, is data handling) – and data analysis itself, which is the means by which information is extracted. Gating is not covered as a specific topic.

The book is intended for biologists using flow cytometers who know about the basic anatomy and physiology of these instruments. Data analysis obviously implies that mathematical ideas and concepts will have to be considered. However, as the book has been written for biologists, an attempt has been made to make this aspect as simple as possible; if you can add, subtract, multiply, divide and, most importantly, think logically then you should have few problems. If you are also familiar with power functions, logarithms, transcendental functions, differentiation, integration and basic statistics then you probably need not be reading this. This book is not intended for highly experienced users and developers with backgrounds in physics, mathematics or statistics who also have struggled with the various problems considered within these pages.

2

Fundamental concepts

Handling and interpreting numbers is not generally a “strong point” for biologists. This applies particularly when there is a large group of numbers that are more easily handled and understood by using some average value as a summary. Indeed, large groups of numbers must be handled by some sort of summary system, because just supplying the raw data – say, 10,000 fluorescence or light-scatter recordings from a flow cytometric analysis run – would be essentially unintelligible. This chapter was included to re-familiarise the reader with some very basic number handling concepts and to set the scene for converting numbers into information.

2.1 Central tendency

If we make a number of measurements on a population, say the weights, shapes and various dimensions of 1000 females between the ages of 15 and 25, we will end up with a large series of numbers. I have chosen this particular example not because I'm a male chauvinist pig, but because I was recently talking with a designer of female undergarments who had the task of converting those numbers into articles for sale on the shelves of a large retail outlet. The end points of the survey were to make the articles as appealing as possible (that was the first consideration), in the minimum number of different shapes and sizes as possible, as cheaply as possible, and to sell all of them all the time. At the outset is was appreciated that some of these aims were mutually exclusive. For example, the minimum number of shapes and sizes is obviously 1, which is the cheapest manufacturing option. A single standard shape and size can be obtained by summing all the measurements in a particular class of measurement, x, and dividing by the number N of females surveyed. This is the arithmetic mean of the population, $\bar{x}$, which is given by the following formula:

$$\bar{x} = \frac{\sum_{k=1}^{k=N}(x_k)}{N},$$

where the symbol $\sum$ represents the summation of all the k x-values from 1 to N.

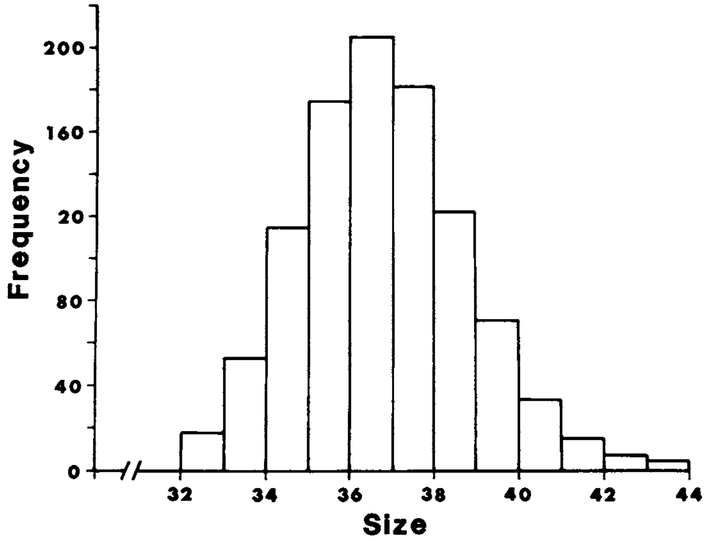

Figure 2.1. Distribution of measurements which exhibits a "skew" to the right.

The problem here is that most of the customers would not be satisfied most of the time, as there are very few who are "average". A further problem, which relates to maturity, is encountered in this age group. There is a relatively greater proportion of smaller sizes than larger, a general phenomenon that applies to the majority of biological observations and measurements. Thus, we cannot represent the whole population by a single set of average measurements as those measurements are distributed; a representation of one set of those data (I'll leave you to guess which) is shown in Figure 2.1, where the distribution is slightly "skewed" to the right. However, we will see later that there are ways of coping with variability of this type.

