Introduction to Statistics – Day 3

Lecture 1
- Probability
- Random variables, probability densities, etc.

Lecture 2
- Brief catalogue of probability densities
- The Monte Carlo method.

→ Lecture 3
- Statistical tests
- Fisher discriminants, neural networks, etc
- Significance and goodness-of-fit tests

Lecture 4
- Parameter estimation
- Maximum likelihood and least squares
- Interval estimation (setting limits)

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

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

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.

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

Selecting events

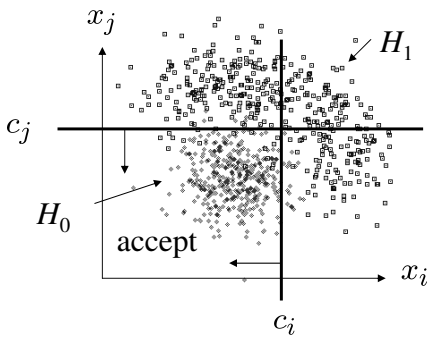
Suppose we have a data sample with two kinds of events, corresponding to hypotheses H_0 and H_1 and we want to select those of type H_0.

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':

$x_i < c_i$

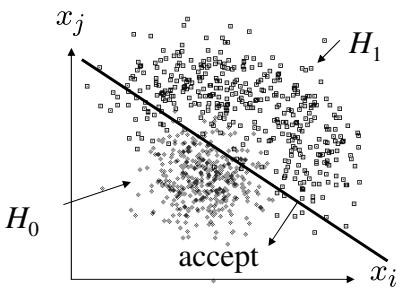
$x_j < c_j$



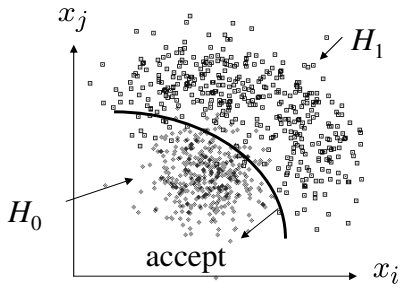
Other ways to select events

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)

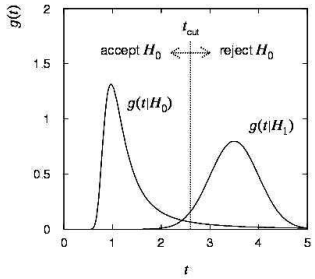
$$t(x_1, \ldots, x_n)$$

Try 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 accept or reject H_0.

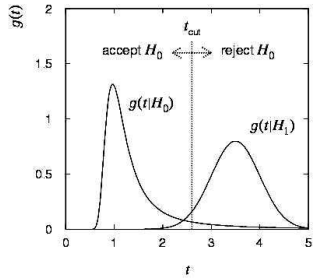


Significance level and power of a test

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

$$\alpha = \int_{t_{\mathsf{cut}}}^{\infty} g(t|H_0)\, dt$$

(significance level)



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

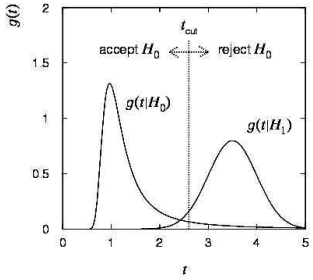
$$\beta = \int_{-\infty}^{t_{\mathsf{cut}}} g(t|H_1)\, dt$$

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

Efficiency of event selection

Probability to accept an event which is signal (signal efficiency):

$$\varepsilon_{\mathsf{s}} = \int_{-\infty}^{t_{\mathsf{cut}}} g(t|\mathsf{s})\, dt = 1 - \alpha$$



Probability to accept an event which is background (background efficiency):

$$\varepsilon_{\mathsf{b}} = \int_{-\infty}^{t_{\mathsf{cut}}} g(t|\mathsf{b})\, dt = \beta$$

Purity of event selection

Suppose only one background type b; overall fractions of signal and background events are π_{s} and π_{b} (prior probabilities).

Suppose we select events with $t < t_{\mathsf{cut}}$. 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:

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

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

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 ε_b for a given ε_s (highest power for a given significance level), choose acceptance region such that

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

where c is a constant which determines ε_s.

Equivalently, optimal scalar test statistic is $\boxed{t(\vec{x}) = \dfrac{f(\vec{x}|\mathsf{s})}{f(\vec{x}|\mathsf{b})}}$

Why Neyman-Pearson doesn't always help

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

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.

