Statistics

Statistics are used in a variety of ways in neuroscience. Perhaps the most familiar example is trying to decide whether some experimental results are reliable, using tests such as the t-test. However, one also often wants to analyse intrinsically random data, such as the number of quanta released in synaptic transmission, the opening and closing of ion channels, and the patterns of neural firing. Perhaps most interestingly, there is increasing realization that the brain itself functions as a statistical analyser of the environment. Our moment-to-moment experiences are ever different, and yet the brain learns to make sense of the turbulence. Actually, the endlessly shifting pattern of inputs produced by our world is far more variable than we realize, and the world appears relatively stable precisely because our brain has learned to understand it. To someone who knows no statistics at all, much of this lecture will seem like a random series of sounds and symbols, while to an expert it will appear to be crude summary of basic concepts.

Random Variables

Often in science, especially biology, if one makes repeated measurements of the same variable one gets different answers, even though the situation appears to be exactly the same. For example, in an electrophysiological experiment one might find that there is trial to trial fluctuation in a synaptic potential, either because of instrumental noise, or because of variation in the number of released vesicles. We say that potential is a "random variable". A random variable arises when measurements vary even though circumstances are apparently similar. In some cases the variation reflects the fact that the circumstances are actually varying in an unknown manner, but in others (quantum mechanics), the variation appears to be essential, and is not due to the operation of "unknown" or "hidden" mechanisms.

A very simple example is the outcome of tossing a die: there are 6 possible values of the variable, the face value of the die. In this case all outcomes are equally likely. Another example is tossing a bunch of n coins, and recording the number of heads. In this case there are n+1 possible values of the random variable, but clearly it is far more likely that half the coins will be heads than that all of them will be (since there are many different patterns of the former outcome than of the latter). Another example would be a variable defined as the product of the face value of throwing a pair of dice. Finally, one could measure the heights of a defined sample of humans (for example, the students in a class).

In the coin and dice tossing examples the random variables are discrete, since they can only take on integral values. In the human height case, the variable can take a continuous range of values, so the random variable "height" is continuous. We will be concerned with both discrete and continuous random variables.

A random variable takes on various possible values, and by making many repeated measurements of the variable in successive "trials" or "collections" one can deduce the relative frequency with which the variable takes particular discrete values (or, in the case of a random variable, lies between particular limits). The relationship between relative

frequency and actual value (or interval of values, in the case of a continuous random variable) is known as the distribution of the random variable. In the limit where the number of observations becomes infinite, we refer to the relative frequency as the probability. For example, in the die-tossing example, the probabilities of observing the 6 possible outcomes are each 1/6, and the random variable is distributed uniformly. In the coin-tossing example, the probability that all n coins are heads (which we could write as $p_{x=n}$) is much less than the probability that half the coins are heads ($p_{x=n/2}$). Here the distribution is nonuniform, and as we will see p_x as a function of x follows the binomial distribution. In the height example, typically the results are distributed around the mean height in a "normal" or "Gaussian" manner. In the Boltzmann distribution, we considered a particle, subject to thermal agitation, in either a high energy state or a low energy state (just like a coin that can be heads or tails). The distribution of p_{hi}/p_{lo} is exp-$\Delta E/kT$, and since $p_{lo} = 1-p_{hi}$ we got $p_{hi} = 1/(1+\exp\Delta E/kT)$.

Some discrete distributions.

Let us consider coin-tossing in more detail. We will consider a generalized version in which coins are biased, so the probability that a coin falls heads (which we will call p) is not necessarily the same as the probability of it falling tails. We want to calculate the probability that we will see x heads if we repeatedly throw a collection of n coins. The key assumptions we will make are (1) all the coins behave identically (same value of p) and (2) each coin behaves independently (the state of one coin does not depend on the others).
The key idea we need is that the probability of a joint event such as one coin being heads when another is also heads is given by the product of the individual probabilities. What is the probability of observing one specific configuration (or arrangement) of x heads in a collection of n coins out of the 2^n possible arrangements(eg HHTHTTH)? Clearly it will be $p^x (1-p)^{n-x}$ (the probability that x of the coins will be heads times the probability that n-x of them will be tails). Note that if p = 0.5, $p^x(1-p)^{n-x} = 0.5^n$, which is the reciprocal of the number of arrangements. However, we are not interested in the probability of a specific configuration, but that of any of the coinfigurations (bad pun) that have the same number of heads (such as HHTHTTH and HHHTTHT). How many configurations will have the same number of heads? There is only one configuration that has all heads (x=n), but there are n configurations that have 1 tail (x=n-1), since the any one of the n coins could be tails. More generally, we want to know how many combinations of x objects out of a total of n can be made, which is n!/(n-x)!x!

