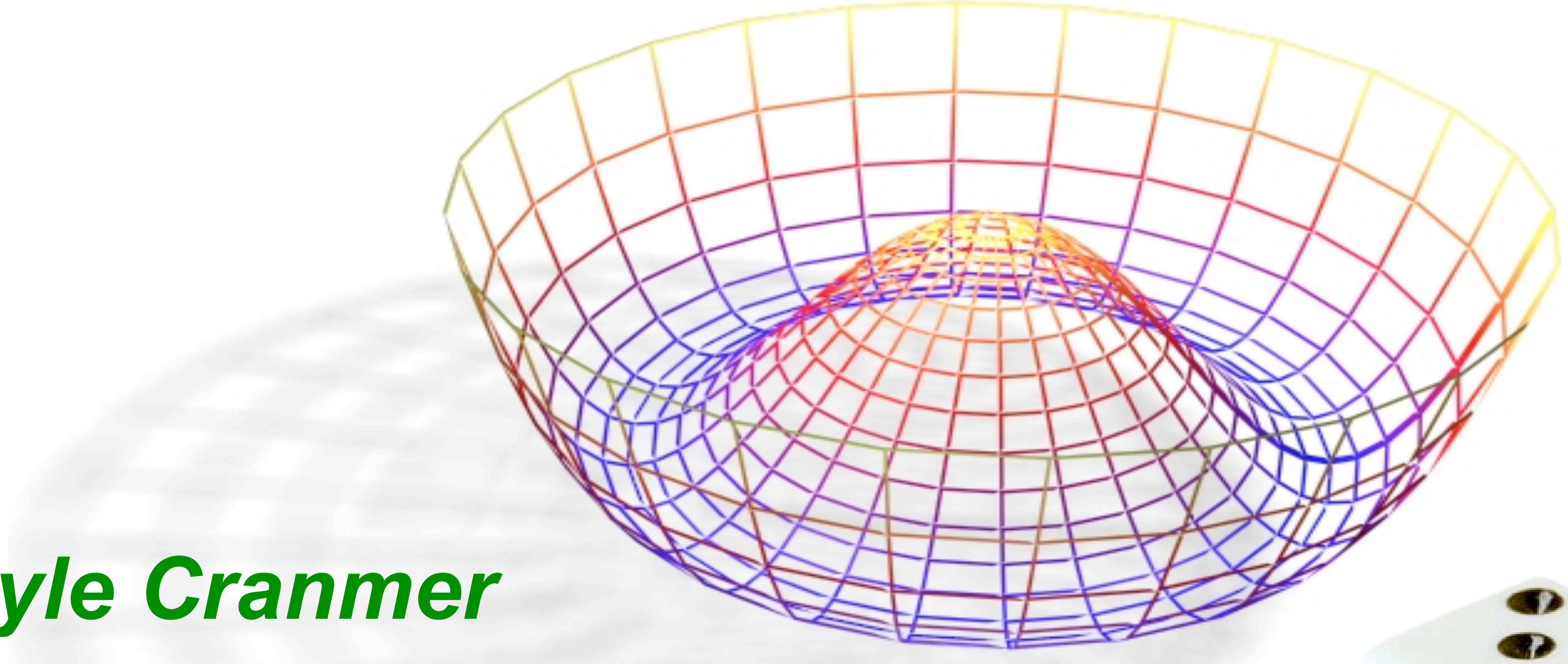
Statistics for Particle Physics



Kyle Cranmer

New York University

LAPP

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 3 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;

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' lectures

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

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

Glen Cowan's lectures

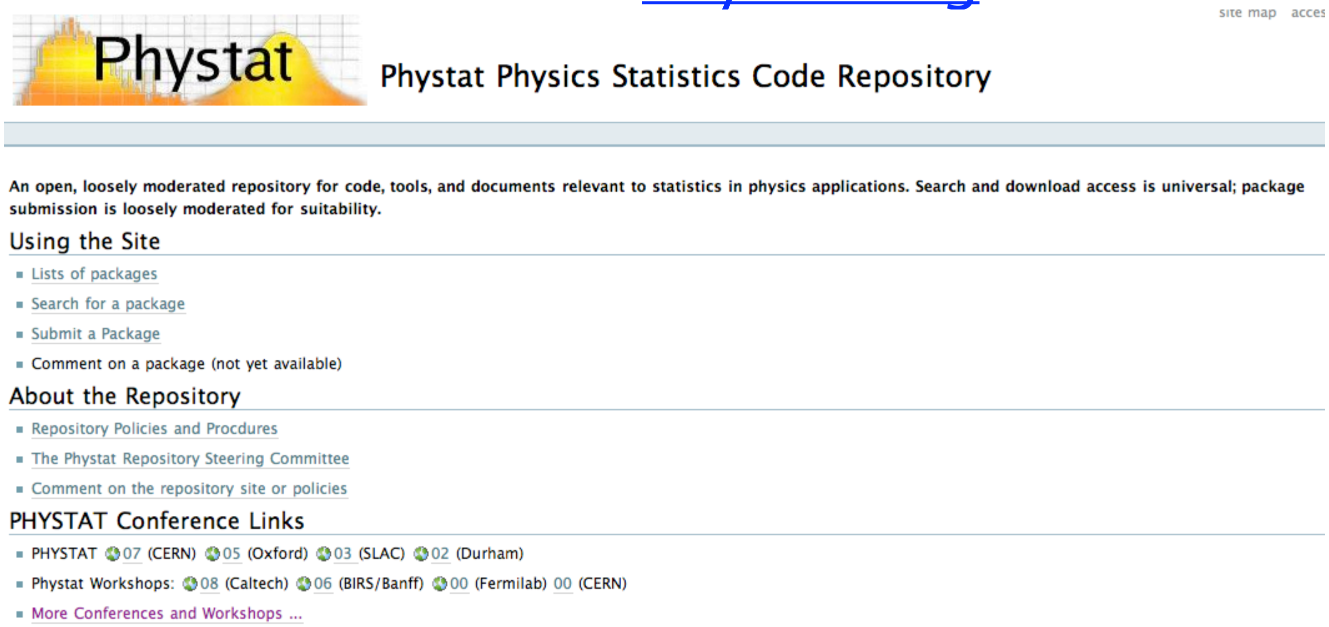
http://www.pp.rhul.ac.uk/~cowan/stat_cern.html

