# Lecture 2

1 Probability
Definition, Bayes' theorem, probability densities and their properties, catalogue of pdfs, Monte Carlo

→ 2 Statistical tests
general concepts, test statistics, multivariate methods, goodness-of-fit tests

3 Parameter estimation
general concepts, maximum likelihood, variance of estimators, least squares

4 Interval estimation
setting limits

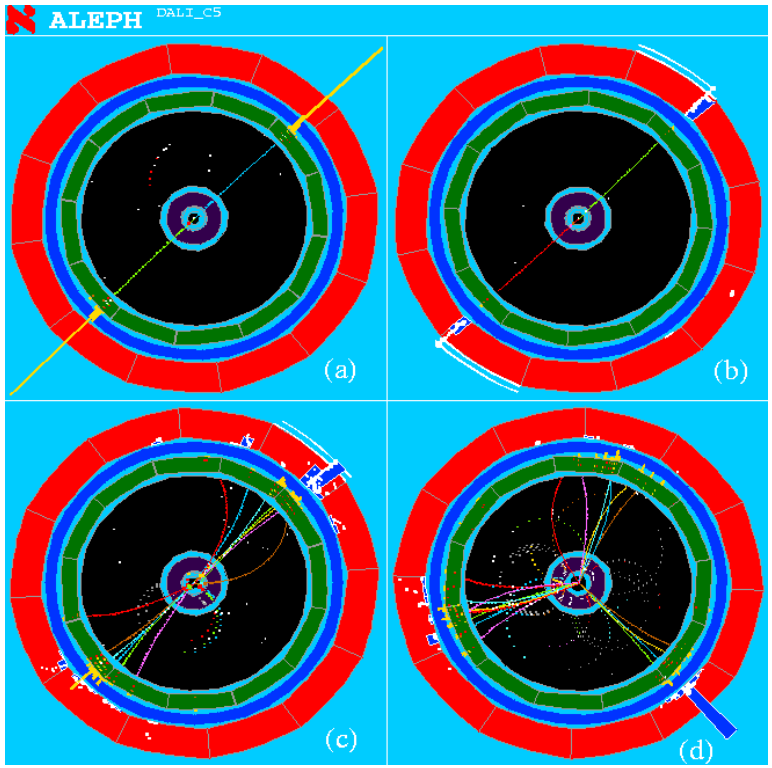
5 Further topics
systematic errors, MCMC

# The data stream

Experiment records a mixture of events of different types, each with different numbers of particles, kinematic properties, ...

We are usually interested in events of a single type, in a search to see if they exist at all and/or to identify them for further study.

$e^+e^- \to e^+e^-$




$e^+e^- \to \mu^+\mu^-$

$e^+e^- \to q\bar{q}$
$\to$ two jets

$e^+e^- \to q\bar{q}g$
$\to$ three jets

# Statistical tests (in a particle physics context)

Suppose the result of a measurement for an individual event is a collection of numbers $\vec{x} = (x_1, \ldots, x_n)$

$x_1$ = number of muons,

$x_2$ = mean $p_t$ of jets,

$x_3$ = missing energy, ...

$\vec{x}$ follows some $n$-dimensional joint pdf, which depends on the type of event produced, i.e., was it

$$pp \to t\bar{t}\ , \qquad pp \to \tilde{g}\tilde{g}\ , \ldots$$

For each reaction we consider we will have a hypothesis for the pdf of $\vec{x}$, e.g., $f(\vec{x}|H_0)$, $f(\vec{x}|H_1)$ , etc.

Often call $H_0$ the signal hypothesis (the event type we want); $H_1$, $H_2$, ... are background hypotheses.

# Selecting events

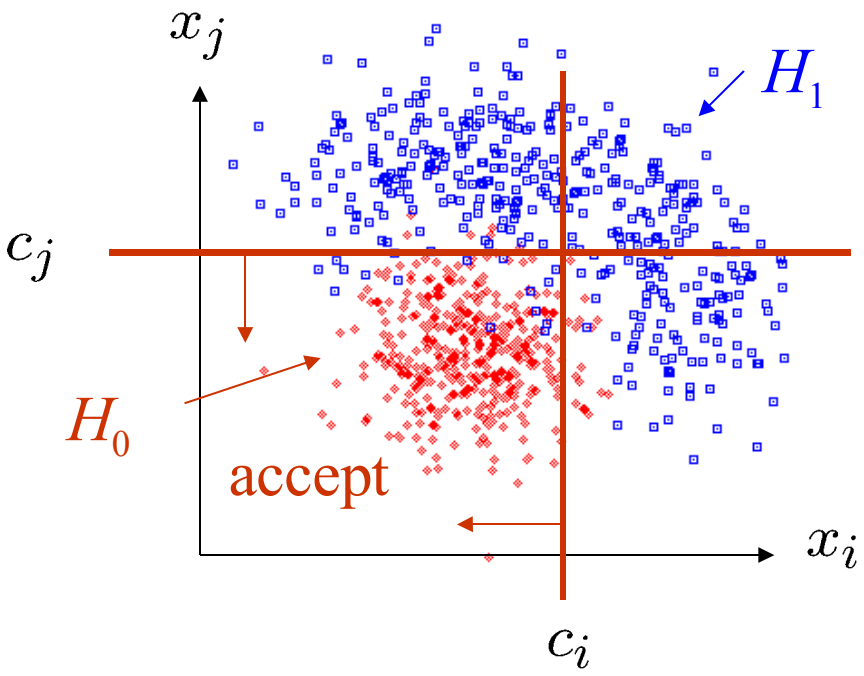
Suppose we have a data sample with two kinds of events, corresponding to hypotheses $H_0$ and $H_1$ and we want to select those of type $H_0$.

Each event is a point in $\vec{x}$ space. What 'decision boundary' should we use to accept/reject events as belonging to event type $H_0$?

