Statistical Methods in HEP

International School of
Theory & Analysis
in Particle Physics

Istanbul, Turkey
31st – 11th February 2011

Jörg Stelzer
Michigan State University, East Lansing, USA

Outline

From probabilities to data samples

Probability, Bayes' theorem

Properties of data samples

Probability densities, multi-dimensional

Catalogue of distributions in HEP, central limit theorem

Data Simulation, random numbers, transformations

From data samples to parameter estimation

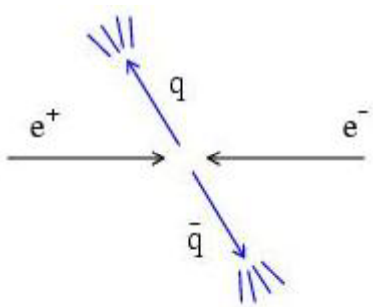
Event classification

Statistical tests, Neyman Pearsen lemma

Multivariate methods

Estimators: general, maximum likelihood, chi-square

Statistical analysis in particle physics



Observe events of a certain type

Measure characteristics of each event

particle momenta, number of muons, energy of jets,...

Theories (e.g. SM) predict distributions of these properties up to free parameters, e.g., α, G_F, M_Z, α_s, m_H, ...

Some tasks of data analysis:

Estimate (measure) the parameters;

Quantify the uncertainty of the parameter estimates;

Test the extent to which the predictions of a theory are in agreement with the data.

What is Probability

S={E_1, E_2,…} set of possible results (events) of an experiment.

E.g. experiment: Throwing a dice.

E_1="throw 1", E_2="throw a 2", E_3="throw an odd number", E_4="throw a number>3",…

E_x and E_y are mutually exclusive if they can't occur at the same time.

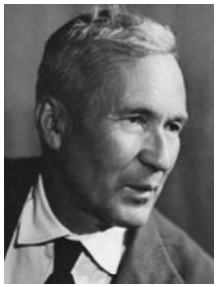
E_1 and E_2 are mutually exclusive, E_3 and E_4 are not

Mathematical probability: For each event *E* exists a *P(E)* with:

$$\text{I}: \quad P(E) \geq 0$$

$$\text{II}: \quad P(E_1 \text{ or } E_2) = P(E_1) + P(E_2) \text{ if } E_1 \text{ and } E_2 \text{ are mutually exclusive}$$

$$\text{III}: \sum P(E_i) = 1, \text{ where the sum is over all mutually exclusive events}$$



A.N. Kolmogorov (1903-1987)

From these axioms we can derive further rules ⇨

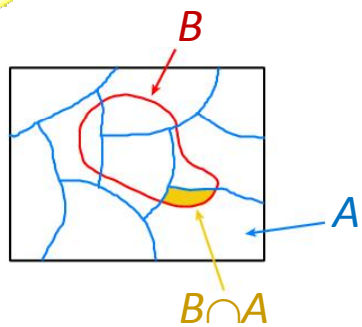
Further properties, conditional probability

We can derive further properties

$$P(\overline{A}) = 1 - P(A)$$
$$P(A \cup \overline{A}) = 1$$
$$P(\varnothing) = 0$$
$$\text{if } A \subset B, \text{ then } P(A) < P(B)$$
$$p(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Conditional probability of A given B

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$



E.g. you are guessing the weekday of birth of a friend: *P*(Sunday) = 1/7.
After the hind it was on a weekday: *P*(Tuesday|weekday) = 1/5
[*P*(Tuesday and weekday) = 1/7, *P*(weekday) = 5/7]

Independent events A and B

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

If your friend hints it was a rainy day:
P(Tuesday|rainday) = 1/7

Axioms can be used to build a complicated theory, but the numbers so far are entirely free of meaning. Different interpretations of probability

Probability as frequency limit

Perform an repeatable experiment N times with outcomes $X_1, X_2, \ldots$ (the ensemble). Count the number of times that X occurs: N_X. Fraction N_X / N tends toward a limit, defined as

the probability of outcome X:

$$P(X) = \lim_{N \to \infty} \frac{N_X}{N}$$

Richard von Mises (1883-1953)

Useful in daily life?

The N outcomes of the experiment are the ensemble. $P(E)$ depends on the experiment and one the <u>ensemble</u> !

The biggest problem when doing demographical studies (shopping behavior, advertisements) is to find the representative ensemble!

Experiment must be <u>repeatable</u>.

German insurance company X finds that 1.1% of their male clients dies between 40 and 41. Does that mean that the probability that Hr. Schmitt, he has a police with X, dies between 40 and 41 is 1.1%? What if the data were collected from a sample of German smoking hang-glider pilots? Likely you would have gotten a different fraction.

Common approach in HEP:

Physical laws are universal and unchanging. Different collider experiments all draw from the same ensemble of particle interactions repeatedly in the same way.

Objective probability – propensity

Examples: throwing a coin, rolling a die, or drawing colored pearls out of a bag, playing roulette.

Probability of a certain event as an intrinsic property of the experiment.

E="Draw a red and then a blue pearl when there are 3 red, 5 blue, and 2 black in the bag". P(E) can be calculated without actually performing the experiment.

Does not depend on any collection of events, it is a single-case probability, often called chance or propensity.

Propensities can not be empirically asserted

If the experiment is being performed, the propensities give rise to frequencies of events. This could be defining the propensity (K.Popper), however problems with the stability of these frequencies arise.

Hence propensities now often defined by the theoretical role they play in science, e.g. based on an underlying physical law.

Bayes Theorem

From conditional probability

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A)$$

follows Bayes' theorem

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

Reverend Thomas Bayes (1702–1761)

Uncontroversial consequence of Kolmogorov's axioms!

Subjective probability

A,*B*, … are hypothesis (statements that are either true or false). Define the probability of hypothesis A:

$$P(A) = \text{degree of belief that } A \text{ is true}$$

(Considered "unscientific" in the frequency definition)

Applied to Bayes' theorem:

Prediction
Probability of a result B assuming hypothesis A is true

(= likelihood function, back later)

Posterior probability
Probability of hypothesis A after seeing the result B

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

Initial degree of belief (prior probability)
Probability of hypothesis A, before seeing the result.
→ this is the **subjective** part

Normalization: total probability of seeing result B.
Involves sum over all possible hypothesis

Experimental evidence or lack thereof modifies initial degree of belief, depending on agreement with prediction.

Interpretation of Bayes theorem

$$P(\text{theory}|\,\text{result}) = \frac{P(\text{result}|\,\text{theory})}{P(\text{result})} P(\text{theory})$$

If a result R forbidden by theory T, $P(R|T) = 0$, then the probability that the theory is correct when the result is observed is 0: $P(T|R)=0$

⇒ An observation of R would disprove T.

If theory T says R is unlikely, $P(R|T) = \searrow$, then the theory T is unlikely under observation of R: $P(T|R)=\searrow$

⇒ An observations of R would lower our belief in T.

If theory T predicts R to have a high probability, $P(R|T) = \nearrow$, then the theory T is likely under observation of R: $P(T|R)=\nearrow$

⇒ An observations of R would strengthen our belief in T.

If the denominator P(R) is large, ie there are many reasons for R to happen, observation of R is not a strong support of T!

- The problem with the background

Law of total probability

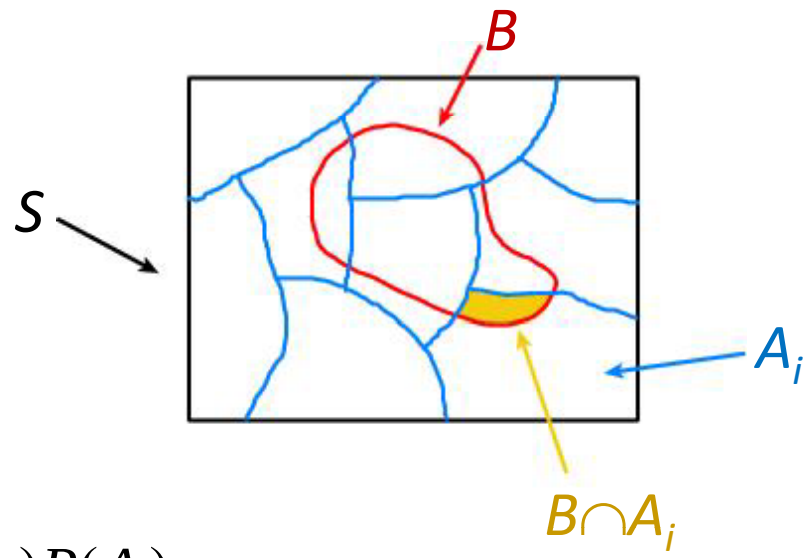
Sample space S with subset B

Disjoint subsets A_i of S: $\bigcup_i A_i = S$

B is made up of disjoint $B \cap A_i$:

$$B = \bigcup_i B \cap A_i$$

$$P(B \cap A_i) = P(B \mid A_i) P(A_i)$$



Law of total probability

$$\boxed{P(B) = \sum_i P(B \cap A_i) = \sum_i P(B \mid A_i) P(A_i)}$$

Bayes' theorem becomes

$$\boxed{P(A \mid B) = \frac{P(B \mid A) P(A)}{\sum_i P(B \mid A_i) P(A_i)}}$$

Example of Bayes' theorem

$$A_1 = \pi = A$$
$$A_2 = \mathrm{K}$$
$$B = \mathrm{signal}$$

Meson beam

Consists of 90% pions, 10% kaons

Cherenkov counter to give signal on pions

95% efficient for pions, 6% fake rate (accidental signal) for kaons

Q1: if we see a signal in the counter, how likely did it come from a pion?

$$p(\pi|\mathrm{signal}) = \frac{p(\mathrm{signal}|\pi)}{p(\mathrm{signal}|\pi)\,p(\pi) + p(\mathrm{signal}|\mathrm{K})\,p(\mathrm{K})}\,p(\pi)$$
$$= \frac{0.95}{0.95\times 0.90 + 0.06\times 0.10}\times 0.90 = 99.3\%$$

$\Rightarrow$ 0.7% chance that the signal came from a kaon.

Q2: if there is no signal, how likely was that a kaon?

$$p(\mathrm{K}|\mathrm{no\ signal}) = \frac{0.94}{0.05\times 0.90 + 0.94\times 0.10}\times 0.10 = 67.6\%$$

Which probability to use?

Frequency, objective, subjective – each has its strong points and shortcomings.

- All consistent with Kolmogorov axioms.

In particle physics frequency approach most often useful.

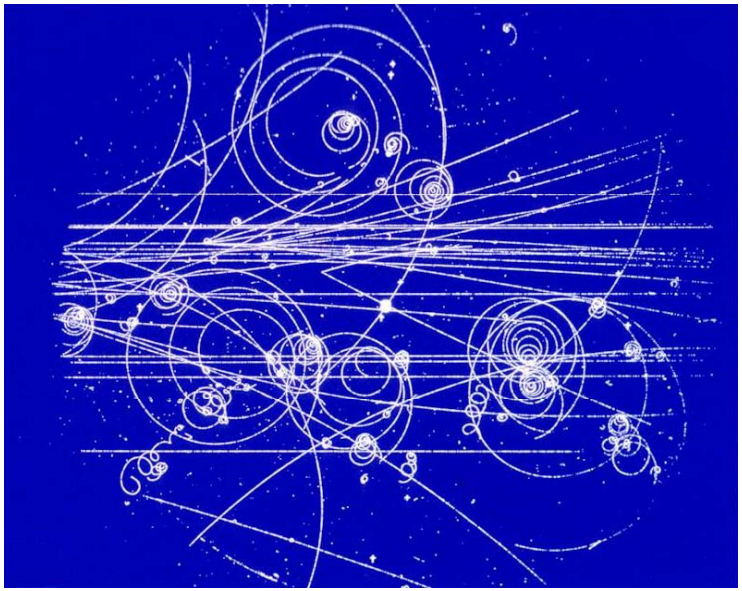
- For instance when deriving results from analyzing many events from a dataset.

Subjective probability can provide you with a more natural way of thinking about and treating non-repeatable phenomena.