[Derivation: we can choose x objects in x steps (the first, the second, finally the xth). In the first step we have our choice of n objects, but in the second step only (n-1) are left to choose, and so on until the last, which can be chosen in (n-x) ways. So we can choose x objects in n(n-1)(n-2)….(n-x) = n!/(n-x)! ways. However, having chosen our x objects, they can be arranged in x! ways, since the first can be chosen in x ways, the second in x-1 ways etc. So of all the ways of choosing x out of n (i.e. n!/(n-x)!x!) , only a fraction 1/x! of them will show up as distinguishable.]

Thus the overall result is

$p_x = [p^x(1-p)^{n-x}]n!/(n-x)!x!$

which is the binomial distribution. For example, if n = 3, then the probabilities of getting 0,1,2, or 3 heads when p = 0.5 (i.e. tossing 3 unbiassed coins) are .125, .375, .375 and .125.

As a class exercise we will calculate p_x for the case n = 6.

The answers are
p0 = .0156 1
p1 = .0938 5
p2 = .2344 12
p3= .3125 16
p4 = .2344 12
p5= .0938 5
p6 = .0156 1

(the third column shows the expected number of heads or tails out of 50 trials)

An obvious question is, what is the average number of heads (m) that will be found? This is given, by $m = \Sigma p_x x$ where the sum runs over all possible outcomes i.e. from x=0 to x=n. This corresponds to a weighted sum, the contribution of each possible outcome being weighted by the probability of that outcome. The result, for the binomial distribution, is

$m = np$

which is intuitively right – it is just the probability that a coin is heads times the number of coins that are tossed.

We would also like to know how much the actual outcome scatters around this mean. The scatter is measured by the variance, σ^2, defined by var $x = \langle(x-m)^2\rangle$ where we use angle brackets to denote "average value". For the binomial distribution it is given by

$\sigma^2 = np(1-p)$.

The relation between σ^2 and m is therefore an inverted parabola. This relation makes sense – if p is very small, then we will almost always get the result x = 0 (very little variation). If p is close to 1, again we will almost always get x = 1. The maximum variance is achieved when p = 0.5.

It is useful to note that one can estimate n and p from σ^2 and m using the above 2 equations.

A practical application of the Binomial distribution

The existence of ion channels in biological membranes was first demonstrated by Katz and Miledi using noise analysis of the fluctuations of the depolarization produced by ACh at the neuromuscular junction. Shortly after, Andersen and Stevens used voltage clamp to provide a more direct estimate of the single channel conductance. The basic experiment is to record the steady depolarization (or the underlying current) at successive times and convert these measurements to the corresponding (slightly fluctuating) conductance change using Ohm's law. The conductance change fluctuates because the number of open channel is varying in a stochastic or probabalistic way. This provides estimates of both the mean conductance change m and the variance of the conductance change σ^2 . If the probability that a given channel is open is p, and this is independent of what the other channels are doing, the fluctuations should follow a binomial distribution, and a plot of σ^2 against m as p is varied (for example, by varying the ACh concentration) should be an inverted parabola. In these experiments the relation was actually linear, corresponding to the foot of the parabola, where $p << 1$. The single channel conductance γ could therefore be calculated using the formula $\sigma^2/m = \gamma^2 np/\gamma np$ and canceling the unknown np. The result was a value of about 30 pS, which is only slightly less than the value 45 pS subsequently directly determine by single channel analysis.

Two limiting cases of the binomial distribution.

Poisson

If n becomes very large, and p becomes very small, such that the product np remains finite, the Binomial distribution reduces to an important special case, the Poisson distribution, defined by

$p_x = e^{-m} m^x/x!$

where $m = np$ (and $\sigma^2 = m$).