Perhaps select events with 'cuts':

$$x_i < c_i$$

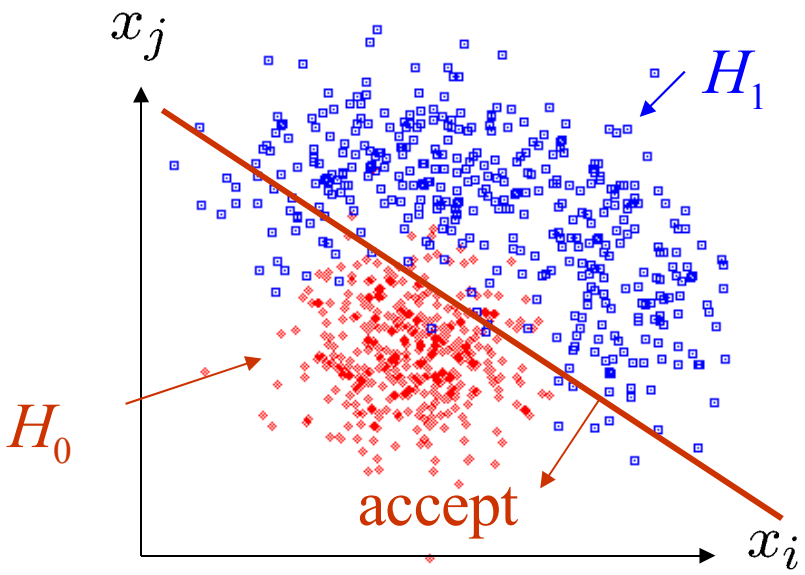
$$x_j < c_j$$



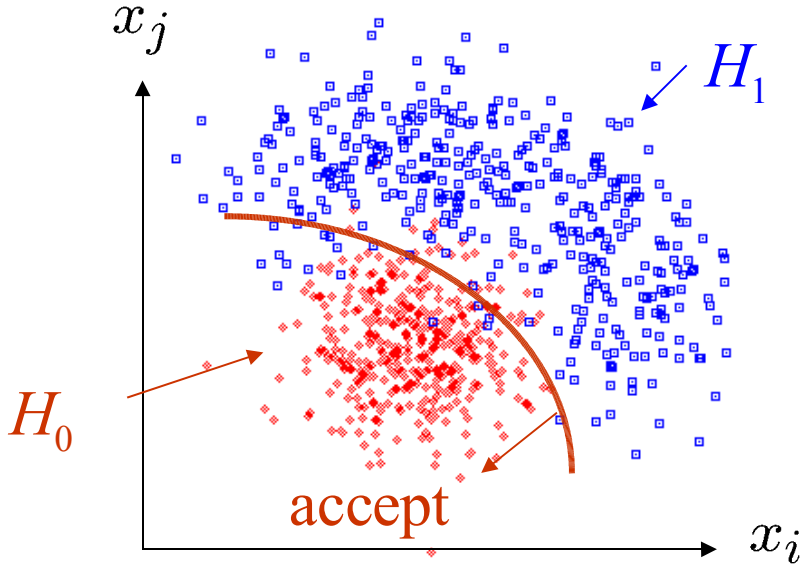
# Other ways to select events

Or maybe use some other sort of decision boundary:

linear




or nonlinear




How can we do this in an 'optimal' way?

# Test statistics

Construct a 'test statistic' of lower dimension (e.g. scalar)

$$t(x_1, \ldots, x_n)$$

Goal is to compactify data without losing ability to discriminate between hypotheses.

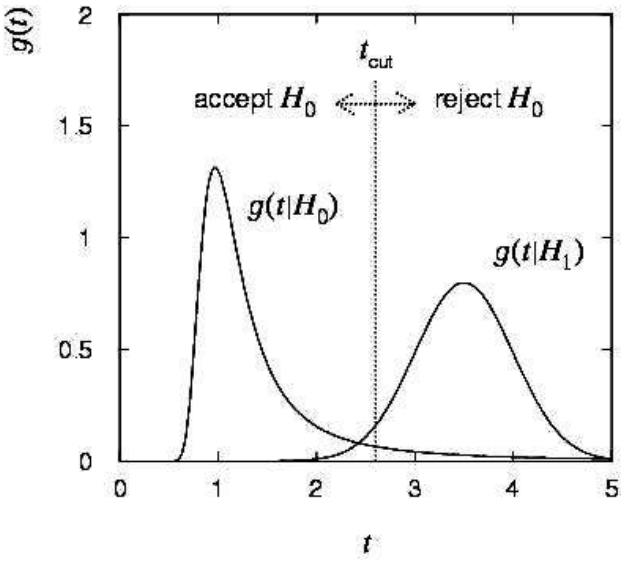
We can work out the pdfs $g(t|H_0),\ g(t|H_1),\ \ldots$

Decision boundary is now a single 'cut' on $t$.

This effectively divides the sample space into two regions where we either:

accept $H_0$ (acceptance region)

or reject it (critical region).



# Significance level and power of a test

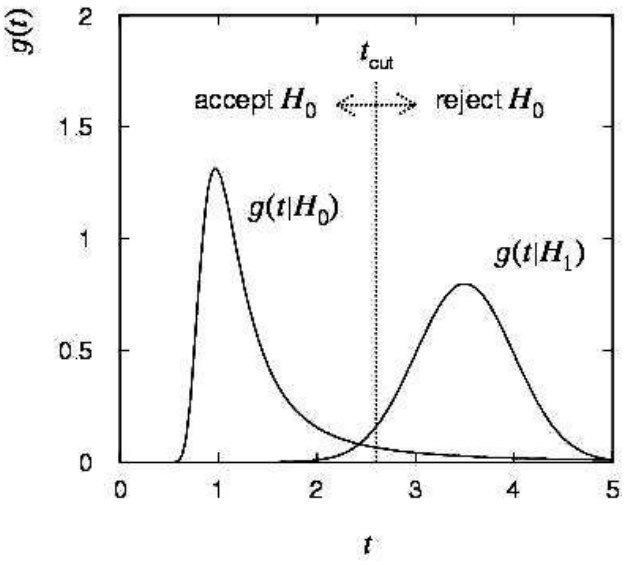
Probability to reject $H_0$ if it is true (error of the 1st kind):

$$\alpha = P(\text{reject } H_0 \,|\, H_0) = \int_{t_{\text{cut}}}^{\infty} g(t|H_0)\, dt \qquad \text{(significance level)}$$

Probability to accept $H_0$ if $H_1$ is true (error of the 2nd kind):

$$\begin{aligned} \beta &= P(\text{accept } H_0 \,|\, H_1) \\ &= \int_{-\infty}^{t_{\text{cut}}} g(t|H_1)\, dt \end{aligned}$$

