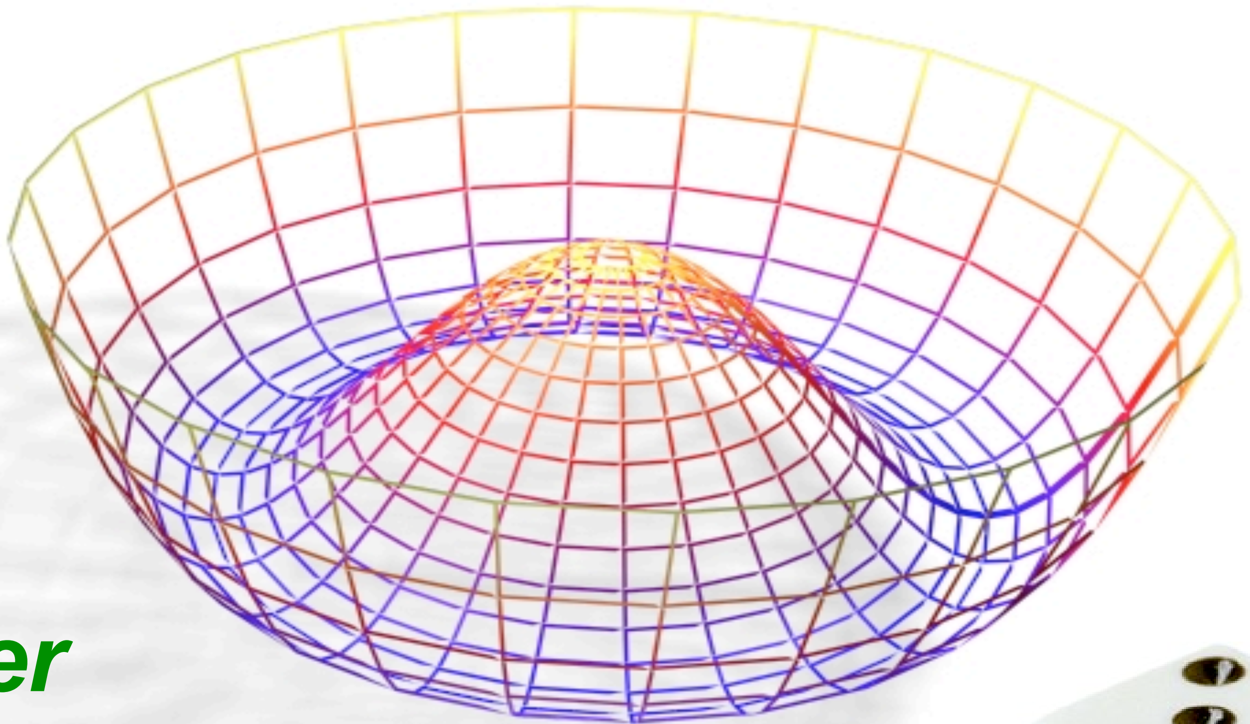
Statistics for Particle Physics



Kyle Cranmer

New York University

Introduction

Statistics plays a vital role in science, it is the way that we:

- quantify our knowledge and uncertainty
- communicate results of experiments

Big questions:

- testing theories, measure or exclude parameters, etc.
- how do we make decisions
- how do we get the most out of our data
- how do we incorporate uncertainties

Statistics is a very big field, and it is not possible to cover everything in 4 hours. In these talks I will try to:

- **explain** some fundamental ideas & prove a few things
- **enrich** what you already know
- **expose** you to some new ideas

I will try to go slowly, because if you are not following the logic, then it is not very interesting.

- Please feel free to ask questions and interrupt at any time

Further Reading

By physicists, for physicists

G. Cowan, Statistical Data Analysis, Clarendon Press, Oxford, 1998.

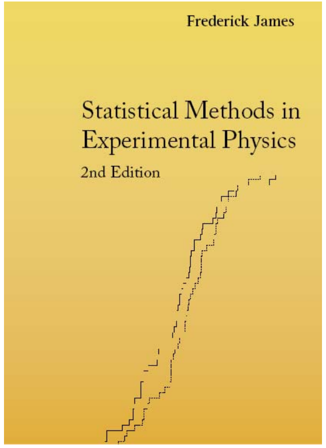
R.J.Barlow, A Guide to the Use of Statistical Methods in the Physical Sciences, John Wiley, 1989;

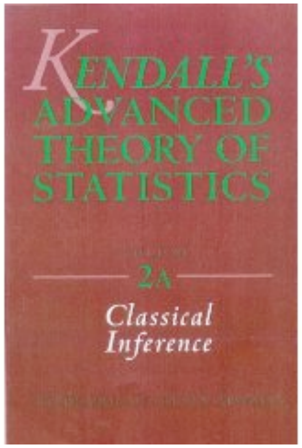
F. James, Statistical Methods in Experimental Physics, 2nd ed., World Scientific, 2006;

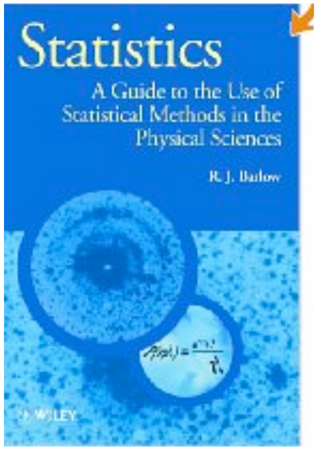
- W.T. Eadie et al., North-Holland, 1971 (1st ed., hard to find);

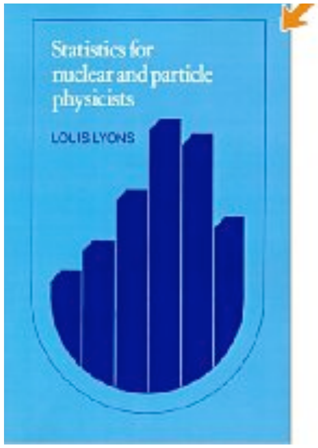
S.Brandt, Statistical and Computational Methods in Data Analysis, Springer, New York, 1998.

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









My favorite statistics book by a statistician:

Stuart, Ord, Arnold. "Kendall's Advanced Theory of Statistics" Vol. 2A *Classical Inference & the Linear Model.*

Other lectures

Fred James's lectures

http://preprints.cern.ch/cgi-bin/setlink?base=AT&categ=Academic_Training&id=AT00000799

http://www.desy.de/~acatrain/

Glen Cowan's lectures

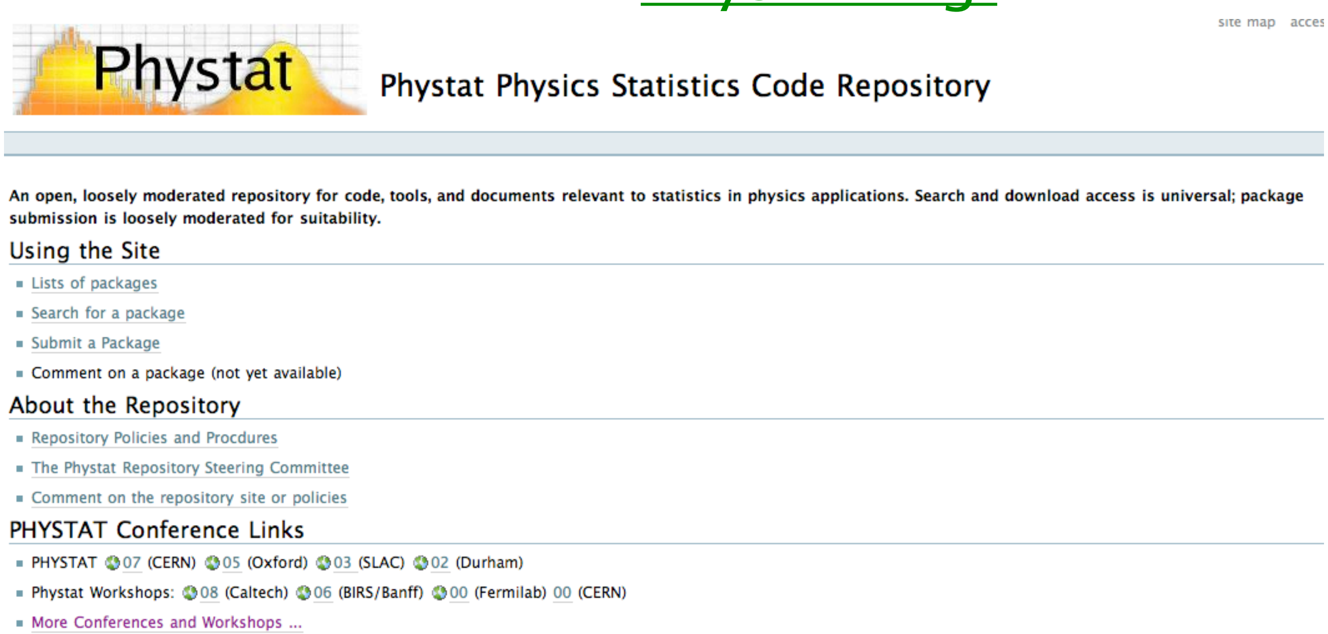
http://www.pp.rhul.ac.uk/~cowan/stat_cern.html

Louis Lyons

http://indico.cern.ch/conferenceDisplay.py?confId=a063350

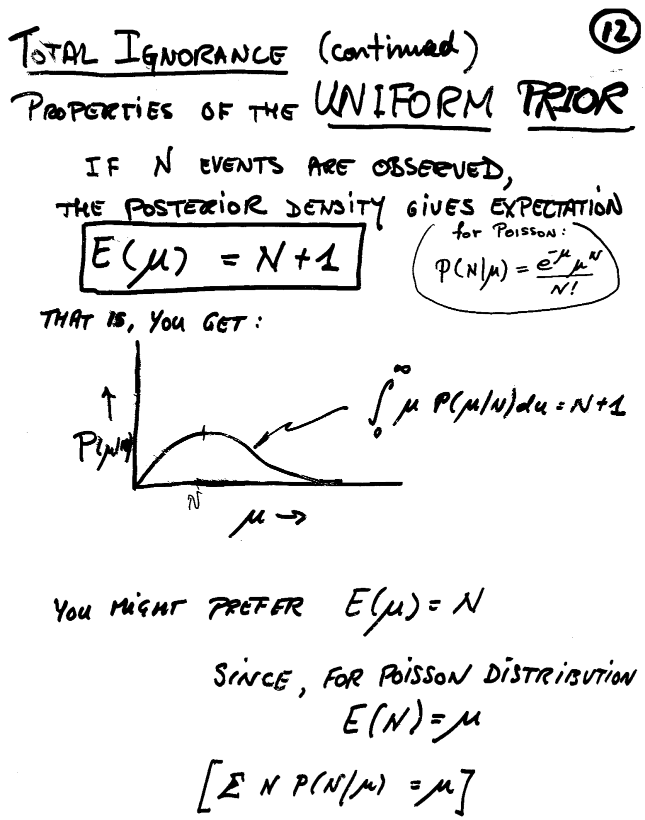
Bob Cousins gave a CMS lecture, may give it more publicly

The PhyStat conference series at PhyStat.org:



Comments on these lectures

Fred James gave a terrific series of lectures. Largely based on principles, focused on comparison of Bayesian & Frequentist



TOTAL IGNORANCE (continued) (12)

PROPERTIES OF THE UNIFORM PRIOR

IF N EVENTS ARE OBSERVED, THE POSTERIOR DENSITY GIVES EXPECTATION

$E(\mu) = N+1$

for Poisson: $P(N|\mu) = \frac{e^{-\mu}\mu^N}{N!}$

THAT IS, YOU GET:

YOU MIGHT PREFER $E(\mu) = N$

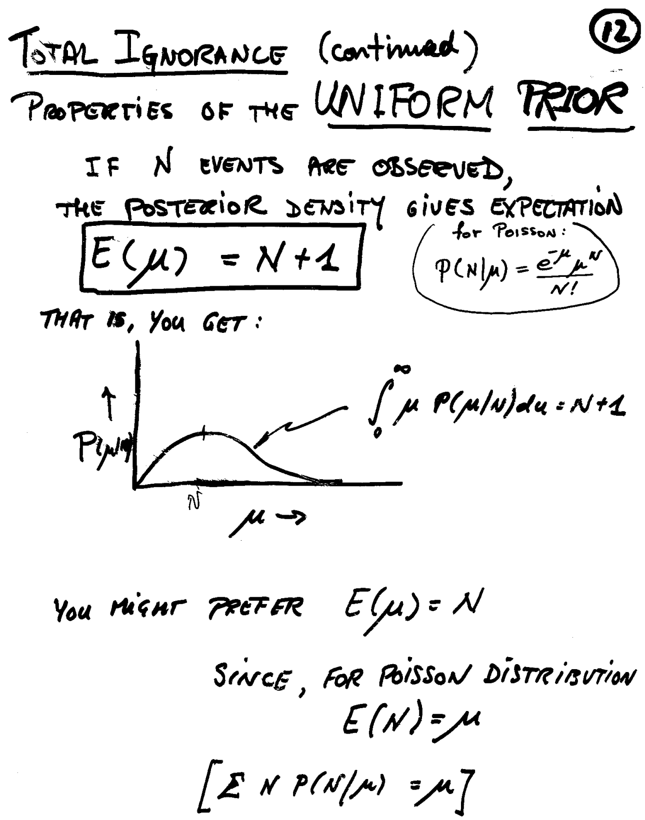
SINCE, FOR POISSON DISTRIBUTION $E(N) = \mu$

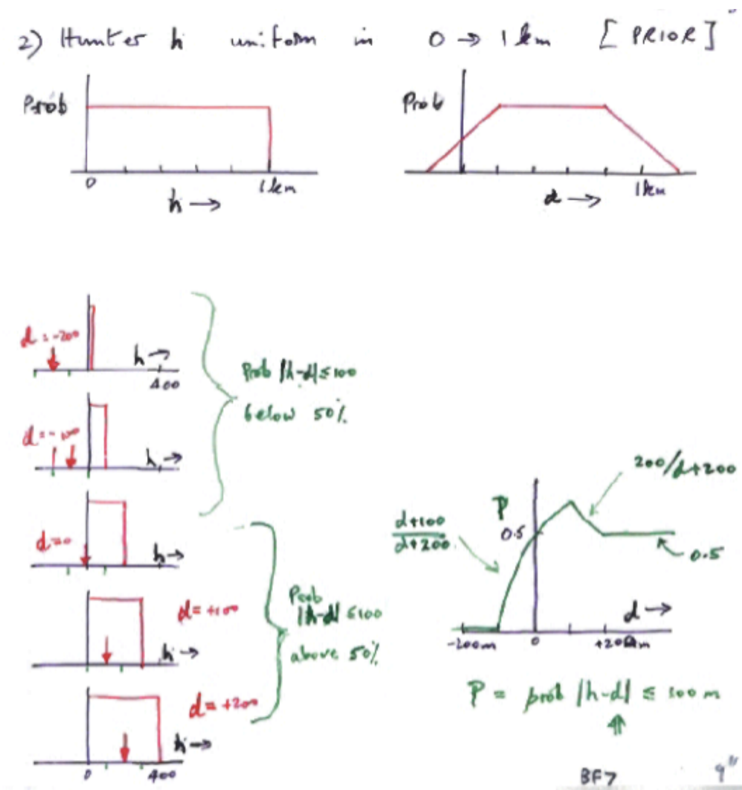
$[\Sigma\, N\, P(N|\mu) = \mu]$

Comments on these lectures

Fred James gave a terrific series of lectures. Largely based on principles, focused on comparison of Bayesian & Frequentist

Louis Lyons, also gave terrific lectures, basic principles plus some useful examples





Comments on these lectures

Fred James gave a terrific series of lectures. Largely based on principles, focused on comparison of Bayesian & Frequentist

Louis Lyons, also gave terrific lectures, basic principles plus some useful examples

Glen Cowan gave lectures for CERN summer school (slightly lower level) & great academic training lectures on multivariate algorithms

Dealing with uncertainty

In particle physics there are various elements of uncertainty:

- theory is not deterministic
  - quantum mechanics

- random measurement errors
  - present even without quantum effects
- things we could know in principle but don't
  - e.g. from limitations of cost, time, ...

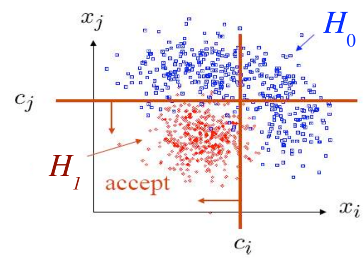
We can quantify the uncertainty using PROBABILITY

Finding an optimal decision boundary

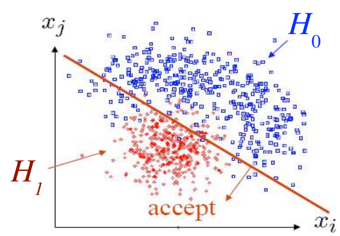
Maybe select events with "cuts":

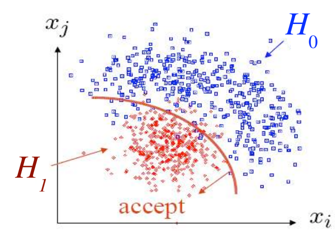
$x_i < c_i$

$x_j < c_j$



Or maybe use some other type of decision boundary:





Goal of multivariate analysis is to do this in an "optimal" way.

Glen Cowan Multivariate Statistical Methods in Particle Physics

Comments on these lectures

Fred James gave a terrific series of lectures. Largely based on principles, focused on comparison of Bayesian & Frequentist

Louis Lyons, also gave terrific lectures, basic principles plus some useful examples

Glen Cowan gave lectures for CERN summer school (slightly lower level) & great academic training lectures on multivariate algorithms

Bob Cousins gave a very comprehensive lecture to CMS on "statistics in theory"

"Statistics in Theory"*:
Prelude to "Statistics in Practice"

Bob Cousins, UCLA
CMS Statistics Tutorial Series
May 8, 2008

***Background for sound work and for avoiding unsound statements.**

Bob Cousins, CMS, 2008 1

Comments on these lectures

So what will be the theme of these lectures?

Definitely not a cook book, I want to convey the fundamental concepts in a fairly general setting.

But I don't want to spend time on special cases or contrived examples.
I want to address the challenges of the LHC.

I also don't want to discuss purely theoretical results if they aren't directly applicable. However, there are many theoretical results that provide an insightful bound.

> In theory, there is no difference between theory and practice; In practice, there is.
> - Chuck Reid

> There is nothing more practical than a good theory.
> - James C. Maxwell

In particular, I'm mainly interested in discovery and measurement, but I will touch on goodness of fit and limits.

I also hope to sprinkle the lectures with advanced topics and expose you to some modern approaches and unsolved problems.

Outline

Lecture 1:

- How we use statistics
- Probability axioms, Bayes vs. Frequentist, from discrete to continuous
- Parametric and non-parametric probability density functions
- Shannon and Fisher Information, correlation, information geometry, Cramér-Rao bound
- A word on subjective and “objective” Bayesian priors

Lecture 2

- Hypothesis testing in the frequentist setting
- The Neyman-Pearson lemma (with a simple proof)
- Decision theory: utility, risk, priors, and game theory
- Contrast hypothesis testing to goodness of fit tests with some warnings
- Related comments on multivariate algorithms
- Matrix element techniques vs. the black box

Outline

Lecture 3:

- The Neyman-Construction (illustrated)
- Inverted hypothesis tests: A dictionary for limits (intervals)
- Coverage as a calibration for our statistical device
- Compound hypotheses, nuisance parameters, & similar tests
- Systematics, Systematics, Systematics

Lecture 4:

- Generalizing our procedures to include systematics
- Eliminating nuisance parameters: profiling and marginalization
- Introduction to ancillary statistics & conditioning
- High dimensional models, Markov Chain Monte Carlo, and Hierarchical Bayes
- The look elsewhere effect and false discovery rate

Lecture 1

How We Use Statistics

Broadly speaking, we use statistical techniques for a few main purposes:

- **Point estimation:** what is the best estimate of a particular parameter
  - eg. measurement of the Z boson mass
- **Confidence Intervals:** regions representing an allowable range of a parameter (in a way to be made precise later)
  - eg. 95% contours, upper-limits, lower-limits
- **Hypothesis Testing:** choosing between two (or more) hypotheses
  - eg. Discover the Higgs, Discover SUSY, reject standard model
- **Goodness-of-fit:** quantify how well the data agrees with a particular model
- **Data reduction:** how to reduce the raw data while loosing minimal information tha is useful for our ultimate goal

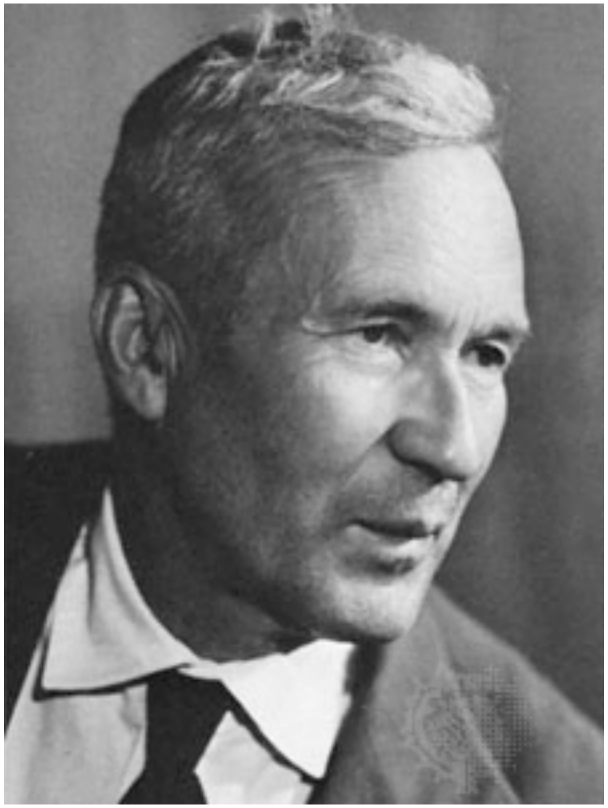
In a broader context, there are related issues:

- **Decision making:** how do we make decisions in the face of uncertainty
- Where does the role of an experimentalist end? How does this impact how we **publish** our results? or how we make decisions?

Axioms of Probability

These Axioms are a mathematical starting point for probability and statistics

1. probability for every element, E, is non-negative $P(E) \geq 0 \quad \forall E \subseteq \mathcal{F} = 2^{\Omega}$
2. probability for the entire space of possibilities is 1 $P(\Omega) = 1.$
3. if elements E_i are disjoint, probability is additive $P(E_1 \cup E_2 \cup \cdots) = \sum_i P(E_i).$



Kolmogorov axioms (1933)

Consequences:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(\Omega \setminus E) = 1 - P(E)$$

Different definitions of Probability

Frequentist

- defined as limit of long term frequency
- probability of rolling a 3 := limit of (# rolls with 3 / # trials)
  - you don't need an infinite sample for definition to be useful
  - sometimes ensemble doesn't exist
    - eg. P(Higgs mass = 120 GeV), P(it will snow tomorrow)
- Intuitive if you are familiar with Monte Carlo methods
- compatible with interpretation of probability in Quantum Mechanics (though some argue this point). Probability to measure spin projected on x-axis if spin of beam is polarized along +z

$$|\langle \rightarrow | \uparrow \rangle|^2 = \frac{1}{2}$$

Subjective Bayesian

- Probability is a degree of belief (personal, subjective)
  - can be made quantitative based on betting odds
  - most people's subjective probabilities are not **coherent** and do not obey laws of probability

http://plato.stanford.edu/archives/sum2003/entries/probability-interpret/#3.1

Bayes' Theorem

Bayes' theorem relates the conditional and marginal probabilities of events A & B

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

- P(A) is the prior probability or marginal probability of A. It is "prior" in the sense that it does not take into account any information about B.
- P($A|B$) is the conditional probability of A, given B. It is also called the posterior probability because it is derived from or depends upon the specified value of B.
- P($B|A$) is the conditional probability of B given A.
- P(B) is the prior or marginal probability of B, and acts as a normalizing constant

Derivation from conditional probabilities

To derive the theorem, we start from the definition of conditional probability. The probability of event A given event B is