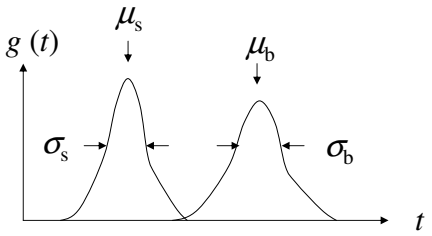
Compromise: make Ansatz for form of test statistic $t(\vec{x})$ with fewer parameters; determine them (e.g. using MC) to give best discrimination between signal and background.

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 $g(t|\mathsf{s})$, $g(t|\mathsf{b})$ have maximum 'separation'. We want:

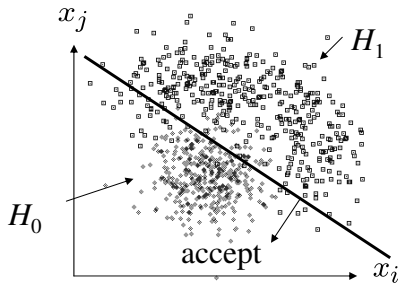
large distance between mean values, small widths



→ Fisher: maximize $J(\vec{a}) = \dfrac{(\mu_s - \mu_b)^2}{\sigma_s^2 + \sigma_b^2}$

Fisher discriminant

Using this definition of separation gives a Fisher discriminant.



Corresponds to a linear decision boundary.

Equivalent to Neyman-Pearson if the signal and background pdfs are multivariate Gaussian with equal covariances; otherwise not optimal, but still often a simple, practical solution.

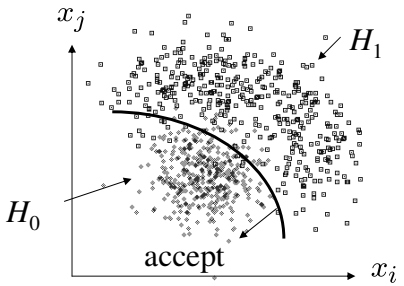
Nonlinear test statistics

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 estimation,
Boosted decision trees, ...



New software for HEP, e.g.,

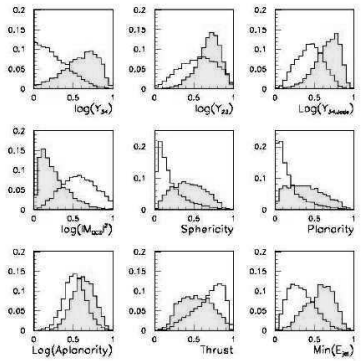
TMVA , Höcker, Stelzer, Tegenfeldt, Voss, Voss, physics/0703039

StatPatternRecognition, I. Narsky, physics/0507143

Neural network example from LEP II

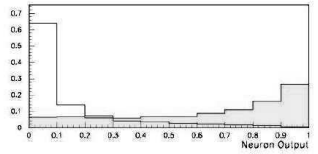
Signal: $e^+e^- \to W^+W^-$ (often 4 well separated hadron jets)

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)

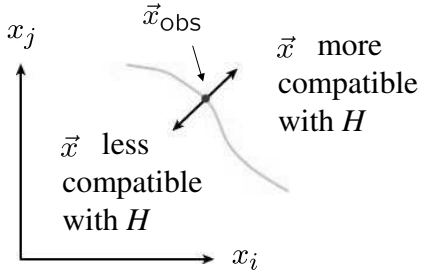
Testing significance/goodness-of-fit

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 = $s + b$:

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

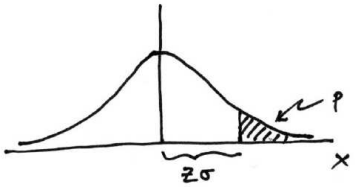
Suppose $b = 0.5$, and we observe $n_{obs} = 5$. Should we claim evidence for a new discovery?

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

$$\begin{aligned} p\text{-value} &= P(n \geq 5; b = 0.5, s = 0) \\ &= 1.7 \times 10^{-4} \neq P(s = 0)! \end{aligned}$$

Significance from *p*-value

Often define significance Z as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same p-value.



$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx = 1 - \Phi(Z)$$ **TMath::Prob**

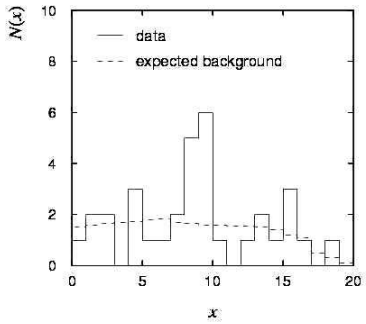
$$Z = \Phi^{-1}(1-p)$$ **TMath::NormQuantile**

E.g. $Z = 5$ (a '5 sigma effect') means $p = 2.87 \times 10^{-7}$

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????