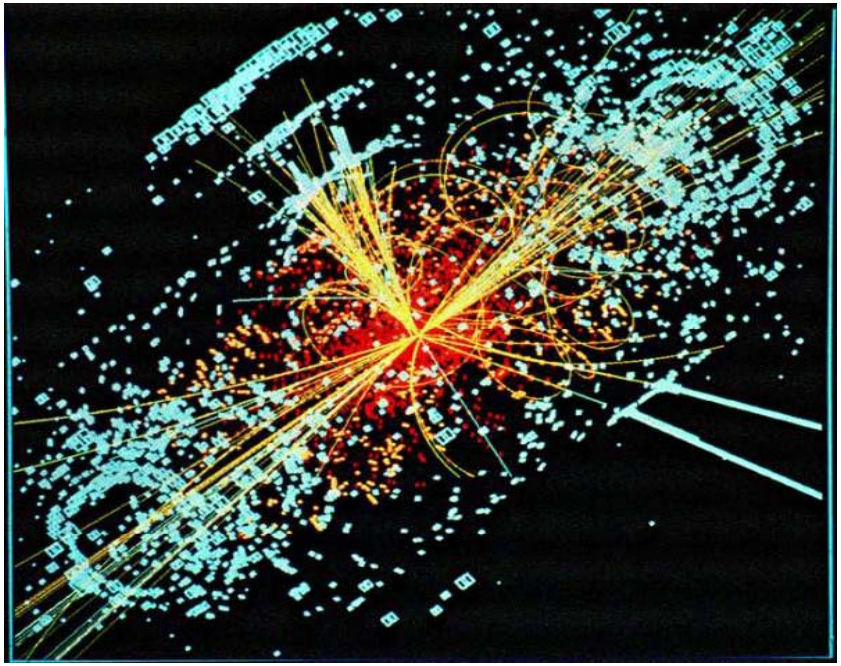
- Treatment of systematic uncertainties, probability that you discovered SUSY or the Higgs in your analysis, probability that parity is violated given a certain measurement, …

Be aware that the naming conventions are not always clear (im particular '*objective*' and '*subjective*'), best bet is to use "Frequentist" and "Bayesian".

Describing data



Tracks in a bubble chamber at CERN as hit by a pion beam



Higgs event in an LHC proton–proton collision at high luminosity

(together with ~24 other inelastic events)

HEP: “events” of particle interactions, measured by complex detectors

Measurements of “random” variables, distribution governed by underlying physics processes

Energy, momentum, angle, number of hits, charge, time(delta)

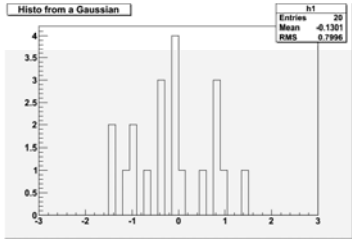
Data sample properties

Data sample (single variable) $x = \{x_1, x_2, \ldots, x_N\}$, can be presented

un-binned

```
0: 0.998933      7: -0.0747045    14: -1.06067
1: -0.434764     8: 0.00791221    15: -1.3883
2: 0.781796      9: -0.410763     16: 0.767397
3: -0.0300528    10: 1.39119      17: -0.73603
4: 0.824264      11: -0.985066    18: 0.579721
5: -0.0567173    12: -0.0489405   19: -0.382134
6: -0.900876     13: -1.44334
```

or binned

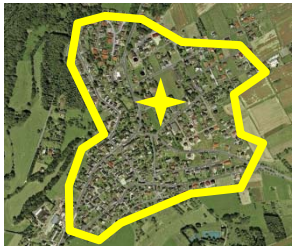


Arithmetic mean: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ or $\bar{x} = \frac{1}{N}\sum_{j=1}^{N_b} n_j x_j$

Variance: $V(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2$

Standard deviation: $\sigma = \sqrt{V(x)} = \sqrt{\overline{x^2} - \bar{x}^2}$

Center of Kleinmai-scheid /Germany



also center of Europe (2005)

More than one variable

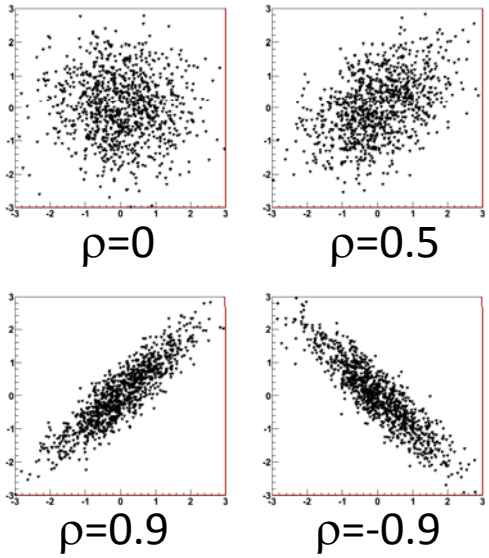
Set of data of two variables $x = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$

```
0: (-1.34361,0.93106)          7: (0.517314,-0.512618)       14: (0.901526,-0.397986)
1: (0.370898,-0.337328)        8: (0.990128,-0.597206)       15: (0.761904,-0.462093)
2: (0.215065,0.437488)         9: (0.404006,-0.511216)       16: (-2.17269,2.31899)
3: (0.869935,-0.469104)       10: (0.789204,-0.657488)       17: (-0.653227,0.829676)
4: (0.452493,-0.687919)       11: (0.359607,-0.979264)       18: (-0.543407,0.560198)
5: (0.484871,-0.51858)        12: (-0.00844855,-0.0874483)   19: (-0.701186,1.03088)
6: (0.650495,-0.608453)       13: (0.264035,-0.559026)
```

There is more information than mean and variance of x and of y !

Covariance:

$$\mathrm{cov}(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \overline{x}\,\overline{y}$$



Correlation:

between -1 and 1

without dimensions

$$\rho = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

<u>Example:</u> group of adults

ρ(height, weight)>0, ρ(weight, stamina)<0, ρ(height, IQ)=0, but ρ(weight, IQ)<0

Probability density function

Suppose outcome of experiment is value v_x for continuous variable x

$$P(A : v_x \text{ found in } [x, x+dx]) = f(x)dx$$

defines the probability density function (PDF): $\boxed{f(x)}$

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

x must be somewhere (axiom III)

Dimensions:

- $P(A)$ is dimensionless (between 0 and 1)
- $f(x)$ has the dimensionality of (1 / dimension of x)

For discrete x, with possible outcomes x_1, x_2, …:

probability mass function: $P(x_i)$ with $\sum_i P(x_i) = 1$

Properties of pdf's

Suppose distribution of variable *x* follows pdf *f(x)*.

Average *x* – the "*expectation value*": $E(x) = \langle x \rangle = \mu = \int_{-\infty}^{\infty} x f(x) dx$

and the variance: $V(x) = \langle x^2 \rangle - \langle x \rangle^2$

Can also be defined for functions of *x*, e.g. *h(x)*: $\langle h \rangle = \int_{-\infty}^{\infty} h(x) f(x) dx$

- $\langle g + h \rangle = \langle g \rangle + \langle h \rangle$
- $\langle gh \rangle \neq \langle g \rangle \langle h \rangle$ unless *g* and *h* are independent

Note: $\langle x \rangle, \langle h \rangle$ are averages over pdf's, $\bar{x}, \bar{h}$ are averages over the real data sample

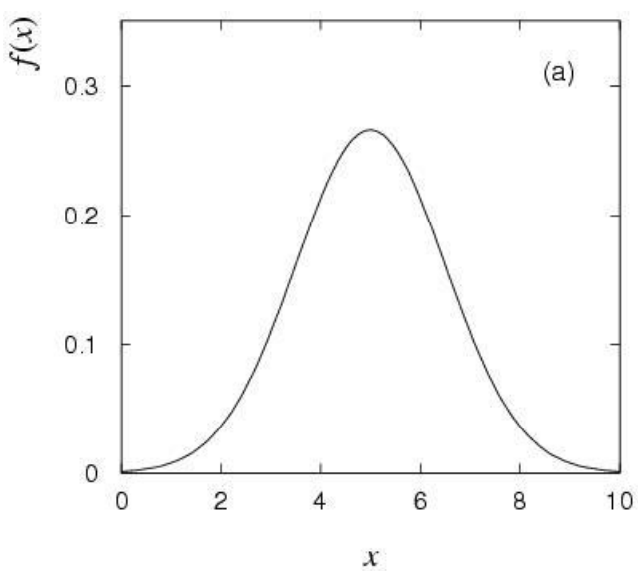
Law of large numbers ensures that $\bar{h} \to \langle h \rangle$

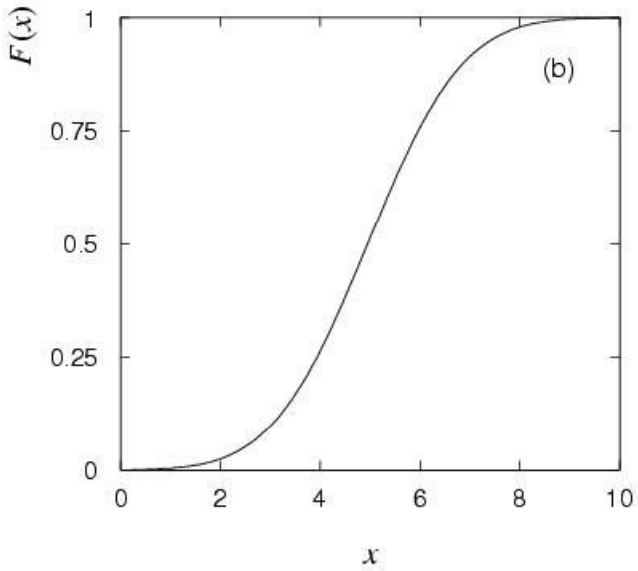
Cumulative distribution function

Probability to have outcome less than or equal to *x* is

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

Monotonously rising function with $F(-\infty)=0$ and $F(\infty)=1$.





Alternatively define pdf with $$f(x) = \frac{\partial F(x)}{\partial x}$$

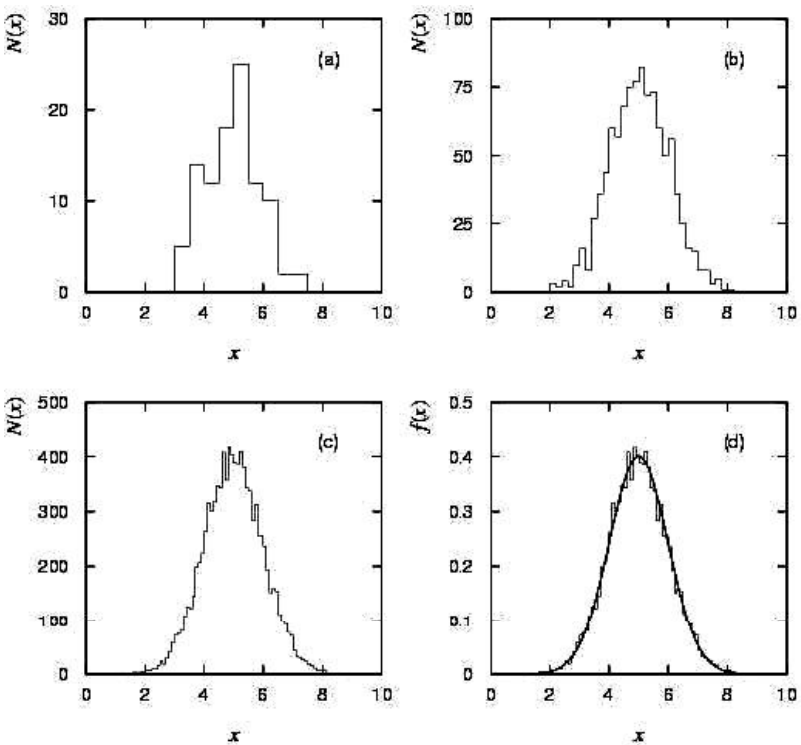
Drawing pdf from data sample

1. Histogram with B bins of width Δx
2. Fill with N outcomes of experiment

   $x_1,\ldots,x_N \Rightarrow H=[n_1,\ldots,n_B]$
3. Normalize integral to unit area

$$\tilde{n}_i = n_i / N \Rightarrow \sum_{i=1}^{B} \tilde{n}_i = 1$$

$$\tilde{n}_i = \text{fraction of } x \text{ found in } [x_i, x_i + \Delta x]$$



PDF

$N \to \infty$ infinite data sample, frequentist approach

$\Delta x \to 0$ zero bin width, step function becomes continuous

Multidimensional pdf's

Outcome of experiment (event) characterized by *n* variables

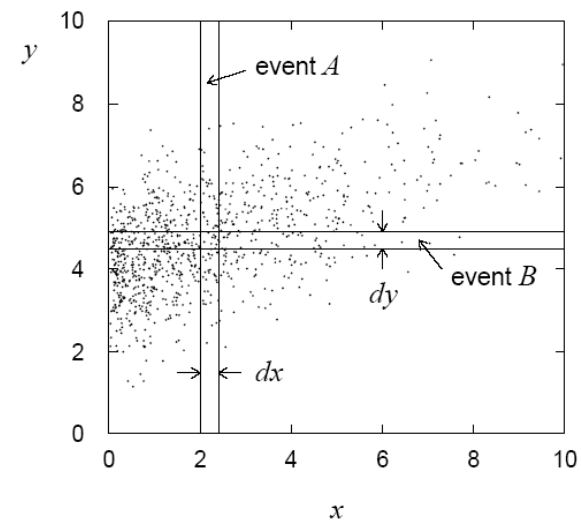
$$\vec{x} = (x^1, x^2, \ldots, x^n)$$

Probability described in *n* dimensions by joint pdf :

$$f(\vec{x}) = f(x^{(1)}, x^{(2)}, \ldots, x^{(n)})$$

$$P(\bigcap_{i=1}^{n} A^{(i)}) = f(\vec{x})d\vec{x}$$

$$= f(x^{(1)}, x^{(2)}, \ldots, x^{(n)})dx^{(1)}dx^{(2)}\cdots dx^{(n)}$$



where

$A^{(i)}$: hypothesis that variable i of event is in interval $x^{(i)}$ and $x^{(i)} + dx^{(i)}$