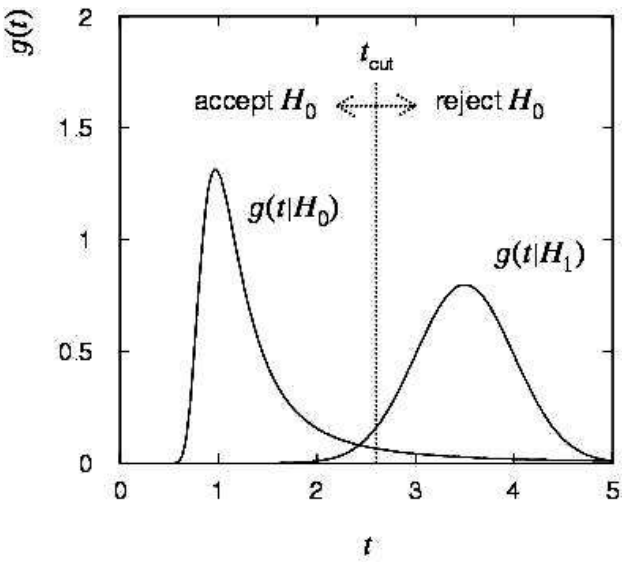
$(1 - \beta = \text{power})$



# Efficiency of event selection

Signal efficiency, i.e., probability to accept event which is signal,

$$\varepsilon_{\mathrm{s}} = P(\text{accept event} \,|\, \mathrm{s}) = 1 - \alpha$$




Background efficiency, i.e., probability to accept background event,

$$\varepsilon_{\mathrm{b}} = P(\text{accept event} \,|\, \mathrm{b}) = \beta$$

Expected number of signal events: $s = \sigma_s \, \varepsilon_s \, L$

Expected number of background events: $b = \sigma_b \, \varepsilon_b \, L$

$\sigma_s$, $\sigma_b$ = signal, background cross sections; $L$ = integrated luminosity

# Purity of event selection

Suppose only one background type b; overall fractions of signal and background events are $\pi_s$ and $\pi_b$ (prior probabilities).

Suppose we select events with $t < t_{\text{cut}}$. What is the 'purity' of our selected sample?

Here purity means the probability to be signal given that the event was accepted. Using Bayes' theorem we find:

$$P(\mathsf{s}|t < t_{\text{cut}}) = \frac{P(t < t_{\text{cut}}|\mathsf{s})\pi_\mathsf{s}}{P(t < t_{\text{cut}}|\mathsf{s})\pi_\mathsf{s} + P(t < t_{\text{cut}}|\mathsf{b})\pi_\mathsf{b}}$$

$$= \frac{\varepsilon_\mathsf{s}\pi_\mathsf{s}}{\varepsilon_\mathsf{s}\pi_\mathsf{s} + \varepsilon_\mathsf{b}\pi_\mathsf{b}}$$

So the purity depends on the prior probabilities as well as on the signal and background efficiencies.

# Constructing a test statistic

How can we select events in an 'optimal way'?

Neyman-Pearson lemma (proof in Brandt Ch. 8) states:

To get the lowest $\varepsilon_b$ for a given $\varepsilon_s$ (highest power for a given significance level), choose acceptance region such that

$$\frac{f(\vec{x}|\mathrm{s})}{f(\vec{x}|\mathrm{b})} > c$$

where $c$ is a constant which determines $\varepsilon_s$.

Equivalently, optimal scalar test statistic is

$$t(\vec{x}) = \frac{f(\vec{x}|\mathrm{s})}{f(\vec{x}|\mathrm{b})}$$

N.B. any monotonic function of this is just as good.

# Purity vs. efficiency — optimal trade-off

Consider selecting $n$ events:

expected numbers $s$ from signal, $b$ from background;

→ $n \sim$ Poisson $(s + b)$

Suppose $b$ is known and goal is to estimate $s$ with minimum relative statistical error.

Take as estimator: $\hat{s} = n - b$ .

Variance of Poisson variable equals its mean, therefore

$$V[\hat{s}] = V[n - b] = V[n] = s + b \quad \rightarrow \quad \frac{\sigma_{\hat{s}}}{s} = \frac{\sqrt{s+b}}{s}$$

So we should maximize $\dfrac{s}{\sqrt{s+b}}$ (or $\varepsilon_s/\sqrt{b}$ if $s \ll b$),

equivalent to maximizing product of signal efficiency × purity.

# Why Neyman-Pearson doesn't always help

The problem is that we usually don't have explicit formulae for the pdfs $f(\vec{x}|\mathrm{s})$, $f(\vec{x}|\mathrm{b})$ .

Instead we may have Monte Carlo models for signal and background processes, so we can produce simulated data, and enter each event into an $n$-dimensional histogram.

Use e.g. $M$ bins for each of the $n$ dimensions, total of $M^n$ cells.

But $n$ is potentially large, → prohibitively large number of cells to populate with Monte Carlo data.

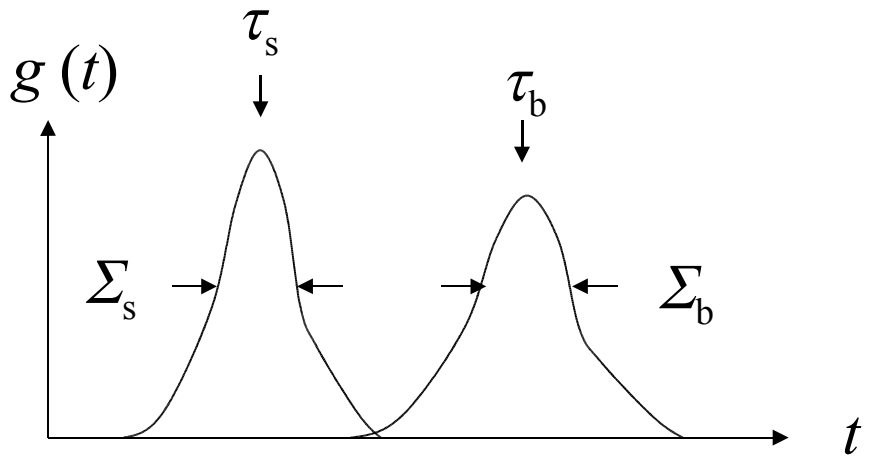
Compromise: make Ansatz for form of test statistic $t(\vec{x})$ with fewer parameters; determine them (e.g. using MC) to give best discrimination between signal and background.

