

# *SLUO Lectures on Statistics and Numerical Methods in HEP*

## *Lecture 8: Signals, Backgrounds and Probabilities*

*Roger Barlow*

The University of Manchester  
29<sup>th</sup> August 2000

### **1. Hypothesis testing**

The rather dry phrase ‘hypothesis testing’ is given a lot of treatment in statistics books for the ‘soft’ sciences. It is relatively neglected in the physical sciences (though accounts can be found in [1] and [2]). So when it is needed, it comes as an unpleasant surprise.

It is the area of statistics that covers questions where the answers are not numerical (0.11 or whatever) but logical - ‘Yes’ or ‘No’. There are only these two possible answers, so the questions have to be carefully framed. And because this is statistics, the answers may be incorrect.

So questions like ‘Is this a pion or a Kaon?’ or ‘Is this signal or background?’ are inadmissible. Admissible versions are ‘Is this a pion?’, and ‘Is this a background event?’.

Statistics can never tell you directly that something is true. Only that something is false. That can lead to the establishment of a true statement, but only with the extra help of logic or other knowledge. (Statistics says a coin is not unbiased; logic then says it is biased. Statistics says that a track cannot be a muon and cannot be a hadron. Physics then says it is an electron.) A particle identification system cannot tell you what a track is, only what it could be and what it could not be.

Information is generally given in the form of probabilities. There is the potential for a lot of confusion here, as a probability refers to an event and an ensemble. (That’s classical probability. In Bayesian probability this problem takes the form of the need to define a prior.) If you are told that the probability for a pion of momentum  $400\text{MeV}/c$  and a Čerenkov ring radius  $2.2m$  is 40%, does this mean the probability that a pion will have this ring radius? Or that a particle of this ring-radius is a pion? Does it mean exactly  $2.2m$ , or  $\geq 2.2m$ , or  $\leq 2.2m$ ? This can be very confusing, but if you keep a clear head the problems go away. This means *always* knowing what you mean by ‘probability’ whenever you use it. Disguising the ambiguity in ‘probability’ by calling it something else (usually ‘Likelihood’) does not help; it makes things worse.

#### **1.1 Some jargon**

We say something and ask ‘Is it true?’: more long-windedly we construct a hypothesis which we call *the hypothesis* or sometimes just  $H$ . If this is unique (‘This is a pion.’, ‘The mass is  $3.1\text{GeV}/c^2$ ,’) then it’s a *simple* hypothesis. If it encompasses several possibilities (‘This is a hadron.’, ‘The mass is between  $3.0\text{GeV}/c^2$  and  $3.2\text{GeV}/c^2$ .’) then it’s a *compound* (or *composite*) hypothesis.

What if it isn’t true? We need to spell out the alternative to the hypothesis, which we call *the alternative hypothesis* or sometimes just  $H_A$ . It also may be simple or compound. (Very often

the hypothesis is simple and the alternative is compound. If this particle is not a pion it could be a Kaon or a muon or a proton or an electron. If this coin is not fair, it is biased to some arbitrary degree.)

This clearly gives two ways for us to get it wrong when we make our decision about  $H$ . We can *reject a true* hypothesis. (For example: throw away a good event from our sample.) This is called a *Type I error*. Or we can *accept a false* hypothesis, i.e. accept  $H$  when in fact  $H_A$  is true. (For example, accept a junk event in our sample.) This is called a *Type II error*. (There are also 2 ways to get it right - but these don't get mentioned.)

The question soft\* scientists most often ask of the data is 'Is there any effect?'. Does the new drug affect recovery rates? Has the new web-site design affected sales? The correct way to ask this question of the data is to propound the hypothesis that the effect is absent. (Recovery rates are unchanged. Sales have changed only though other factors.) This hypothesis that nothing has happend is called the *null hypothesis* or just  $H_0$ , and is a simple hypothesis. The alternative is compound: there is an effect, but we're totally unspecific about its magnitude . (Though we may be able to be specific about the sign: is it on the cards that the drug makes recovery rates worse? Or that the website has driven customers away?)

If the data refute the null hypothesis then that suggests the effect exists. (Recovery rates for patients who took the drug are so different from normal that it must be doing something. Sales have gone up significantly more than those of similar products.)

