

Data Analysis Methods

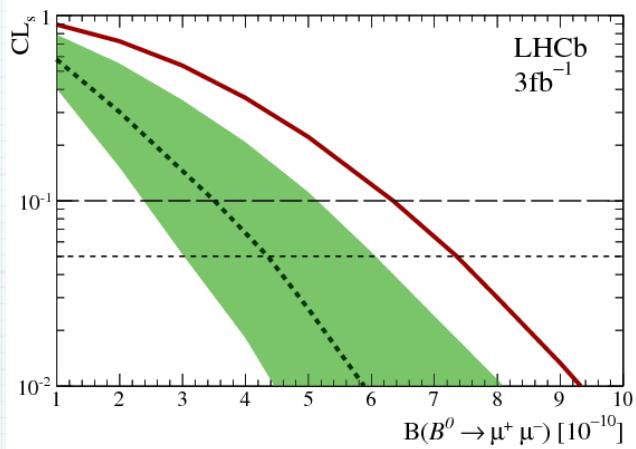
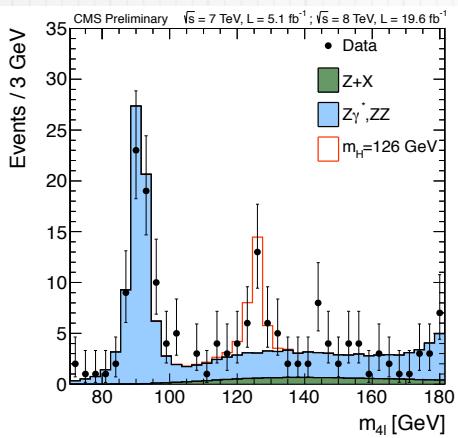


José A. Hernando
Departamento de Física de Partículas

20/01/2015

1

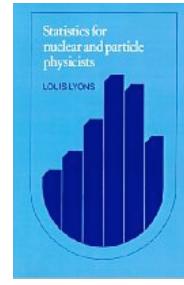
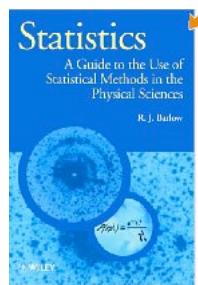
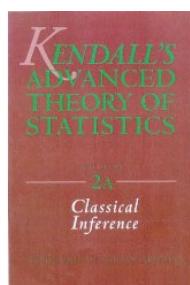
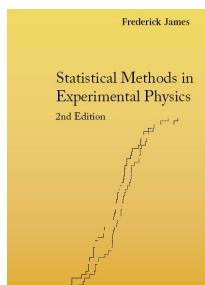
Objectives



- Is there a new signal?
 - (In this case the Higgs boson)
- What is the limit on the production of a given process?
 - In this case the Branching ratio of $B \rightarrow \mu^+ \mu^-$

Bibliography

- Excellent lectures given by experts and related with HEP, that I use extensively here:
 - Glen Cowan: Freiburg lectures (2011)
 - http://www.pp.rhul.ac.uk/~cowan/stat_freiburg.html
 - Kyle Cranmer: CERN training lectures (2009)
 - Mike Williams: IDPASC Santiago de Compostela lectures (2013)
- Excellent books:



3

Objectives

- Introduction to Statistical Data Analysis Methods in High Energy Physics
 - 1. Classification of events into known categories
 - 2. Hypothesis testing: Is this incompatible with the null hypothesis? Do we reject the alternative hypothesis?
 - 3. Setting Confidence Level intervals or Upper Limits to physics parameters
 - 4. Parameter estimation (regression) and treatment of errors (statistical and systematic)

4

Objectives

- The physicist goals, depending on the experiment, are:
 - choose between different hypothesis (is there a Higgs boson?)
 - provide an upper limit or a C.L. interval of an observable:
 $B(B_s \rightarrow \mu\mu)$
 - estimate the parameters of model (ϕ_s from $B_s \rightarrow J/\psi \phi$)
 - Estimate the statistical and systematic uncertainties
- To present or publish your result, you:
 - should indicate the methods used to obtain your measurements
 - use them correctly to not introduce bias or “incorrect” results
 - know what your measurement is and what is not!

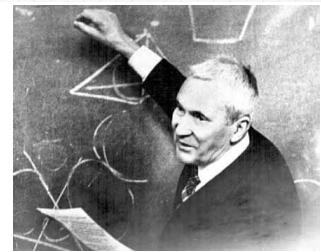
5

Outlook

- Introduction and reminder
 - 1. probability density functions (PDFs)
 - 2. Bayes’ theorem
- Classification
 - 3. Hypothesis testing
 - Neyman-Pearson Lemma
 - MVA-methods
 - 4. Goodness of the fit
 - 5. C.L intervals
- Regression
 - 6. Parameter estimation
 - 7. Uncertainties

6

Probability and Bayes



◆ Kolmogorov axioms:

1. probability of an event is non-negative $P(E) \geq 0$
2. probability for the entire space of events is 1
3. if elements are disjoint, probability is additive $P(E1) + P(E2)$

◆ Corollaries

1. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
2. $P(\neg A) = 1 - P(A)$

7

Bayes's theorem

- Marginal $P(A)$ and conditional probability $P(A | B)$
- Bayes' theorem

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$



e

Check bayes': two dices, probability they add to six if one is four, and that one is four if they add six.

$$\frac{1}{3} \frac{1}{6} = \frac{2}{5} \frac{5}{36} = \frac{1}{18}$$

8

Bayesians vs Frequentist

- The impossible divide:

- **Frequentists:** they are objective, they obtain **P(data | theory)**, they are deductive, they assume an experiment can be repeated n-large times in the same conditions
- **Bayesians:** they are “subjective” (with some reasonable assumptions), obtain **P(theory | data)**, they are inductive, they assume some prior knowledge and use Bayes’ theorem

e

Does the Higgs exists? What a frequentist will say about the Higgs?

e

Does this Higgs exists? I expect 2 background events, if there will be a Higgs I expected 8 events, and I observe 6 events. (Consider now that instead of the Higgs is a very rare, exotic, bizarre Z' that a crazy theoretician has proposed...)

$$P(H|n) = \frac{P(n|H)P(H)}{P(n|H)P(H) + P(n|!H)P(!H)}$$

9

Bayesians

e

An underground neutrino detector, it fails 10% of the times, and the reactor stops 30% of the time, in this moment it does not detect, what is the probability that the reactor is not running?

e

An underground solar neutrino detector, it fails 10% of the times. It does not detect now, what is the probability that the Sun is off?