# Linear test statistic

Ansatz: $$t(\vec{x}) = \sum_{i=1}^{n} a_i x_i$$

Choose the parameters $a_1$, ..., $a_n$ so that the pdfs $g(t|\mathrm{s})$, $g(t|\mathrm{b})$ have maximum 'separation'. We want:

large distance between mean values, small widths




→ Fisher: maximize $$J(\vec{a}) = \frac{(\tau_s - \tau_b)^2}{\Sigma_s^2 + \Sigma_b^2}$$

# Determining coefficients for maximum separation

We have
$$(\mu_k)_i = \int x_i f(\vec{x}|H_k)\, d\vec{x}$$
$$(V_k)_{ij} = \int (x-\mu_k)_i (x-\mu_k)_j f(\vec{x}|H_k)\, d\vec{x}$$

where
$$k = 0, 1 \quad \text{(hypothesis)}$$
$$i, j = 1, \ldots, n \quad \text{(component of } \vec{x}\text{)}.$$

In terms of mean and variance of $t(\vec{x})$ this becomes

$$\tau_k = \int t(\vec{x}) f(\vec{x}|H_k) d\vec{x} = \vec{a}^T \vec{\mu}_k \,,$$
$$\Sigma_k^2 = \int (t(\vec{x}) - \tau_k)^2 f(\vec{x}|H_k)\, d\vec{x} = \vec{a}^T V_k \vec{a} \,.$$

# Determining the coefficients (2)

The numerator of $J(\boldsymbol{a})$ is

$$(\tau_0 - \tau_1)^2 = \sum_{i,j=1}^{n} a_i a_j (\mu_0 - \mu_1)_i (\mu_0 - \mu_1)_j$$

'between' classes

$$= \sum_{i,j=1}^{n} a_i a_j B_{ij} = \vec{a}^T B \vec{a} \,,$$

and the denominator is

'within' classes

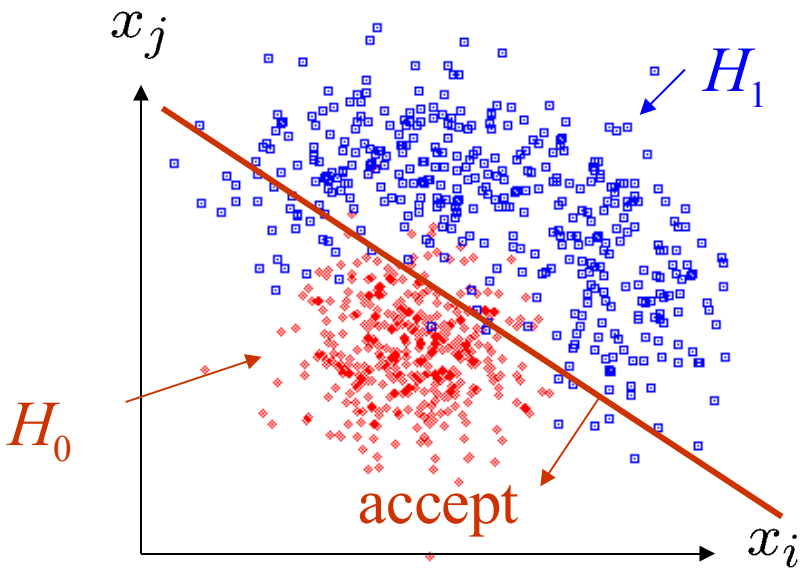
$$\Sigma_0^2 + \Sigma_1^2 = \sum_{i,j=1}^{n} a_i a_j (V_0 + V_1)_{ij} = \vec{a}^T W \vec{a}$$

→ maximize $J(\vec{a}) = \dfrac{\vec{a}^T B \vec{a}}{\vec{a}^T W \vec{a}} = \dfrac{\text{separation between classes}}{\text{separation within classes}}$

# Fisher discriminant

Setting $\dfrac{\partial J}{\partial a_i} = 0$ gives Fisher's linear discriminant function:

$$t(\vec{x}) = \vec{a}^T \vec{x}, \qquad \text{with } \vec{a} \propto W^{-1}(\vec{\mu}_0 - \vec{\mu}_1)$$




Corresponds to a linear decision boundary.

# Fisher discriminant for Gaussian data

Suppose $f(\vec{x}|H_k)$ is multivariate Gaussian with mean values

$$E_0[\vec{x}] = \vec{\mu}_0 \text{ for } H_0\,, \qquad E_1[\vec{x}] = \vec{\mu}_1 \text{ for } H_1\,,$$

and covariance matrices $V_0 = V_1 = V$ for both.

For this case we can show that the Fisher discriminant is equivalent to using the likelihood-ratio, and thus gives the maximum purity for a given efficiency.

For non-Gaussian data this no longer holds, but linear discriminant function may be simplest practical solution.

Can try to transform data so as to better approximate Gaussian before constructing Fisher discrimimant.

# Nonlinear test statistics

The optimal decision boundary may not be a hyperplane,

→ nonlinear test statistic $t(\vec{x})$

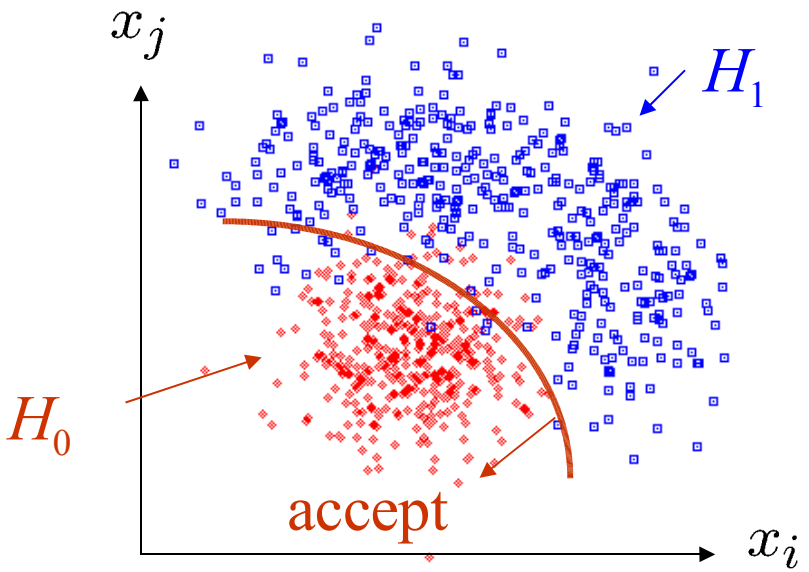
Multivariate statistical methods are a Big Industry:

Neural Networks,

Support Vector Machines,

Kernel density methods,

...