Normalization: $\int \cdots \int f(x^{(1)}, x^{(2)}, \ldots, x^{(n)})dx^{(1)}dx^{(2)}\cdots dx^{(n)} = 1$

Marginal pdf's, independent variables

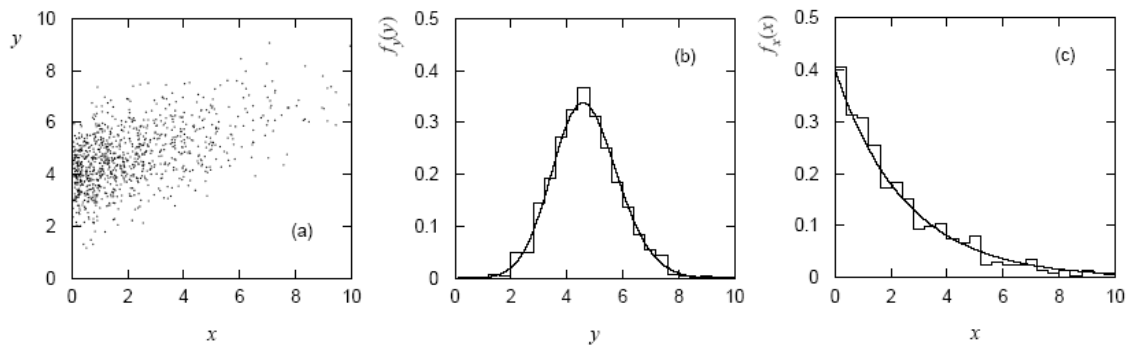
PDF of one (or some) of the variables, integration of all others

$\Rightarrow$ marginal PDF:

$$f_{X^j}(x^{(j)}) = \int f(x^{(1)}, x^{(2)}, \ldots, x^{(n)}) dx^{(1)} dx^{(2)} \cdots dx^{(j-1)} dx^{(j+1)} \cdots dx^{(n)}$$

Marginal PDFs are projections of joint PDFs on individual axis

Note that $\int f_{X^i}(x^{(i)}) dx^{(i)} = 1$



Variables $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ are independent from each other if-and-only-if they factorize:

$$f(\vec{x}) = \prod_i f_{X^i}(x^{(i)})$$

Conditional pdf's

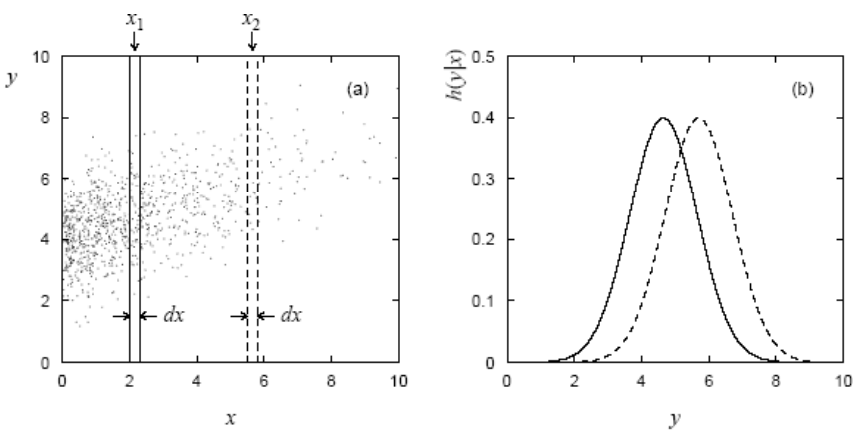
Sometimes we want to consider some variables of joint pdf as constant.

Let's look at two dimensions, start from conditional probability:

$$P(B\,|\,A) = \frac{P(A \cap B)}{P(A)} = \frac{f(x,y)\,dxdy}{f_x(x)\,dx} \equiv h(y\,|\,x)\,dy$$

Conditional pdf, distribution of y for fix $x=x_1$:

$$h(y\,|\,x = x_1) = \frac{f(x = x_1, y)}{f_x(x = x_1)}$$



- In joint pdf treat some variables as constant and evaluate at fix point (e.g. $x=x_1$)
- Divide the joint pdf by the marginal pdf of those variables being held constant evaluated at fix point (e.g. $f_x(x=x_1)$)
- $h(y|x_1)$ is a slice of $f(x,y)$ at $x=x_1$ and has correct normalization $\int h(y\,|\,x = x_1)\,dy = 1$

Some Distributions in HEP

Binomial	Branching ratio
Multinomial	Histogram with fixed N
Poisson	Number of events found in data sample
Uniform	Monte Carlo method
Exponential	Decay time
Gaussian	Measurement error
Chi-square	Goodness-of-fit
Cauchy (Breit-Wigner)	Mass of resonance
Landau	Ionization energy loss

Other functions to describe special processes:

Crystal Ball function, Novosibirsk function, …

Binomial distribution

Outcome of experiment is 0 or 1 with $p=P(1)$ (Bernoulli trials). r : number of 1's occurring in n independent trials.

Probability mass function:

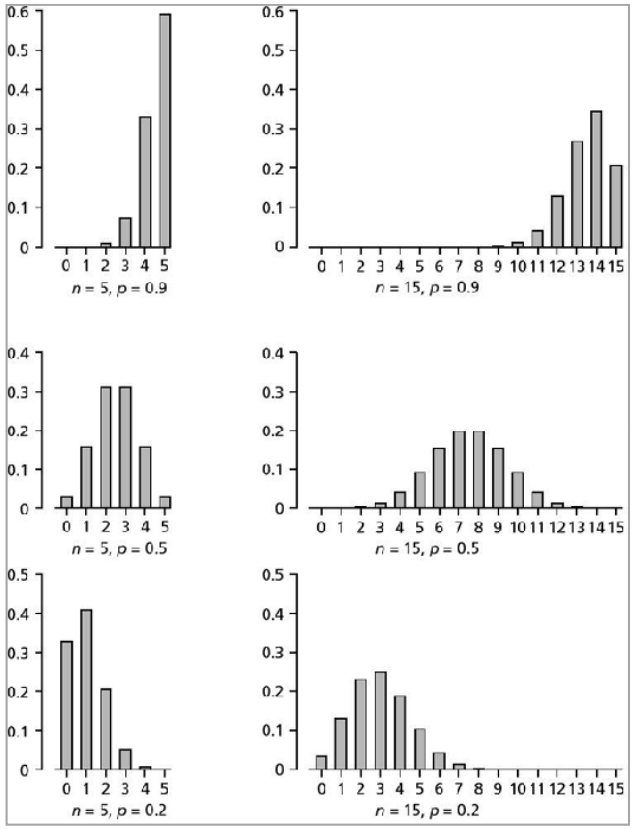
$$P(r;p,n) = p^r (1-p)^{n-r} \frac{n!}{r!(n-r)!}$$

***r* times "1", *n-r* times "0"** **combinatoric term**

Properties:

$$\langle r \rangle = np$$

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



Expectation: A coin with p("head")=0.45 you expect to land on its head np=45 out of n=100 times.

Example: spark chamber 95% efficient to detect the passing of a charged particle. How efficient is a stack of four spark chambers if you require at least three hits to reconstruct a track?

$$P(3;0.95,4) + P(4;0.95,4) = 0.95^3 \times 0.05 \times 4 + 0.95^4 \times 1 = 0.171 + 0.815 = 98.6\%$$

Poisson distribution (law of small numbers)

Discrete like binomial distribution, but no notion of trials. Rather λ, the mean number of (rare) events occurring in a continuum of fixed size, is known.

Derivation from binomial distribution:

- Divide the continuum into *n intervals*, in each interval assume *p="probability that event occurs in interval"*. Note that λ=*np* is the known and constant.
- Binomial distribution in the limit of large n (and small p) for fixed r

$$P(r;p,n) = p^r (1-p)^{n-r} \frac{n!}{r!(n-r)!} \to p^r (1-\lambda/n)^n \frac{n^r}{r!}$$

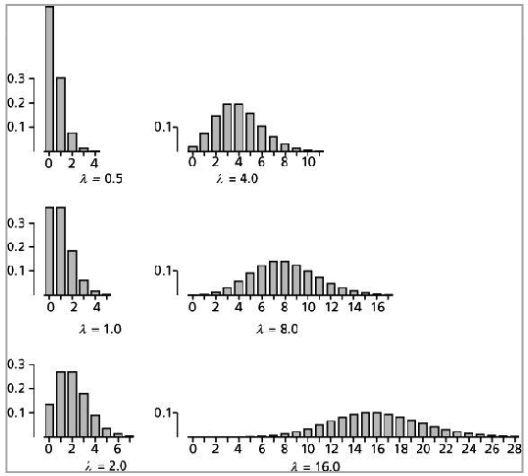
Probability mass function:

$$P(r;\lambda) = \frac{\lambda^r e^{-\lambda}}{r!}$$

Properties:

$$\langle r \rangle = \lambda$$

$$V(r) = \sigma^2 = \lambda$$



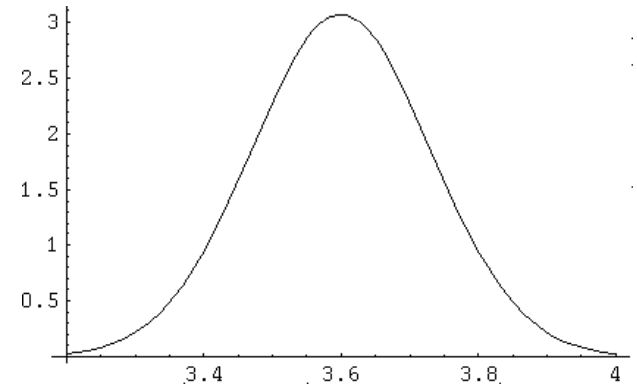
Famous example: Ladislaus Bortkiewicz (1868-1931). The number of soldiers killed by horse-kicks each year in each corps in the Prussian cavalry: 122 fatalities in 10 corps over 20 years. λ=122/200=0.61 deaths on average per year and corp.

Deaths	Prediction	Cases
0	108.7	109
1	66.3	65
2	20.2	22
3	4.1	3
4	0.6	1

Probability of no deaths in a corp in a year: $P(0;0.61) = 0.5434$

Gaussian (normal) distribution

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$



Properties:

$$\langle x \rangle = \mu$$
$$V(x) = \sigma^2$$

Note that μ and σ also denote mean and standard deviation for any distribution, not just the Gaussian. The fact that they appear as parameters in the pdf justifies their naming.

Standard Gaussian

transform x→x'=(x-μ)/σ

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Cumulative distribution $\Phi(x) = \int_{-\infty}^{x} \varphi(x')dx'$ can not be calculated analytically.