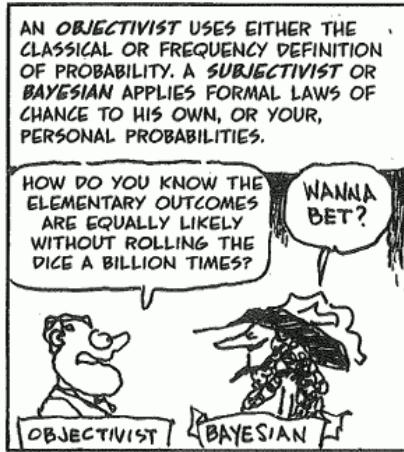
$$P(!R|!D) = \frac{P(!D|R)P(!R)}{P(!D|R)P(!R) + P(!D|R)P(R)}$$

as $P(!D|R) = 1$, we have:

$$P(!R|!D) = \frac{0.3}{0.3 + 0.7 \times 0.1} = 0.81$$

Frequentist vs Bayesians

- The impossible divide (*L. Lyons* quote?):
 - **Frequentists:** use impecable logic to answers questions that nobody cares about!
 - **Bayesians** address the questions everyone is interested on using assumptions that nobody believes



- Finally it is not so dramatic: enough data speaks by itself!

11

Bayesian posterior probability

- Bayesians obtain a *posterior* probability from data and a *prior* probability

$$P(h|x) = \frac{P(x|h)\pi(h)}{\int P(x|h)\pi(h)dh}$$

e

Consider dices of 4, 6, 12, 24 sides. Take one of them, rolling it 6 times we get: 3,5,6,1,2,2, what is the probability that has 4, 6 12 or 24 sides? What is the dependence with your priors?

1. hypothesis: $n = 4, 6, 12$ and 24 sides
2. probability $P(i|n) = 1/n$ (if $i < n$), 0 otherwise
3. start with prior probabilities, suggestion: $\pi(n) = 1/4$
4. change the priors: $\pi(4) = 1/2$, $\pi(6) = 1/8$, $\pi(12) = 1/8$, $\pi(24) = 1/4$

12

PDFs

- Probability (mass) density functions and random variables (rvs)

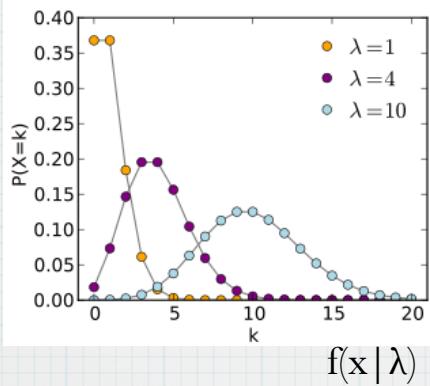
$$\int_{-\infty}^{\infty} f(x) dx = 1$$

- Cumulative (mass) density functions

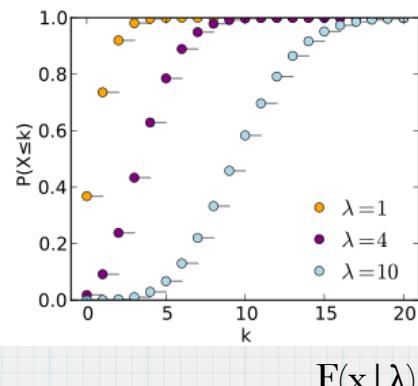
$$F(x) = \int_{-\infty}^x f(x) dx$$

e

Generate n-large number of poisson rvs with $\lambda=10$ and get its cumulative distribution



$f(x | \lambda)$



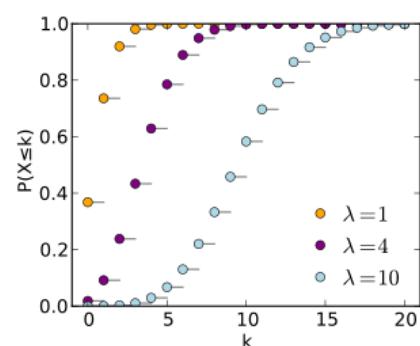
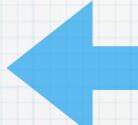
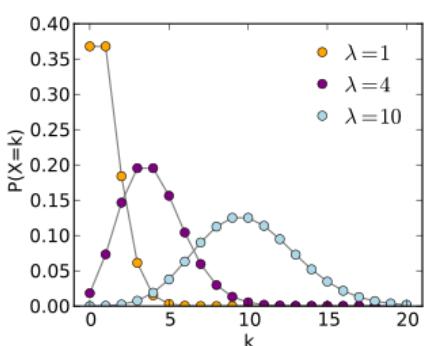
$F(x | \lambda)$

13

PDFs

- cumulative distributions of a pdf are flat!

- Use it to generate random numbers
- u random uniform [0,1], $x = \text{cdf}^{-1}(u)$



14

PDFs

- Expectation values:

$$E[g(x)] \equiv \int g(x) f(x) dx$$

The *mean* (or average value), is the expected value of x :

$$E[x] = \mu \equiv \int x f(x) dx$$

The *variance*, the expected value of $(x - \mu)^2$:

$$V[x] = \sigma^2 = E[(x - \mu)^2] = E[x^2] - \mu^2 \equiv \int (x - \mu)^2 f(x) dx$$

We call *standard deviation* to:

$$\sigma = \sqrt{V[x]}$$

Finally, the *expected value* is the x value with the higher $f(x)$, and the *median*, is the x value that divides the PDF distribution in half, $\int_0^x f(x) dx = 0.5$. For the symmetric PDFs, the mean and median are the same, this is the case of the Gaussian PDF.



Compute the mean and standard deviation of a uniform distribution between 0. and d . What is the “resolution” of a digital detector with pitch = d ?

15

PDFs

- The common HEP PDFs

name	PDF	example
Binomial	$f(n N, p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$	Branching ratio
Multinomial	$f(\mathbf{n} N, \mathbf{p}) = \frac{N!}{\prod_{i=1}^m n_i!} \prod_{i=1}^m p_i^{n_i}$	Histogram with N fixed entries
Poisson	$f(n \nu) = \frac{\nu^n}{n!} e^{-\nu}$	Number of events observed
Uniform	$f(x a, b) = \frac{1}{(b-a)}$	Monte Carlo method
Exponential	$f(x \tau) = \frac{1}{\tau} e^{-x/\tau}$	Decay time
Gaussian	$f(x \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	Measurement error
χ_n^2	$f(x n) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}$	Goodness of fit
Breit-Wigner	$f(x \Gamma, x_0) = \frac{1}{\pi} \frac{\Gamma/2}{\Gamma^2/4 + (x-x_0)^2}$	Mass of a resonance
Landau	$f(\Delta E \beta)$	Ionization energy loss

- Their mean and variances

name	average	variance
Binomial	Np	$Np(1-p)$
Multinomial	Np_i	$Np_i(1-p_i)$
Poisson	ν	ν
Uniform	$\frac{b-a}{2}$	$\frac{(b-a)^2}{12}$
Exponential	τ	τ^2
Gaussian	μ	σ^2
χ_n^2	n	$2n$
Breit-Wigner		∞
Landau	∞	∞