Particle Physics can benefit from progress in Machine Learning.

# Introduction to neural networks

Used in neurobiology, pattern recognition, financial forecasting, ...
Here, neural nets are just a type of test statistic.

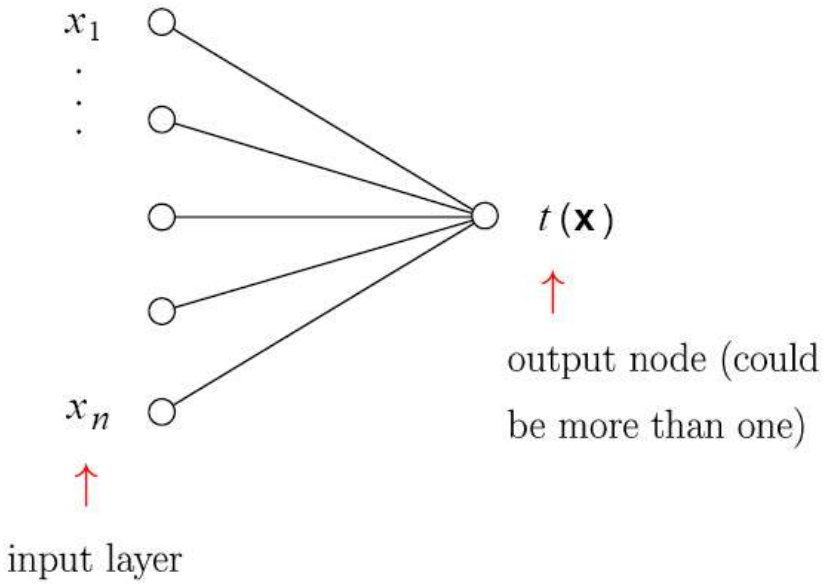
Suppose we take $t(\boldsymbol{x})$ to have the form

logistic sigmoid

$$t(\vec{x}) = s\left(a_0 + \sum_{i=1}^{n} a_i x_i\right) \quad , \quad \text{where } s(u) \equiv (1+e^{-u})^{-1} \, .$$

This is called the
single-layer perceptron.

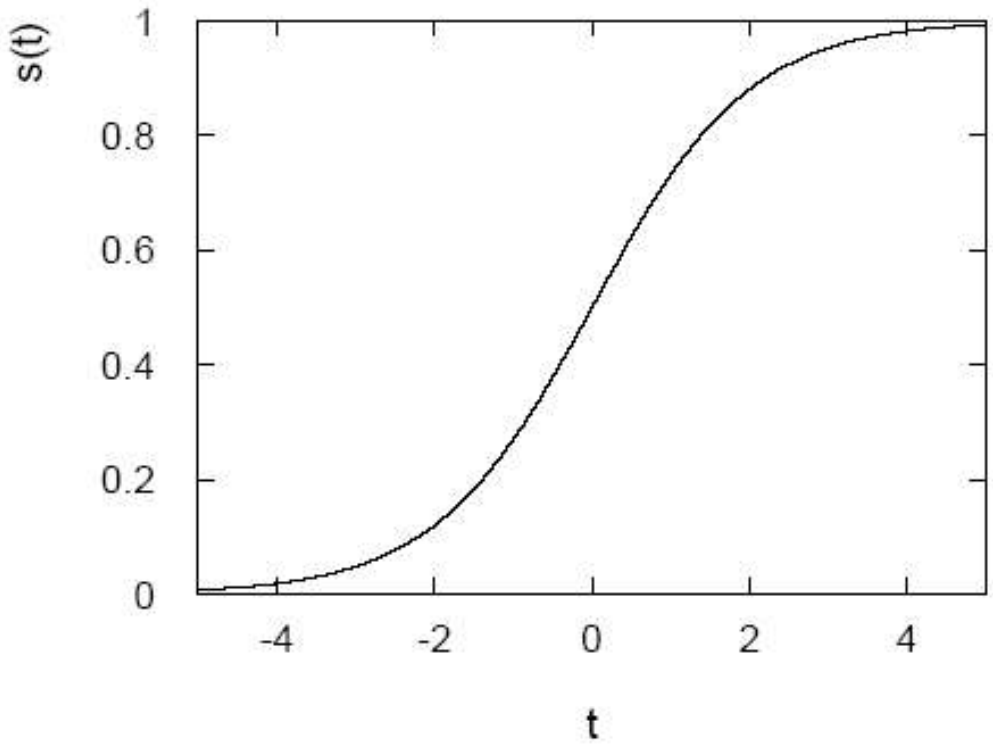
$s(\cdot)$ is monotonic
→ equivalent to linear $t(\boldsymbol{x})$



# The activation function

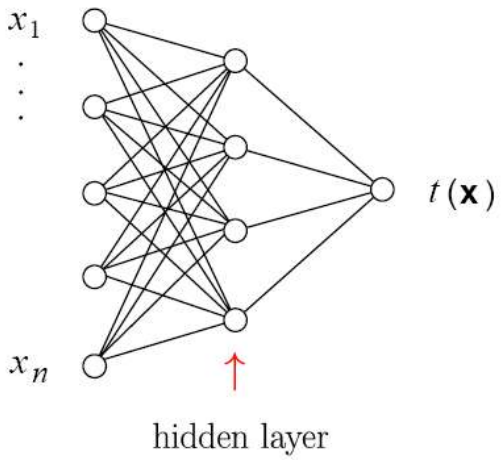
The activation function function $s(t)$ is often taken to be a logistic sigmoid:

$$s(t) = \frac{1}{1 + e^{-t}}$$



# The multi-layer perceptron

Generalize from one layer
to the multilayer perceptron:




The values of the nodes in the intermediate (hidden) layer are

$$h_i(\vec{x}) = s\left(w_{i0} + \sum_{j=1}^{n} w_{ij}x_j\right) ,$$

and the network output is given by

$$t(\vec{x}) = s\left(a_0 + \sum_{i=1}^{n} a_i h_i(\vec{x})\right) .$$

$a_i, w_{ij} =$ weights (connection strengths)

# Neural network discussion

Easy to generalize to arbitrary number of layers.