If n becomes very large, then x can take on an enormous range of possible numbers. Although this is also a discrete distribution, it can be applied to events that occur randomly in time, at some constant average frequency f (per unit time), such as clicks of a Geiger counter. First, we can ask what is the probability of observing 0,1,2… clicks in a given time interval t (say 1 sec). In theory it is eventually *possible* that a very large number of events could occur, by chance, within a given time interval, though this would have a vanishingly small probability. Thus if we consider the number of events x out of the total possible n events, we expect a discrete Poisson distribution. A famous example is the annual number of fatal horsekicks in the 19^{th} century Prussian cavalry. Fortunately the probability of fatal horsekicks was very low, but because there were many soldiers, there were sometimes 1, 2 or even 3 deaths per year. It should be noted that if one tries to estimate n and p separately from data that are Poisson distributed, one will get measningless answers (very large values of n and small values of p, which change

enormously if only a few more observations are made). This is because the Poisson distribution depends only on the product np.

However, we could also consider the *continuous* distribution of the gap durations that are longer than a given time t. By definition within such a gap no events occur, so we need to consider the Poisson distribution for x = 0. This leads to

$p_0 = e^{-m} = e^{-ft}$

where p_0 refers to the probability of observing a gap lasting greater than t seconds.

Clearly, all the gaps between events must be greater than zero. So for t = 0, $p_0 = 1$. Also, if the events are occurring at a nonzero rate, we will never observe a gap that lasts forever. These 2 extremes are clearly correctly predicted by the formula. This formula is the continuous Poisson distribution for randomly occurring events. The "time constant" (or 1/e time) for the exponential is simply the reciprocal of the event frequency. It is equal to the mean gap duration. Note that because the distribution is nonsymmetric (skewed) the median gap duration (0.693 f) is less than the mean. It is a *cumulative* continuous distribution, because it asks what is the probability that a random variable (in this case the time between events) is *greater* than a given quantity. The cumulative distribution of the probability that the time is less than a given interval is $1-e^{-ft}$.

Let us now examine the distribution of the duration of gaps between events. In this case, it doesn't really make sense to ask what is the probability of finding a gap of a specific duration. Instead, we ask what is the probability of finding a gap of duration lying between t and Δt, as a function of t. Δt is a "bin size". If we divide this probability by the time step Δt, we get a "probability density" (just as we get a density of a substance by dividing the mass by the volume). We then take the limit Δt approaching zero to define a probability density function, which gives the distribution of the gaps between events. This "pdf" is given by the derivative of the cumulative distribution of gaps that are NOT longer than t. So

pdf (gap duration) = $d(1-e^{-ft})/ dt = f e^{-ft}$

It has exactly the same shape as the previous cumulative distribution, and its "time constant" is also equal to the reciprocal of the frequency. Once again the time constant is equal to the reciprocal frequency. In practice, one plots a discrete approximation to the pdf, by using finite time bins Δt.

The Bus Paradox

An interesting application is to consider how long one must expect to wait for a bus if one is equally likely to arrive at the bus-stop at any time, given that the buses are equally likely to arrive at any time, with an average frequency of 1 per hour. Intuitively one thinks that one is equally likely to arrive just before the bus arrives as just after, therefore typically needing to wait 1/2 hour. However, this would only be true if the bus arrives at

exactly 1 hour intervals. But because the bus is arriving randomly, it is much more likely to arrive during long waits than during short waits (or conversely, one is much more likely to arrive during long interbus intervals). Thus the bus arrived on average 1 hour *before* you arrived at the bus-stop, and the next bus will on average arrive in 1 hour. In a way it makes sense – whenever you arrive, the bus is not there, and the next bus arrives independently of when the last bus came. The key assumption here is that the events are independent – the probability of an event does NOT depend on whether other events have already happened, or when.

Normal.

If we let n grow without constraining p, then there are more and more possible values of x. In the limit of large n, x becomes effectively a continuous number, and we need the pdf of x. It can be shown that in this case the Binomial distribution approaches the Normal or Gaussian distribution:

$pdf(x) = [\exp-(x-m)^2/2\sigma^2]/ \sigma \text{ root } 2\pi$

where m and σ^2 the same significance as in the full Binomial distribution (i.e. np and np(1-p)). This is a symmetrical bell-shaped curve. The area under the curve is one. As σ^2 increases, the curve get wider and flatter, preserving the area. The cumulative normal distribution is also interesting: it is the integral of the pdf, and is given by the "error function" of x. It plots the probability that the variable takes on a value less than x, as a function of x.