16

PDFs

- **Central Limit Theorem:** adding rvs with given pdfs, the addition follows a gaussian pdf

e

Generate n=100 rvs following a uniform distribution in [0.,1.] and take the addition, repeat m-large times, and get the distribution.

e

Generate n-large events with a binomial pmf with p=0.01 and N=100, what is the distribution of the yes events?

e

Generate n-large events with a poison pmf with $\lambda=10$

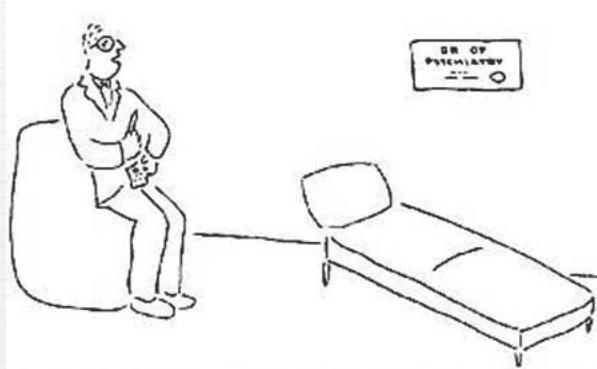
e

Generate n-large (100) events with a gaussian pdf (μ, σ) , compute for each one: $(x-\mu)^2/\sigma^2$, what is the distribution that you get?

17

Hypothesis testing

- **Hypothesis testing:** Comparing a well known hypothesis (null) **H₀** with respect an hypothetical hypothesis (**H₁**), with one observation, **when to exclude the H₁ and to reject H₀?**

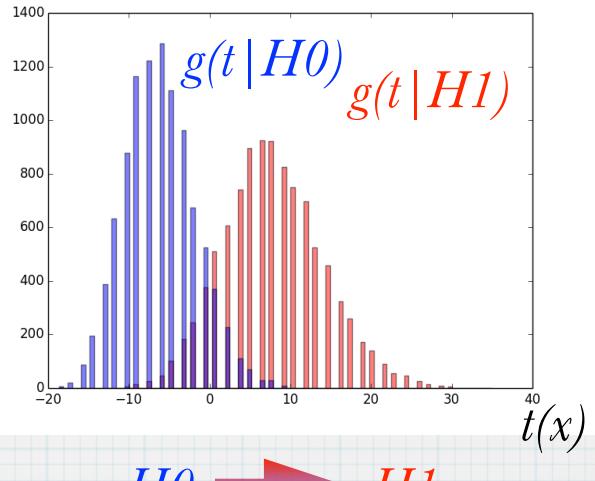


18

Hypothesis testing

- hypothesis pdfs: $f(x|H_0), f(x|H_1)$,
- test-statistics $t(x)$ escalar with ordering of events (i.e $H_0 < H_1$)
 - $g(t|H_0)$ and $g(t|H_1)$, the pdfs of t for H_0 and H_1
 - Likelihood Ratio, $\Lambda(x)$, is the “optimal” $t(x)$

$$\Lambda = \frac{\mathcal{L}(H_1)}{\mathcal{L}(H_0)} = \frac{f(x|H_1)}{f(x|H_0)}$$



19

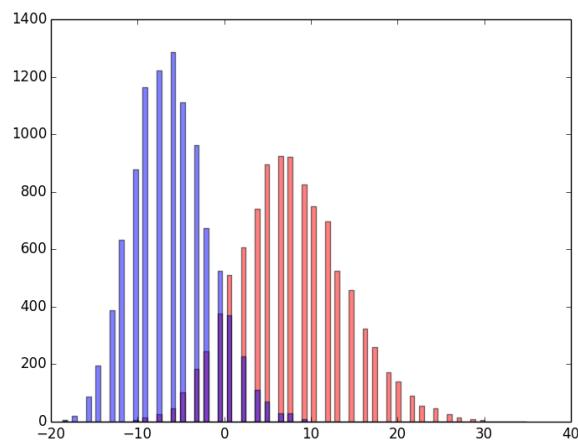
Hypothesis testing

e

Consider H_0 , a poisson distribution $\nu = 10$, and H_1 , a poisson $\nu = 20$, construct the test-statistics as the likelihood ratio, Λ

e

Consider n-bins, each one independent, H_0 , follows a poisson with $\nu = [2,1,0.5]$, and H_1 , a poisson $\nu = [1,3,1]$, construct a test-statistics from the likelihood ratio, Λ (suggestion use $-2\log \Lambda$)



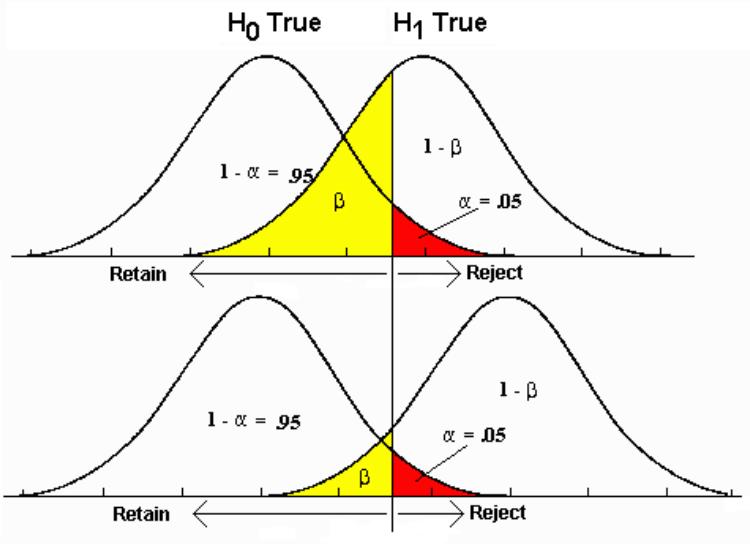
20

Hypothesis testing

- Data \mathbf{x} , observation $t_0(\mathbf{x})$
- Errors:
 - Type I, α , false positives
 - Type II, β , false negatives
- Power $1-\beta$

$$\alpha = \int_{t_c}^{\infty} t(x|H_0) dx,$$

$$\beta = \int_{-\infty}^{t_c} t(x|H_1) dx.$$



- p -value: fraction of events of H equal or less similar to H that the observation $t_0(\mathbf{x})$
 - α , p -value of H_0 , β , p -value of H_1