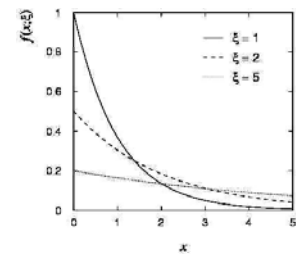
Tables provide: $\Phi(1) = 68.27\%, \quad \Phi(2) = 95.45\%, \quad \Phi(3) = 99.73\%$

Central role in the treatment of errors: central limit theorem

Other distributions

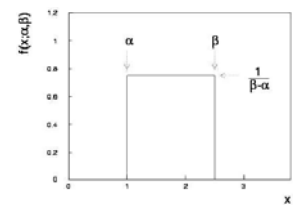
Gaussian, poisson, and binomial are by far the most common and useful. For the description of physical processes you encounter

Exponential $$f(x;\xi)=\begin{cases}\frac{1}{\xi}e^{-x/\xi} & x\geq 0\\ 0 & x<0\end{cases},\quad \langle x\rangle=\xi,\quad V(x)=\xi^2$$

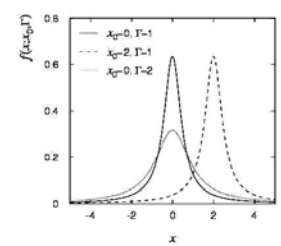


Decay of an unstable particle with mean life-time ξ.

Uniform $$f(x;\alpha,\beta)=\begin{cases}\frac{1}{\beta-\alpha} & \alpha\leq x\leq\beta\\ 0 & \text{otherwise}\end{cases},\quad \langle x\rangle=(\alpha+\beta)/2,\quad V(x)=(\beta-\alpha)^2/12$$

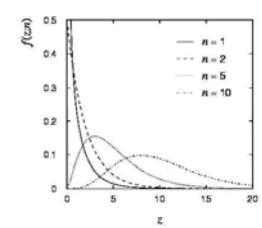


Breit-Wigner $$f(x;\Gamma,x_0)=\frac{1}{\pi}\frac{\Gamma/2}{(\Gamma/2)^2+(x-x_0)^2},\quad \langle x\rangle\text{not well defined},\quad V(x)\to\infty$$



Mass of resonance, e.g. K*, ϕ, ρ. Full width at half maximum, Γ, is the decay rate, or the inverse of the lifetime.

Chi-square $$f(x;n)=\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2},\quad \langle x\rangle=n,\quad V(x)=2n$$



Goodness-of-fit test variable with method of least squares follows this. Number of degrees of freedom *n*.

Central limit theorem

A variable Y that is produced by the cumulative effect of many independent variables X_i, $Y = \sum_{i=1}^{N} X_i$, with mean μ_i and variance σ_i^2 will be approximately Gaussian.

Expectation value $$\langle Y \rangle = \sum_{i=1}^{N} \langle X_i \rangle = \sum_{i=1}^{N} \mu_i$$

Variance $$V(Y) = \sum_{i=1}^{N} V(X_i) = \sum_{i=1}^{N} \sigma_i^2$$

Becomes Gaussian as $N \to \infty$

<u>Examples</u>

- E.g. human height is Gaussian, since it is sum of many genetic factors.
- Weight is not Gaussian, since it is dominated by the single factor food.

Half-time summary

Part I

Introduced probability

- Frequency, subjective. Bayes theorem.

Properties of data samples

- Mean, variance, correlation

Probability densities – underlying distribution from which data samples are drawn

- Properties, multidimensional, marginal, conditional pdfs
- Examples of pdfs in physics, CLT

Part II

HEP experiment: repeatedly drawing random events from underlying distribution (the laws of physics that we want to understand). From the drawn sample we want to estimate parameters of those laws

- Purification of data sample: statistical testing of events
- Estimation of parameters: maximum likelihood and chi-square fits
- Error propagation

Intermezzo: Monte Carlo simulation

Looking at data, we want to infer something about the (probabilistic) processes that produced the data.

Preparation:

- tuning signal / background separation to achieve most significant signal
- check quality of estimators (later) to find possible biases
- test statistical methods for getting the final result

all of this requires data based on distribution with known parameters

<u>Tool: Monte Carlo simulation</u>

Based on sequences of random numbers simulate particle collisions, decays, detector response, …

Generate random numbers

Transform according to desired (known) PDF

Extract properties

Random numbers

Sequence of random numbers uniformly distributed between 0 and 1

- True random numbers in computers use special sources of entropy: thermal noise sources, sound card noise, hard-drive IO times, …
  - Simulation has to run on many different types of computers, can't rely on these
- Most random numbers in computers are pseudo-random: algorithmically determined sequences

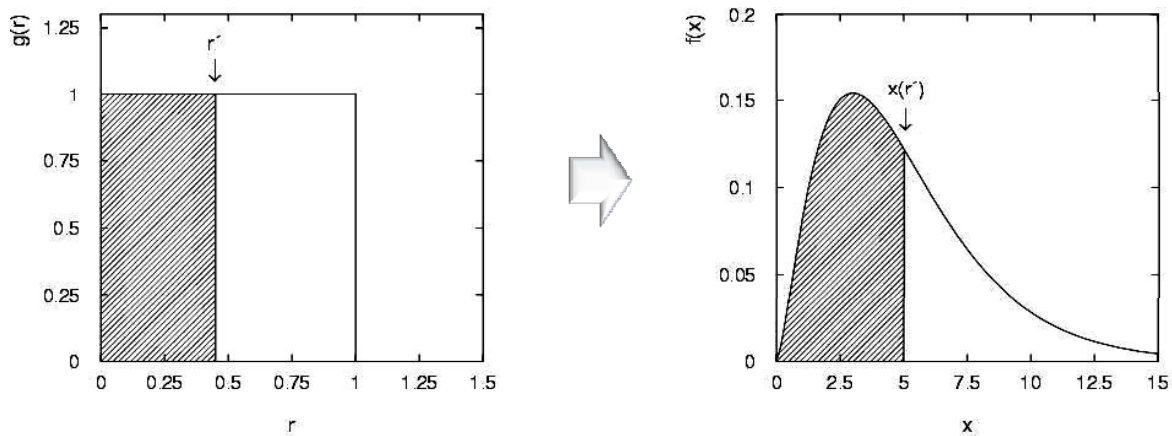
Many different methods, e.g. 4 in root

- TRandom $x_{n+1} = (ax_n + c) \bmod m$ with $a = 1103515245, c = 12345,$ and $m = 2^{31}$
  - Same as BSD rand() function. Internal state 32bit, short period ~10^9.
- TRandom1
  - Based on mathematically proven Ranlux. Internal state 24 x 32bit, period ~10^{171}. 4 luxury levels. Slow. Ranlux is default in ATLAS simulation.
- TRandom2
  - Based on maximally equi-distributed combined Tausworthe generator. Internal state 3 x 32bit, period ~10^{26}. Fast. Use if small number of random numbers needed.
- TRandom3
  - Based on Mersenne and Twister algorithm. Large state 624 x 32bit. Very long period ~10^{6000}. Fast. Default in ROOT.
- Seed: Seed 0 uses random seed, anything else gives you reproducible sequence.

Transformation method – analytic

Given $r_1, r_2, \ldots, r_n$ uniform in [0, 1], find $x_1, x_2, \ldots, x_n$ that follow $f(x)$ by finding a suitable transformation $x(r)$.

Require $$P(r \le r') = P(x \le x(r'))$$



this means $$\int_{-\infty}^{r'} u(r)\,dr = r' = \int_{-\infty}^{x(r')} f(x')\,dx' = F(x(r'))$$

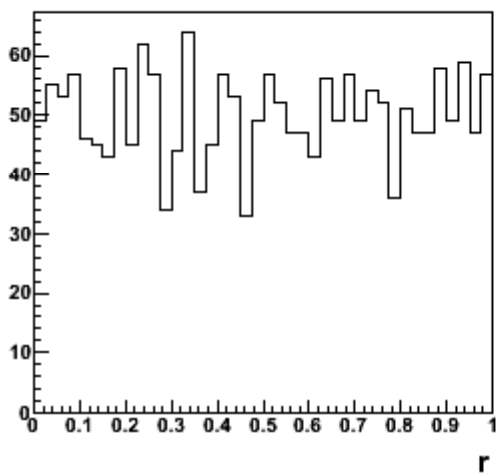
so set $F(x) = r'$ and solve for $x(r')$.

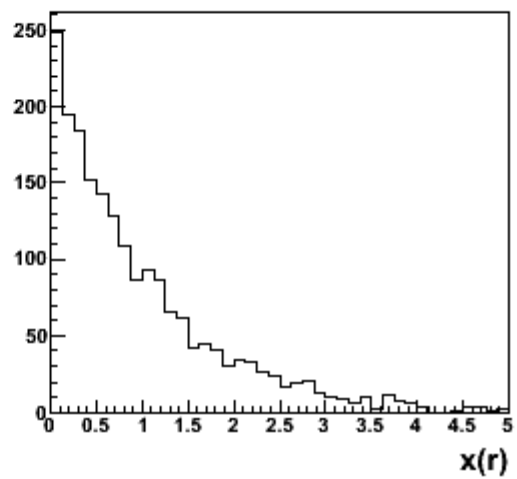
Example

Exponential pdf: $f(x;\xi) = \frac{1}{\xi} e^{-x/\xi}, \quad \text{with } x \geq 0$

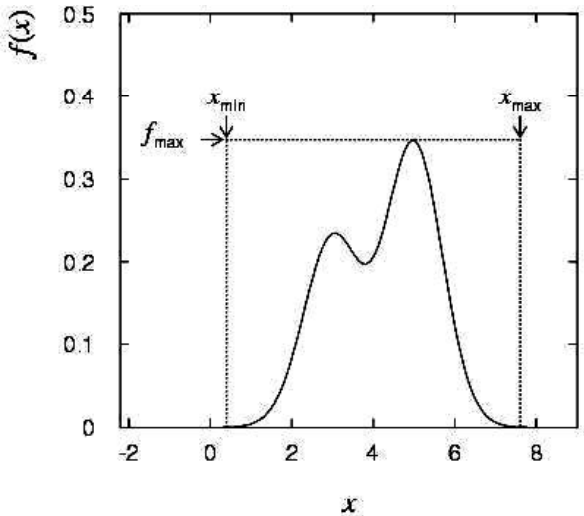
So set $F(x) = \int_0^x \frac{1}{\xi} e^{-x'/\xi}\, dx' = r$ and solve for $x(r)$

This gives the transformation $x(r) = -\xi \ln(1-r)$





Accept – reject method



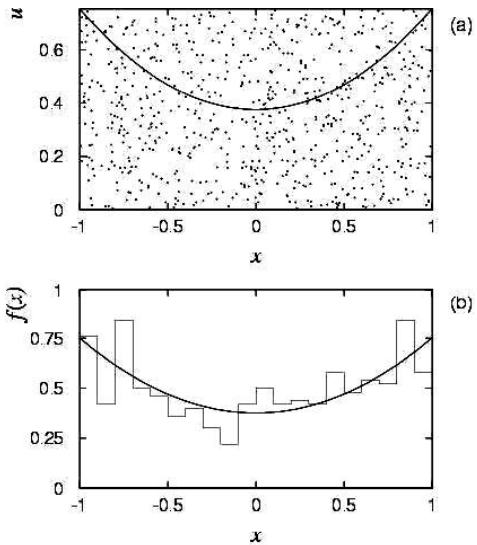
Enclose the pdf in a box

$[\, x_{min} , x_{max} \,] \times [\, 0 , f_{max} \,]$

Procedure to select x according to f(x)

1) Generate two random numbers
   1) x, uniform in $[x_{min}, x_{max}]$
   2) u, uniform in $[0, f_{max}]$

1) If $u<f(x)$, then accept x

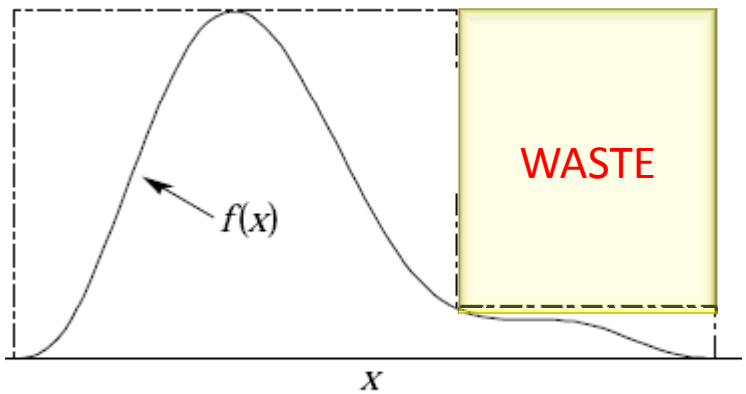


"If dot below curve, use x value in histogram"

Improving accept – reject method

In regions where $f(x)$ is small compared to f_{max} a lot of the sampled points are rejected.