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



Phystat Physics Statistics Code Repository

An open, loosely moderated repository for code, tools, and documents relevant to statistics in physics applications. Search and download access is universal; package submission is loosely moderated for suitability.

Using the Site

- Lists of packages
- Search for a package
- Submit a Package
- Comment on a package (not yet available)

About the Repository

- Repository Policies and Procdures
- The Phystat Repository Steering Committee
- Comment on the repository site or policies

PHYSTAT Conference Links

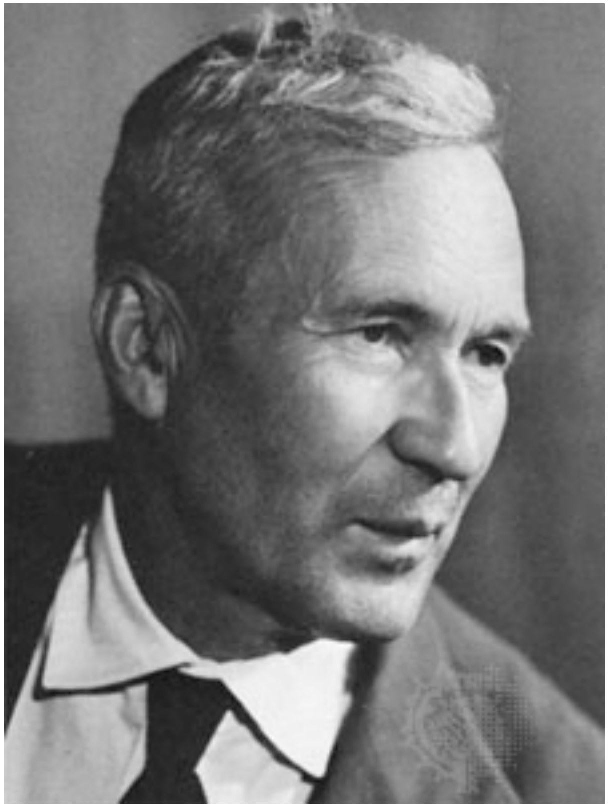
- PHYSTAT 07 (CERN) 05 (Oxford) 03 (SLAC) 02 (Durham)
- Phystat Workshops: 08 (Caltech) 06 (BIRS/Banff) 00 (Fermilab) 00 (CERN)
- More Conferences and Workshops ...

Lecture 1

Axioms of Probability

These Axioms are a mathematical starting point for probability and statistics

1. probability for every element, E, is non-negative $P(E) \geq 0 \qquad \forall E \subseteq F$
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)
  - not very practical, sometimes ensemble doesn't exist
  - eg. probability Higgs mass is 120 GeV, weather tomorrow
- basis of 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 is aligned along +z $|\langle \rightarrow \mid \uparrow \rangle|^2 = \frac{1}{2}$

Subjective Bayesian

- Probability is a degree of belief (personal, subjective)
  - can be made more rigorous 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