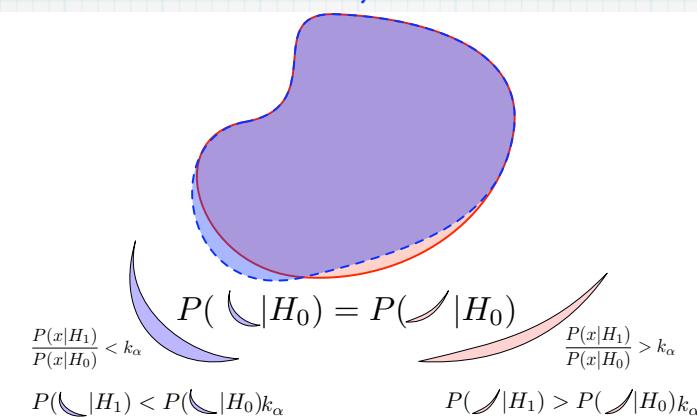
21

Hypothesis testing

- Neyman-Pearson Lemma:
 - For a fix α , called size of the test, the likelihood ratio, Λ , is the test-statistics that maximizes the power, $1-\beta$, that is it is optimal, fulfills:

$$\Lambda = \frac{\mathcal{L}(H_1)}{\mathcal{L}(H_0)} = \frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)} > k_\alpha$$

- The power is the efficiency on H_1 of the cut k



22

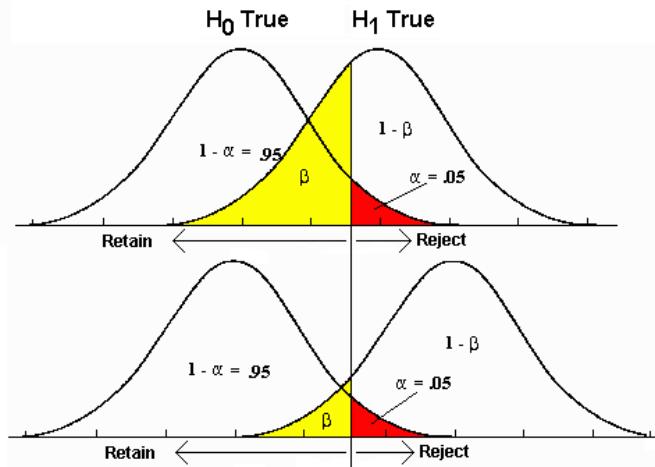
Hypothesis testing

- Tradition in HEP:

- discovery** of H_1 , if α , *p-value* of H_0 , is smaller 2.87×10^{-7} , “5- σ ”
- exclusion** of H_1 at 90 (95)% Confidence Leve (CL), if β , *p-value* of H_1 , is smaller than 0.1 (0.05)
- sensitivity** of the experiment: the ROC ($1-\beta$ vs α) curve

$$\alpha < 2.87 \times 10^{-7},$$

$$\beta < 0.1 \text{ (0.05)}$$

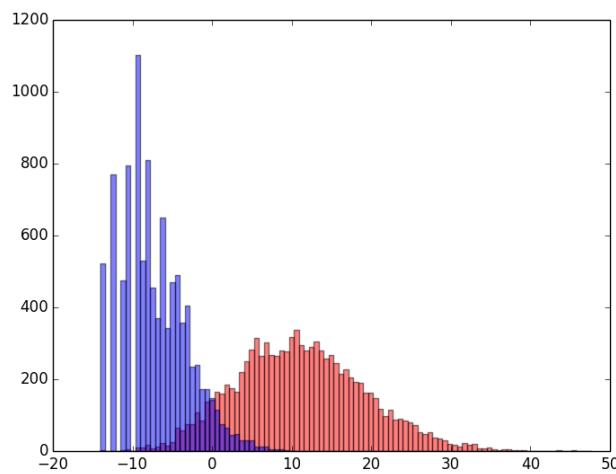


23

Hypothesis testing

e

Consider n-bins, each one independent, H_0 , follows a poisson with $v = [1.5, 1, 0.5]$, and H_1 , a poisson $v = [3.5, 4, 2.5]$, construct a test-statistics from the likelihood ratio (suggestion $-2 \log A$), we observe $[2, 1, 1]$, what is the *p-value* for H_1 ? at what confidence level H_1 is excluded? what is the *p-value* of H_0 ?



e

Compute the ROC curve ($1-\beta$ vs α)

24

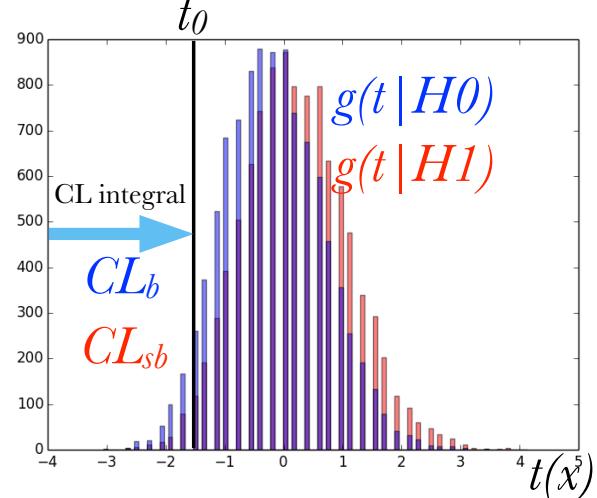
Hypothesis testing - CLs

- CLs method

- Can we exclude when there is no sensitivity?
- Use $CL_s = CL_{sb}/CL_b$ not proper coverage, but commonly used)

$$CL_s = \frac{CL_{sb}}{CL_b} = \frac{p_{sb}}{1-p_b} = \frac{\beta}{1-\alpha}.$$

notice $CL_b (1-\alpha) > CL_{sb} (\beta)$



e

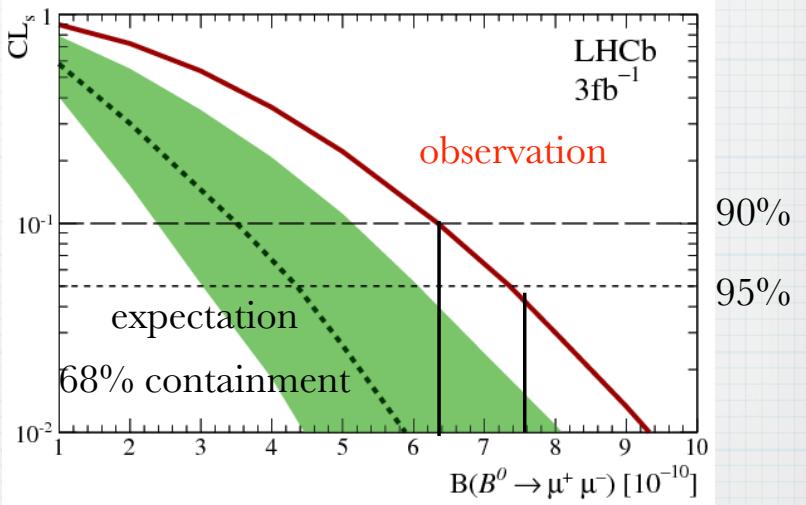
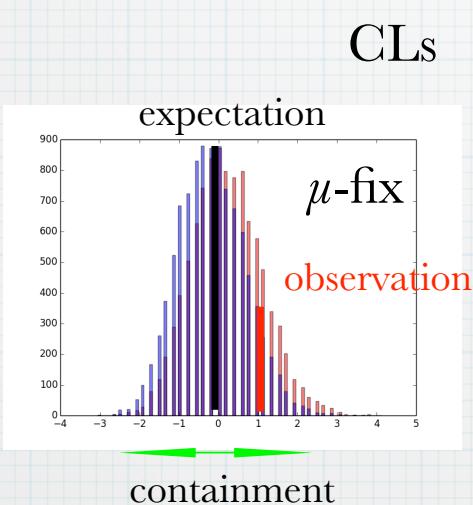
Consider n-bins, each one independent, H_0 , follows a poisson with $\nu = [10,5]$, and H_1 , a poisson $\nu = [12,6]$, the observation is $[6,1]$, what is the *p-value* for H_1 ? what is the $1-\alpha$? What is the CLs?