**Example 1:** Fisher says that 'Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.' Comment.

## 2. The $\chi^2$ distribution

Very often the PDF(s) you have to handle follow a  $\chi^2$  distribution, so let's deal with that here.  $\chi^2$  is defined as  $\sum \left(\frac{d_i}{\sigma_i}\right)^2$ , the sum of squares of some deviations scaled by the standard errors. (The more general  $\tilde{\mathbf{d}}\mathbf{V}^{-1}\mathbf{d}$  can be rotated into this diagonal form.) Assuming that these deviations are Gaussian, then the probability of a particular  $d_1, d_2...d_N$  is proportional to

$$e^{-d_1^2/2\sigma_1^2} e^{-d_2^2/2\sigma_2^2} \dots e^{-d_N^2/2\sigma_N^2} = e^{-\chi^2/2}$$

Now there are a whole lot of points in  $\mathbf{d}$  space that give the same  $\chi$ : the probability density of  $\chi$  is the product of this phase-space factor proportional to  $\chi^{N-1}$  (just as the perimeter of a circle is proportional to  $r$ , the surface area of a sphere to  $r^2$ ). This gives the PDF for  $\chi$ : to convert to  $\chi^2$  we divide by  $\frac{d\chi^2}{d\chi}$ . Including the normalisation constant gives

$$P(\chi^2; N) = \frac{2^{-N/2}}{\Gamma(N/2)} \chi^{2\frac{N}{2}-1} e^{-\chi^2/2}$$

---

\* The term 'soft' is not meant in any derogatory sense. Soft science is hard! Effects are subtle, and controlled experiments are very difficult. Whatever the properties of  $b$  quarks, they do not fail to return questionnaires (sociology), get eaten by rabbits (agriculture) or sue you for malpractice (medicine). Be happy you're a physicist.

Now suppose that the  $d_i$  are still distributed according to a multidimensional Gaussian, but are also subject to a homogeneous linear constraint

$$C_1d_1 + C_2d_2 \dots + C_Nd_N = 0$$

This constraint defines an  $N - 1$  dimensional subspace of the original. but the probability density is still proportional to  $e^{-\chi^2}$  everywhere. The only difference is that when you integrate over the subspace you pick up a factor of  $\chi^{N-2}$  instead of  $\chi^{N-1}$ . (The intersection of a plane and a sphere gives a circle.) The  $\chi^2$  distribution for  $N$  terms but with a constraint is the same as the unconstrained  $\chi^2$  distribution for  $N - 1$  terms. 2 constraints give  $P(\chi^2; N - 2)$ , etc.

When parameters are adjusted so that the predictions fit the data better, this often takes the form of a homogeneous linear constraint. For example, normalising predicted numbers of events in a histogram to the total in the actual data,  $\sum f_i = N$ , is equivalent to  $\sum f_i - n_i = 0$ . The equations obtained by setting the differentials of  $\chi^2$  wrt. fitted parameters to zero have this form if the prediction is a linear function of the parameters and the errors are constant. Thus if you have  $N$  terms in the sum, but have adjusted  $M$  parameters to ‘fit the data’ (i.e. minimise  $\chi^2$ ) then the resulting  $\chi^2$  follows the distribution  $P(\chi^2; N - M)$ .  $N - M$  is known as the number of *Degrees of Freedom*,  $N_D$ .

This is a pretty unique property of the  $\chi^2$  as a measure of agreement. Others (for example the Kolmogorov test statistic) don’t have this nice easy way of incorporating the fact that the predictions have been adjusted to fit the data so the agreement is bound to be improved. Strictly speaking, if the fitting is not represented by a set of homogeneous linear constraints the ‘degrees of freedom’ business is invalid, but people don’t worry about this.

The distribution has a mean of  $N$  – not surprising, each term in the sum contributes 1, on average. Thus the  $\chi^2$  per degree of freedom should be roughly 1. The standard deviation of that is  $\sqrt{2N}$ , but as the distribution is not a good Gaussian unless  $N$  is very large, this is not that useful. The distribution is broad and (legitimate) large  $\chi^2$  values are not uncommon. A quantitative way of evaluating the ‘goodness’ of a  $\chi^2$  is to use the probability  $Prob(\chi^2; N)$  discussed in the next section.