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

Equivalently, the probability of event B given event A is

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

Rearranging and combining these two equations, we find

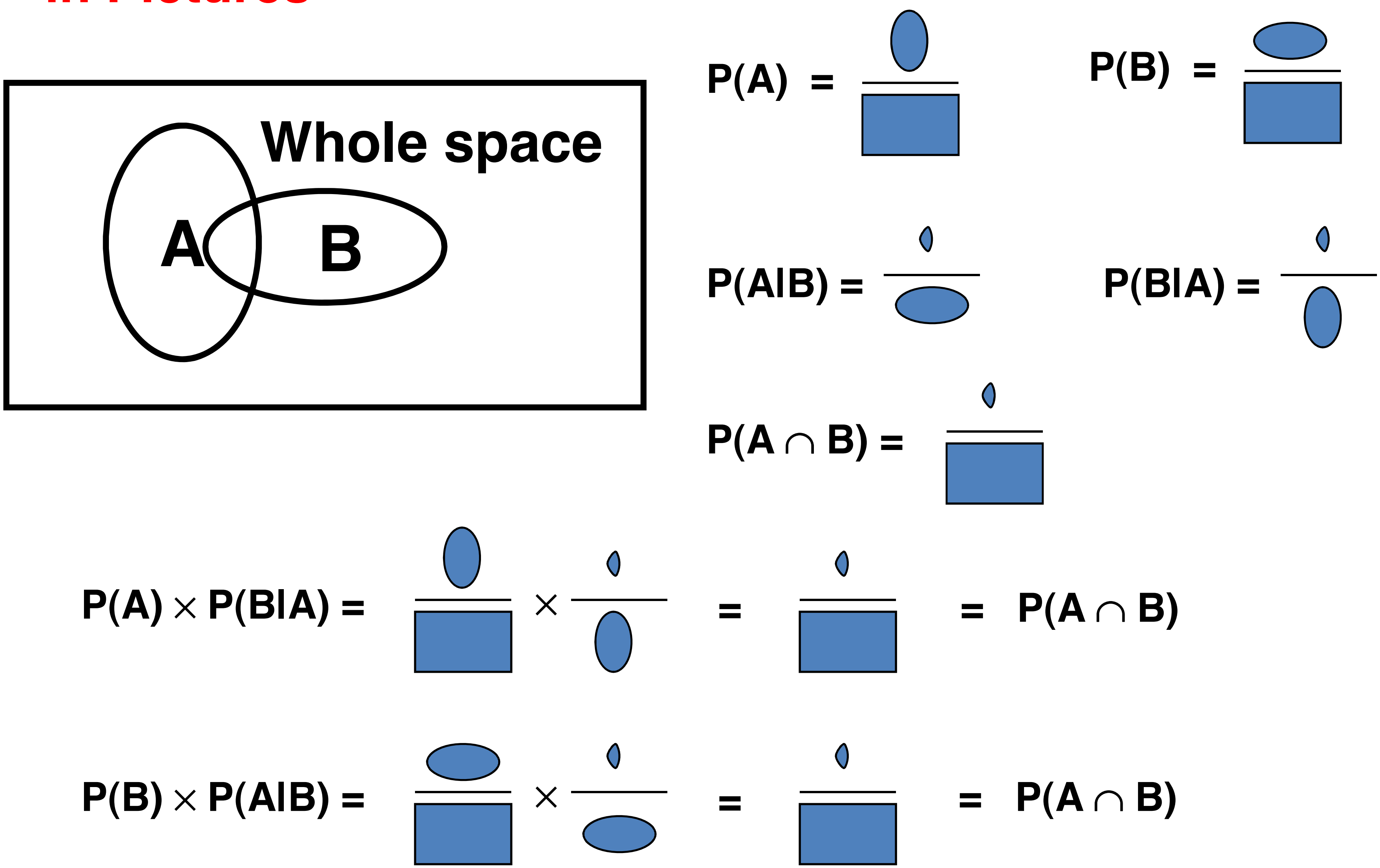
$$P(A|B)\,P(B) = P(A \cap B) = P(B|A)\,P(A).$$

This lemma is sometimes called the product rule for probabilities. Dividing both sides by P(B), providing that it is non-zero, we obtain Bayes' theorem:

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

... in pictures (from Bob Cousins)

P, Conditional P, and Derivation of Bayes' Theorem in Pictures

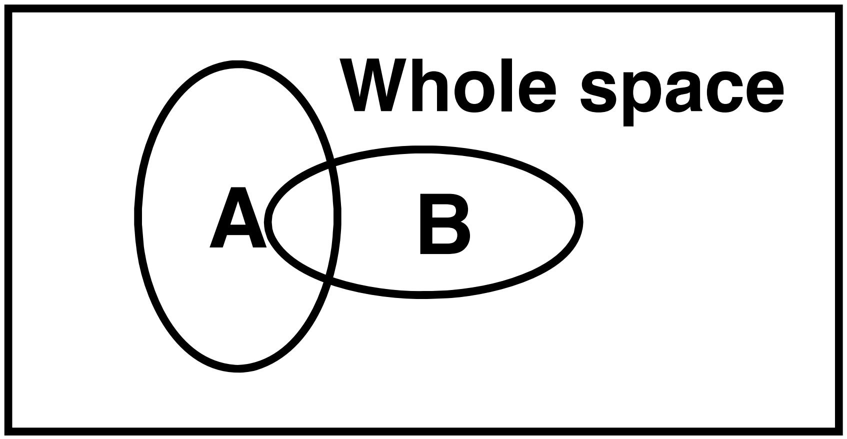


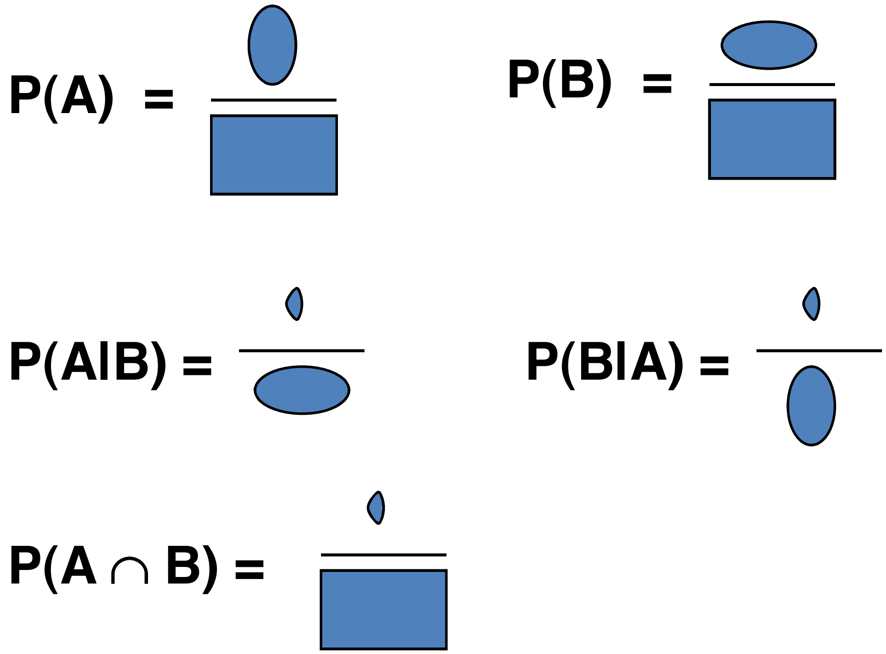
Bob Cousins, CMS, 2008

$\Rightarrow$ **P(B|A) = P(A|B) × P(B) / P(A)**

... in pictures (from Bob Cousins)

P, Conditional P, and Derivation of Bayes' Theorem in Pictures





Don't forget about "Whole space" Ω. I will drop it from the notation typically, but occasionally it is important.

Bob Cousins, CMS, 2008

$\Rightarrow$ **P(B|A) = P(A|B) × P(B) / P(A)**

Louis's Example

P (Data;Theory) $\neq$ P (Theory;Data)

Theory = male or female

Data = pregnant or not pregnant

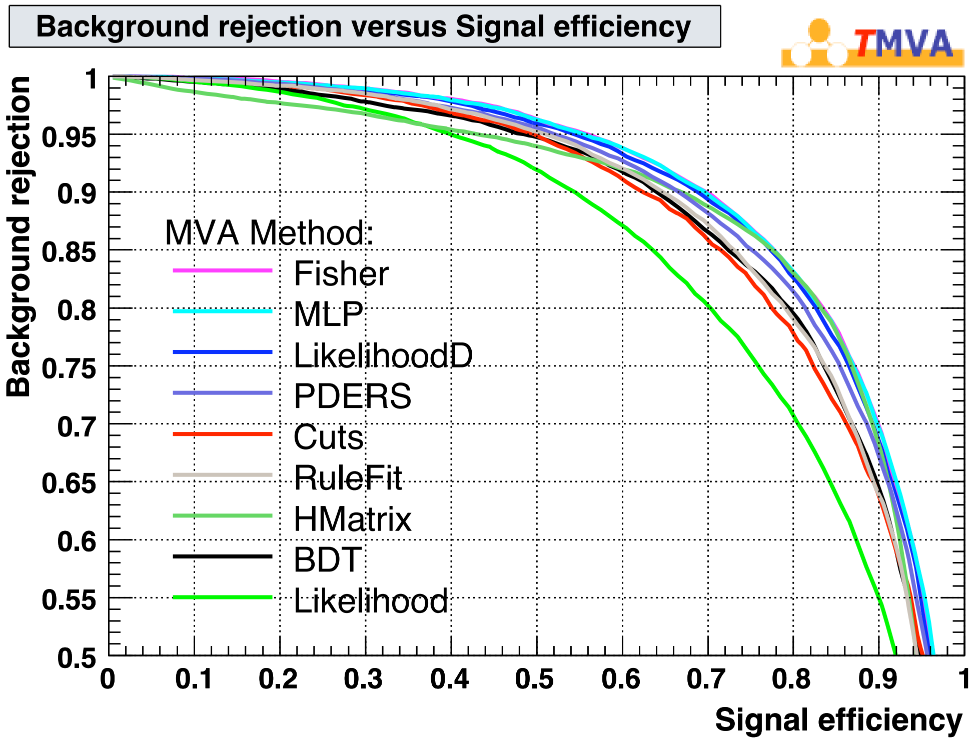
P (pregnant ; female) ~ 3%

but

P (female ; pregnant) >>>3%

Bob's Example

A b-tagging algorithm gives a curve like this



One wants to decide where to cut and to optimize analysis

- For some point on the curve you have:
  - P(btag| b-jet), i.e., efficiency for tagging b's
  - P(btag| not a b-jet), i.e., efficiency for background

Bob's example of Bayes' theorem

Now that you know:

- P(btag| b-jet), i.e., efficiency for tagging b's
- P(btag| not a b-jet), i.e., efficiency for background

Question: Given a selection of jets tagged as b-jets, what fraction of them are b-jets?

- I.e., **what is P(b-jet | btag) ?**

Bob's example of Bayes' theorem

Now that you know:

- P(btag| b-jet), i.e., efficiency for tagging b's
- P(btag| not a b-jet), i.e., efficiency for background

Question: Given a selection of jets tagged as b-jets, what fraction of them are b-jets?

- I.e., **what is P(b-jet | btag) ?**

Answer: Cannot be determined from the given information!

- Need to know **P(b-jet)**: fraction of all jets that are b-jets.
- Then Bayes' Theorem inverts the conditionality:
  - $P(\text{b-jet} \mid \text{btag}) \propto P(\text{btag}|\text{b-jet})\, P(\text{b-jet})$

Bob's example of Bayes' theorem

Now that you know:

- P(btag| b-jet), i.e., efficiency for tagging b's
- P(btag| not a b-jet), i.e., efficiency for background

Question: Given a selection of jets tagged as b-jets, what fraction of them are b-jets?

- I.e., **what is P(b-jet | btag) ?**

Answer: Cannot be determined from the given information!