25

Hypothesis testing - CLs

- Composite hypothesis

- H_1 can depends on certain parameters:
 - μ strength of the signal $H_1 = B + \mu S$, $H_0 = B$, B is background, S is signal
 - scan on μ !
- The $B \rightarrow \mu^+ \mu^-$ case (we did not know the Br!)

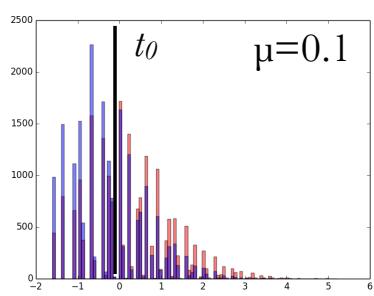
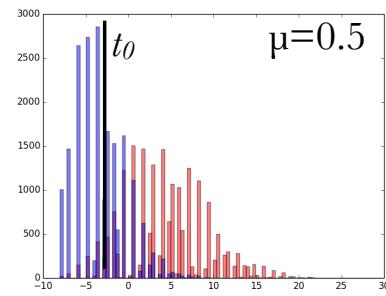
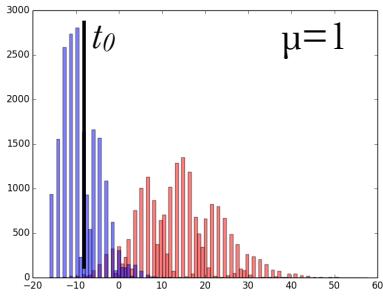


26

$\mu = \text{Br}/\text{Br}^{\text{SM}}$

Hypothesis testing - CLs

- Scan on μ
 - From the $g(t|B)$, pdf of t for $H0=B$, get the median (expectation) and a containment (i.e. 68%) t -region
 - Get the $t_0(\mu)$ and compute $CLs(\mu)$
- Exclude μ values at given confidence level (CL) [see confidence intervals]



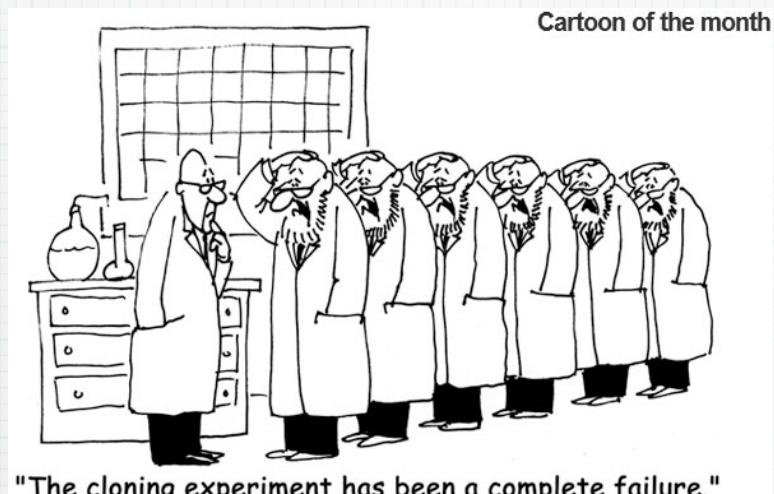
e

Consider n-bins, each one independent, $H0=B$, follows a poisson with $B = [1.5, 1, 0.5]$, and $H1=B+\mu S$, a poisson $S = [2, 4, 2]$, the observation is $[1, 1, 1]$, compute the expectation vs μ ? What is excluded μ region at 90 % CL with CLs?

27

Goodness of the fit

- How good does the hypothesis fits the data?
- When two data samples are identical?



Goodness of the fit

e

Flipping n=20 coins, m=10 times, we got 12,10,9,6,11,10,9,10,12,11 heads, is this fair? Flipping a coin m=10 times, we got 9 heads, is this fair?

e

Do a text to a physics student class in Quantum Physics, now repeat the text to a group of Spanish literature students, how you compare both samples?

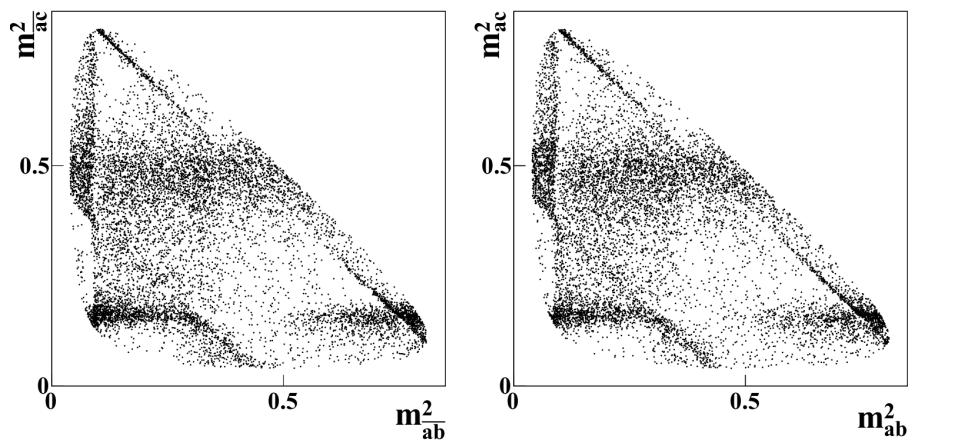
- Goodness of the fit test

- Several tests: Chi2, Kolmogorov, Energy, k-nearest neighbor, etc...
- use the p-value
 - Define a test-statistics $t(x)$, get the pdf, $g(t)$, and the cumulative cdf, $G(t)$
 - a suggestion: $t(x) = -\log(f(x))$
 - Compute the p-value of t_0 , if $p < 0.05$ reject agreement at 95 %CL

29

Goodness of the fit

- Are these samples equal?
 - If not, there is CP-violation!