It is worth pointing out that for  $N_D > 2$ , the probability function is 0 at  $\chi^2 = 0$ . Although the most likely value for each deviation is zero, it is unlikely that they will all be small. Persistently small  $\chi^2$  values indicate a problem with overgenerous errors. Fortunately this problem is rare (at least, much rarer than persistently large  $\chi^2$  values.).

### 3. What you can do with one PDF

We need some raw material for the statistics to work on. Let’s suppose you have some sort of measurement which tells you whether a track likely to be a pion. In a basic form, it tells you a number which you know tends to be (say) low for pions, and high for non-pions. In general this is called a *discriminating variable* or *test statistic* or *indicator* and we’ll call it  $x$ . The figure shows a probability density for  $x$  under the (simple) hypothesis  $H$ . (If it’s not a simple hypothesis, you can’t draw the PDF.) In some cases (e.g. if  $x$  is the number of hits in the muon detector) this can be a probability distribution rather than a probability density.

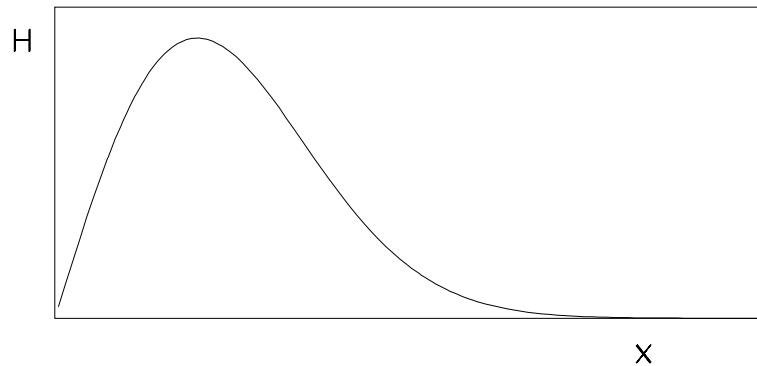


Figure 1: PDF for a discriminator under the hypothesis  $H$

To make a yes/no decision, we choose an *acceptance region* for  $x$ . We will accept  $H$  if an event falls in this region. Everywhere else is known as the *rejection region*

There are 3 possible topologies.

- i The regions can be divided by a single cut.
- ii The acceptance region can be defined by an upper and lower cut. If the PDF is symmetric, and if the Alternative Hypothesis has a symmetric PDF then these cuts will presumably be symmetric, but this is not true in general.
- iii The acceptance region may be split into several sections by many cuts.

If the discriminator  $x$  is multidimensional there are even more possibilities.

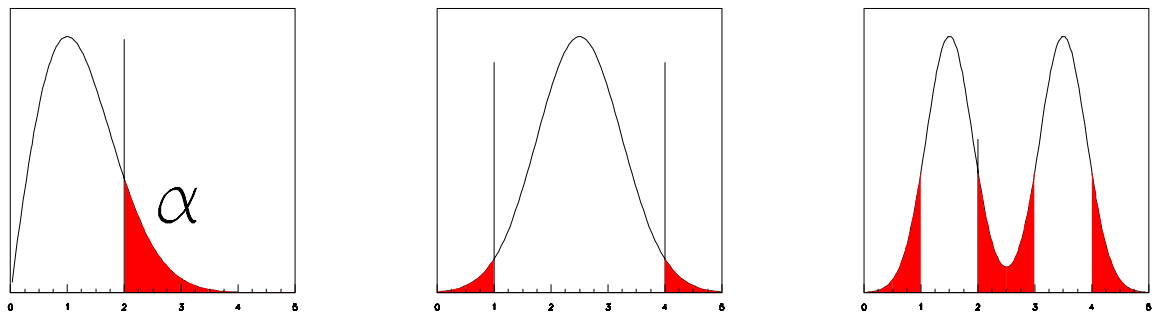


Figure 2: Possible accept/reject algorithms

Where and how we put the cut(s) is a matter of choice. There is no ‘best’ prescription (expect in the restricted case of the Neyman-Pearson Lemma, discussed later.)

Let’s see what we can define.