- Need to know **P(b-jet)**: fraction of all jets that are b-jets.
- Then Bayes' Theorem inverts the conditionality:
  - $P(\text{b-jet} \mid \text{btag}) \propto P(\text{btag}|\text{b-jet})\, P(\text{b-jet})$

Note, this use of Bayes' theorem is fine for Frequentist

Bayesian vs. Frequentist

In short, Frequentist are always restricted to statements related to

- P(Data | Theory) (deductive reasoning)
- the data is considered random
- each point in the "Theory" space is treated independently
  - (no notion of probability in the "Theory" space)

Bayesian vs. Frequentist

In short, Frequentist are always restricted to statements related to

- P(Data | Theory) (deductive reasoning)
- the data is considered random
- each point in the "Theory" space is treated independently
  - (no notion of probability in the "Theory" space)

Bayesians can address questions of the form:

- P(Theory | Data) $\propto$ P(Data | Theory) P(Theory)
  - intuitively what we want to know (inductive reasoning)
- but it requires a prior on the Theory
  - [short discussion subjective vs. empirical Bayes goes here]

Bayesian vs. Frequentist

In short, Frequentist are always restricted to statements related to

- P(Data | Theory) (deductive reasoning)
- the data is considered random
- each point in the "Theory" space is treated independently
  - (no notion of probability in the "Theory" space)

Bayesians can address questions of the form:

- $P(\text{Theory} \mid \text{Data}) \propto P(\text{Data} \mid \text{Theory})\, P(\text{Theory})$
  - intuitively what we want to know (inductive reasoning)
- but it requires a prior on the Theory
  - [short discussion subjective vs. empirical Bayes goes here]

Later I will discuss the "Likelihood Principle" and Likelihood-based analysis: it's a third approach to statistical inference

An different example of Bayes' theorem

An analysis is developed to search for the Higgs boson

- background expectation is 0.1 events
  - you know P(N | no Higgs)
- signal expectation is 10 events
  - you know P(N | Higgs)

An different example of Bayes' theorem

An analysis is developed to search for the Higgs boson

- background expectation is 0.1 events
  - you know P(N | no Higgs)
- signal expectation is 10 events
  - you know P(N | Higgs)

Question: one observes 8 events, **what is P(Higgs | N=8) ?**

An different example of Bayes' theorem

An analysis is developed to search for the Higgs boson

- background expectation is 0.1 events
  - you know P(N | no Higgs)
- signal expectation is 10 events
  - you know P(N | Higgs)

Question: one observes 8 events, **what is P(Higgs | N=8) ?**

Answer: Cannot be determined from the given information!

- Need in addition: P(Higgs)
  - no ensemble! no frequentist notion of P(Higgs)
  - prior based on degree-of-belief would work, but it is subjective. This is why some people object to Bayesian statistics for particle physics

A Joke

“Bayesians address the question everyone is interested in, by using assumptions no-one believes”

“Frequentists use impeccable logic to deal with an issue of no interest to anyone”

- P. G. Hamer

Some personal history

Archbishop of Canterbury Thomas Cranmer (born: 1489, executed: 1556) author of the “Book of Common Prayer”

Two centuries later (when this Book had become an official prayer book of the Church of England) Thomas Bayes was a non-conformist minister (Presbyterian) who refused to use Cranmer’s book

a little on Information Theory

Information Theory

How much information in this message?

$$\underbrace{1000110101001011}_{\text{16 entries}}$$

Information Theory

How much information in this message?

$$\underbrace{1000110101001011}_{\text{16 entries}}$$

What about this one

$$\underbrace{0101010101010101}_{\text{16 entries}}$$

Information Theory

How much information in this message?

$$\underbrace{1000110101001011}_{\text{16 entries}}$$

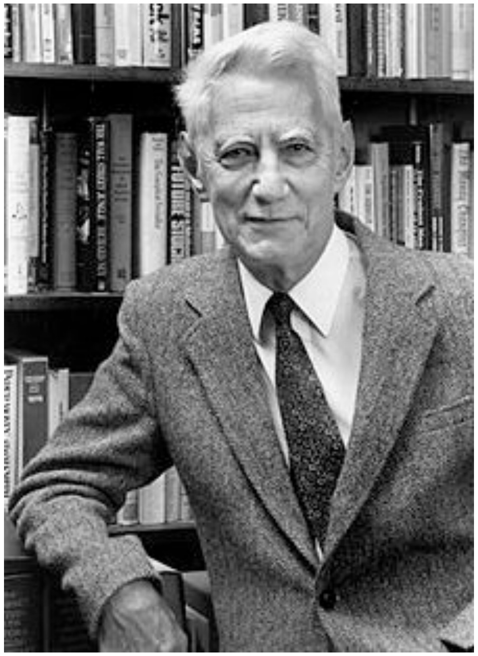
What about this one

$$\underbrace{0101010101010101}_{\text{16 entries}}$$

... and this one?

$$\underbrace{abcdabcdabcdabcd}_{\text{16 entries}}$$

Information Theory



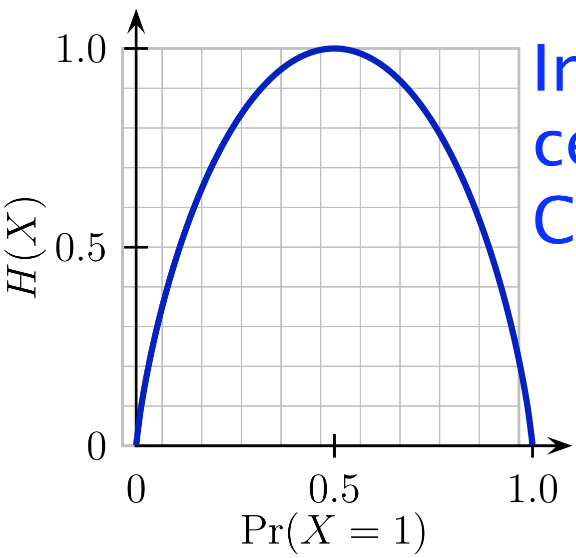
How much information in this message?

$$\underbrace{1000110101001011}_{16\ \text{entries}}$$

- 16 bits? (**bit** is unit when log is base 2)
- it depends on probabilities of 0,1

In 1870's Boltzman and Gibbs defined entropy:

$$S = -k_B \sum_i p_i \ln p_i$$



In 1948, Calude Shannon uses entropy as a centerpiece of his "Mathematical Theory of Communication" eg. information theory

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

- information maximized when p_i all equal

Probability Density Functions

When dealing with continuous random variables, need to introduce the notion of a **Probability Density Function** (PDF... not parton distribution function)

$$P(x \in [x, x+dx]) = f(x)dx$$

Note, $f(x)$ is NOT a probability

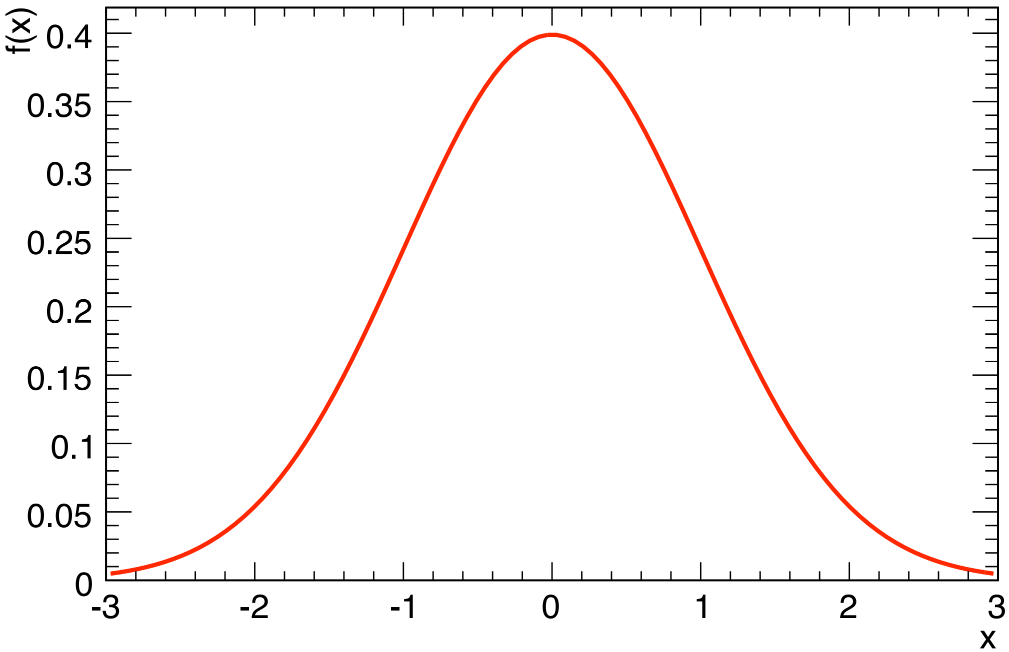
Equivalent of second axiom...

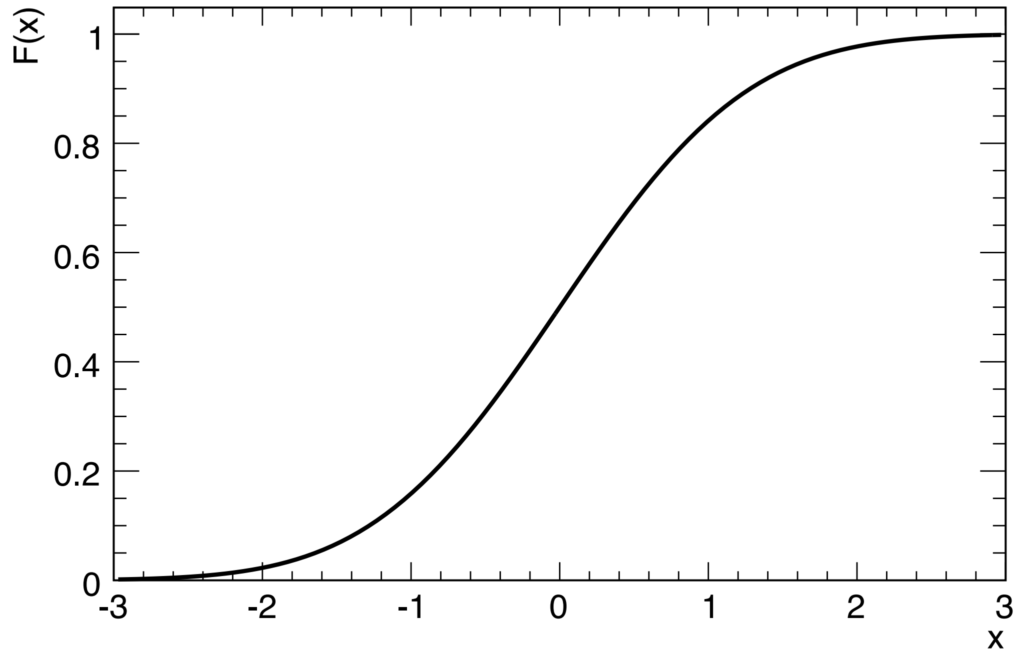
$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Cumulative Density Functions

Often useful to use a cumulative distribution:

- in 1-dimension: $$\int_{-\infty}^{x} f(x')dx' = F(x)$$

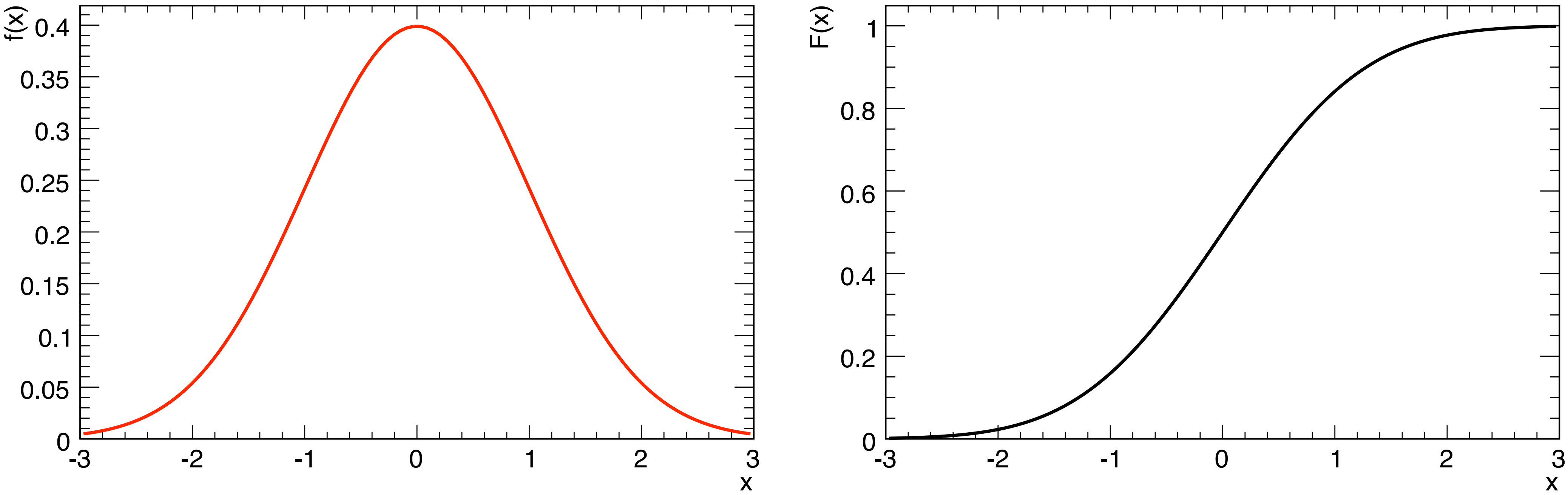




Cumulative Density Functions

Often useful to use a cumulative distribution:

- in 1-dimension: $$\int_{-\infty}^{x} f(x')dx' = F(x)$$



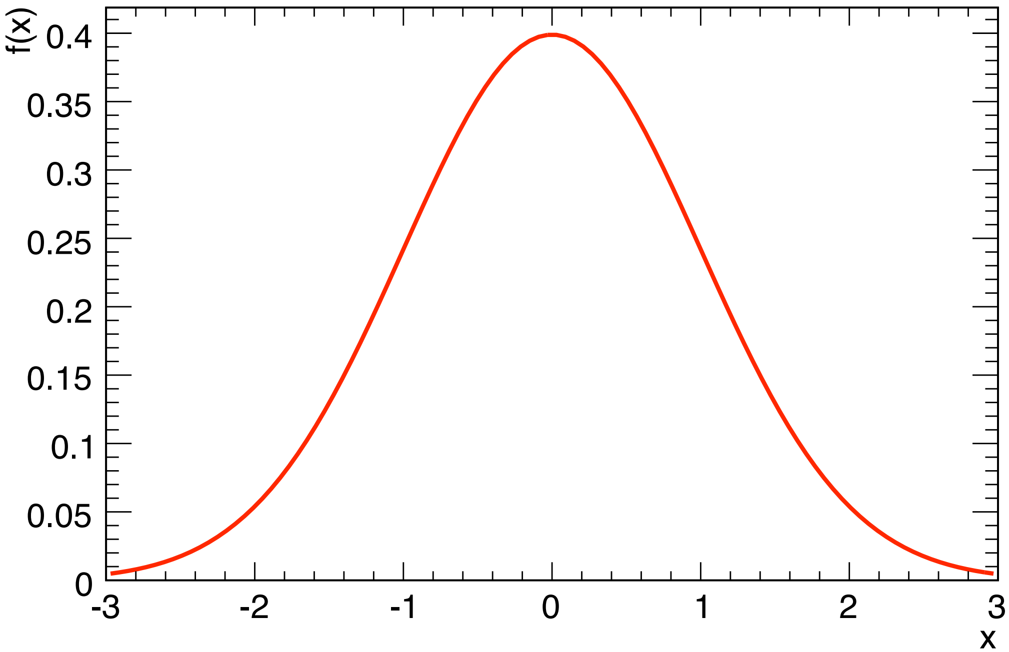
- alternatively, define density as partial of cumulative:

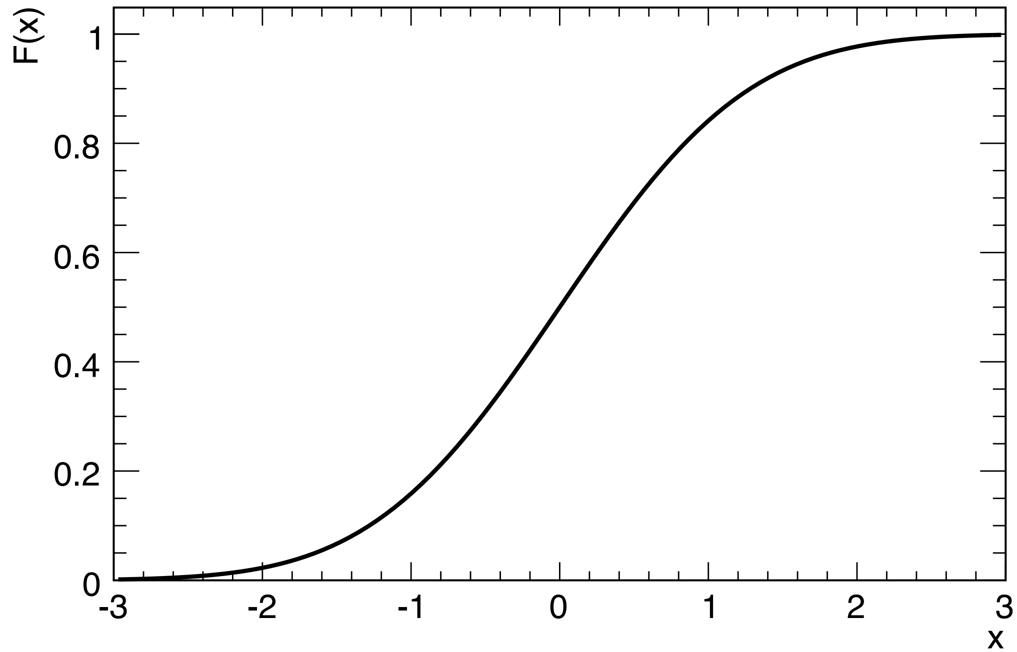
$$f(x) = \frac{\partial F(x)}{\partial x}$$

Cumulative Density Functions

Often useful to use a cumulative distribution:

- in 1-dimension: $$\int_{-\infty}^{x} f(x')dx' = F(x)$$





- alternatively, define density as partial of cumulative:

$$f(x) = \frac{\partial F(x)}{\partial x}$$

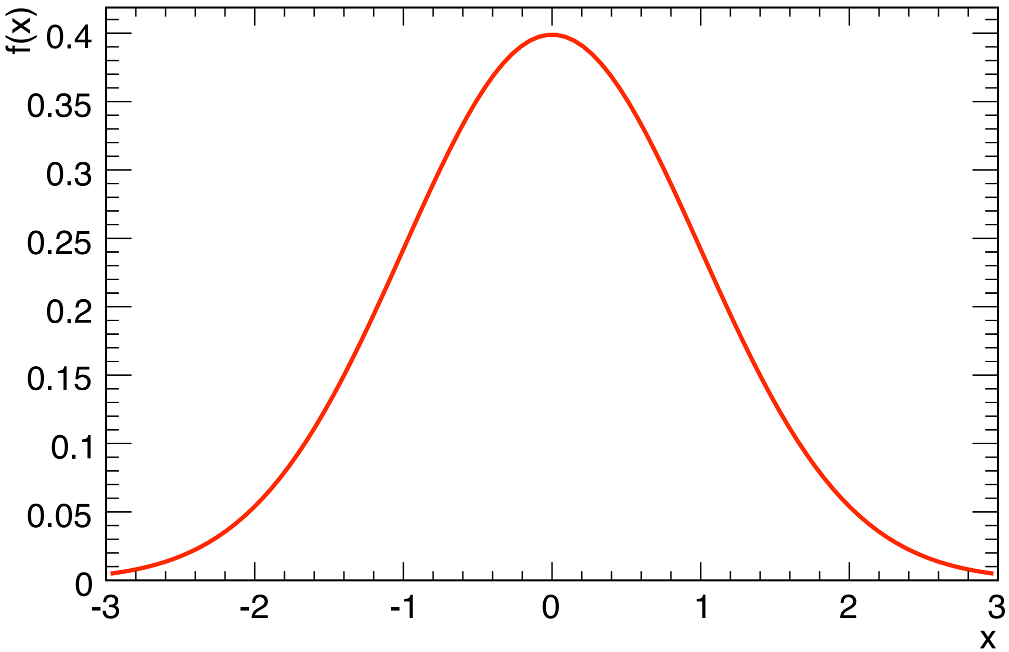
- similar to relationship of total and differential cross section:

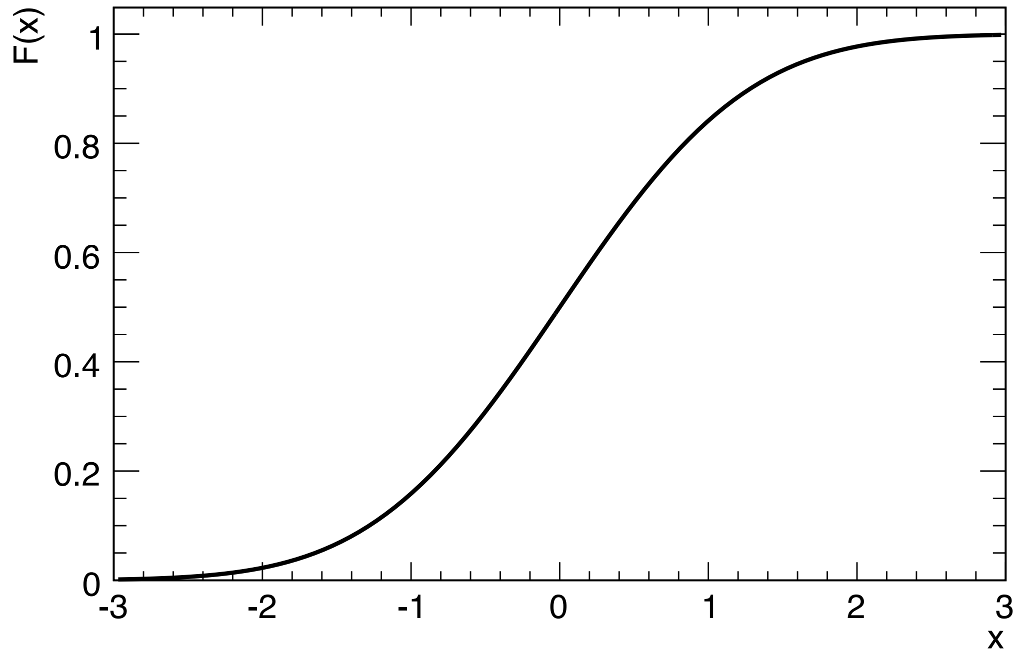
$$f(E) = \frac{1}{\sigma}\frac{\partial \sigma}{\partial E}$$

Cumulative Density Functions

Often useful to use a cumulative distribution:

- in 1-dimension: $$\int_{-\infty}^{x} f(x')dx' = F(x)$$





- alternatively, define density as partial of cumulative:

$$f(x) = \frac{\partial F(x)}{\partial x}$$

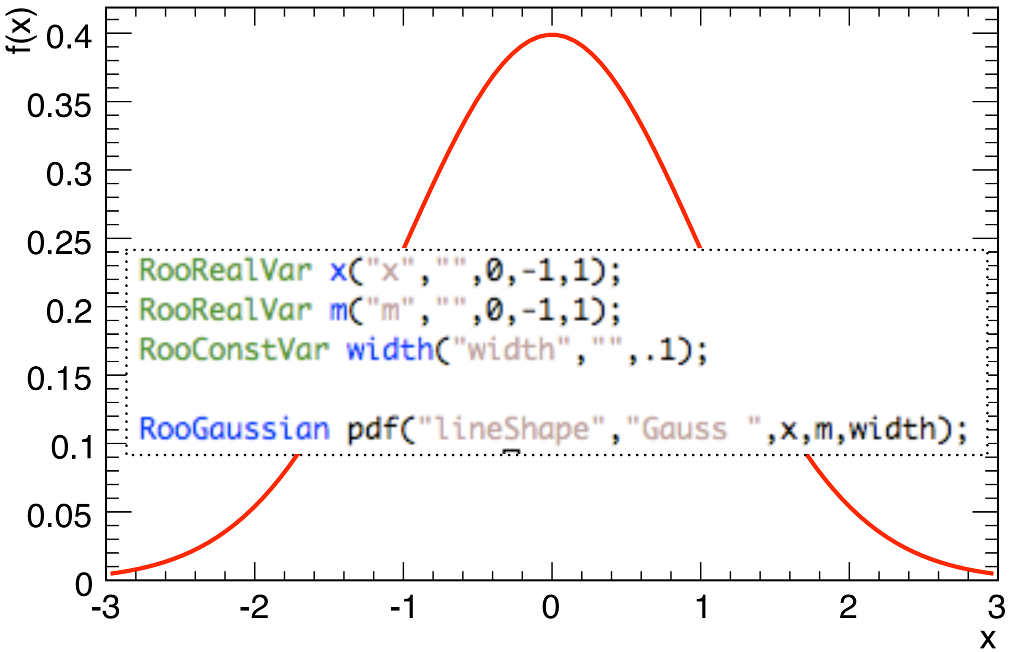
- similar to relationship of total and differential cross section:

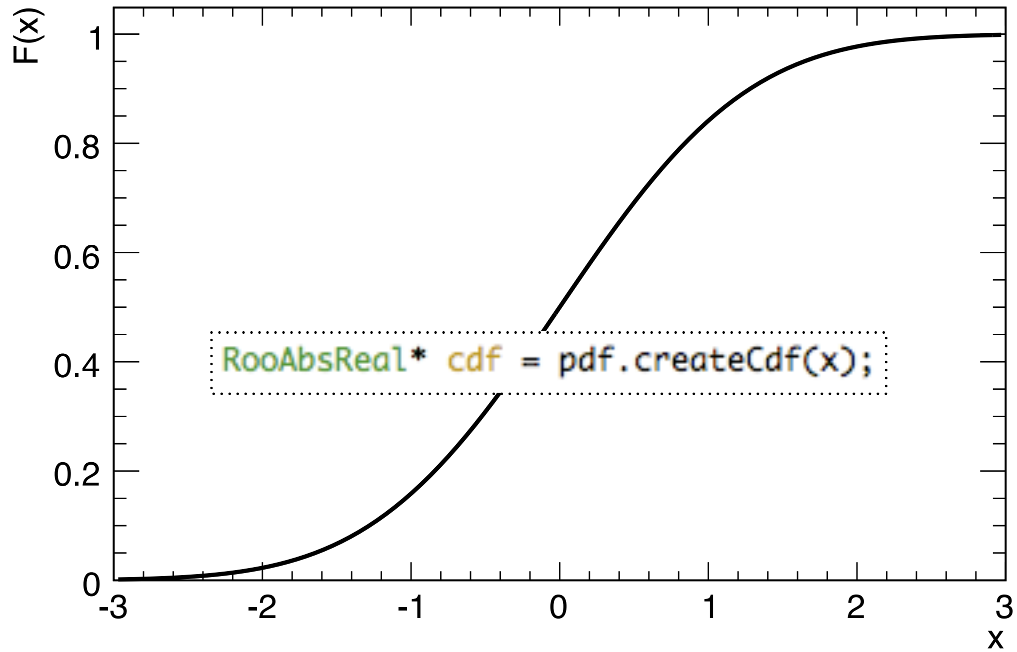
$$f(E,\eta) = \frac{1}{\sigma}\frac{\partial^2 \sigma}{\partial E \partial \eta}$$

Cumulative Density Functions

Often useful to use a cumulative distribution:

- in 1-dimension: $$\int_{-\infty}^{x} f(x')dx' = F(x)$$





- alternatively, define density as partial of cumulative:

$$f(x) = \frac{\partial F(x)}{\partial x}$$

- similar to relationship of total and differential cross section:

$$f(E,\eta) = \frac{1}{\sigma}\frac{\partial^2 \sigma}{\partial E \partial \eta}$$

Bayesian Statistics – general philosophy

In Bayesian statistics, use subjective probability for hypotheses:

probability of the data assuming hypothesis *H* (the likelihood)

prior probability, i.e., before seeing the data

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

posterior probability, i.e., after seeing the data

normalization involves sum over all possible hypotheses

Bayes' theorem has an "if-then" character: If your prior probabilities were $\pi(H)$, then it says how these probabilities should change in the light of the data.

No unique prescription for priors (subjective!)

Parametric vs. Non-Parametric PDFs

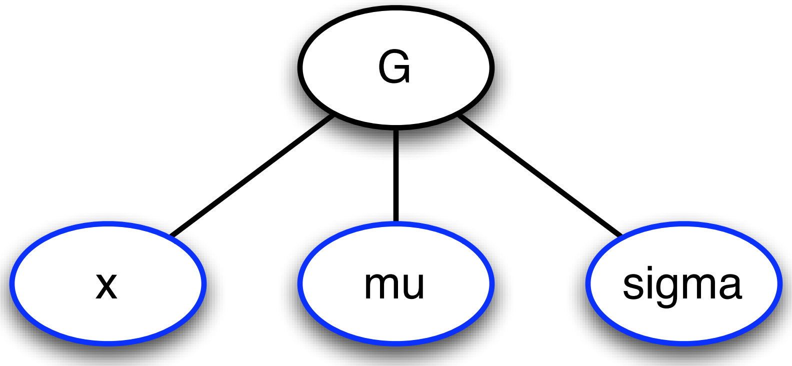
Many familiar pdfs are considered **parametric**

- eg. a Gaussian $G(x|\mu,\sigma)$ is parametrized by (μ,σ)
- defines a family of functions
- allows one to make inference about parameters
- some examples have very complicated parametric pdfs

Parametric vs. Non-Parametric PDFs

Many familiar pdfs are considered **parametric**

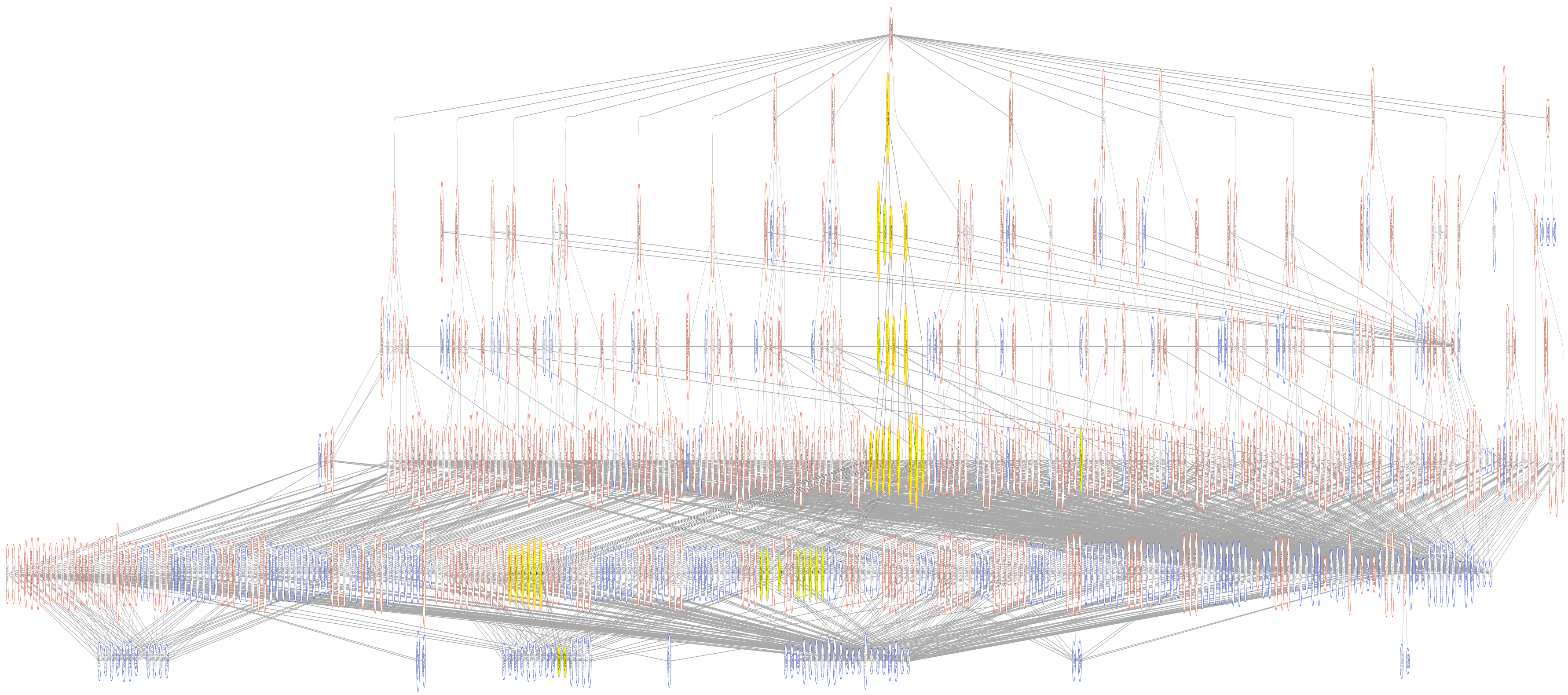
- eg. a Gaussian $G(x|\mu, \sigma)$ is parametrized by (μ, σ)
- defines a family of functions
- allows one to make inference about parameters
- some examples have very complicated parametric pdfs



Parametric vs. Non-Parametric PDFs

Many familiar pdfs are considered **parametric**

- eg. a Gaussian $G(x|\mu,\sigma)$ is parametrized by (μ,σ)
- defines a family of functions
- allows one to make inference about parameters
- some examples have very complicated parametric pdfs