Feed-forward net: values of a node depend only on earlier layers, usually only on previous layer ("network architecture").

More nodes → neural net gets closer to optimal $t(\boldsymbol{x})$, but more parameters need to be determined.

Parameters usually determined by minimizing an error function,

$$\mathcal{E} = E_0[(t - t^{(0)})^2] + E_1[(t - t^{(1)})^2] \, ,$$

where $t^{(0)}$ , $t^{(1)}$ are target values, e.g., 0 and 1 for logistic sigmoid.

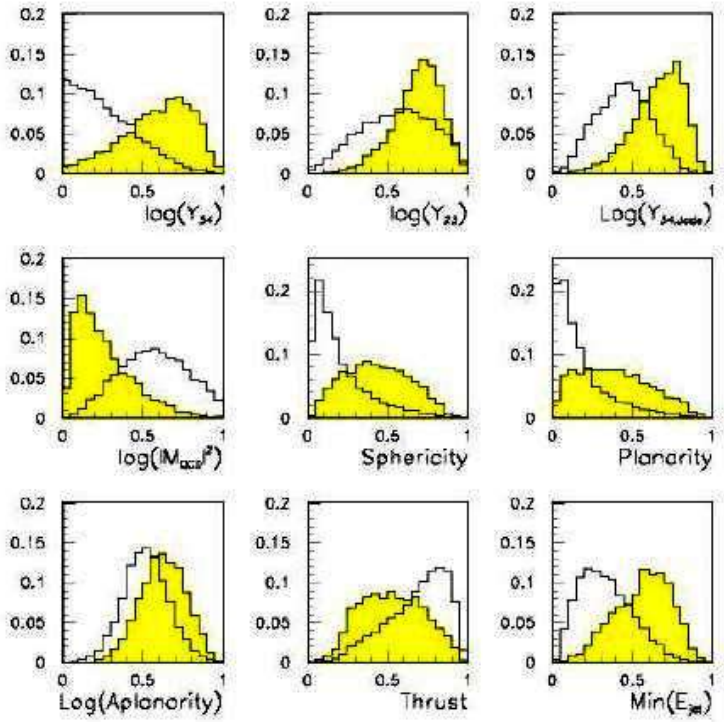
Expectation values replaced by averages of training data (e.g. MC).

In general training can be difficult; standard software available.

# Neural network example from LEP II

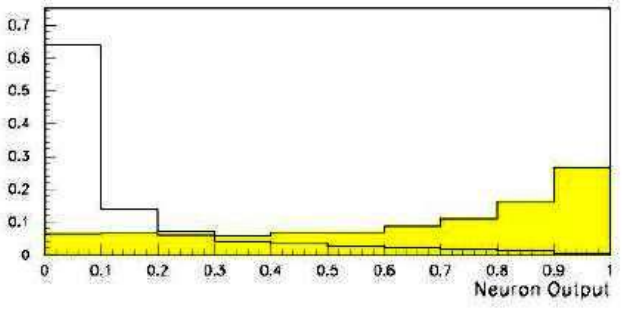
Signal: $e^+e^- \to W^+W^-$ (often 4 well separated hadron jets)

Background: $e^+e^- \to qqgg$ (4 less well separated hadron jets)




← input variables based on jet structure, event shape, ...
none by itself gives much separation.

Neural network output does better...




(Garrido, Juste and Martinez, ALEPH 96-144)

# Neural network discussion (2)

Why not use all of the available input variables?

Fewer inputs → fewer parameters to be adjusted,
→ parameters better determined for finite training data.

Some inputs may be highly correlated → drop all but one.

Some inputs may contain little or no discriminating power between the hypotheses → drop them.

NN exploits higher moments (nonlinear features) of joint pdf $f(\boldsymbol{x}|H)$, but these may not be well modeled in training data.

Better to have simper $t(x)$ where you can 'understand what it's doing'.

# Neural network discussion (3)

Recall that the purpose of the statistical test is usually to select objects for further study; e.g. select WW events, then measure their properties (e.g. particle multiplicity).

Need to avoid input variables that are correlated with the properties of the selected objects that you want to study. (Not always easy; correlations may be poorly known.)

Some NN references:


L. Lönnblad et al., *Comp. Phys. Comm.*, 70 (1992) 167;

C. Peterson et al., *Comp. Phys. Comm.*, 81 (1994) 185;

C.M. Bishop, *Neural Networks for Pattern Recognition*, OUP (1995);

John Hertz et al., *Introduction to the Theory of Neural Computation*, Addison-Wesley, New York (1991).

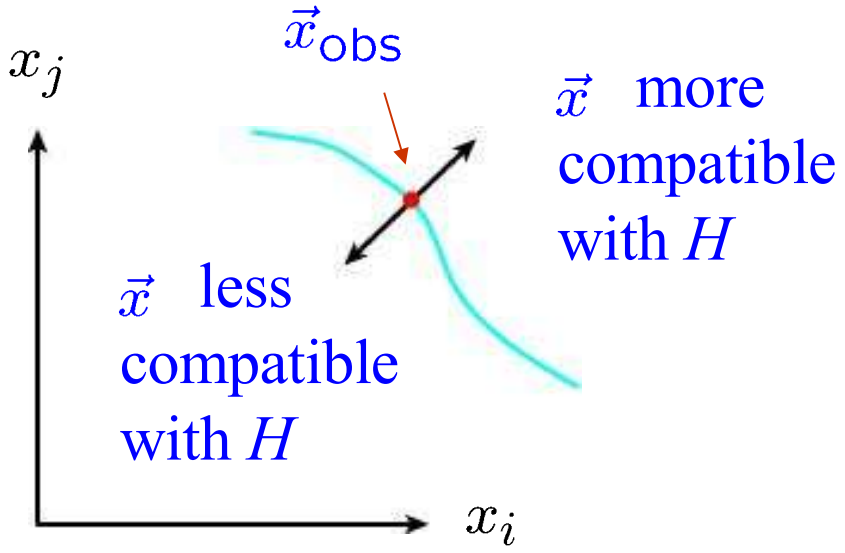
# Testing goodness-of-fit

Suppose hypothesis $H$ predicts pdf $f(\vec{x}|H)$ for a set of observations $\vec{x} = (x_1, \ldots, x_n)$ .