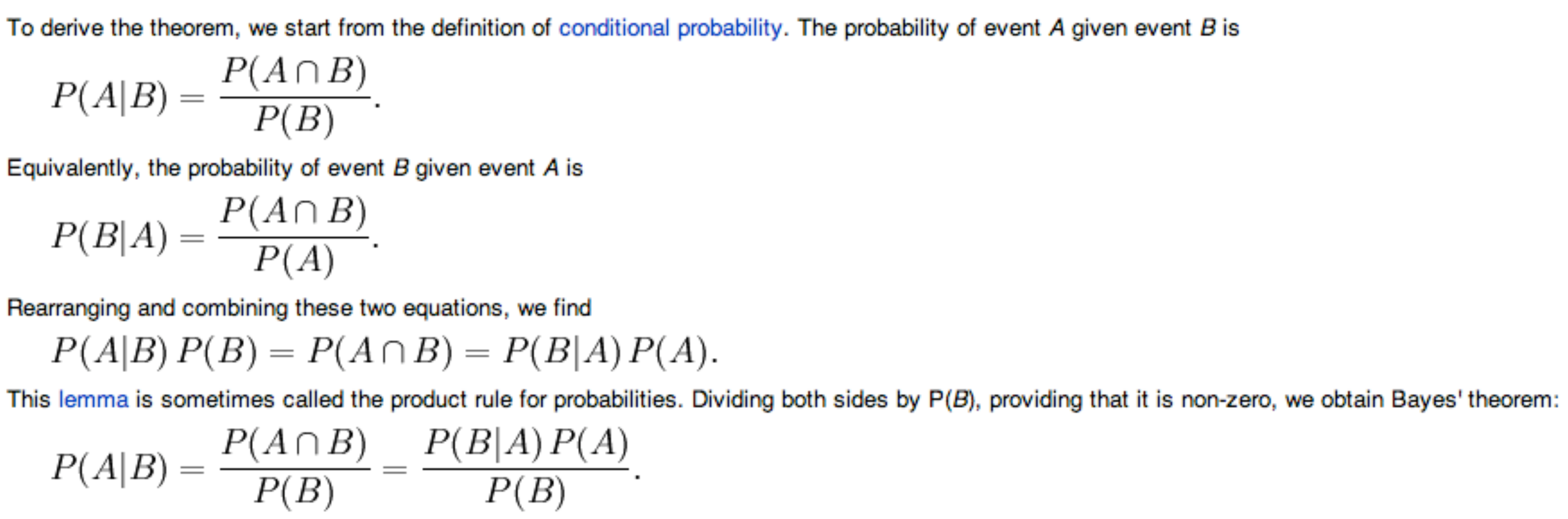
Bayes' Theorem

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*l*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*l*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

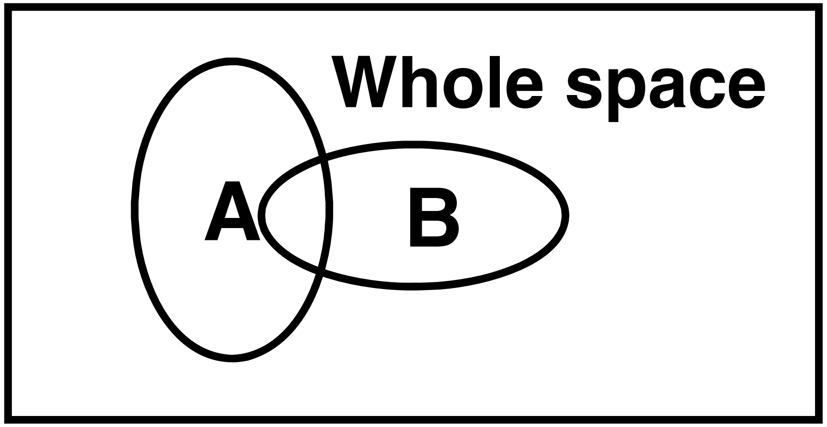
$$P(A|B)\,P(B) = P(A \cap B) = P(B|A)\,P(A).$$

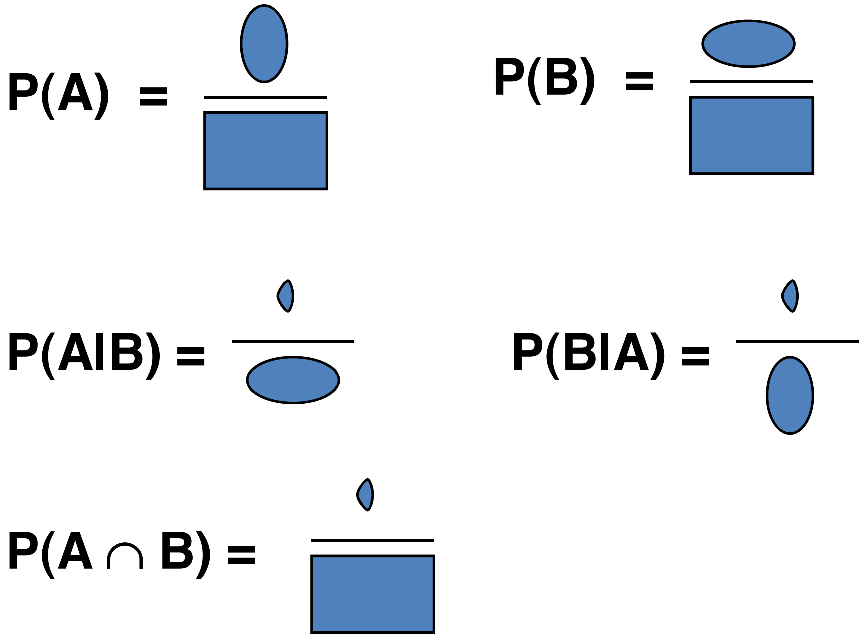
This lemma is sometimes called the product rule for probabilities. Dividing both sides by P(*B*), providing that it is non-zero, we obtain Bayes' theorem:

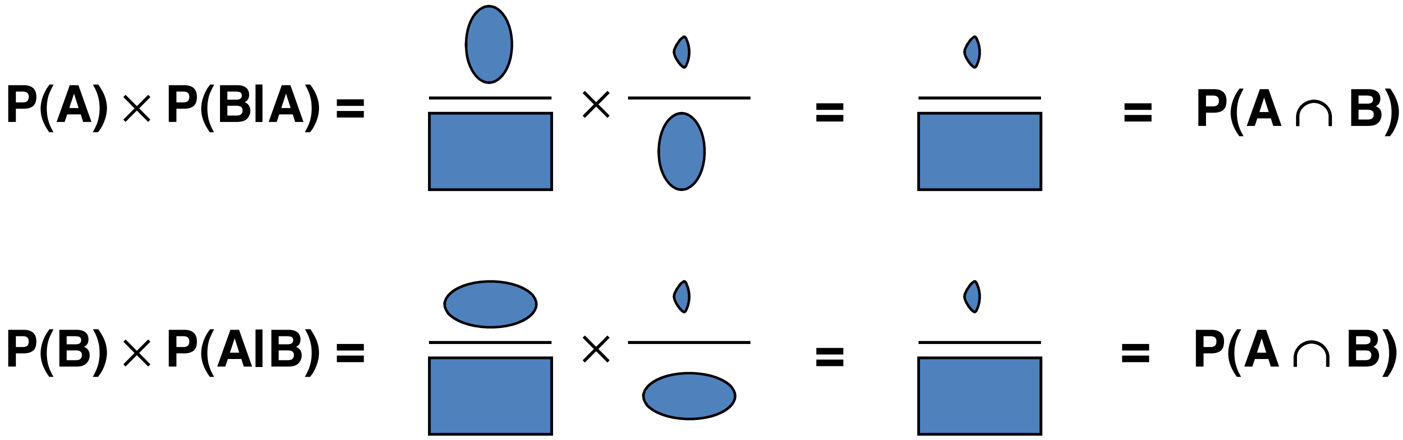
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)}.$$

... in pictures (from Bob Cousins)

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







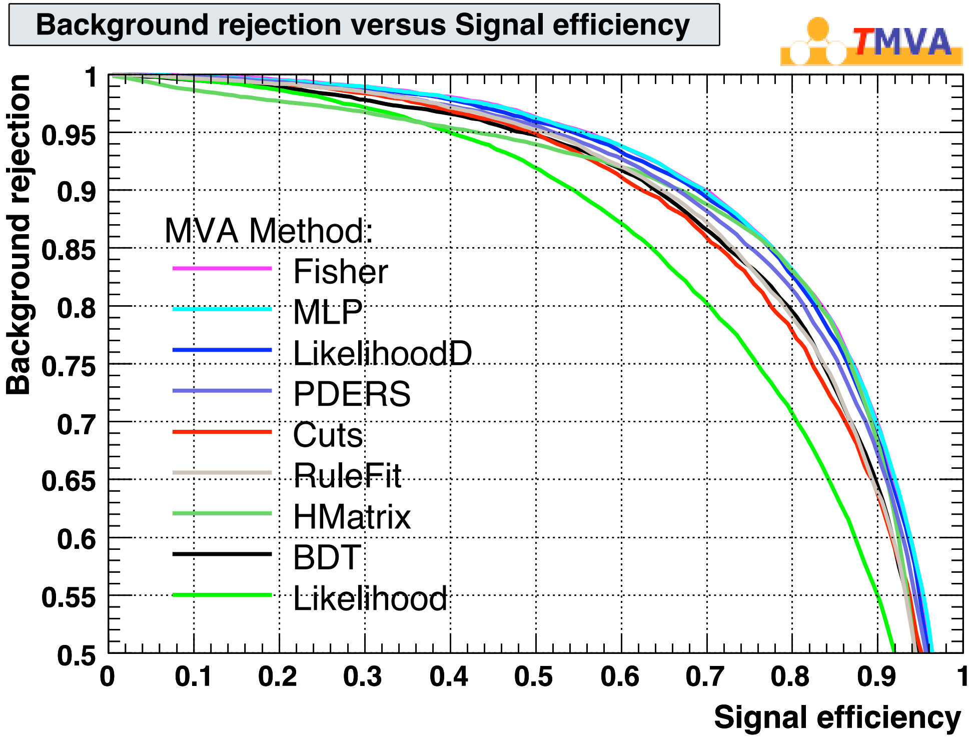
⇒ P(BIA) = P(AIB) × P(B) / P(A)

Bob Cousins, CMS, 2008

7

An example of Bayes' theorem

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

An 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) ?**

An 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:
  - $P(\text{b-jet} \mid \text{btag}) \propto P(\text{btag} \mid \text{b-jet})\, P(\text{b-jet})$

An 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

An 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

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

Bayesian vs. Frequentist

In short, Frequentist are always restricted to statements related to

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

Bayesians can address questions like:

- P(Theory | Data) $\propto$ P(Data | Theory) P(Theory)
  - intuitively what we want to know
- 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

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?

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

Information Theory

How much information in this message? $\underbrace{1000110101001011}_{16 \text{ entries}}$

What about this one $\underbrace{0101010101010101}_{16 \text{ entries}}$

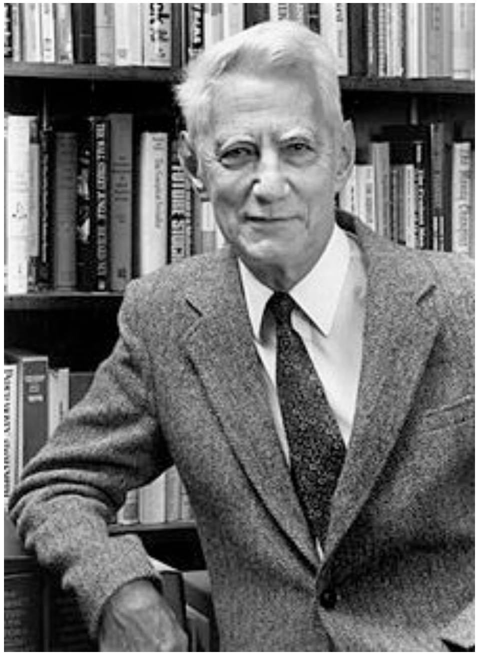
Information Theory

How much information in this message? $\underbrace{1000110101001011}_{16 \text{ entries}}$