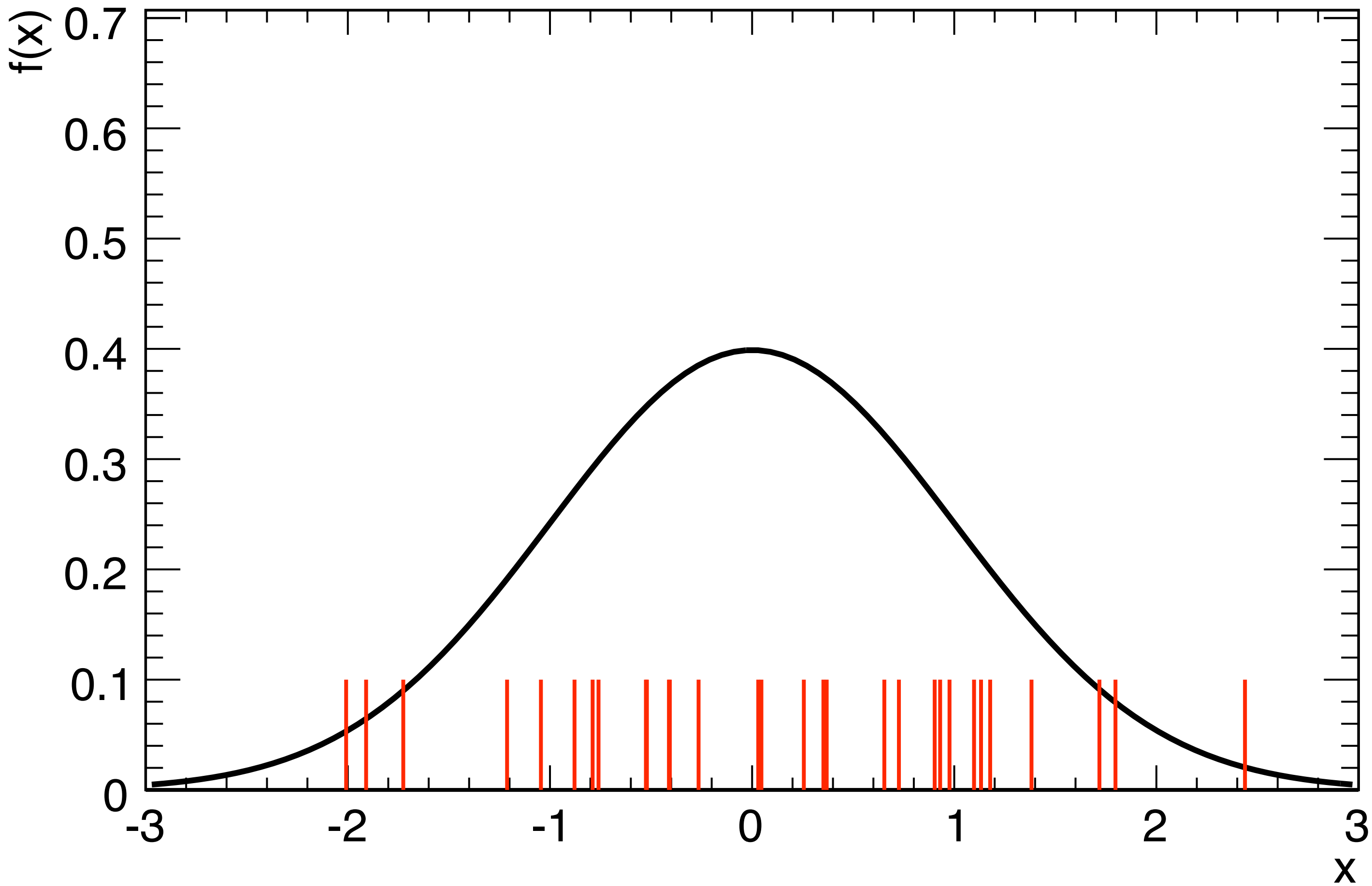


Parametric vs. Non-Parametric PDFs

Alternatively, one can consider **non-parametric** pdfs

From empirical data, one has empirical PDF

$$f_{emp} = \frac{1}{N}\sum_i^N \delta(x - x_i)$$

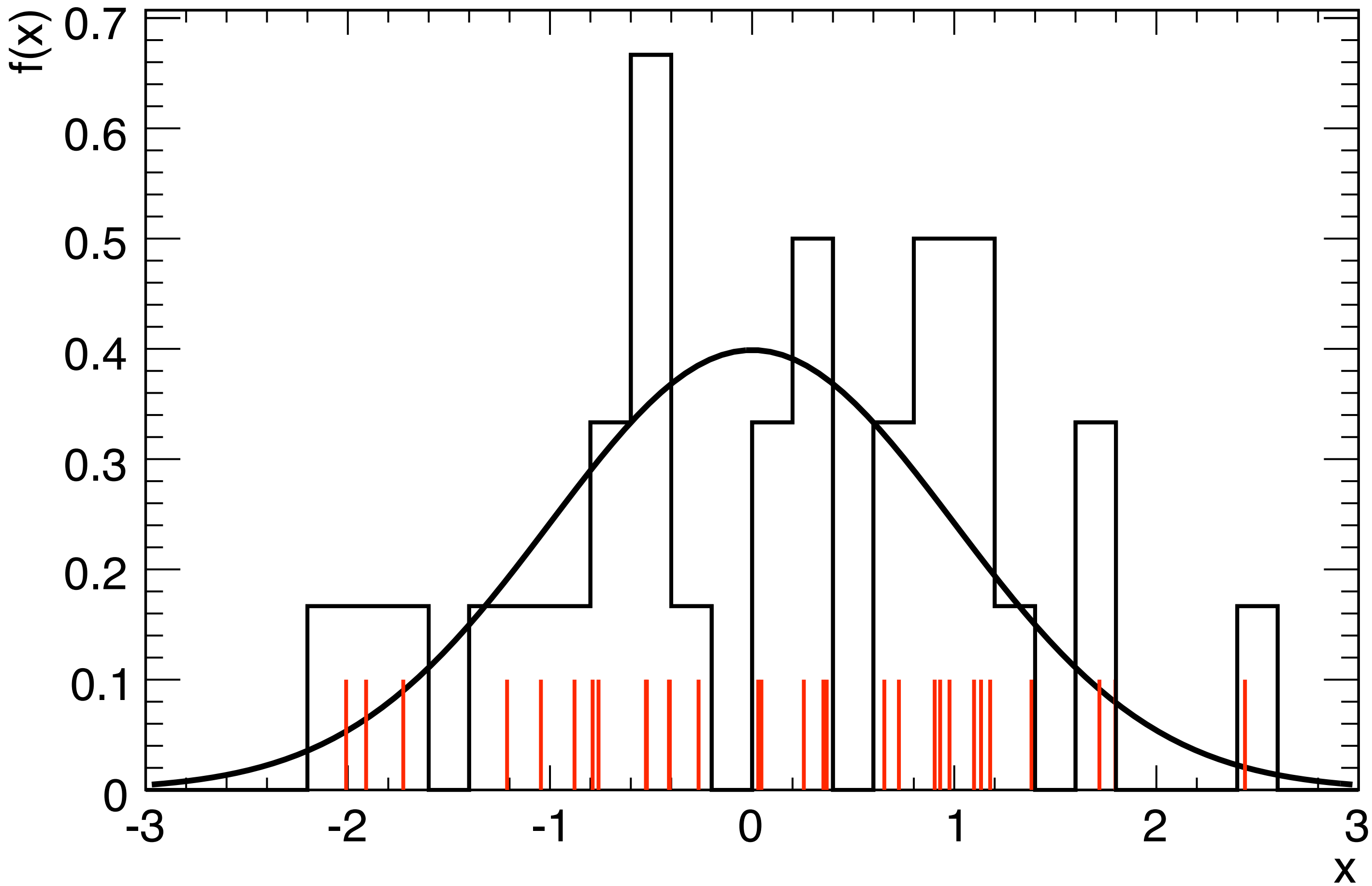


Parametric vs. Non-Parametric PDFs

Alternatively, one can consider **non-parametric** pdfs

or, one can make a histogram

$$f^{w,s}_{hist}(x) = \frac{1}{N}\sum_i h_i^{w,s}$$

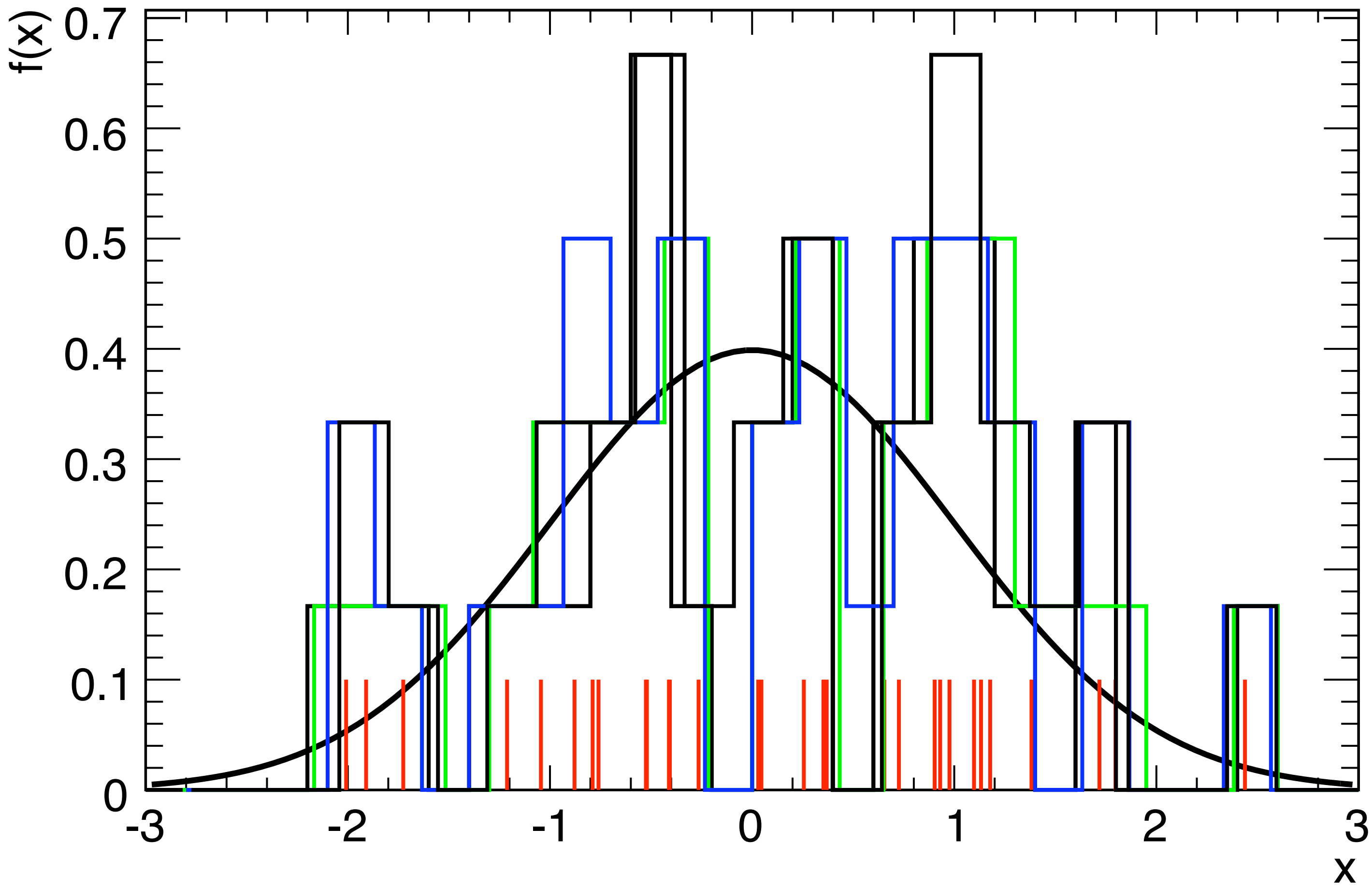


Parametric vs. Non-Parametric PDFs

Alternatively, one can consider **non-parametric** pdfs

but they depend on bin width and starting position

$$f^{w,s}_{hist}(x) = \frac{1}{N}\sum_i h_i^{w,s}$$

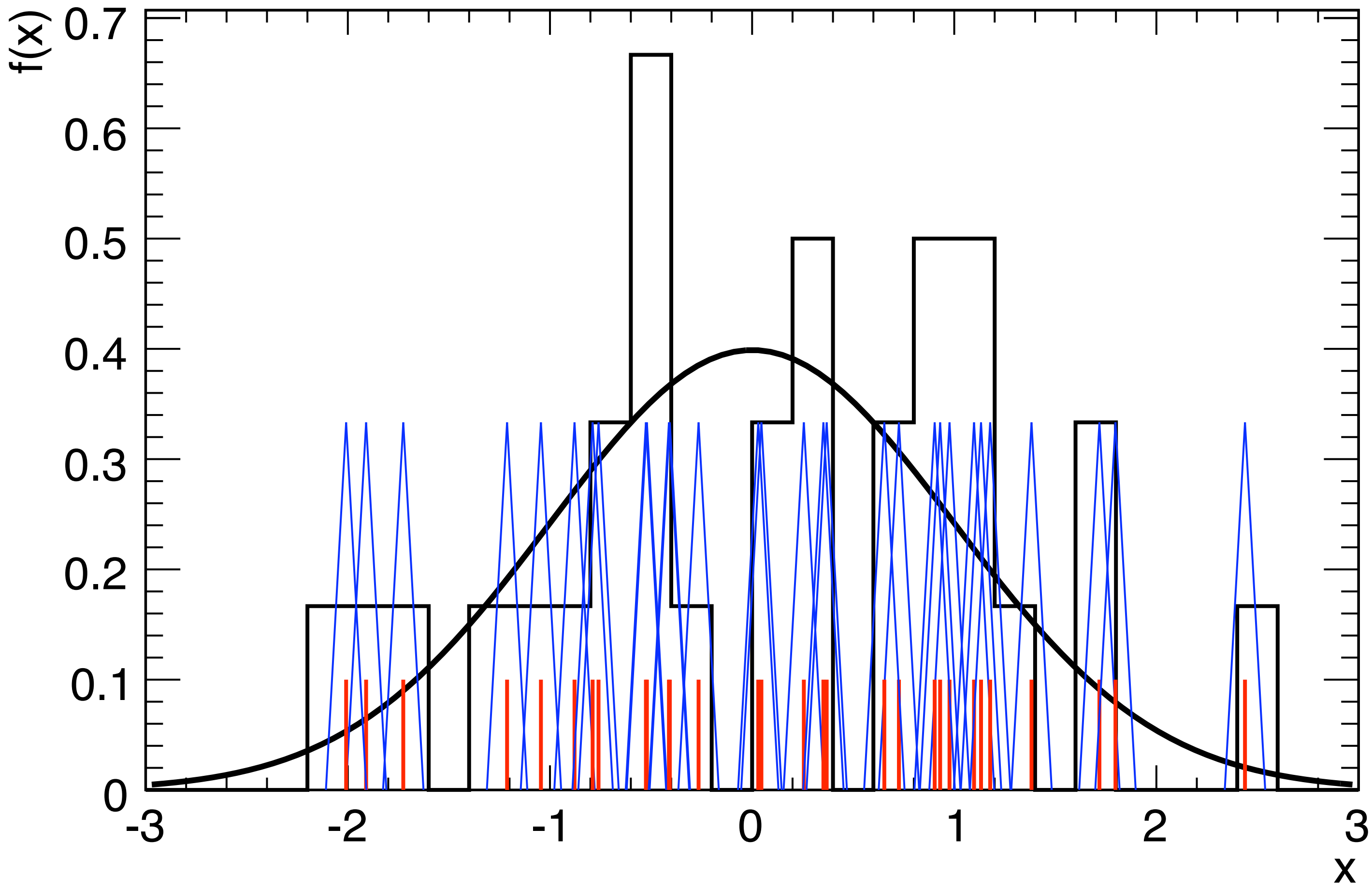


Parametric vs. Non-Parametric PDFs

Alternatively, one can consider **non-parametric** pdfs

"Average Shifted Histogram" minimizes effect of binning

$$f^{w}_{ASH}(x) = \frac{1}{N}\sum_{i}^{N} K^{w}(x - x_i)$$

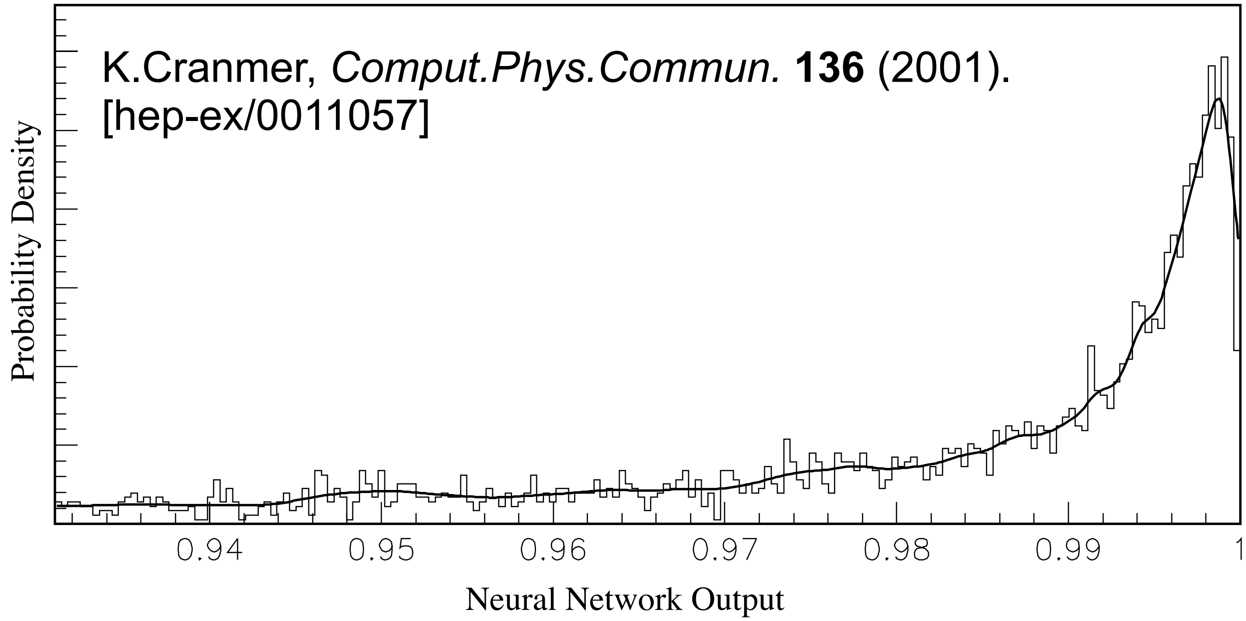


Kernel Estimation

Kernel estimation is the generalization of Average Shifted Histograms

$$\hat{f}_1(x) = \sum_i^n \frac{1}{nh(x_i)} K\left(\frac{x - x_i}{h(x_i)}\right)$$

$$h(x_i) = \left(\frac{4}{3}\right)^{1/5} \sqrt{\frac{\sigma}{\hat{f}_0(x_i)}}\, n^{-1/5}$$



"the data is the model"

Adaptive Kernel estimation puts wider kernels in regions of low probability

Used at LEP for describing pdfs from Monte Carlo (KEYS)

Multivariate PDFs

Kernel Estimation has a nice generalizations to higher dimensions

- practical limit is about 5-d due to curse of dimensionality

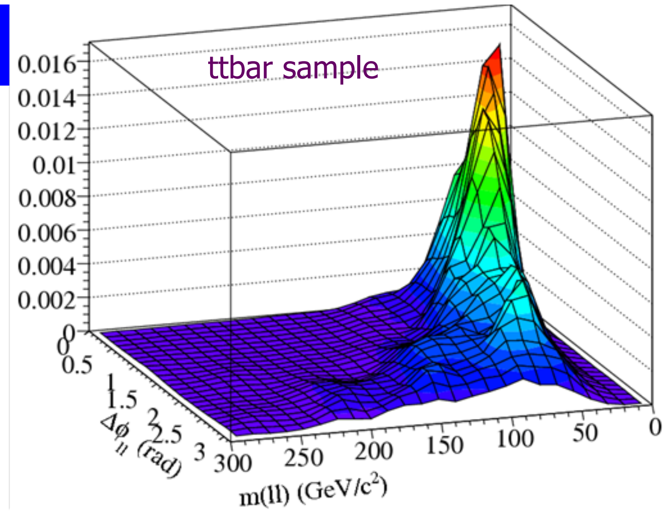
Max Baak has coded N-dim KEYS pdf described in Comput.Phys.Commun. **136** (2001) in RooFit.

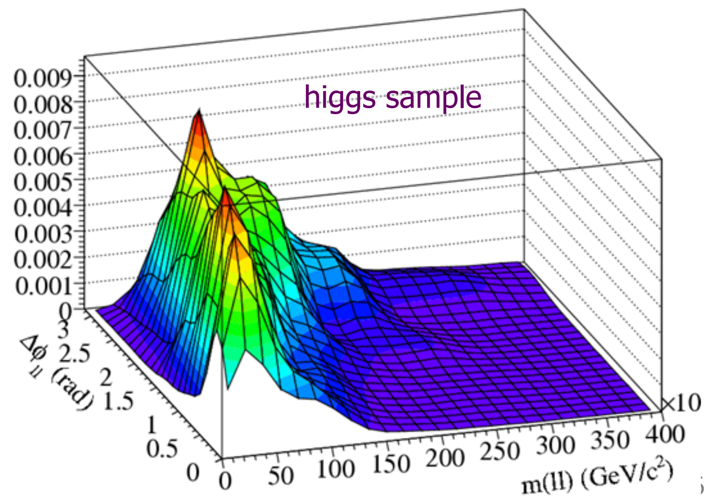
These pdfs have been used as the basis for a multivariate discrimination technique called "PDE"

$$D(\vec{x}) = \frac{f_s(\vec{x})}{f_s(\vec{x}) + f_b(\vec{x})}$$

Correlations

- 2-d projection of pdf from previous slide.
- RooNDKeys pdf automatically models (fine) correlations between observables ...





Max Baak

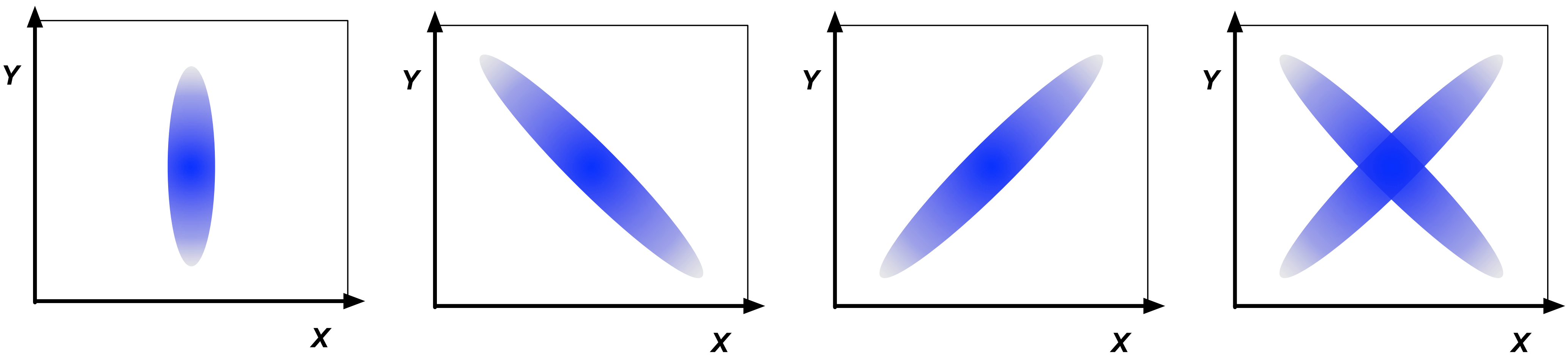
Correlation / Covariance

Correlation is a common way to describe how one variable depends on another

- however, it only captures the lowest order of dependence between variables, and