When to publish

HEP folklore: claim discovery when p-value of background only hypothesis is 2.87×10^{-7}, corresponding to significance $Z = 5$.

This is very subjective and really should depend on the prior probability of the phenomenon in question, e.g.,

phenomenon	reasonable p-value for discovery
$D^0\bar{D}^0$ mixing	~0.05
Higgs	~ 10^{-7} (?)
Life on Mars	~10^{-10}
Astrology	~10^{-20}

Wrapping up lecture 3

We looked at statistical tests and related issues:
discriminate between event types (hypotheses),
determine selection efficiency, sample purity, etc.

Some modern (and less modern) methods were mentioned:
Fisher discriminants, neural networks,
support vector machines,...

We also talked about significance and goodness-of-fit tests:
p-value expresses level of agreement between data and hypothesis

Next we'll turn to the second main part of statistics:
parameter estimation

Extra slides

Probability Density Estimation (PDE) techniques

Construct non-parametric estimators of the pdfs $f(\vec{x}|H_0), f(\vec{x}|H_1)$:

and use these to construct the likelihood ratio

$$t(\vec{x}) = \frac{\hat{f}(\vec{x}|H_0)}{\hat{f}(\vec{x}|H_1)}$$

(*n*-dimensional histogram is a brute force example of this.)

More clever estimation techniques can get this to work for (somewhat) higher dimension.

See e.g. K. Cranmer, *Kernel Estimation in High Energy Physics*, CPC **136** (2001) 198; hep-ex/0011057;
T. Carli and B. Koblitz, *A multi-variate discrimination technique based on range-searching*,
NIM A **501** (2003) 576; hep-ex/0211019

Kernel-based PDE (KDE, Parzen window)

Consider *d* dimensions, *N* training events, $\boldsymbol{x}_1$, ..., $\boldsymbol{x}_N$, estimate $f(\boldsymbol{x})$ with

$$\hat{f}(\vec{x}) = \frac{1}{Nh^d} = \sum_{i=1}^{N} K\left(\frac{\vec{x} - \vec{x}_i}{h}\right)$$

kernel

bandwidth (smoothing parameter)

Use e.g. Gaussian kernel: $K(\vec{x}) = \frac{1}{(2\pi)^{d/2}} e^{-|\vec{x}|^2/2}$

Need to sum *N* terms to evaluate function (slow);
faster algorithms only count events in vicinity of $\boldsymbol{x}$
(*k*-nearest neighbor, range search).

Product of one-dimensional pdfs

First rotate to uncorrelated variables, i.e., find matrix *A* such that

for $\vec{x}' = A\vec{x}$ we have $\mathrm{cov}[x'_i, x'_j] = \delta_{ij}\sigma_i^2$.

Estimate the *d*-dimensional joint pdf as the product of 1-d pdfs,

$$\hat{f}(\vec{x}) \approx \prod_{i=1}^{d} \hat{f}_i(x_i) \qquad \text{(here } \boldsymbol{x} \text{ decorrelated)}$$

This does not exploit non-linear features of the joint pdf, but simple and may be a good approximation in practical examples.

Decision trees

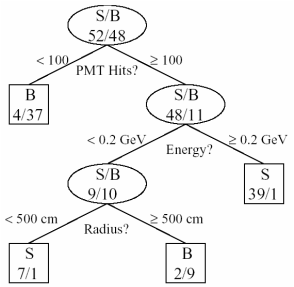
A training sample of signal and background data is repeatedly split by successive cuts on its input variables.

Order in which variables used based on best separation between signal and background.

Iterate until stop criterion reached, based e.g. on purity, minimum number of events in a node.

Resulting set of cuts is a 'decision tree'.

Tends to be sensitive to fluctuations in training sample.



Example by Mini-Boone, B. Roe et al., NIM A **543** (2005) 577

Boosted decision trees

Boosting combines a number classifiers into a stronger one; improves stability with respect to fluctuations in input data.

To use with decision trees, increase the weights of misclassified events and reconstruct the tree.