What about this one $\underbrace{0101010101010101}_{16 \text{ entries}}$

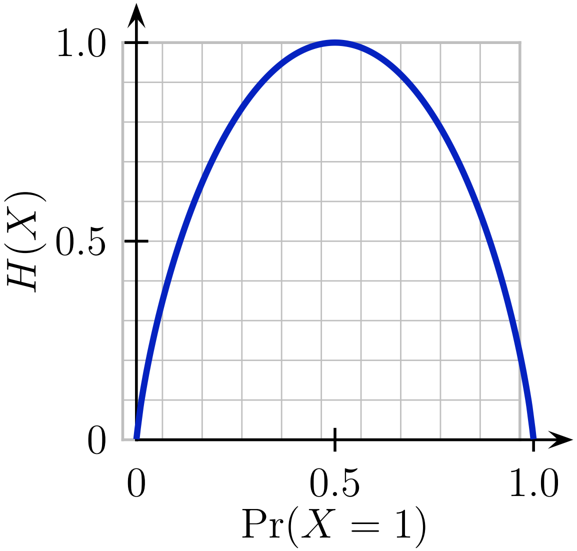
... and this one? $\underbrace{abcdabcdabcdabcd}_{16 \text{ entries}}$

Information Theory



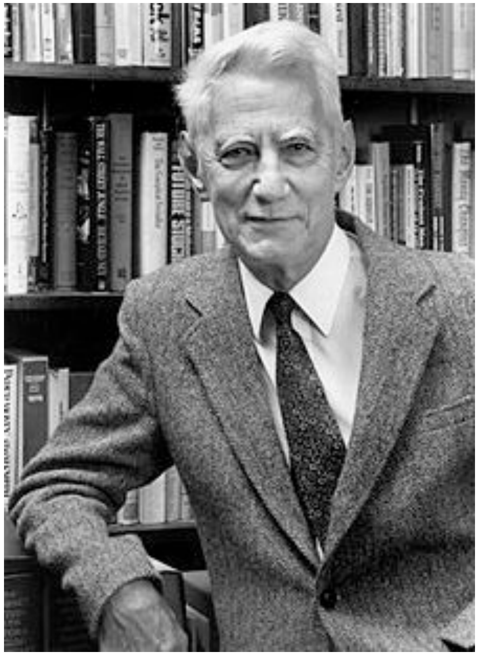
$$\underbrace{1000110101001011}_{\text{16 entries}}$$

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



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

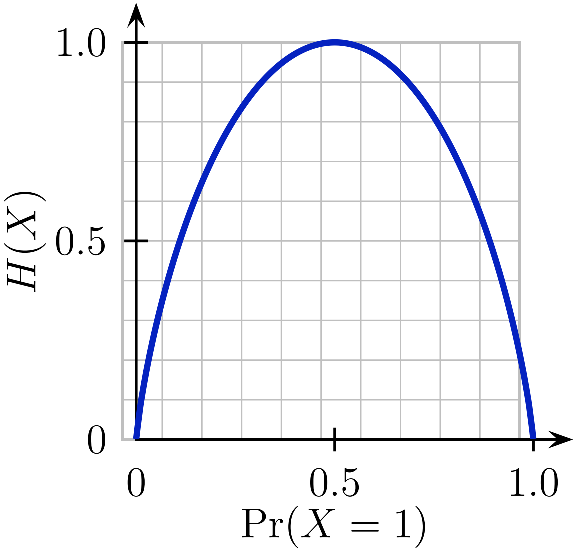
Information Theory



How much information in this message?

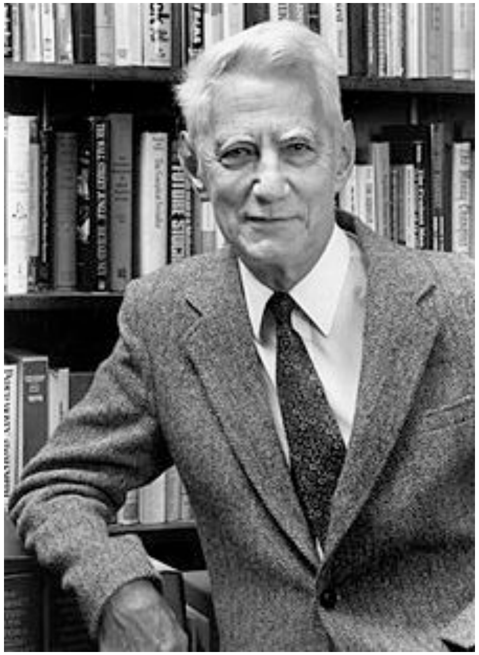
$$\underbrace{1000110101001011}_{\text{16 entries}}$$