The fact that the Binomial approaches the Normal as n increases is a special case of a remarkable general principle. According to the *central limit theorem* the sum of many independent random variables drawn from identical distributions of *any* shape also approaches the Normal. In the case of the binomial, we are actually dealing with the sum of the values of n random variables (the face values of individual coins), each of which have the distribution $p_0 = p$ and $p_1 = (1-p)$. We explicitly used the independence of the individual random variables to deduce the distribution. In many cases in nature, such as height, it is suspected that the measured variable is the result of the addition of many separate independent variables (such as nutrition, exercise, genetics etc) each of which may have specific, and perhaps even rather unusual, distributions.

The normal distribution also arises in diffusion problems- it describes the probability of finding a particle at a given distance x from the starting point, and the variance reflects the diffusion coefficient. This is because the particle undergoes a random walk from the origin. In 1 dimension, at every moment the particle "tosses a coin" to decide whether to go backwards or forwards, so the outcome is described by a very large number of coin tosses.

2 or more random variables

If measurements of 2 different fluctuating quantities are made (for example the heights x and and weights y of students in a class), we can define a joint probability density function, the probability that x has a given value given that y has a given value, a 3D plot which can be drawn in 2D as a contour plot, or as a scatter plot showing individual observations. It is often useful to first subtract the relevant mean from the 2 random variables, so the distribution is centered on the origin. As well as the variance of the individual variables, we can also consider their covariance which is the average value of the product of the 2 zero-mean variables. It is a measure of how one variable increases as the other increases. If for example the points scatter equally about one of the axes, the covariance will be zero – we say the variables are uncorrelated.

In linear regression analysis, we try to fit the best straight line to the scattered data. Best is defined as the line that minimizes the sum of the squared distances of the data from the line, the distance being measured by a line through the point that is perpendicular to the proposed straight regression line. We could consider these vertical intersections on the regression line as representing a new random variable which is given by a weighted sum of the 2 original variables, the weights depending on the slope of the regression line. For example, in the case of the height and weight of the students in a class, the new variable could be called “size”, a measurement which takes both weight and height into account. If we could only assign one variable to each student, we might choose this new size variable rather than one of the original variables, because provided the original variables are correlated, the new variable gives information about both the original variables.

If the 2 variables, say x and y, are both normally distributed, then the closeness with which the data cluster around the fitted regression line is measured by Pearson’s correlation coefficient r, which is given by

R = cov (y,x)/ root var x var y.

If all the points lie exactly on the line, then R = 1.

The joint distribution of n random variables is n dimensional. A simple way to characterise the relation between these variables is to construct the covariance matrix, which is made of all the covariances (i.e. average pairwise products) of the zero-mean variables. The entries along the diagonal are the variances, and as expected if the offdiagonal elements (the covariances) are zero, the variables are uncorrelated. It can be shown that the direction of the least squares line that best fits the data is also the direction of the leading eigenvector of the covariance matrix (the eigenvectors of a matrix are sets of variables values which are not rotated by multiplication by the matrix). These directions are also known as principal components. It can be shown that neurons with Hebbian synapses perform principle component analysis, and these neurons provide statistically optimal representation of their inputs if these inputs have a normal distribution. However, Gaussian inputs are quite rare in the real world, and the brain probably performs even more powerful statistical analysis than least squares.

The Brain and the World

The task of the brain is to understand the world, i.e. to answer all questions such as given x,y,z…. what is X,Y,Z…? This essentially involves constructing the entire joint pdf for all input and output variables (including at least 100 million retinal pixel values). Even if this pdf is only determined to within 10% accuracy, this still involves well over $10^{\text{hundred million}}$ quantities, which is clearly impossible. Fortunately, the task is made easier by 2 important considerations (1) the pdf seems to be smooth (if yellow bananas are good, slightly green or brown bananas are at least edible) (2) it is mostly empty (there are no blue or orange bananas). It is an interesting question whether a Hebb rule can be used to estimate this redundant but still immensely complex pdf.