B-bar

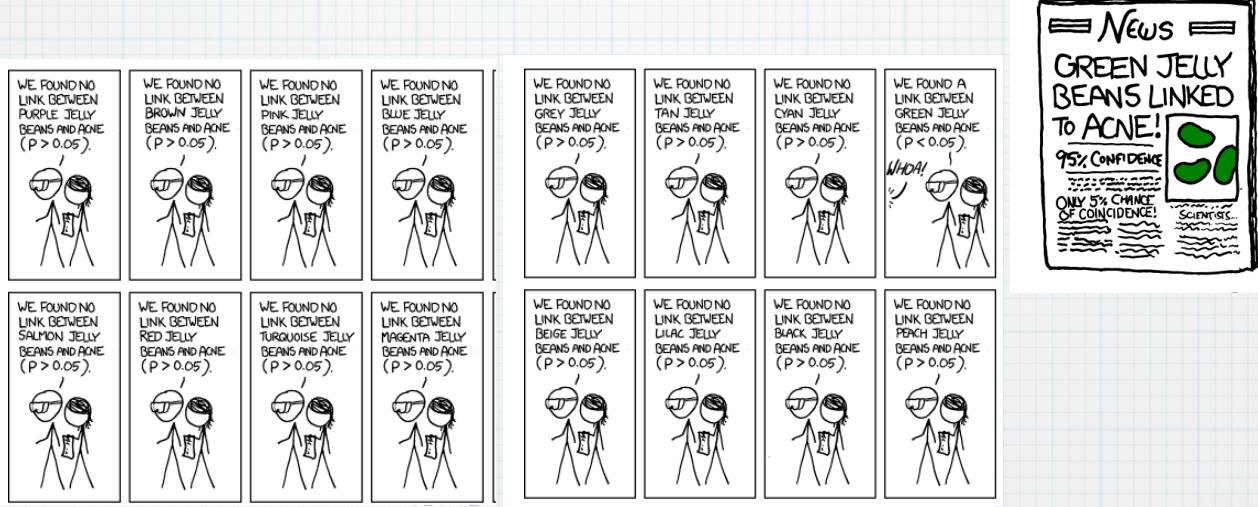
B

30

Goodness of the fit - p -value

- Look Elsewhere effect

- Looking for a rare (small p -value) event in a m-large number of places could result in a discovery, how to “normalize” the search?
- move from local to global p -value



31

Confidence Intervals

- Confidence Intervals:

- Looking for a phenomena, if not found, put limits at a given confidence level (CL) on the strength parameter μ
- Or to set an interval of a CL for a parameter μ



32

Confidence Intervals

- What does it means a CL interval?
 - **Frequentist:** the interval *covers* the true parameter μ a CL (i.e 95%) of the cases.
 - If you will perform m -large number of experiments and get m -intervals, in a CL (i.e 95%) fraction of the experiments, the interval covers the true μ .
 - **Bayesians:** talk about *credible* intervals, the probability that the true μ is inside the interval is CL (i.e. 95%)
 - $P(\mu \in \text{Interval}) = \text{CL}$
- Construction of frequentist intervals:
 - Classical or **Feldman-Cousins** construction

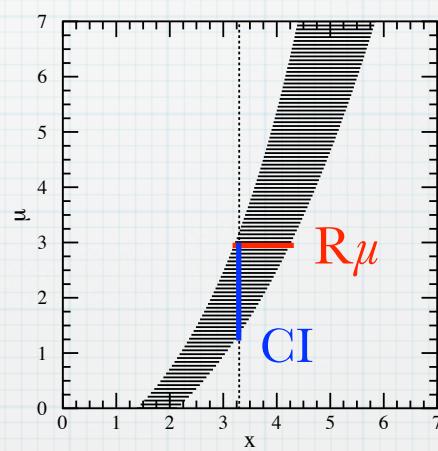
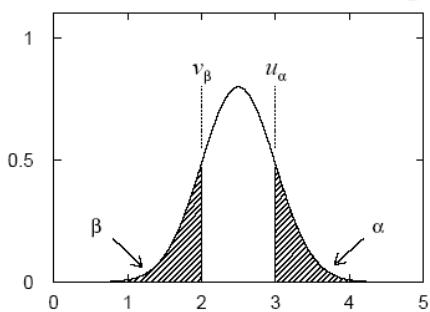
33

Confidence Intervals

- Classical - Neyman's construction
 - For a given value of μ , find the region $R\mu$ of x at a given CL
 - ambiguity: select central, upper, lower interval
 - For a given x_0 measurement, find the interval of μ , where x in in the $R\mu$

$$\alpha = \int_{x_l}^{\infty} f(x) dx, \quad \beta = \int_{-\infty}^{x_u} f(x) dx;$$

upper	$\alpha = 1 - CL$	$\beta = 0$
lower	$\alpha = 0$	$\beta = 1 - CL$
central	$\alpha = (1 - CL)/2$	$\beta = (1 - CL)/2$



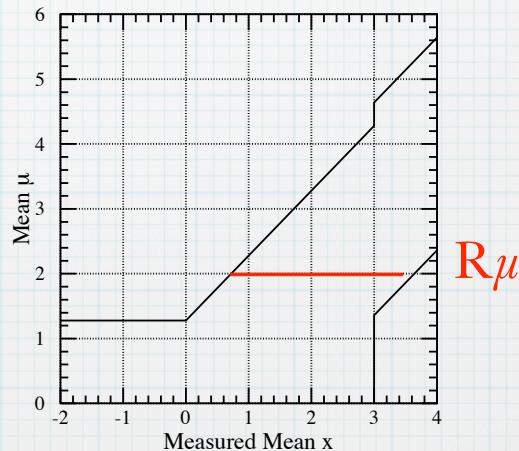
34

Confidence Intervals

- Flip-flopping problem of naive physicist before searching for a new phenomena:
 - “If I get a negative result, I publish as if I got nothing; if I do not “observe” ($<3\sigma$), I publish an CL upper limit; If I “observe” something ($>3\sigma$) I publish a central CL interval!
 - The publication depend on the result! worrisome... it does not have the correct coverage either...

e

Is the coverage R_μ correct for $\mu=2$?



35

Confidence Intervals

- Alternative test-statistic ML- μ (à la Feldman-Cousins):
 - Use likelihood ratio with respect the **maximum likelihood** ML (**physical**) μ estimate (hat μ)

$$q(\mu) = \begin{cases} -2 \log \frac{\mathcal{L}(\mu)}{\mathcal{L}(\hat{\mu})} & \text{if } \hat{\mu} > 0 \\ -2 \log \frac{\mathcal{L}(\mu)}{\mathcal{L}(\mu=0)} & \text{if } \hat{\mu} \leq 0 \end{cases}$$

q-ordering is:
 $q(\mu) < q(\text{ML-}\mu)$

- Remember that the likelihood is:

$$\mathcal{L}(\mu) = \mathcal{L}(B + \mu S) = f(x|B + \mu S), \quad \mathcal{L}(\mu = 0) = \mathcal{L}(B) = f(x|B)$$

- ML- μ must be physical, $S < 0$ is not allowed!
 - Will you publish a negative signal search after building your experiment?