The *efficiency* of your selection is just the probability of avoiding a Type I error. It is the integral over the acceptance region, and is well defined in all cases above. It is a property of the PDF and the cuts. You can establish it using a sample of pure data for which  $H$  is true. This can be done using Monte Carlo data, or (better) by using some real data which has been selected by some means totally disconnected from your detector (a *control sample*).

Now let’s take a measurement:

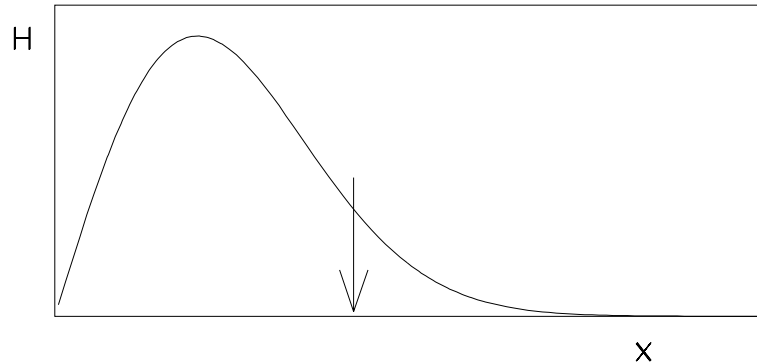


Figure 3: Still life: PDF and Data Value

For a given set of cuts we can say ‘yes’ or ‘no’, but what more can we say?

We have the probability density value  $P(x; H)$ : the probability that if  $H$  is true it would give an  $x$  like this. This by itself is pretty meaningless.

If the acceptance and rejection regions are divided by a single cut, then we can consider what would happen if we had placed that cut at this  $x$ . We can draw a figure like the first of Fig 2, except that this cut has come out of the data, not imposed from outside.

In the scenario as drawn, low  $x$  values are ‘good’ and high  $x$  values are ‘bad’. The fraction of the integrated PDF above  $x$  is referred to, especially if that curve is a  $\chi^2$  distribution, as the ‘probability’.  $Prob(x)$  or  $Prob(\chi^2, N)$ . It is the probability that, if the hypothesis is true, a measurement would give a value of  $x$  this bad or worse.

$$Prob(x) = CL(x) = SL(x) = \int_x^\infty P(x'; H) dx'$$

The language may be unfortunate, but it is deep rooted. The PDG has replaced the misleading term ‘probability’ with the term ‘confidence level’ which is probably worse, especially as they define it to mean the opposite of what any reasonable person (such as the authors of reference [3]) would take it to mean.

This fraction outside is also called the *significance*. and given the symbol  $\alpha$ . This is more soft science jargon. The null hypothesis is denied if the discriminator falls outside the cuts. There is a probability  $\alpha$  that this will happen even if the null hypothesis is true (Type I error.) Hence the expression ‘significant at the 5% level’ means there is (only) a 5% chance that data this weird could have arisen by a random fluctuation from the null hypothesis. (Small significances are good. The smaller the better. This seems to me counter-intuitive.)

If the acceptance regions is defined by two cuts, and these are symmetric, then it is still OK to define probabilities in this way, as we still have the concept ‘worse than  $x$ ’

$$Prob(x) = CL(x) = SL(x) = \int_{-\infty}^{\mu-x} P(x'; H) dx' + \int_{\mu+x}^\infty P(x'; H) dx' = 1 - \int_{\mu-x}^{\mu+x} P(x'; H) dx'$$

If you want to define it with more complicated topologies, it can only be done if all the cuts can be determined by a single parameter, for example rejecting everywhere the PDF falls below a certain value. Then ‘worse than  $x$ ’ means ‘lower PDF than  $P(x)$ ’ and the region can be

integrated over. This may make you feel good, but it is not a universal prescription. When you make cuts you should do so with at least one eye to the alternative Hypothesis.

If you take a number of these measurements for which  $H$  is true and plot the distribution in  $Prob(x)$ , it should be flat. (10% of the data should lie in the top 10% of the PDF. Another 10% should lie between the 20% line and the 10% line.) This is a very revealing distribution to plot. It will not look right until you've got everything right – unlike pull distributions which can look good to the casual glance even when they're off centre/too wide/too narrow/non Gaussian.