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



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

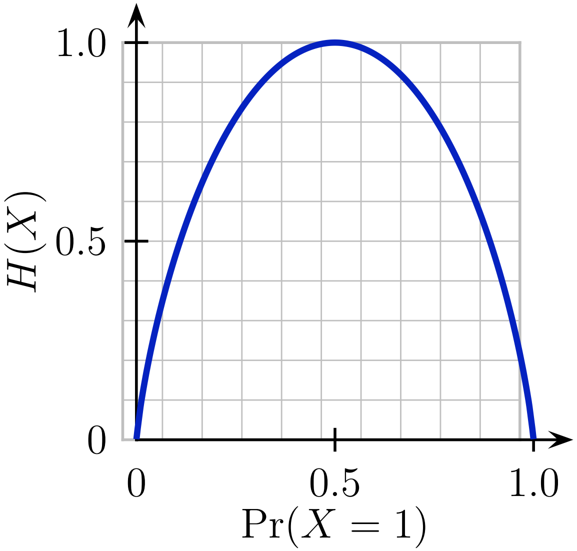
Information Theory



How much information in this message?

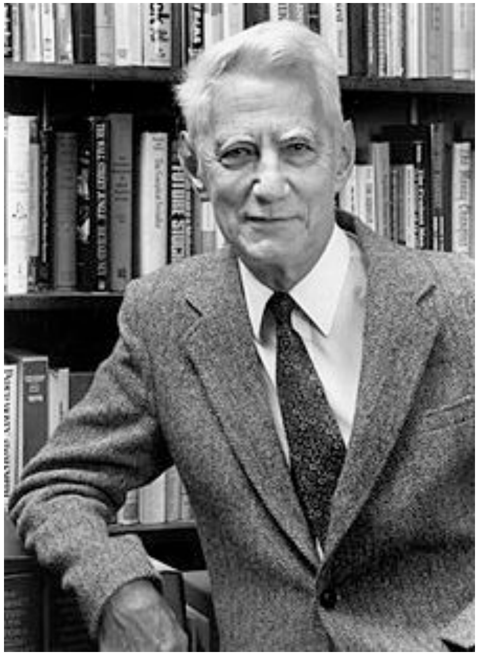
$$\underbrace{1000110101001011}_{16\text{ entries}}$$

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



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

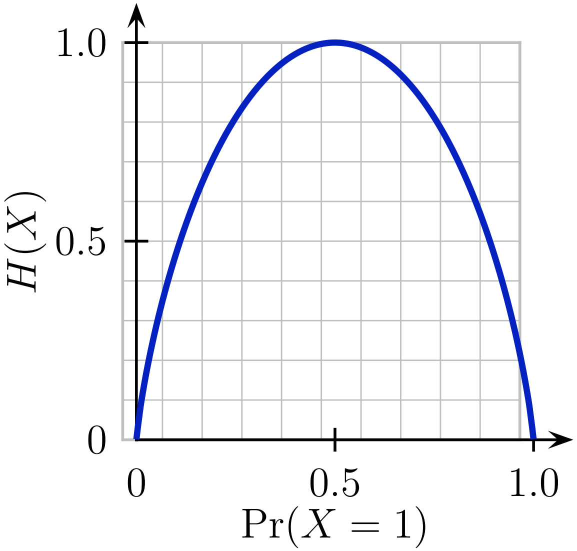
Information Theory



How much information in this message?

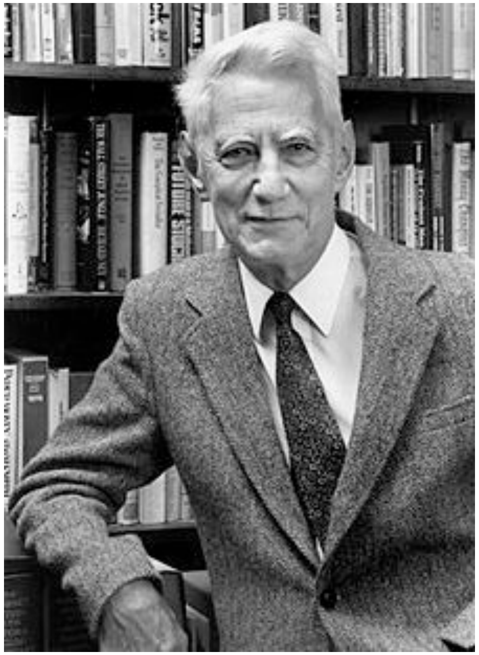
$$\underbrace{1000110101001011}_{\text{16 entries}}$$

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



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

Information Theory

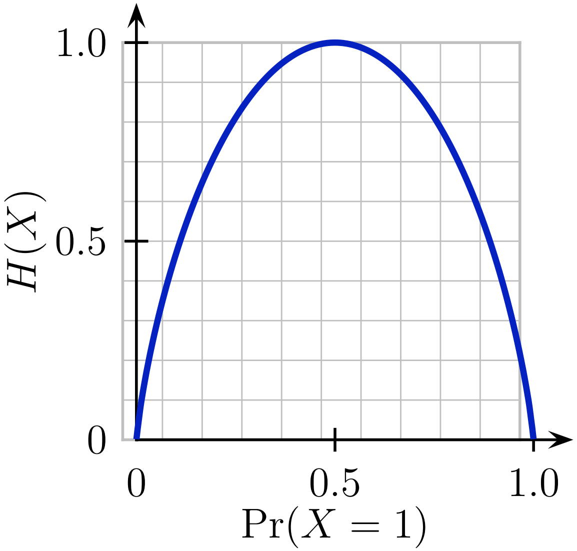


How much information in this message?