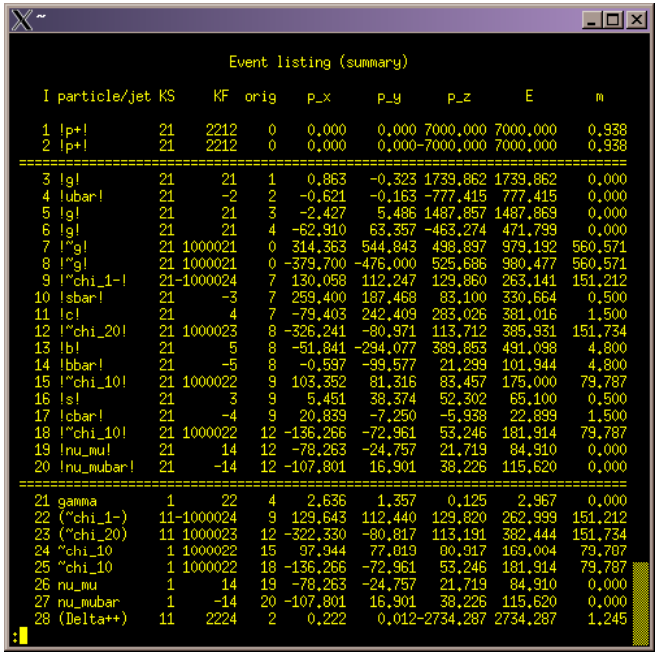
Serious waste of computing power, simulation in HEP consists of billions of random numbers, so this does add up!



Split $[x_{min}, x_{max}]$ in regions (i), each with its own $f_{max}^{(i)}$, and simulate pdf separately. Proper normalization $N^{(i)} \propto A^{(i)} = (x_{max}^{(i)} - x_{min}^{(i)}) \times f_{max}^{(i)}$

More general: find enveloping function around $f(x)$, for which you can generate random numbers. Use this to generate x.

MC simulation in HEP



Event listing (summary)

I	particle/jet	KS	KF	orig	p_x	p_y	p_z	E	m
1	!p+!	21	2212	0	0,000	0,000	7000,000	7000,000	0,938
2	!p+!	21	2212	0	0,000	0,000	-7000,000	7000,000	0,938
3	!g!	21	21	1	0,863	-0,323	1739,862	1739,862	0,000
4	!ubar!	21	-2	2	-0,621	-0,163	-777,415	777,415	0,000
5	!g!	21	21	3	-2,427	5,486	1487,857	1487,869	0,000
6	!g!	21	21	4	-62,910	63,357	-463,274	471,799	0,000
7	!~g!	21	1000021	0	314,363	544,843	498,897	979,192	560,571
8	!~g!	21	1000021	0	-379,700	-476,000	525,686	980,477	560,571
9	!~chi_1-!	21	-1000024	7	130,058	112,247	129,860	263,141	151,212
10	!sbar!	21	-3	7	259,400	187,468	83,100	330,664	0,500
11	!c!	21	4	7	-79,403	242,409	283,026	381,016	1,500
12	!~chi_20!	21	1000023	8	-326,241	-80,971	113,712	385,931	151,734
13	!b!	21	5	8	-51,841	-294,077	389,853	491,098	4,800
14	!bbar!	21	-5	8	-0,597	-99,577	21,299	101,944	4,800
15	!~chi_10!	21	1000022	9	103,352	81,316	83,457	175,000	79,787
16	!s!	21	3	9	5,451	38,374	52,302	65,100	0,500
17	!cbar!	21	-4	9	20,839	-7,250	-5,938	22,899	1,500
18	!~chi_10!	21	1000022	12	-136,266	-72,961	53,246	181,914	79,787
19	!nu_mu!	21	14	12	-78,263	-24,757	21,719	84,910	0,000
20	!nu_mubar!	21	-14	12	-107,801	16,901	38,226	115,620	0,000
21	gamma	1	22	4	2,636	1,357	0,125	2,967	0,000
22	(~chi_1-)	11	-1000024	9	129,643	112,440	129,820	262,999	151,212
23	(~chi_20)	11	1000023	12	-322,330	-80,817	113,191	382,444	151,734
24	~chi_10	1	1000022	15	97,944	77,819	80,917	169,004	79,787
25	~chi_10	1	1000022	18	-136,266	-72,961	53,246	181,914	79,787
26	nu_mu	1	14	19	-78,263	-24,757	21,719	84,910	0,000
27	nu_mubar	1	-14	20	-107,801	16,901	38,226	115,620	0,000
28	(Delta++)	11	2224	2	0,222	0,012	-2734,287	2734,287	1,245

<u>Event generation:</u> PYTHIA, Herwig, ISAJET,…

general purpose generators

for a large variety of reactions:

$e^+e^- \rightarrow \mu^+\mu^-, \tau^+,\tau^-$, hadrons, ISR, ...

pp → hadrons, Higgs, SUSY,...

Processes: hard production, resonance decays, parton showers, hadronization, normal decays, …

Get a long list of colliding particles:

intermediated resonances, short lived particles, long lived particles and their momentum, energy, lifetimes

<u>Detector response:</u> GEANT

multiple Coulomb scattering (scattering angle)

particle decays (lifetime)

ionization energy loss (ΔE)

electromagnetic, hadronic showers

production of signals, electronics response, ...

Get simulated raw data

<u>Data reconstruction:</u>

Same as for real data but keep truth information

Clustering, tracking, jet-finding

Estimate efficiencies

found / # generated

= detector acceptance x reconstruction efficiencies x event selection

Test parameter estimation

Nature,
Theory

Probability

Data
simulated or real

Given these distributions, how will the data look like ?

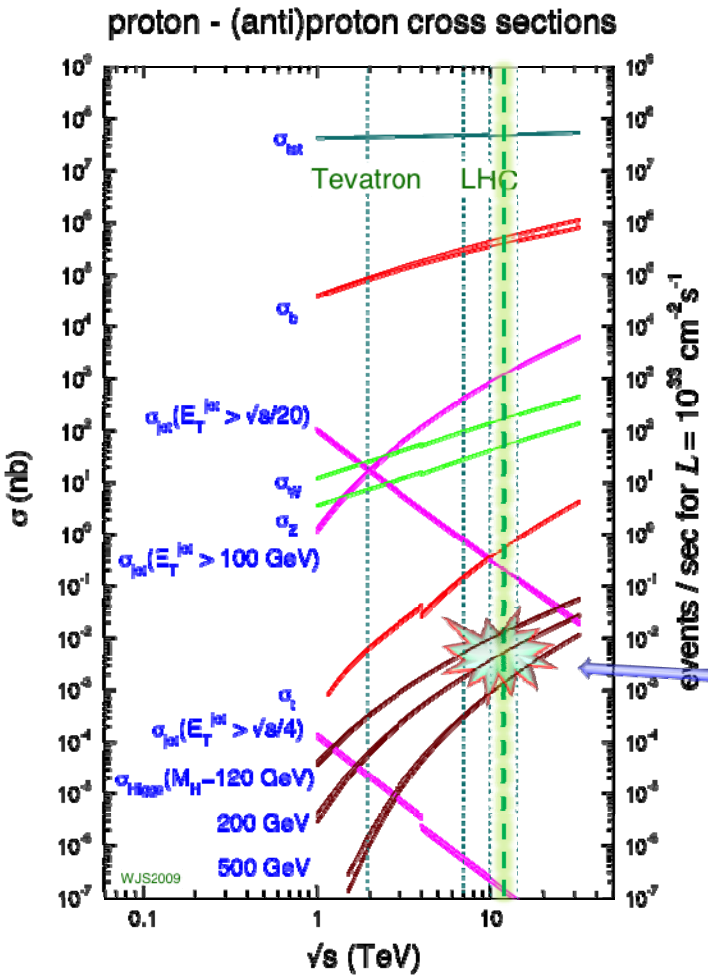
Nature,
Theory

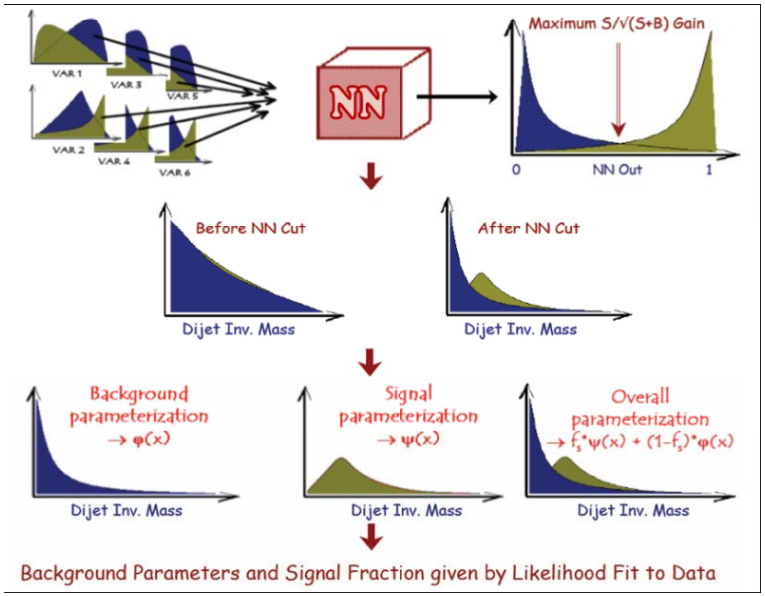
Statistical inference

Data
simulated or real

Given these data, what can we say about the correctness, paramters, etc. of the distribution functions ?

Typical HEP analysis





Signal ~10 orders below total cross-section

1. Improve significance: Discriminate signal from background. Multivariate analysis, using all available information.
   Event(W/SUSY), cone(τ,jet), object level (PID)

2. Parameter estimation
   Mass, CP, size of signal

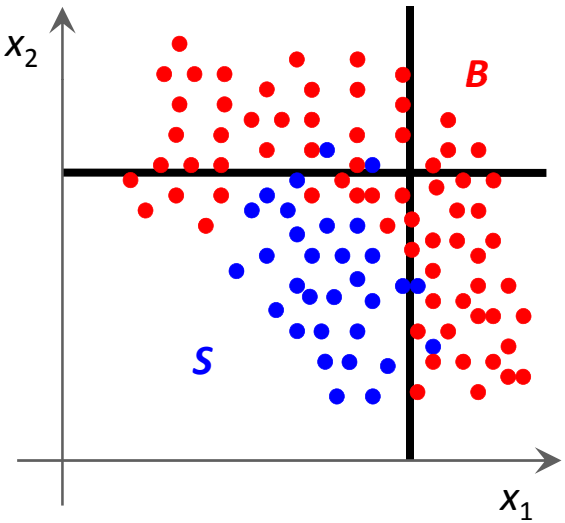
Event Classification

Suppose data sample with two types of events: *Signal* ***S***, *Background* ***B***

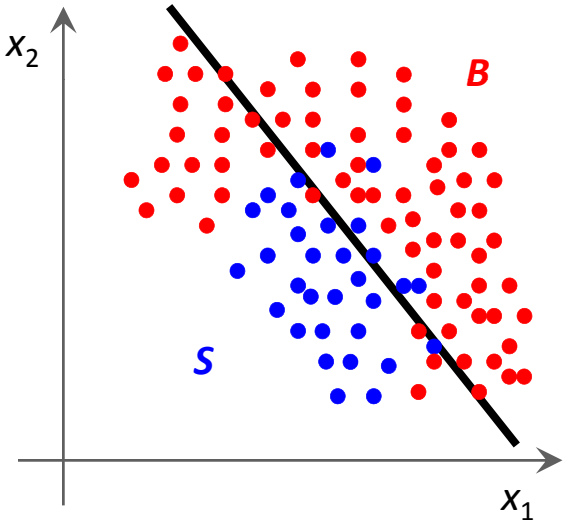
Suppose we have found discriminating input variables x_1, x_2, …

What decision boundary should we use to select signal events (type ***S***)?