$$cov[x,y] = V_{xy} = E[(x-\mu_x)(y-\mu_y)]$$

$$\rho_{xy} = \frac{\mathrm{cov}[x,y]}{\sigma_x \sigma_y}$$



Propagation of errors

The Covariance matrix plays a central rôle in propagation of errors from x to y

$$\sigma_y^2 \approx \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij}$$

but remember, that this is only the first-order in the Taylor expansion

$$y(\vec{x}) \approx y(\vec{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i)$$

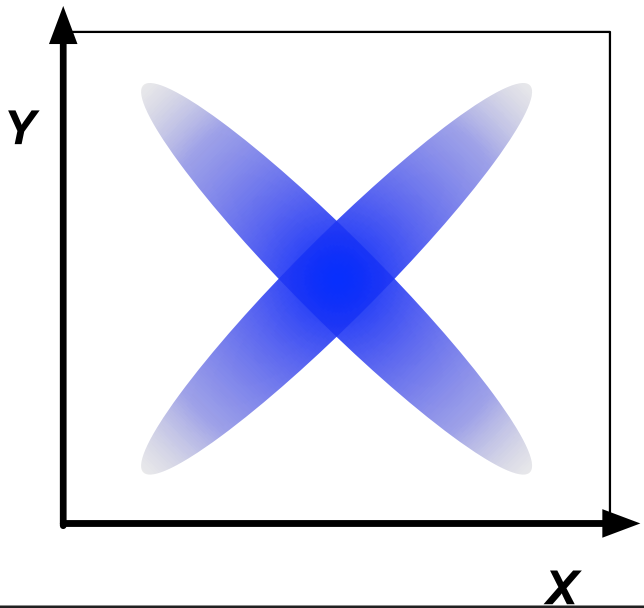
Mutual Information

A more general notion of 'correlation' comes from **Mutual Information**:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p_1(x)\, p_2(y)} \right),$$

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

- it is symmetric: I(X;Y) = I(Y;X)
- if and only if X,Y totally independent: I(X;Y)=0
- possible for X,Y to be uncorrelated, but not independent



Mutual Information doesn't seem to be used much within HEP, but it seems quite useful

Remaining topics for "Lecture 1"

Lecture 1:

- How we use statistics
- Probability axioms, Bayes vs. Frequentist, from discrete to continuous
- Parametric and non-parametric probability density functions
- Shannon and Fisher Information, correlation, information geometry, Cramér-Rao bound
- A word on subjective and "objective" Bayesian priors

Next Time

Lecture 2

- Hypothesis testing in the frequentist setting
- The Neyman-Pearson lemma (with a simple proof)
- Decision theory: utility, risk, priors, and game theory
- Contrast hypothesis testing to goodness of fit tests with some warnings
- Related comments on multivariate algorithms
- Matrix element techniques vs. the black box

Lecture 3:

- The Neyman-Construction (illustrated)
- Inverted hypothesis tests: A dictionary for limits (intervals)
- Coverage as a calibration for our statistical device
- Compound hypotheses, nuisance parameters, & similar tests
- Systematics, Systematics, Systematics

Lecture 4:

- Generalizing our procedures to include systematics
- Eliminating nuisance parameters: profiling and marginalization
- Introduction to ancillary statistics & conditioning
- High dimensional models, Markov Chain Monte Carlo, and Hierarchical Bayes
- The look elsewhere effect and false discovery rate