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

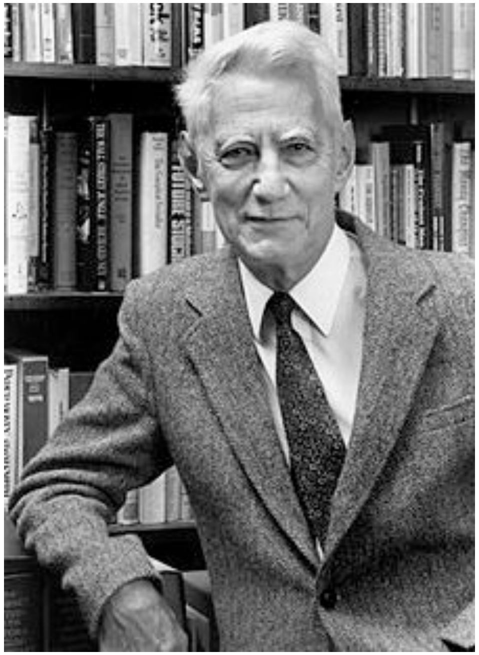
- 16 bits? (**bit** is unit when log is base 2)

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



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

Information Theory

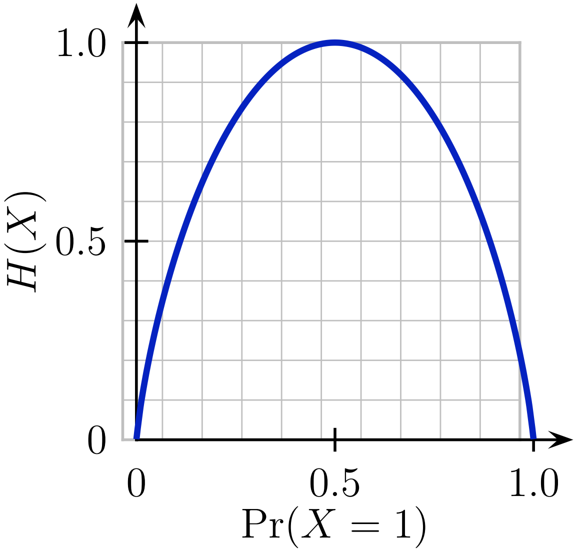


How much information in this message?

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

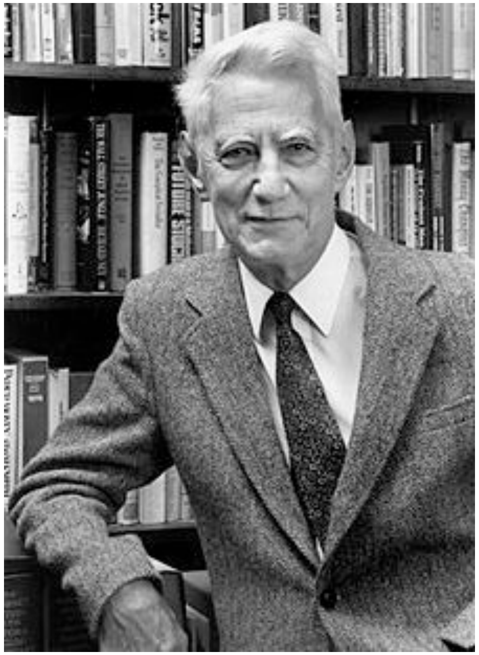
- 16 bits? (**bit** is unit when log is base 2)
- it depends on probabilities of 0,1

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



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

Information Theory

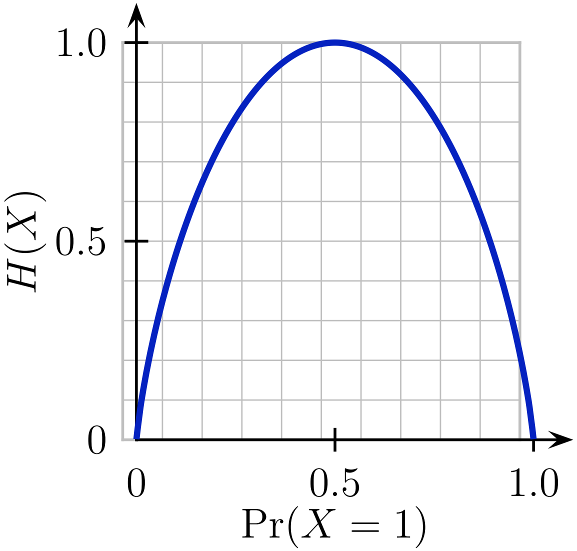


How much information in this message?