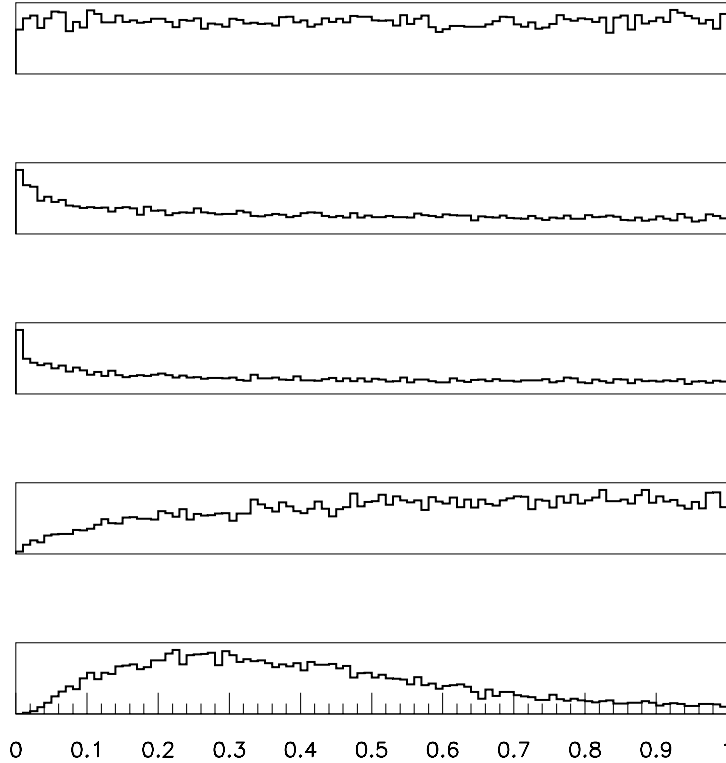


Figure 5: Some probability distributions.

The figure shows some examples using  $Prob(\chi^2; 1)$ . The top plot is what a probability distribution should look like. The second has low probabilities (high  $\chi^2$ ) as the predictions are wrong (by  $0.7 \sigma$ .) The third looks similar, but is in fact due to  $\sigma$  being underestimated by 20%. Overestimating  $\sigma$  (by 25%) leads to the rising slope of the 4th plot. In the final plot a prediction discrepancy of  $2\sigma$  has been ‘masked’ by increasing the error estimate to get the right mean  $\chi^2$ , but the distribution is seriously unflat.

When working with real data you will see a spike at low probability. This is inevitable from background events. It may also come from non-Gaussian tails of signal events.

### 3.1 Serious language warning

*Likelihood:* This is dangerous and misleading. Technically it is used to mean combined probability of a set of results (as in ‘Maximum Likelihood estimate’.) It is the probability (density) for the set of results  $x_1 \dots x_N$  under the hypothesis  $H$ , e.g. that a parameter has a particular value

$$L(x_1, x_2 \dots x_N; H) = P(x_1; H)P(x_2; H) \dots P(x_N; H)$$

(if they are uncorrelated: the extension to correlated values is not a problem.)

There is a widespread and evil desire to pervert this form by writing it backwards and claiming it tells you something about  $H$  given  $x_1...x_N$

$$L(H; x_1...x_N)$$

Thus Reference [3] defines  $L(\pi^+; p_{obs}, x_{obs})$  as the Likelihood for a pion of a track with measured momentum  $p_{obs}$  and PID-response  $x_{obs}$ . Their Likelihood is the same as the probability

$$\mathcal{L}(H; p, x) \equiv \mathcal{P}(x; p, H)$$

(Ref [3], Equation 1). They say the difference is ‘subtle’. It is also pointless and misleading.

After all, for many PID variables the signatures of a pion and a muon are pretty much the same. But a construct which tells you that the ‘pion likelihood’ and the ‘muon likelihood’ are the same is illogical in a world where pions are many times more plentiful than muons..