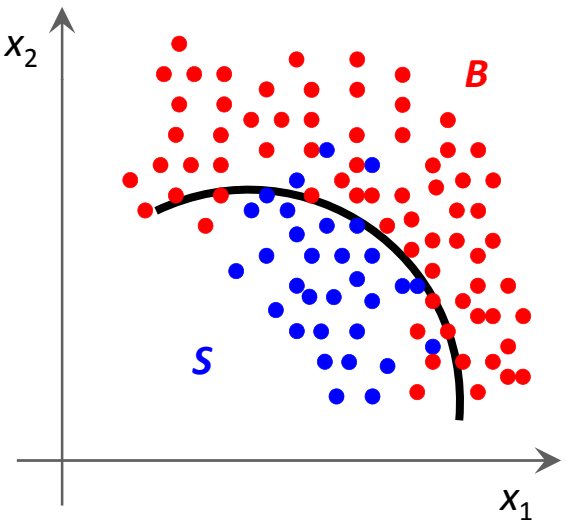
Rectangular cuts?



Linear boundary?



A nonlinear one?



How can we decide this in an optimal way?

Multivariate event classification. → Machine learning

Multivariate classifiers

Input:

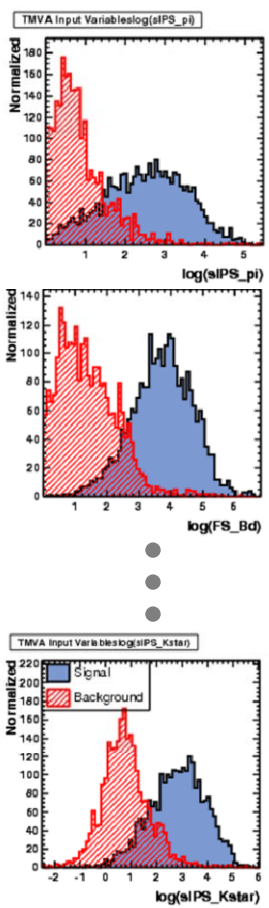
n variables as event measurement described by *n*-dimensional joint pdf, one for each event type: ***f(x|S)*** and ***f(x|B)***

Classifier:

Maps *n*-dimensional input space $\vec{x} = (x^1, x^2, \ldots, x^n) \in \Re^n$ to one-dimensional output $y(\vec{x}) \in \Re$.

Output:

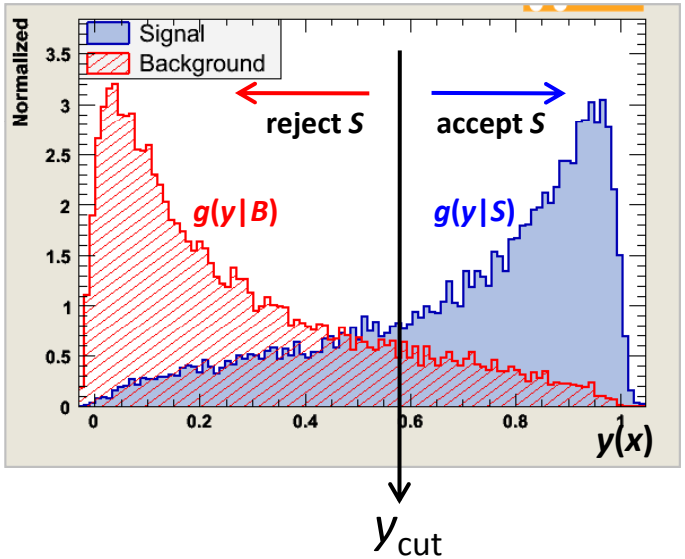
Distributions have maximum ***S***/***B*** separation.



$\Re^n$

$$y : \Re^n \rightarrow \Re$$

Classifier output distribution



Decision boundary can now be defined by single cut on the classifier output $y_{\text{cut}} = y(\vec{x})$, which divides the input space into the rejection (critical) and acceptance region. This defines a test, if event falls into critical region we reject ***S*** hypothesis.

Convention

In literature one often sees

- Null-hypothesis H_0, the presumed “default stage”
- Alternative hypothesis H1

In HEP we usually talk about signal and background and it is common to assign

Background $B = H_0$

Signal $S = H_1$

Definition of a test

Goal is to make some statement based on the observed data ***x*** as to the validity of the possible hypotheses, e.g. signal hypothesis *S*.

A test of $H_0=B$ is defined by specifying a critical region W_S (the signal region) of the data space such that there is an (only small) probability, α, for an event ***x*** of type $H_0=B$, to be observed in there, i.e.,

$$P(\vec{x} \in W_S \mid H_0) \leq \alpha$$

Events that are in critical region W_S: reject hypothesis H_0 = accept as signal. α is called the *size* or *significance level* of the test. Note that all α larger than $P(x \in W_S | H_0)$ are called significance of this test. Let's think of α now as the smallest significance.

<u>Errors:</u>

Reject H_0 for background events $\Rightarrow$ Type-I error α

Accept H_0 for signal events $\Rightarrow$ Type-II error β

$$P(\vec{x} \notin W \mid \$) = \beta$$

Accept as: / Truly is:	Signal	Back-ground
Signal	☺	Type-2 error
Back-ground	Type-1 error	☺

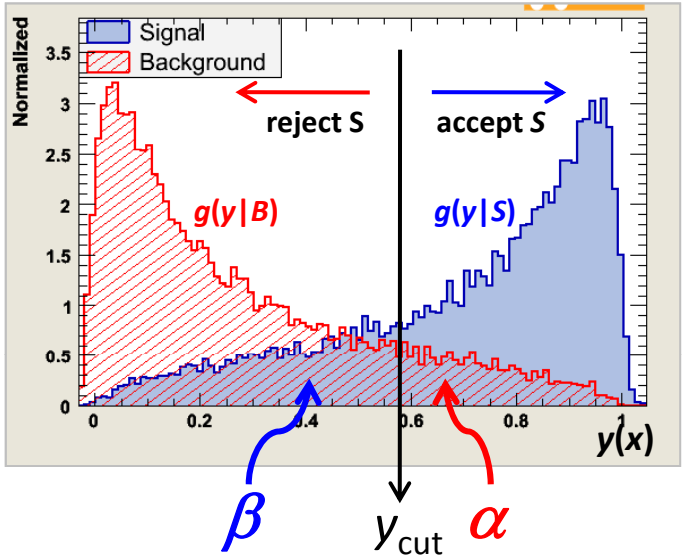
Efficiencies

<u>Signal efficiency:</u>

Probability to accept signal events as signal

$$\varepsilon_S = \int_{y_{\text{cut}}}^{\infty} g(y \mid S)\, dy = 1 - \beta$$

1-β also called "*the power*"



<u>Background efficiency:</u>

Probability to accept background events as signal

$$\varepsilon_B = \int_{y_{\text{cut}}}^{\infty} g(y \mid B)\, dy = \alpha$$

Neyman – Pearson test

Design test in n-dimensional input space by defining critical region W_S. Selecting event in W_S as signal with errors α and β:

$$\alpha = \int_{W_S} f_B(\vec{x})\, d\vec{x} = \varepsilon_B \quad \text{and} \quad \beta = 1 - \int_{W_S} f_S(\vec{x})\, d\vec{x} = 1 - \varepsilon_S$$

A good test makes both errors small, so chose W_S where f_B is small and f_S is large, define by likelihood ratio

$$\frac{f_S(\vec{x})}{f_B(\vec{x})} \geq c$$

Any particular value of c determines the values of α and β.

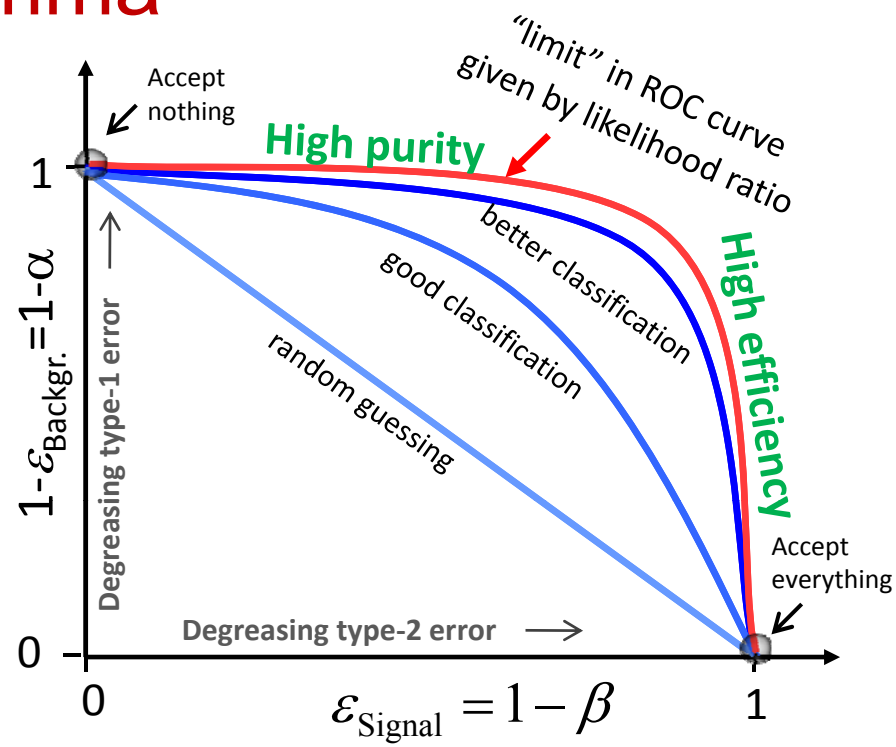
Neyman – Pearson Lemma

Likelihood ratio $y_r(x) = \frac{P(\vec{x}\,|\,S)}{P(\vec{x}\,|\,B)} = \frac{f_S(\vec{x})}{f_B(\vec{x})}$

"The *likelihood-ratio test* as selection criteria gives for each selection efficiency the best background rejection."

It maximizes the area under the *ROC*-curve

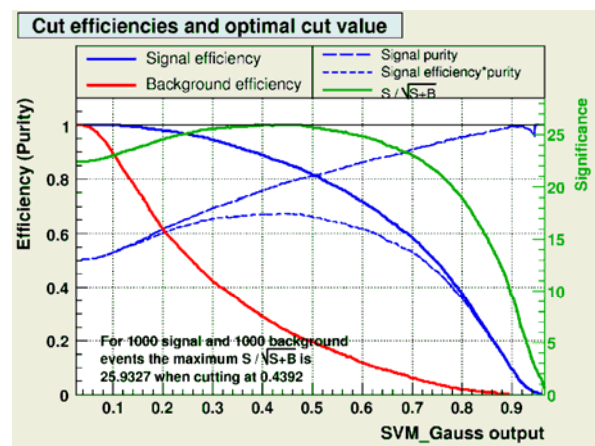
"*Receiver Operating Characteristics*" (ROC) curve plots (1-) the minimum type-II error as a function of (1-) the type-I error. The better the classifier the larger the area under the ROC curve.



From the ROC of the classifier chose the working point

➡ need expectation for *S* and *B*

Cross section measurement:	maximum of S/√(S+B) or equiv. √(ε·*p*)
Discovery of a signal:	maximum of S/√(B)
Precision measurement:	high purity (*p*)
Trigger selection:	high efficiency (ε)



Realistic event classification

Neyman-Pearson lemma doesn't really help us since true densities are typically not known!