36

Confidence Intervals

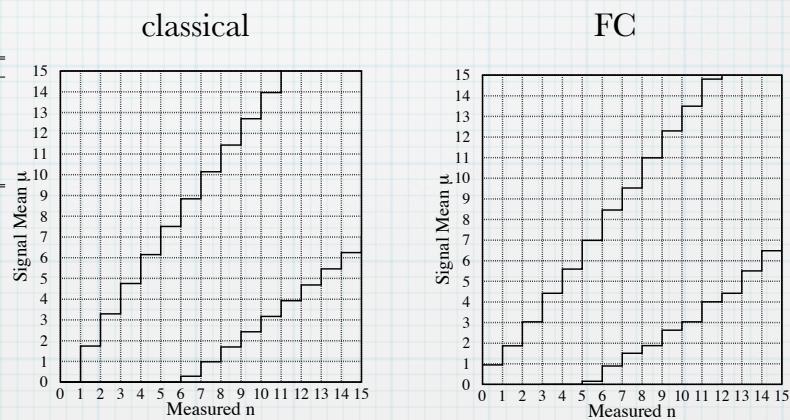
- Feldman-Cousins construction:

- Use $q(\mu)$ ordering to define $R\mu$!
- unambiguous! not needed to decide upper, lower, central, no flip-flopping dilemma! and correct (or conservative) coverage

e In a single experiment, you expect $b=3$ background events, and you look for an unknown signal, s , you observe n , what is the classical and FC confidence intervals? What happens if you measure $n=0$? compare the classical vs FC methods in that case.

n	0	1	2	3	4	5	6	7	8
$f(n s = 0.5)$	0.03	0.106	0.185	0.216	0.189	0.132	0.077	0.038	0.017
classical order	5	3	1	2	4	6	7		
\hat{s}	0	0	0	0	1	2	3	4	5
$f(\hat{s} n)$	0.050	0.149	0.224	0.224	0.195	0.175	0.161	0.149	140
$\frac{f(n s=0.5)}{f(n \hat{s})}$	0.607	0.708	0.826	0.963	0.966	0.753	0.480	0.259	
FC order	6	5	3	2	1	4	7		

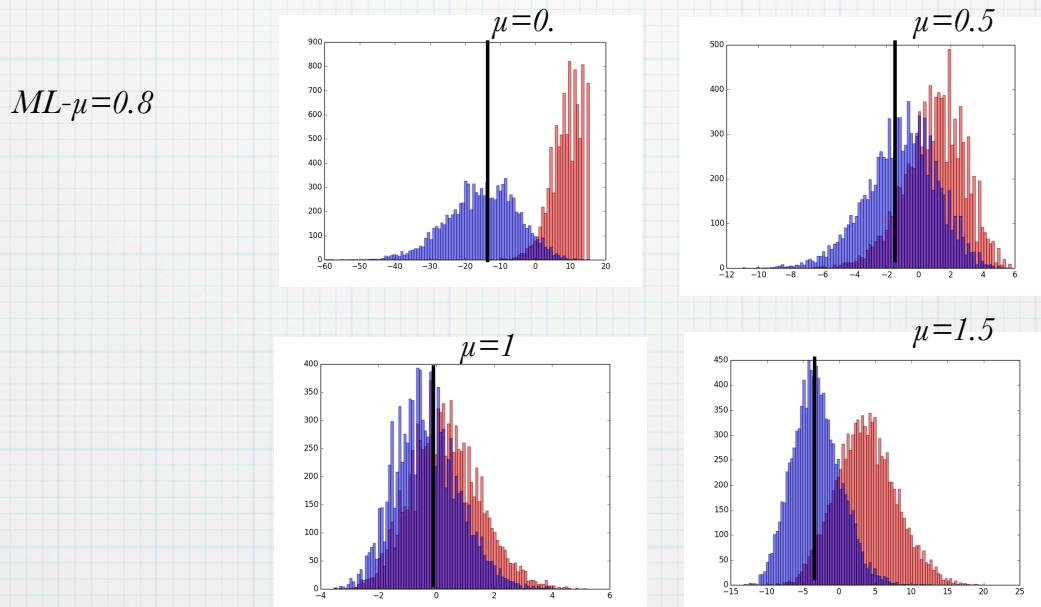
$s=0.5$



37

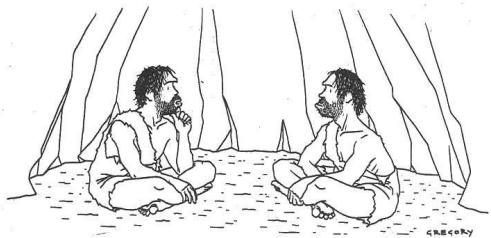
Confidence Intervals

e Consider a n-bins independent experiment, with background expectation $b=[1.5,1,0.5]$ and signal $s=[2,6,2]$ for strength $\mu=1$, the observation is $n=[3,6,2]$, what is the ML- μ ? Obtain the FC interval at 90 % CL for μ . Plot the CLs value vs μ .



38

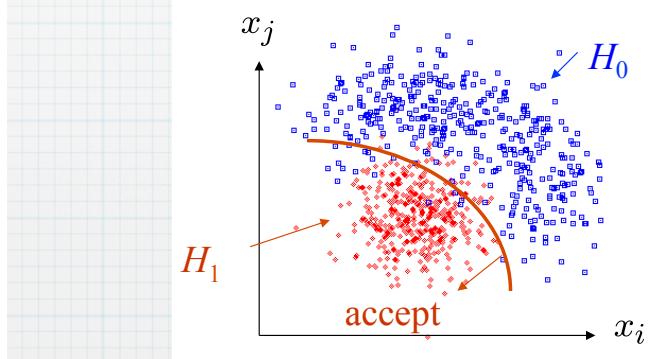
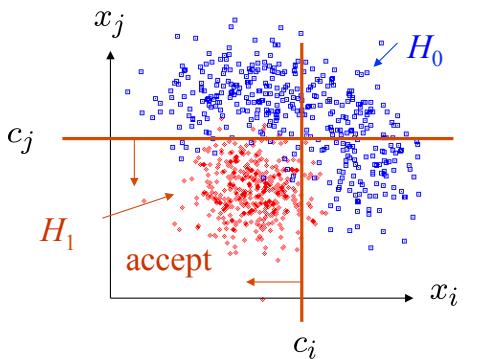
MVA



Something's just not right -- our air is clean, our water is pure, we all get plenty of exercise, everything we eat is organic and free-range, and yet nobody lives past thirty.