There are two further methods of expressing central tendency, the mode and the median of the distribution, but neither will help us with the undergarment problem. The mode is the point in the distribution where the frequency is at a maximum, and the median is the point where half the area lies to each side. In symmetrical distributions the mean, mode and median are the same point, but this is not true for skewed distributions; the relationship between these parameters is shown in Figure 2.2.

2.2 Absolute distance between points

Measuring the absolute distance between two points is not as immediately straightforward as it might seem, even though we are dealing with Euclidean as opposed to relativistic distances. Consider two points marked on a sheet of paper. We can place a ruler between these points so that the "zero" mark is adjacent to one point and read off the number on the scale adjacent to the

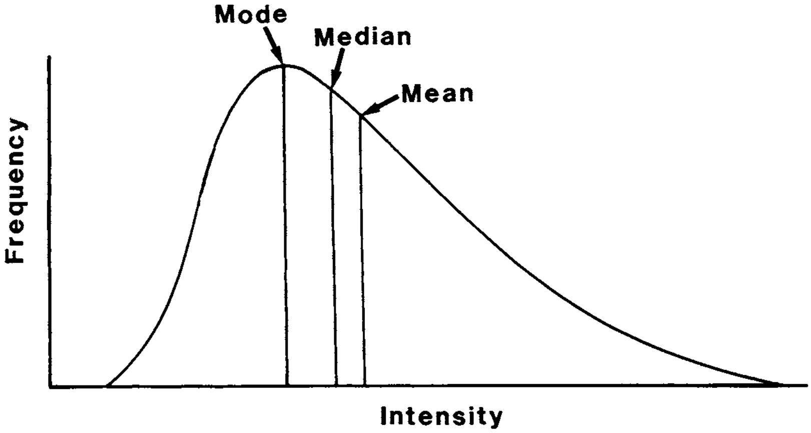

Figure 2.2. Relationships between the mean, median and mode.

second point. Let us suppose the answer is 50.8. If we now take a different ruler we might obtain a distance of 2.0. Which is the correct answer? Obviously, both are correct and you will immediately recognise a calibration difference: the first ruler was metric and the second was imperial. Thus, in order to avoid any possibility of misinterpretation, always state which units you are using. Now, rotate the sheet of paper 180° and repeat the exercise. You will get a very similar, but not an identical, answer. This begs two further questions. Where is ZERO, and why might the results be slightly different? The answer to the first question is that ZERO is exactly where you want it to be. There is no such reality as ZERO; it is a convention just as the whole of the measuring system being discussed is a convention. An answer (but not necessarily *the* answer) to the second question is in Section 2.4.

At first sight the introduction of a discussion about where ZERO might be may seem a little esoteric, unnecessary or even futile, but it is very important in flow cytometry. We do not directly measure whatever commodity we are measuring, we measure light. This strikes the photocathode in the photomultiplier, which then emits electrons in direct proportion to the number of photons striking the cathode. This is the point at which the light flux is transduced (changed) into the electronic signal. The electron flux emitted from the photocathode is then amplified through a dynode chain within the photomultiplier to give a current, thence to produce a voltage which is amplified. Modern electronics are capable of amplifying a signal by many orders of magnitude depending on the various settings you happen to choose, and instruments can have a total dynamic range, in terms of number of molecules measurable, of from 10^2–10^{16} per cell. This range of fourteen orders of magnitude must be viewed through a "window" which, even with log-amplifiers, is still only 10^4 wide. Hence, ZERO on the particular viewing window you are using is an arbitrary point dictated by the assay you are performing, which is somewhere within the 10^2 to 10^{16} total possible dynamic range available to you.

How then can we measure the absolute distance between two points within this arbitrary measurement system, particularly as we do not know where ZERO happens to be located? The square of a positive number is positive and the square of a negative number is also positive. Thus, we obtain the absolute distance between two points by subtracting one from the other (it doesn't matter which from which), squaring the result then taking the square root. This gets around the problem of where ZERO might happen to find itself; it could be on the left or the right or up or down, or in or out, or anywhere, it is no longer of any consequence. This relationship, which is extremely simple and of fundamental importance, is as follows:

$$D_a = \sqrt{(x_1 - x_2)^2},$$

where D_a is the absolute distance between points and where x_1 and x_2 are the two measurements in a 1-dimensional data space.