Need a way to describe them approximately:

MC simulated events

Control samples derived from data (even better but generally more difficult to get)

Use these "training" events to

- Try to estimate the functional form of $f_{S/B}(x)$ from which the likelihood ratio can be obtained

  e.g. D-dimensional histogram, Kernel density estimators, MC-based matrix-element methods, …

- Find a "discrimination function" y(x) and corresponding decision boundary (i.e. affine hyperplane in the "feature space": y(x) = const) that optimally separates signal from background

  e.g. Linear Discriminator, Neural Networks, Boosted Decision, Support Vector Machines, …

$\Rightarrow$ Supervised Machine Learning (two basic types)

Machine Learning

Computers do the hard work (number crunching) but it's not all magic. Still need to …

- Choose the discriminating variables, check for correlations
- Choose the class of models (linear, non-linear, flexible or less flexible)
- Tune the "learning parameters"
- Check the generalization properties (avoid overtraining)
- Check importance of input variables at the end of the training
- Estimate efficiency
- Estimate systematic uncertainties (consider trade off between statistical and systematic uncertainties)

Let's look at a few:

Probability density estimation (PDE) methods

Boosted decision trees

PDE methods

Construct non-parametric estimators $\hat{f}$ of the pdfs $f(\vec{x}\,|\,S)$ and $f(\vec{x}\,|\,B)$ and use these to construct the likelihood ratio:

$$y_r(\vec{x}) = \frac{\hat{f}(\vec{x}\,|\,S)}{\hat{f}(\vec{x}\,|\,B)}$$

Methods are based on turning the training sample into PDEs for signal and background and then provide fast lookup for $y_r(\vec{x})$

Two basic types

- Projective Likelihood Estimator (Naïve Bayes)

- Multidimensional Probability Density Estimators
  - Parcel the input variable space in cells. Simple example: n-dimensional histograms
  - Kernels to weight the event contributions within each cell.
  - Organize data in search trees to provide fast access to cells

Projective Likelihood Estimator

Probability density estimators for each input variable (marginal PDF) combined in overall likelihood estimator, much liked in HEP.

likelihood ratio for event *i*

$$y(\vec{x}_i) = \frac{\prod_{k \in \{\text{variables}\}} f^k_{\text{Signal}}(x^k_i)}{\sum_{U \in \{\text{Signal,Background}\}} \left(\prod_{k \in \{\text{variables}\}} f^k_U(x^k_i) \right)}$$

PDE for each variable *k*

Normalize with *S+B*

Naïve assumption about independence of all input variables

- Optimal approach if correlations are zero (or linear → decorrelation)
- Otherwise: significant performance loss

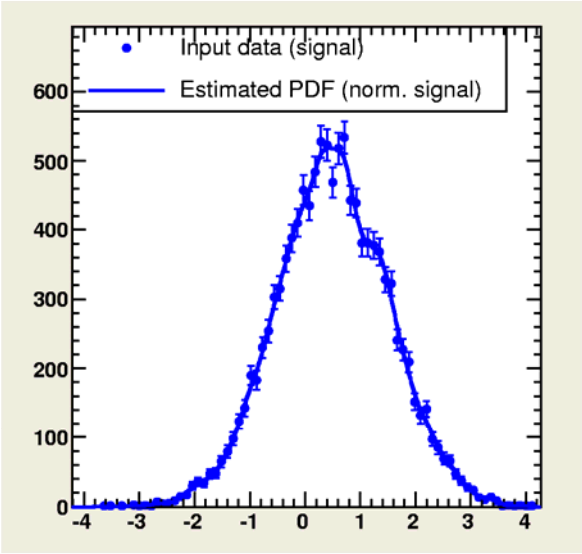
Advantages:

- independently estimating the parameter distribution alleviates the problems from the "*curse of dimensionality*"
- Simple and robust, especially in low dimensional problems

Estimating the input PDFs from the sample

Technical challenge, three ways:

- Parametric fitting: best
  - but variable distribution function must be known. Cannot be generalized to a-priori unknown problems.
  - Use analysis package RooFit.
- Non-parametric fitting: ideal for machine learning
  - Easy to automate
  - Can create artifacts (edge effects, outliers) or hide information (smoothing)
  - Might need tuning.
- Event counting: unbiased PDF (histogram)
  - Automatic
  - Sub-optimal since it exhibits details of the training sample



Nonparametric fitting

Binned (uses histograms)

- shape interpolation using spline functions or adaptive smoothing

Unbinned (uses all data)

- adaptive kernel density estimation (KDE) with Gaussian smearing

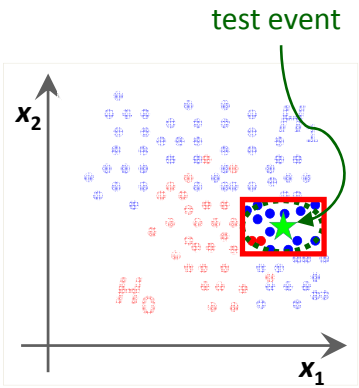
Validation of goodness-of-fit afterwards

Multidimensional PDEs

Incorporates variable correlations, suffers in higher dimensions from lack of statistics!

PDE Range-Search

Count number of reference events (signal and background) in a rectangular volume around the test event



k-Nearest Neighbor

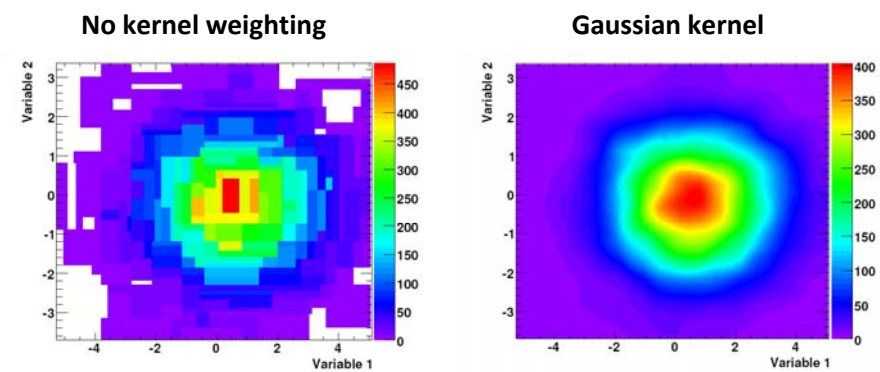
Better: count adjacent reference events till statistically significant number reached (method intrinsically adaptive)

PDE-Foam

Parcel input space into cells of varying sizes, each cell contains representative information (the average reference for the neighborhood)

Advantage: limited number of cells, independent of number of training events

- Fast search: binary search tree that sorts objects in space by their coordinates
- Evaluation can use kernels to determine response



Curse of Dimensionality

Problems caused by the exponential increase in volume associated with adding extra dimensions to a mathematical space:

Volume in hyper-sphere becomes negligible compared to hyper-cube

All the volume is in the corners

$$\lim_{D\to\infty} \frac{V_{\text{sphere}}}{V_{\text{cube}}} = \lim_{D\to\infty} \frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)} = 0$$

Distance functions losing their usefulness in high dimensionality.

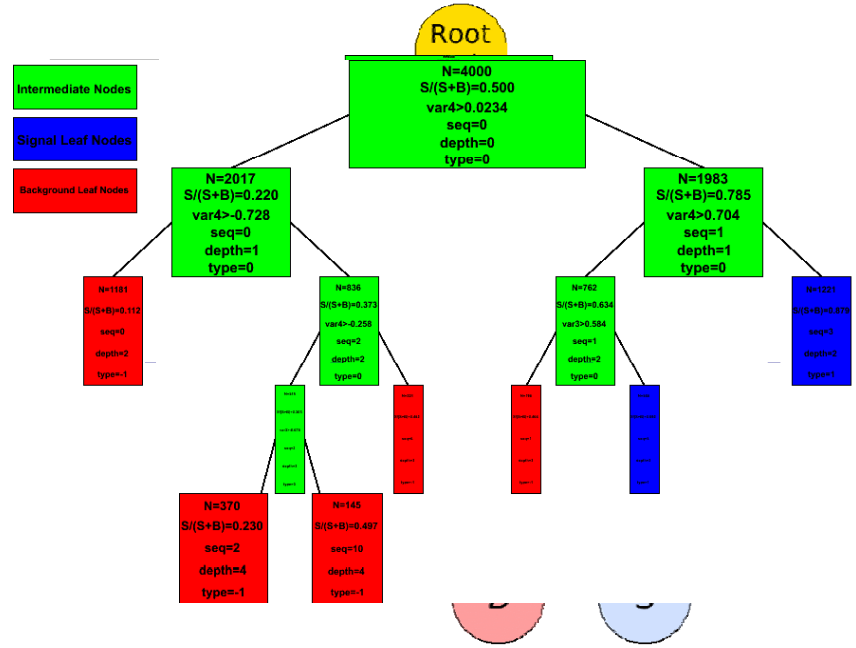
$$\lim_{D\to\infty} \frac{d_{\max} - d_{\min}}{d_{\min}} = 0$$

⇒ Finding local densities in a many-dimensional problem requires a lot of data. Nearest neighbor methods might not work well.

Especially if non-significant variables are included.

⇒ In many dimensions it is better to find the separation borders not by using the likelihood ratio.

Boosted Decision Tree



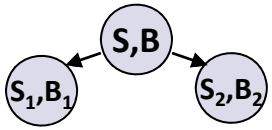
DecisionTree (DT)

Series of cuts that split sample set into ever smaller subsets

Growing

Each split try to maximizing gain in separation ΔG

$$\Delta G = NG - N_1 G_1 - N_2 G_2$$



Gini- or inequality index:

$$G_{\text{node}} = \frac{S_{\text{node}} B_{\text{node}}}{(S_{\text{node}} + B_{\text{node}})^2}$$

Leafs are assigned either *S* or *B*

Event classification

Following the splits using test event variables until a leaf is reached: *S* or *B*

Pruning

Removing statistically insignificant nodes

- Bottom-up
- Protect from overtraining

DT dimensionally robust and easy to understand but alone not powerful !

2) Boosting method *Adaboost*

Build forest of DTs:

1. Emphasizing classification errors in DT_k: increase (boost) weight of incorrectly classified events
2. Train new tree DT_{k+1}

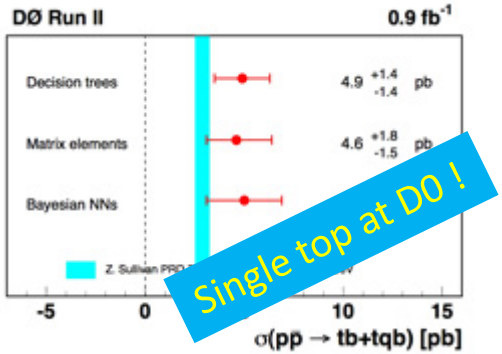
Final classifier linearly combines all trees

- DT with small misclassification get large coefficient

Good performance and stability, little tuning needed.

Popular in HEP (Miniboone, single top at D0)

Multivariate summary

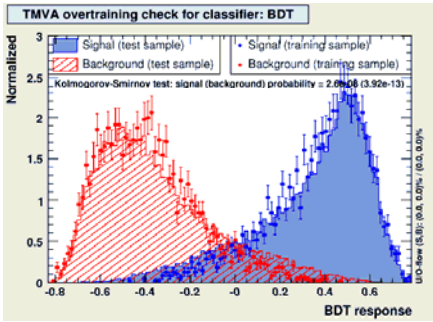


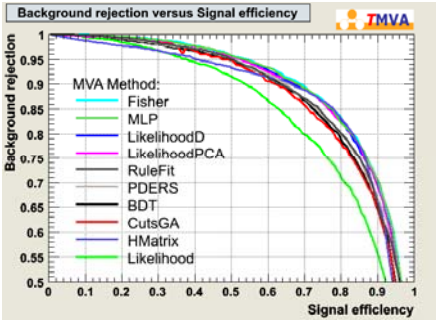
Multivariate analysis packages:

- StatPatternRecognition: I.Narsky, arXiv: physics/0507143
  ⇒ http://www.hep.caltech.edu/~narsky/spr.html
- TMVA: Hoecker, Speckmayer, Stelzer, Therhaag, von Toerne, Voss, arXiv: physics/0703039
  ⇒ http://tmva.sf.net or every ROOT distribution
- WEKA: ⇒ http://www.cs.waikato.ac.nz/ml/weka/

Huge data analysis library available in "R": ⇒ http://www.r-project.org/

Support training, evaluation, comparison of many state-of-the-art classifiers





How to proceed: chose most suitable method, then:

Use MVA output distribution, fit to estimate number of signal and background events.

or

Chose working point for enhance signal selection. Use an independent variable to estimate parameters of underlying physics of signal process.

Parameter estimation ⇨

Estimation of variable properties

<u>Estimator:</u>

A procedure applicable to a data sample *S* which gives the numerical value for a …

a) property of the parent population from which *S* was selected

b) property or parameter from the parent distribution function that generated *S*

Estimators are denoted with a hat ^ over the parameter or property

Estimators are judged by their properties. A good estimator is

consistent $\lim_{N \to \infty} \hat{a} = a$

unbiased $\langle \hat{a} \rangle = a$

For large N any consistent estimator becomes unbiased!

Efficient $V(\hat{a})$ is small

More efficient estimators a more likely to be close to true value. There is a theoretical limit of the variance, the minimum variance bound, MVB. The efficiency of an estimator is MVB$/V(\hat{a})$.