Iterate → forest of trees (perhaps > 1000). For the *m*th tree,

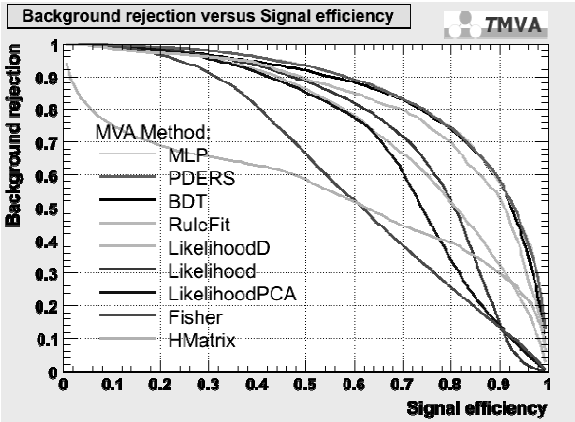
$$T_m(\vec{x}) = \begin{cases} 1 & \vec{x} \text{ in signal acceptance region} \\ -1 & \text{otherwise} \end{cases}$$

Define a score α_m based on error rate of *m*th tree.

Boosted tree = weighted sum of the trees: $T(\vec{x}) = \sum_m \alpha_m T_m(\vec{x})$

Algorithms: AdaBoost (Freund & Schapire), ε-boost (Friedman).

Comparing multivariate methods (TMVA)



Choose the best one!

Multivariate analysis discussion

For all methods, need to check:

Sensitivitiy to statistically unimportant variables (best to drop those that don't provide discrimination);

Level of smoothness in decision boundary (sensitivity to over-training)

Given the test variable, next step is e.g., select *n* events and estimate a cross section of signal: $\hat{\sigma}_s = (n - b)/\varepsilon_s L$

Now need to estimate systematic error...

If e.g. training (MC) data ≠ Nature, test variable is not optimal, but not necessarily biased.

But our estimates of background *b* and efficiencies would then be biased if based on MC. (True also for 'simple cuts'.)

Multivariate analysis discussion (2)

But in a cut-based analysis it may be easier to avoid regions where untested features of MC are strongly influencing the decision boundary.

Look at control samples to test joint distributions of inputs.

Try to estimate backgrounds directly from the data (sidebands).

The purpose of the statistical test is often to select objects for further study and then measure their properties.

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 multivariate analysis references

Hastie, Tibshirani, Friedman, *The Elements of Statistical Learning*, Springer (2001);

Webb, *Statistical Pattern Recognition*, Wiley (2002);

Kuncheva, *Combining Pattern Classifiers*, Wiley (2004);

Specifically on neural networks:

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).

General comments on Bayesian Higgs analysis

The main idea in a Bayesian analysis is to evaluate the probability of a hypothesis, where here the probability is interpreted as a (subjective) degree of belief:

$$P(\theta|x) = \frac{P(x|\theta)\pi(\theta)}{\int P(x|\theta)\pi(\theta)\,d\theta}$$

The probability of hypothesis H_0 relative to its complementary alternative H_1 is often given by the posterior odds:

$$\frac{P(H_0|x)}{P(H_1|x)}$$

no Higgs

Higgs

Bayes factors

The posterior odds is

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(x|H_0)}{P(x|H_1)} \times \frac{\pi(H_0)}{\pi(H_1)}$$

Bayes factor B_{01} prior odds

The Bayes factor is regarded as measuring the weight of evidence of the data in support of H_0 over H_1.

In its simplest form the Bayes factor is the likelihood ratio.

Interchangeably use $B_{10} = 1/B_{01}$

Bayes factors with undetermined parameters

If H_0, H_1 (no Higgs, Higgs) are composite, i.e., they contain one or more undetermined parameters λ, then

$$B_{10} = \frac{\int P(x|s+b,\lambda)\pi(\lambda)\,d\lambda}{\int P(x|b,\lambda)\pi(\lambda)\,d\lambda}$$

$\pi(\lambda)$ = prior, could be based on other measurement or could be "purely subjective", e.g., a theoretical uncertainty.

So the Bayes Factor is a ratio of "integrated likelihoods"

(the likeihood ratio uses maximized likelihoods).

Assessing Bayes factors

One can use the Bayes factor much like a p-value (or Z value).

There is an "established" scale, analogous to our 5σ rule:

B_{10}	Evidence against H_0
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Very strong

Kass and Rafferty, *Bayes Factors*, J. Am Stat. Assoc 90 (1995) 773.

Not clear how useful this scale is for HEP.