We observe a single point in this space: $\vec{x}_{\text{obs}}$

What can we say about the validity of $H$ in light of the data?

Decide what part of the data space represents less compatibility with $H$ than does the point $\vec{x}_{\text{obs}}$ . (Not unique!)



# *p*-values

Express 'goodness-of-fit' by giving the *p*-value for *H*:

*p* = probability, under assumption of *H*, to observe data with equal or lesser compatibility with *H* relative to the data we got.

This is not the probability that *H* is true!

In frequentist statistics we don't talk about *P*(*H*) (unless *H* represents a repeatable observation). In Bayesian statistics we do; use Bayes' theorem to obtain

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

where $\pi(H)$ is the prior probability for *H*.

For now stick with the frequentist approach;
result is *p*-value, regrettably easy to misinterpret as *P*(*H*).

# *p*-value example: testing whether a coin is 'fair'

Probability to observe $n$ heads in $N$ coin tosses is binomial:

$$P(n; p, N) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

Hypothesis $H$: the coin is fair ($p = 0.5$).

Suppose we toss the coin $N = 20$ times and get $n = 17$ heads.

Region of data space with equal or lesser compatibility with $H$ relative to $n = 17$ is: $n = 17, 18, 19, 20, 0, 1, 2, 3$. Adding up the probabilities for these values gives:

$$P(n = 0, 1, 2, 3, 17, 18, 19, \text{ or } 20) = 0.0026 .$$

i.e. $p = 0.0026$ is the probability of obtaining such a bizarre result (or more so) 'by chance', under the assumption of $H$.

# The significance of an observed signal

Suppose we observe $n$ events; these can consist of:

$n_b$ events from known processes (background)

$n_s$ events from a new process (signal)

If $n_s$, $n_b$ are Poisson r.v.s with means $s$, $b$, then $n = n_s + n_b$ is also Poisson, mean = $s + b$:

$$P(n; s, b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

Suppose $b = 0.5$, and we observe $n_{obs} = 5$. Should we claim evidence for a new discovery?

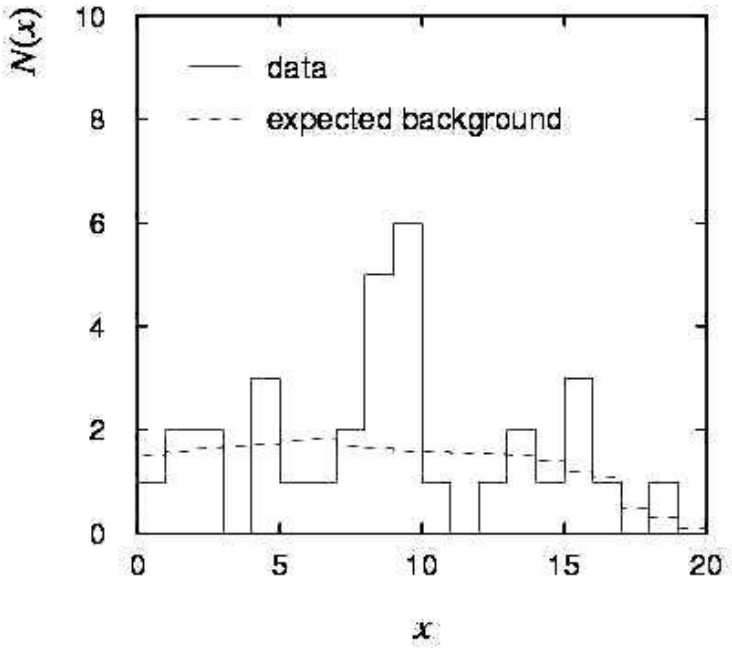
Give $p$-value for hypothesis $s = 0$:

$$p\text{-value} = P(n \geq 5; b = 0.5, s = 0)$$

$$= 1.7 \times 10^{-4} \neq P(s = 0)!$$

# The significance of a peak

Suppose we measure a value $x$ for each event and find:




Each bin (observed) is a Poisson r.v., means are given by dashed lines.

In the two bins with the peak, 11 entries found with $b = 3.2$.
The $p$-value for the $s = 0$ hypothesis is:

$$P(n \geq 11; b = 3.2, s = 0) = 5.0 \times 10^{-4}$$

# The significance of a peak (2)

But... did we know where to look for the peak?

→ give $P(n \geq 11)$ in any 2 adjacent bins

Is the observed width consistent with the expected $x$ resolution?

→ take $x$ window several times the expected resolution

How many bins × distributions have we looked at?

→ look at a thousand of them, you'll find a $10^{-3}$ effect

Did we adjust the cuts to 'enhance' the peak?

→ freeze cuts, repeat analysis with new data

How about the bins to the sides of the peak... (too low!)

Should we publish????

# Making a discovery

Often compute $p$-value of the 'background only' hypothesis $H_0$ using test variable related to a characteristic of the signal.

$p$-value = Probability to see data as incompatible with $H_0$, or more so, relative to the data observed.

Requires definition of 'incompatible with $H_0$'

HEP folklore: claim discovery if $p$-value equivalent to a $5\sigma$ fluctuation of Gaussian variable (one-sided) $\approx 2.85 \times 10^{-7}$

Actual $p$-value at which discovery becomes believable will depend on signal in question (subjective)

Why not do Bayesian analysis?

Usually don't know how to assign meaningful prior probabilities

# Pearson's $\chi^2$ statistic

Test statistic for comparing observed data $\vec{n} = (n_1, \ldots, n_N)$ ($n_i$ independent) to predicted mean values $\vec{\nu} = (\nu_1, \ldots, \nu_N)$ :

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\sigma_i^2} \,, \quad \text{where } \sigma_i^2 = V[n_i] \,.$$

(Pearson's $\chi^2$ statistic)

$\chi^2$ = sum of squares of the deviations of the $i$th measurement from the $i$th prediction, using $\sigma_i$ as the 'yardstick' for the comparison.

For $n_i \sim$ Poisson($\nu_i$) we have $V[n_i] = \nu_i$, so this becomes

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} \,.$$

# Pearson's $\chi^2$ test

If $n_i$ are Gaussian with mean $\nu_i$ and std. dev. $\sigma_i$, i.e., $n_i \sim \mathrm{N}(\nu_i\,,\,\sigma_i^2)$, then Pearson's $\chi^2$ will follow the $\chi^2$ pdf (here for $\chi^2 = z$):

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

If the $n_i$ are Poisson with $\nu_i >> 1$ (in practice OK for $\nu_i > 5$) then the Poisson dist. becomes Gaussian and therefore Pearson's $\chi^2$ statistic here as well follows the $\chi^2$ pdf.