Let us now suppose we are working in a 2-dimensional data space of (x, y)-coordinates. The absolute distance between any two points with respective coordinates (x_1, y_1) and (x_2, y_2) is given by

$$D_a = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

This should have a familiar ring to it, as Pythagoras of Samos did it first in 600 B.C. or thereabouts (it's in the reference list!). It should now be intuitively obvious that we can generalize this relationship for any number of sets of different measurements. Moving up to the 3-dimensional coordinates of the (x, y, z) data space, we have

$$D_a = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}.$$

When we go beyond 3-dimensional coordinates we enter Euclidean hyperspace, which is just a fancy term (but it sounds impressive) to describe sets of four or more coordinates. It is not something that is easily visualized, as we live in and are familiar with only 3-dimensional space; however, extension of the Pythagorean relationship to hyperspace is perfectly valid and was used to solve part of the multidimensional undergarment problem.

2.3 Dispersion and its representation

Just as we must have some way of summarising an average, say the arithmetic mean of a set of numbers, in order to make those numbers "handlable", so we must also have some way of measuring and representing dispersion in that set of numbers. Clearly, we cannot give each individual dispersion, so we must represent this by some sort of average dispersion value. Such a measurement of "spread" is required whenever data are distributed, as there must be some way of expressing how far the overall characteristics of the population differ from the average.

2.3.1 Mean deviation

The simplest measure of dispersion is given by the mean deviation d, where

$$d = \Sigma(x-\bar{x})/N.$$

The expression $(x-\bar{x})$ represents the subtraction of the mean $\bar{x}$ from a given point x to give the deviation of this point from the mean, and the symbol Σ represents a summation of the deviations of all the points from the mean. The summation is then divided by N, the number of observations, to obtain the mean deviation. You will recognise that the mean deviation can be zero if the sum of the deviations above the mean is equal to the sum of deviations below the mean. We will see in Sections 4.2.3, 4.4.1 and 5.2 how this can be used in significance testing.

2.3.2 Mean squared deviation, variance

The mean deviation just described gives equal relative weight to all deviations from the mean, whatever their magnitudes. Another measure of spread, one that gives greater weight to the larger deviations, is given by the mean squared deviation or *variance* σ^2:

$$\sigma^2 = \Sigma(x-\bar{x})^2/N.$$

Thus, the variance is defined as the overall average squared distance (deviation) of all the points from the mean of all the points.

There are a number of ways in which variance is defined, depending on the type of problem and distribution being considered. This particular definition of variance relates to observations or measurements where deviations from the mean are likely to be equal on either side of the mean.

2.3.3 Standard deviation

The standard deviation s is obtained from the variance by taking the square root; hence, it is the root-mean-squared or RMS deviation:

$$s = \sqrt{\Sigma(x-\bar{x})^2/N}.$$

If we now take the square root of the top and bottom of this expression separately, we may write this as

$$s = \frac{\sqrt{\Sigma(x-\bar{x})^2}}{\sqrt{N}}.$$

We can now see that the term $\sqrt{\Sigma(x-\bar{x})^2}$ is a Pythagorean type of expression, where

$$\sqrt{\Sigma(x-\bar{x})^2} = \sqrt{\sum_{i=1}^{i=N}(x_i-\bar{x})^2} = \sqrt{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_N-\bar{x})^2}.$$

If we now regard this as an n-dimensional Euclidean hyperspace relationship, where each of the N points represents a "dimension" within a 1-dimensional distribution, we can see that $\sqrt{\Sigma(x-\bar{x})^2}$ is the total absolute distance of all the points $(x_1, x_2, x_3, \ldots, x_N)$ from the mean of all the points. We can further rearrange the expression for the standard deviation as

$$s = \sqrt{N} \times \left[\frac{\sqrt{\Sigma(x-\bar{x})^2}}{N}\right].$$

I've written it out like this to demonstrate that the term within the square brackets is the average absolute distance of all the points from the mean of all the points, which, when multiplied by $\sqrt{N}$, gives the standard deviation.