#### 4. What you can do with Two PDFs :The Neyman Pearson Lemma

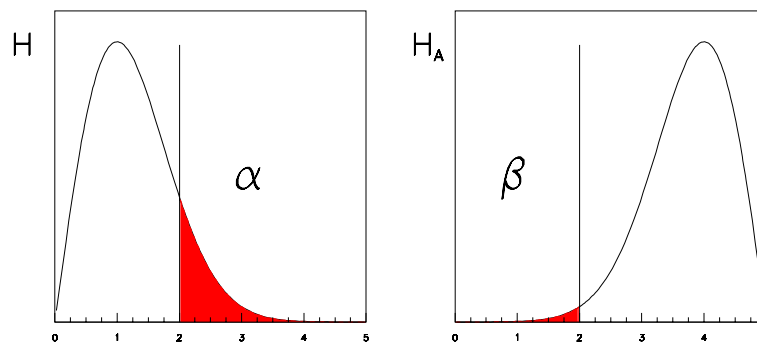


Figure 5: PDFs for  $H$  and  $H_A$

If the alternative is simple, then we can plot its pdf for  $x$ , and for any specified cut(s) evaluate the probability of an unwanted event making it through. This is known as  $\beta$ , and  $1 - \beta$  is (in the language of null-hypothesis-refutation) called the *power*. In our language,  $\beta$  is the *contamination probability*: the probability for a false event to end up in your sample  $N_{accepted}^{junk}/N_{total}^{junk}$ . It's the probability of a Type II error when given an event produced by  $H_A$ .

In this most basic case – both hypotheses simple: it's a pion or a kaon, it's  $B\bar{B}$  signal or  $udsc$  background – we can say what  $\alpha$  and  $\beta$  will be for any set of cuts. In this case one can even ask what the ‘best’ cut(s) are, and get the reply (Neyman Pearson Lemma): for a given  $\alpha$ , the rejection region with the smallest  $\beta$  is the region of greatest  $P_{H_A}(x)/P_H(x)$  i.e. the region is defined to be those  $x$  values for which  $P_{H_A}(x)/P_H(x) > c$ , where  $c$  is chosen to get the desired  $\alpha$ .

This is easy to see. Suppose you set up such a rejection region. If slice at  $x_r$  of width  $\Delta$  were to be moved from the rejection region to the acceptance region, it must be replaced (to preserve  $\alpha$ ) by a slice from some  $x_a$  of width  $\frac{P_H(x_r)}{P_H(x_a)}\Delta$ . The change in  $\beta$  is then  $P_{H_A}(x_r)\Delta - \frac{P_H(x_r)P_{H_A}(x_a)}{P_H(x_a)}\Delta$  which is positive (i.e. worse) because by construction  $P_{H_A}(x_r)/P_H(x_r) > P_{H_A}(x_a)/P_H(x_a)$ .

A very significant feature of this Lemma is the restricted conditions under which it applies: only for simple versus simple hypotheses. If the alternative hypothesis is composite, one worries about the worst possible case.

**4.1 Efficiency and Purity**

Notice that even in this ideal case, you do *not* have a handle on the *purity* of your sample. It is given by

$$p = \frac{\alpha S}{\alpha S + \beta B}$$

where  $S$  and  $B$  are the numbers of signal and background events, and you *don't necessarily know* their relative size. (After all, if you do, why are you making this measurement?)

**Example 2:** ACME corporation is selling a muon detector with an 80% efficiency for muons and a 5% contamination probability from hadrons. Should we buy one?

People often talk about the ‘efficiency and purity’ of a sample in one breath. They are very different animals. To know your efficiency you (only) need a pure signal sample. To know your purity you need pure background sample(s) *and* you need to know the amount of background there in the first place.

This is one case where Bayes theorem is a great help. Suppose you know that there are 5 times as many pions in your data as kaons (and no other types of particle.). You get an  $x$  value. Then

$$P(\pi; x) = \frac{5P(x; \pi)}{5P(x; \pi) + P(x; K)}$$

Information from many detectors can be combined in this way (assuming they’re independent.) Reference [3] calls this a ‘relative probability’ but the word ‘relative’ is unhelpful. It is the perfectly straightforward probability that a particle with a particular value of the indicator variable  $x$  is a pion. (They’re also called posterior probabilities, or conditional probabilities, but that refers to the way they were derived, not what they mean.)

Just be careful that this 5:1 ratio really applies to your dataset. It may be true for the total event sample, but if you’re looking at some particular channel for which the ratio is different, that invalidates the algebra. The fact that your ensemble is part of the larger ensemble is irrelevant. If you determine it for your data sample, fine. If you take it from somewhere else, be prepared to defend it. You can vary it by the uncertainty and evaluate a systematic error.