The $\chi^2$ value obtained from the data then gives the $p$-value:

$$p = \int_{\chi^2}^{\infty} f_{\chi^2}(z;N)\,dz\ .$$

# The '$\chi^2$ per degree of freedom'

Recall that for the chi-square pdf for $N$ degrees of freedom,

$$E[z] = N\,, \quad V[z] = 2N\,.$$

This makes sense: if the hypothesized $\nu_i$ are right, the rms deviation of $n_i$ from $\nu_i$ is $\sigma_i$, so each term in the sum contributes ~ 1.

One often sees $\chi^2/N$ reported as a measure of goodness-of-fit.
But... better to give $\chi^2$ and $N$ separately. Consider, e.g.,

$$\chi^2 = 15,\ N = 10 \rightarrow p - \text{value} = 0.13\,,$$
$$\chi^2 = 150,\ N = 100 \rightarrow p - \text{value} = 9.0 \times 10^{-4}\,.$$

i.e. for $N$ large, even a $\chi^2$ per dof only a bit greater than one can imply a small $p$-value, i.e., poor goodness-of-fit.

# Pearson's $\chi^2$ with multinomial data

If $n_{\text{tot}} = \sum_{i=1}^{N}$ is fixed, then we might model $n_i \sim$ binomial

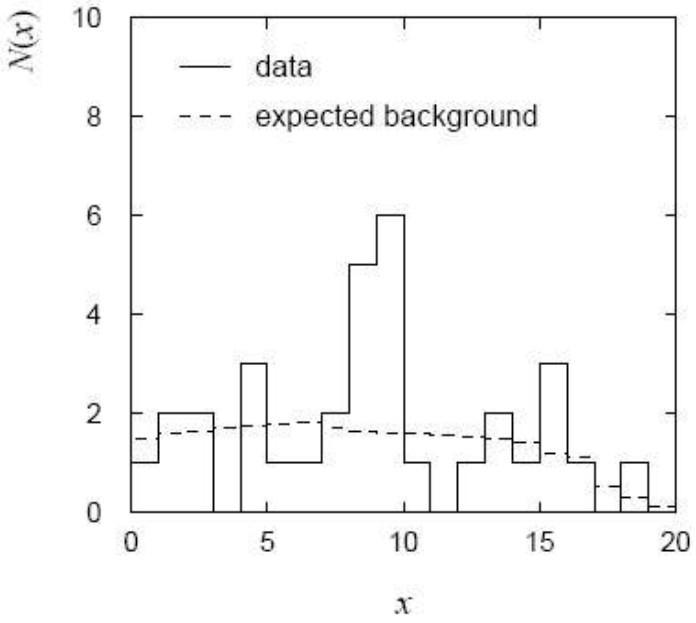
with $p_i = n_i / n_{\text{tot}}$. I.e. $\vec{n} = (n_1, \ldots, n_N)$ $\sim$ multinomial.

In this case we can take Pearson's $\chi^2$ statistic to be

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - p_i n_{\text{tot}})^2}{p_i n_{\text{tot}}}$$

If all $p_i\, n_{\text{tot}} >> 1$ then this will follow the chi-square pdf for $N$–1 degrees of freedom.

# Example of a $\chi^2$ test




← This gives

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} = 29.8$$

for $N = 20$ dof.

Now need to find $p$-value, but... many bins have few (or no) entries, so here we do not expect $\chi^2$ to follow the chi-square pdf.

# Using MC to find distribution of $\chi^2$ statistic

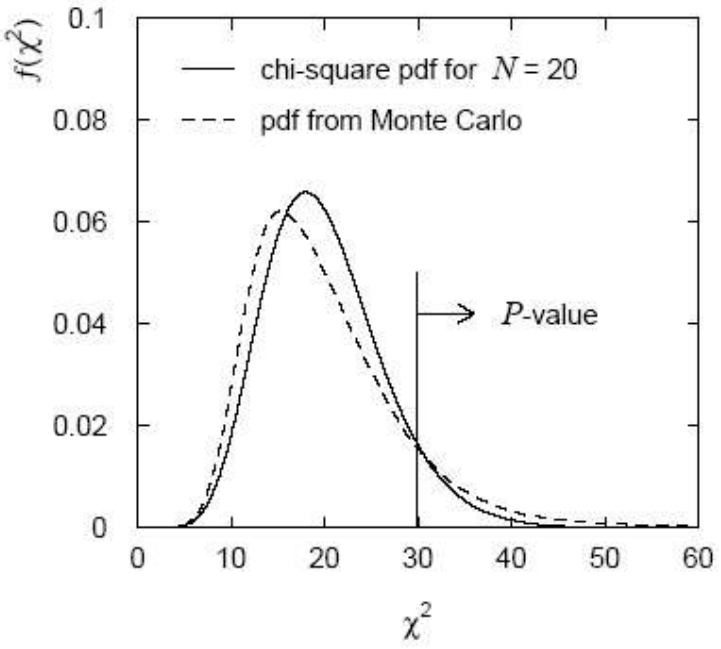
The Pearson $\chi^2$ statistic still reflects the level of agreement between data and prediction, i.e., it is still a 'valid' test statistic.

To find its sampling distribution, simulate the data with a Monte Carlo program: $n_i \sim \text{Poisson}(\nu_i)$ , $i = 1, N$ .

Here data sample simulated $10^6$ times. The fraction of times we find $\chi^2 > 29.8$ gives the $p$-value:

$$p = 0.11$$

If we had used the chi-square pdf we would find $p = 0.073$.



# Wrapping up lecture 2

Main ideas of statistical tests and related issues for HEP:

Discriminate between event types (hypotheses),
determine selection efficiency, sample purity, etc.

Some methods for constructing a test statistic

Linear: Fisher discriminant

Nonlinear: Neural networks

Goodness-of-fit tests

$p$-value (not same as $P(H_0)$!),

$\chi^2 = \Sigma$ (data – prediction)$^2$ / variance.

Often $\chi^2 \sim$ chi-square pdf $\rightarrow$ use to get $p$-value.

Next we turn to: parameter estimation