- Context:

- In hypothesis testings: how to get a test-statistics $t(x)$ when we do not know the pdfs $f(x)$? Specially if we \mathbf{x} is n-dim!
- How to separate different populations in a n-dim space?



39

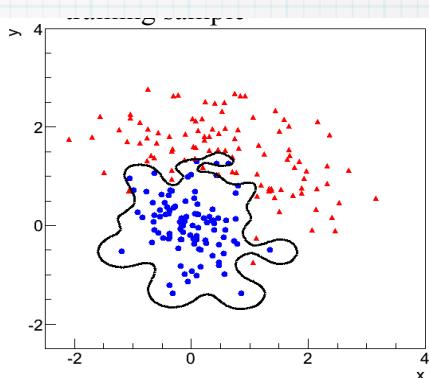
MVA

- Multi Variate Analysis (MVA)

- Provide a test-statistics $t(\mathbf{x})$ to separate population or hypothesis
- Most of them are learning machine methods: they are trained with samples with known populations
 - Be careful to not overtrain!
 - Divide training sample in (at least) two!
 - Use for train and test
- Performance via the ROC curve

- Non exhaustive list:

- Fisher Discriminant
- Neural Network (NN): similar to nervous system neurons
- (Boosted) Decision Trees: classification into branches
- Probability Density Estimation (based on non-parametric pdfs)

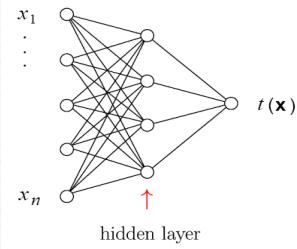


40

MVA-NN

- Neural Networks

- They are training machines
- They have neurons (nodes) and synapses
- As many entries and input variables, \mathbf{x}
- One final node that provides $t(\mathbf{x})$
- Several hidden layers



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

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

- Training based on minimization of an error function

event data \mathbf{x}_i
 class label $y_i = +1, -1$
 weight w_i

$$E[\sum_i w_i (t(\mathbf{x}_i) - y_i)^2]$$

41

MVA-BDT

- Decision trees

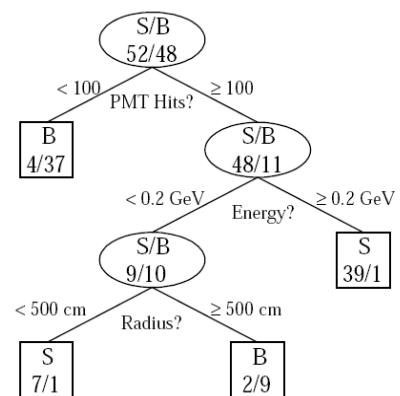
- Classify data in branches
 - using a cut on a variable
 - optimizing the purity

$$\eta_S = \frac{\sum_{i \in S} w_i}{\sum_{i \in S} w_i + \sum_{i \in B} w_i},$$

- Boost

- Create K-classifiers $h_k(\mathbf{x})$ in a iterative process
 - re-weight events that were badly classified
 - obtain a new tree

- Add them with different weights



$$t(\mathbf{x}) = \sum_{k=1}^K \alpha_k h_k(\mathbf{x})$$

42

MVA-PDE

- PDE

- define non-parametric PDFs

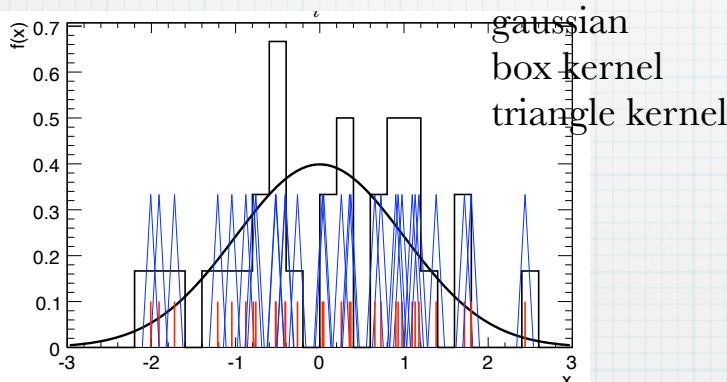
- use kernel functions:

- box, triangle, gaussian

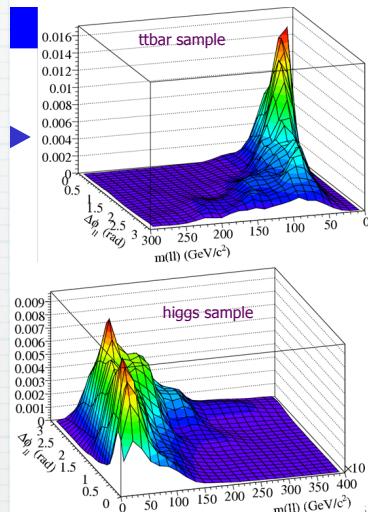
$$\hat{f}(\mathbf{x}) = \frac{1}{Mw^d} \sum_{i=1}^M k\left(\frac{\mathbf{x} - \mathbf{x}_i}{w}\right)$$

$$k(\mathbf{u}) = \frac{1}{(2\pi)^{n/2}} e^{-\mathbf{u}^2/2}$$

- methods to smooth the n-dim



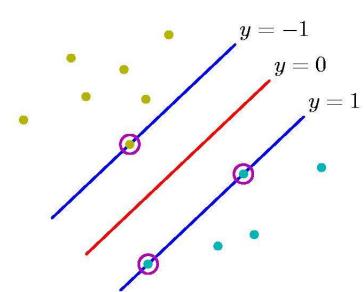
43



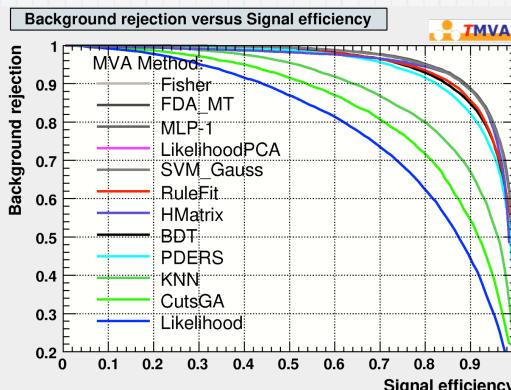
MVA

- SV

- transform \mathbf{x} n-dim into \mathbf{y} m-dim space ($m > n$)
 - “kernel” transformation
- define a *hyper-plane* (frontier) in m-dim



 Use the data file that contains several variables for signal and background categories and generate a test-statistics using different MVA methods.
Suggestion: use the TMVA package in ROOT. Compute the ROC curves.



44