A mean estimator example

Estimators for the mean of a distribution

1) Sum up all x and divide by N
2) Sum up all x and divide by N-1
3) Sum up every second x and divide by int(N/2)
4) Throw away the data and return 42

	Consistent	Unbiased	Efficient
1	√	√	√
2	√	×	√
3	√	√	×
4	×	×	×

Law of large numbers

1)

$$\hat{\mu} \equiv \frac{x_1 + x_2 + \ldots + x_N}{N} = \bar{x} \rightarrow \langle x \rangle = \mu$$

$$\langle \hat{\mu} \rangle \equiv \left\langle \frac{x_1 + x_2 + \cdots + x_N}{N} \right\rangle = \frac{\langle x \rangle + \langle x \rangle + \cdots + \langle x \rangle}{N} = \mu$$

2)

$$\hat{\mu} \equiv \frac{x_1 + x_2 + \ldots + x_N}{N-1} = \frac{N}{N-1}\bar{x} \rightarrow \langle x \rangle = \mu$$

$$\langle \hat{\mu} \rangle \equiv \left\langle \frac{x_1 + x_2 + \cdots + x_N}{N-1} \right\rangle = \frac{N}{N-1}\mu \neq \mu$$

3) is less efficient than 1) since it uses only half the data. Efficiency depends on data sample S.

Note that some estimators are always consistent or unbiased. Most often the properties of the estimator depend on the data sample.

Examples of basic estimators

Estimating the mean: $\hat{\mu} = \bar{x}$

Consistent, unbiased, maybe efficient: $V(\hat{\mu}) = \frac{\sigma^2}{N}$ (from central limit theorem)

Estimating the variance, …

a) when knowing the true mean μ: $\widehat{V(x)} = \frac{1}{N}\sum(x_i - \mu)^2$

This is usually not the case!

b) when not knowing the true mean: $\widehat{V(x)} = s^2 = \frac{1}{N-1}\sum(x_i - \bar{x})^2$

Note the correction factor of $N/(N-1)$ from the naïve expectation. Since $\bar{x}$ is closer to the average of the data sample S than the mean μ, the result would underestimate the variance and introduce a bias!

A more general estimator for a parameter a and a data sample $\{x_1, x_2, \ldots, x_N\}$ is based on the likelihood function

$$L(x_1, x_2, \ldots, x_N; a) = \prod P(x_i; a)$$

Maximum likehood estimator

Variable *x* distributed according to pdf *P* which depends on *a*: $P(x;a)$

Sample *S* of data drawn from according to *P*: $S = \{x_1, x_2, \ldots, x_N\}$

Probability of *S* being drawn: Likelihood $L(x_1, x_2, \ldots, x_N; a) = \prod_{i=1}^{N} P(x_i; a)$

For different $a_1, a_2, \ldots$ we find different likelihoods $L(S;a_1), L(S;a_2), \ldots$

ML principle: a good estimator $\hat{a}(S;a)$ of *a* for sample *S* is the one with the highest likehood for *S* being drawn:

$$\left.\frac{\mathrm{d}\ln L(S;a)}{\mathrm{d}\,a}\right|_{a=\hat{a}} = 0$$

In practice use ln *L* instead of *L* $\Rightarrow$ easier

This is called the Maximum likelihood (ML)-estimator

Properties of the ML estimator

Usually consistent

Invariant under parameter transformation: $\widehat{f(a)} = f(\hat{a})$

Peak in likelihood function: $\left.\frac{\mathrm{d}\ln L}{\mathrm{d}a}\right|_{a=\hat{a}} = \left.\frac{\mathrm{d}\ln L}{\mathrm{d}f}\right|_{f=\hat{f}=f(\hat{a})} \left.\frac{\mathrm{d}f}{\mathrm{d}a}\right|_{a=\hat{a}} = 0$

Price to pay: ML estimators are generally biased !

Invariance between two estimators is incompatible with both being unbiased !

Not a problem when sample size N is large! Remember, consistent estimators become unbiased for large N.

At large N an ML estimator becomes efficient !

Error on an ML estimator for large N

Expand $\ln L$ around its maximum $\hat{a}$. We have seen $\left.\frac{\mathrm{d}\ln L(x_1,\ldots,x_N;a)}{\mathrm{d}\,a}\right|_{a=\hat{a}} = 0$

Second derivative important to estimate error: $\frac{\mathrm{d}^2 \ln L}{\mathrm{d}\,a^2}$

One can show for any unbiased and efficient ML estimator (e.g. large N)

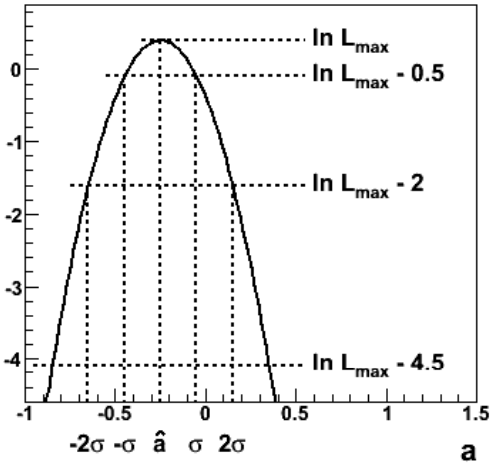
$$\frac{\mathrm{d}\ln L(x_1,\ldots,x_N;a)}{\mathrm{d}\,a} = A(a)\left(\hat{a}(x_1,\ldots,x_N) - a\right)$$, with proportionality factor $A(a) = -\left\langle \frac{\mathrm{d}^2 \ln L}{\mathrm{d}\,a^2} \right\rangle$

The CLT tells us that the probability distribution of $\hat{a}$ is Gaussian. For this to be (close to be) true *A* must be (relatively) constant around $a = \hat{a}$

$$\Rightarrow \quad L(x_1, x_2, \ldots, x_N; a) \propto e^{\frac{A[a-\hat{a}(x_1,x_2,\ldots,x_N)]^2}{2}}$$

For large N the likelihood function becomes Gaussian, the log-likelihood a parabola

The errors in your estimation you can read directly of the ln *L* plot.



About ML

Not necessarily best classifier, but usually good to use. You need to assume the underlying probability density $P(x;a)$

Does not give you the most likely value for a, it gives the value for which the observed data is the most likely !

$$\left.\frac{\mathrm{d}\ln L(S;a)}{\mathrm{d}a}\right|_{a=\hat{a}} = 0$$

Usually can't solve analytically, use numerical methods, such as MINUIT. You need to program you $P(x;a)$

Below the large N regime, LH not a Gaussian (log-LH not a parabola)

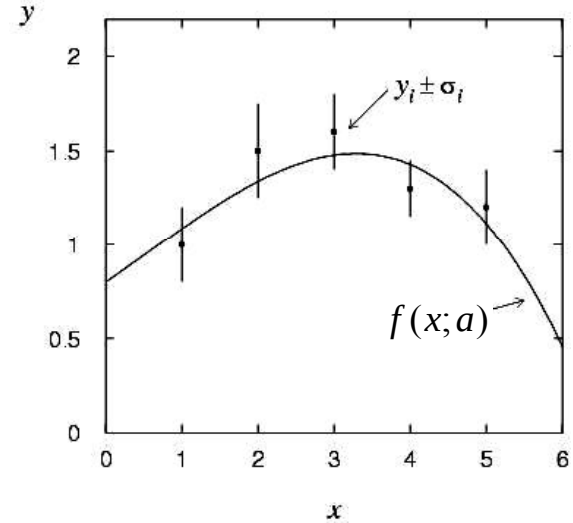
- MC simulation: generate U experiments, each with N events. Find and plot MLE. Use graphical solution: plot ln$L$ and find the points where it dropped by 0.5, 2, 4.5 to find $\pm\sigma$, $\pm 2\sigma$, $\pm 3\sigma$
- Perhaps use transformation invariance to find a estimator with Gaussian distribution
- Quote asymmetric errors on your estimate

No quality check: the value of $\ln L(S;\hat{a})$ will tell you nothing about how good your $P(x;a)$ assumption was

Least square estimation

Particular MLE with Gaussian distribution, each of the sample points y_i has its own expectation $f(x_i;a)$ and resolution σ_i

$$P(y_i;a) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-[y_i - f(x_i;a)]^2/2\sigma_i^2}$$



To maximize LH, minimize $\chi^2 = \sum_i \left(\frac{y_i - f(x_i;a)}{\sigma_i} \right)^2$

Fitting binned data:

Proper χ^2:

$$\chi^2 = \sum_j \frac{(n_j - f_j)^2}{f_j}$$

Simple χ^2:
(simpler to calculate)

$$\chi^2 = \sum_j \frac{(n_j - f_j)^2}{n_j}$$

n_j content of bin i follows poisson statistics

f_j expectation for bin i, also the squared error

Advantages of least squares

Method provides goodness-of-fit

The value of the χ^2 at its minimum is a measure of the level of agreement between the data and fitted curve.

χ^2 statistics follows the chi-square distribution $f(\chi^2; n)$

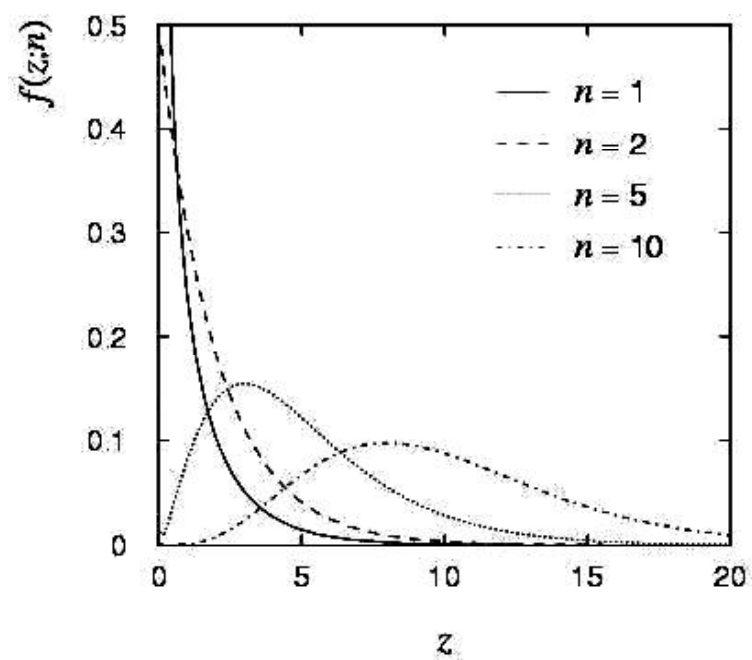
Each data point contributes $\chi^2 \approx 1$, minimizing χ^2 makes it smaller by ≈1 per free variable

Number of degrees of freedom $n = N_{\text{bin}} - N_{\text{var}}$

χ^2 has mean *n* and variance 2*n*

If χ^2/n much larger than 1 something might be wrong

n should be large for this test. Better to use $\sqrt{2\chi^2}$ which has mean $\sqrt{2n-1}$ and variance 1, and becomes Gaussian at n~30.



Error Analysis

Statistical errors:

How much would result fluctuate upon repetition of the experiment

Also need to estimate the systematic errors: uncertainties in our assumptions

Uncertainty in the theory (model)

Understanding of the detector in reconstruction (calibration constants)

Simulation: wrong simulation of detector response (material description)

Error from finite MC sample (MC statistical uncertainty)

$\Rightarrow$ requires some of thinking and is not as well defined as the statistical erro

Literature

Statistical Data Analysis
G. Cowan, Clarendon, Oxford, 1998, see also www.pp.rhul.ac.uk/~cowan/sda

Statistics, A Guide to the Use of Statistical in the Physical Sciences
R.J. Barlow, Wiley, 1989 see also hepwww.ph.man.ac.uk/~roger/book.html

Statistics for Nuclear and Particle Physics
L. Lyons, CUP, 1986

Statistical and Computational Methods in Experimental Physics, 2nd ed.
F. James., World Scientific, 2006

Statistical and Computational Methods in Data Analysis
S. Brandt, Springer, New York, 1998 (with program library on CD)

Review of Particle Physics (Particle Data Group)
Journal of Physics G 33 (2006) 1; see also pdg.lbl.gov sections on probability statistics, Monte Carlo

TMVA - Toolkit for Multivariate Data Analysis
A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss,
PoS ACAT 040 (2007), arXiv:physics/0703039