If you don’t know these *a priori* numbers then similar ratios formed without them are not very informative, and potentially dangerous.

**5. Combining information**

Suppose you have a simple hypothesis  $A$  and a simple alternative  $B$ , and a whole set of variables  $x_1, x_2 \dots x_N$  all of which tell you something about whether an event belongs to one type or the other. You have full theoretical understanding (or large Monte Carlo datasets or control samples) of the PDFs.

Let the PDFs  $P^A(x_1 \dots x_N)$  and  $P^B(x_1 \dots x_N)$  have means  $\mu_1^A \dots \mu_N^A$  and  $\mu_1^B \dots \mu_N^B$  and the same variance matrix  $\mathbf{V}$ . Then the best discriminator that can be made out of a linear sum of the  $x_i$  is the *Fisher Discriminant* [4]

$$y = (\tilde{\mu}^A - \tilde{\mu}^B)\mathbf{V}^{-1}\mathbf{x}$$

Which is elegant. (There is an extension to different variances which is rather less elegant.) But I should warn you that it’s usually not very useful. It can’t use information such that, for one

of the  $x_i$ ,  $A$  has a broad distribution while  $B$  has a sharp one. It is only sensitive to the difference. Maybe you can find a transformation which turns the difference in shape into a difference in mean.

But you're almost certainly better off using a neural network[5]. They are now a standard part of the Particle Physics toolkit, and are also much used by highly-paid financial analysts. A number of nodes  $n_j$  each add up the  $x_i$  discriminator values, scaling them by a weight  $W_{ij}$ . These totals are fed through some sigmoid-like function which is linear near the centre but saturates at large values. The outputs (one per node) are then fed to another layer of nodes which do the same thing, and so on until a final output layer gives a value which is the yes/no answer. The difficult part is getting the weight factors right – this is known as ‘training’.

### 6. Statistical separation

Suppose you have a discriminator variable  $x$  for which the two hypotheses (call them  $S$  and  $B$ ) have well-established behaviour. To study a sample (or a set of samples - perhaps histogram bins in another more interesting variable) you can weight the events, giving those with more signal-like  $x$  a high weight, and the implausible ones a low weight. This by-passes the whole efficiency/purity problem, though you never actually classify individual events as  $S$  or  $B$ .

If you weight by

$$w(x) = \frac{1}{c} \left( \frac{B(x)}{aS(x) + B(x)} - b \right)$$

where  $b = \int \frac{B(x)^2}{aS(x)+B(x)} dx$  and  $c = \int \frac{S(x)B(x)}{aS(x)+B(x)} dx - b$  then in a sample of mixed  $S$  and  $B$  this will project out the  $S$  events (as the expectation value for  $B$  is 0 and for  $S$  is 1.)

If the value chosen for  $a$  is the actual signal to background number in the sample, then this projection is as efficient as doing a maximum likelihood fit to the  $x$  distribution[6]. If not (and after all you don't know this ratio, you're trying to find it) it makes the method slightly less efficient but does not bias it. Likewise if  $S(x)$  and  $B(x)$  are not exactly right, this makes the method slightly less optimal but not invalid, provided the quantities  $b$  and  $c$  are evaluated accurately (probably by finding the mean value of  $\frac{B}{aS+B}$  for reliable signal and background Monte Carlo samples.)

### 7. References

[1] ‘Statistics’ .R. Barlow, Wiley 1989 Chapter 8  
 [2] ‘Statistical Data Analysis’. G. Cowan, OUP 1998, Chapter 4.  
 [3] ‘Some Statistics for Particle Identification’. J. Izen, A. Snyder, M. Sokoloff and R. Waldi, BaBar Note 422  
 [4] ‘The Advanced Theory of Statistics’. M.G. Kendall and A. Stuart, 4th 3 Volume edition, Charles Griffin & Co., 1983. Section 44.  
 [5] ‘JETNET 3.0: A versatile artificial neural network package’. C. Peterson, T. Rognvaldsson, and L. Lonnblad, Comput. Phys. Commun. 81 (1994) 185  
 [6] ‘Event Classification using Weighting methods’. R. Barlow, Journal of Computational Physics. 72 (1987) 202