Thus, the term "deviation" in the definition of standard deviation is a measure of the absolute distance of all points from the mean of those points, irrespective of whether they are less than or greater than the mean.

2.4 Probability

Probability is all around us from conception to the grave. It is familiar to everyone and used in everyday life and language. And yet by its very nature it has a somewhat ephemeral ring to it, and its definition does not easily trip off the tongue. The Chambers Twentieth Century Dictionary (1976) doesn't help very much, where we find "quality of being probable" and "that which is probable", neither of which is very enlightening. But next there is "chance or likelihood of something happening", which looks more hopeful. When we turn to the definition of *chance* we have "that which falls out or happens fortuitously, or without assignable cause": "an unexpected event": "possibility of something happening": "probability". We seem to be going 'round in circles, so let's try *likelihood* where we find "similitude": "semblance": "resemblance": "probability"; we *are* going 'round in circles.

Perhaps we should start with intuitive logic and something we can really hang our hats on. Death is the only certain outcome of life. Indeed, there are those jaded individuals who regard life merely as a terminal disease, the ultimate outcome of which has a p-value of unity. On the other hand the fact that an individual exists as that particular entity must have a relatively low p-value, which probably (it's that word creeping in again) tends to zero but is clearly not zero because of that particular existence. I say a relatively low p-value because that particular ovum (one of about 400 released at random) had to have been fertilised by that particular sperm (one of legion which is also random) to produce that particular individual. A different ovum or a different sperm would have produced a different individual.

In the last example *p* was low, but some things have a probability of zero, as they are clearly impossible. I well remember the first time in my life I was perplexed. My doting Grandmother would read me the following nursery rhyme:

Hey diddle diddle, the cat and the fiddle
The cow jumped over the moon
The little dog laughed to see such fun
And the dish ran away with the spoon.

Apparently, I was about 2½ years old and wasn't too concerned about the cat or the fiddle or about the cow jumping over the moon. At that stage I wasn't too conversant with Newton's laws of motion, gravity, propulsion or the distance between the earth and the moon, though clearly here $p=0$. I liked the idea of the little dog laughing as the lady next door had a dog that laughed, but I just could not cope with the dish running away with the spoon. Even at that early age I argued that this was a recognisable impossibility ($p=0$) as neither the dish nor the spoon had legs. Thus, probability has a range from 0.0, absolute impossibility, to 1.0, absolute certainty.

Very few events occur with probabilities of 0.0 or 1.0; most fall somewhere in the middle, some closer to 0.0, others closer to 1.0. It will not have gone unnoticed that there are two fundamental types of humans, male and female. There are other subclassifications, but these are of no importance. Contrary to popular belief the two varieties are very similar. The most readily appreciated differences are anatomical but these are just quantitative, not qualitative. Embryologically they are identical, and there isn't anything I've got (I happen to be male) that the next female I bump into hasn't got and vice versa. It is also obvious that the two types are about equally prevalent. Each time a new human is conceived there is a 50% chance that it will be female. It can't be both (or not very often) so there is an equal chance that it will be male. Thus, the probability either of being female or being male is 0.5.

This also applies when we spin a coin – it will end up either heads or tails. This assumes, of course, that we have a regular coin with both a head and a tail and that it did not come to rest on its edge, which is not impossible but extremely unlikely. The chance of the latter occurrence can be calculated with a number of assumptions about the angular velocity and momentum of the spin, the thickness of the edge compared with the radius, the elasticity of the coin (remember it's going to bounce) and a number of other things. Just for fun I did this calculation for a U.K. 2p coin and got a probability of 10^{-14}. (Extraordinary what some people do for fun isn't it, but I've been in medicine for some time now and *nothing* surprises me!) This result is probably not in error by more than three orders of magnitude either way. However, if 10^{-14} is correct and if the coin were spun once every 4.75 seconds, then you could expect it to land and come to rest on its edge just 1000 times in the whole lifetime of the universe to date (15,000,000,000 years). This is what is meant by the