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

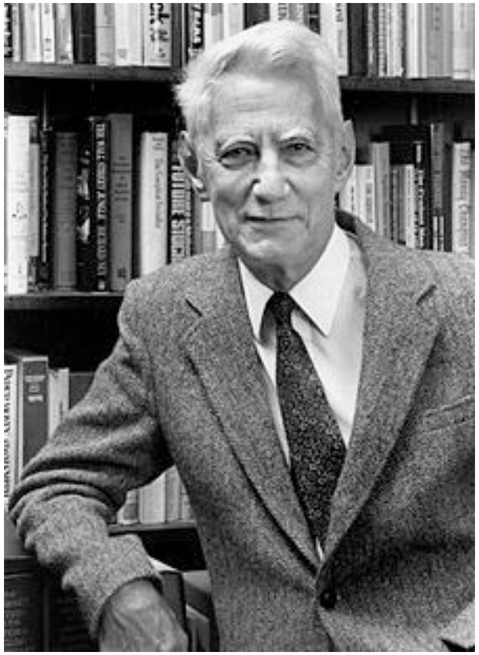
- 16 bits? (**bit** is unit when log is base 2)
- it depends on probabilities of 0,1

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



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

Information Theory



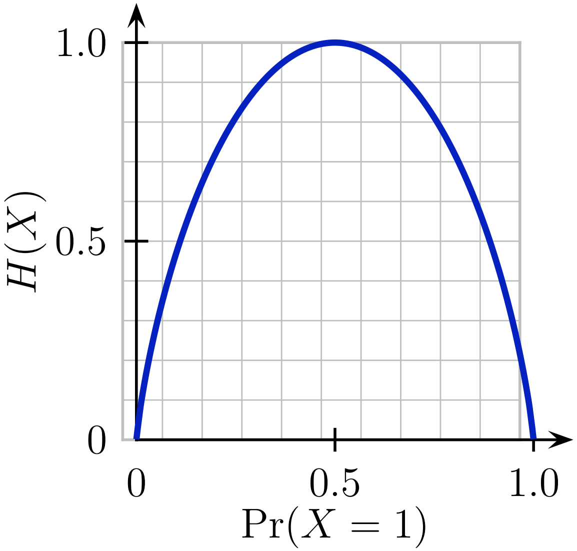
How much information in this message?

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

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

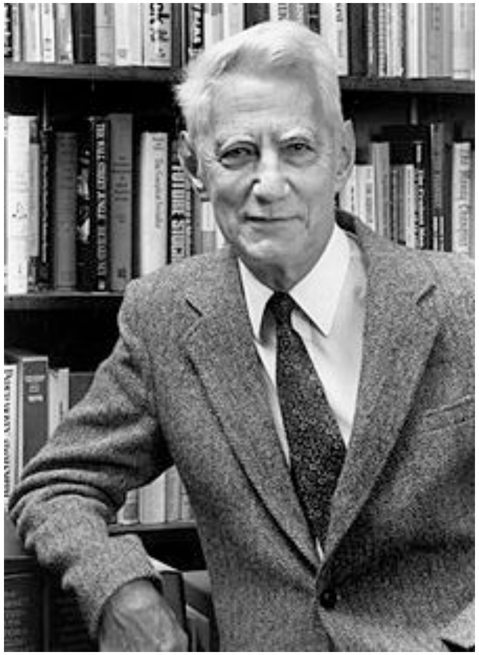
In 1870's Boltzman and Gibbs defined entropy:

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



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

Information Theory



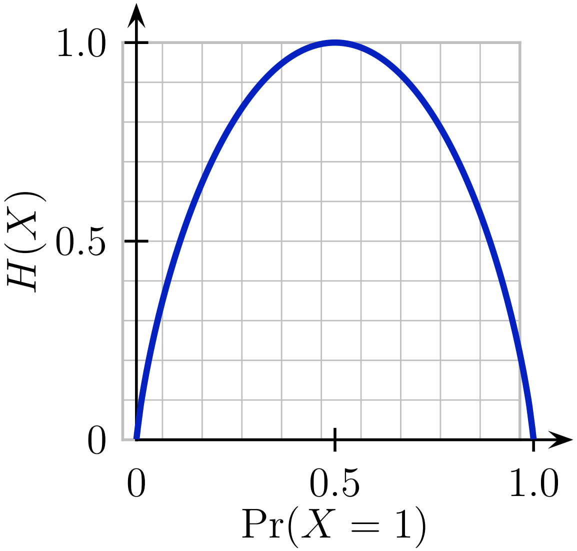
How much information in this message?

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

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

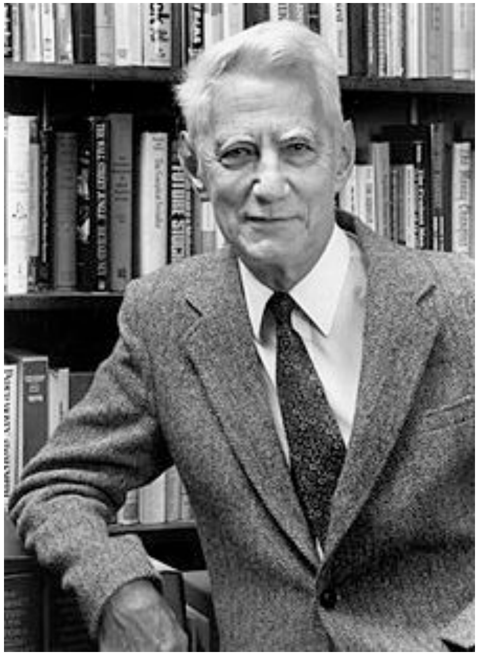
In 1870's Boltzman and Gibbs defined entropy:

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



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

Information Theory



How much information in this message?

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

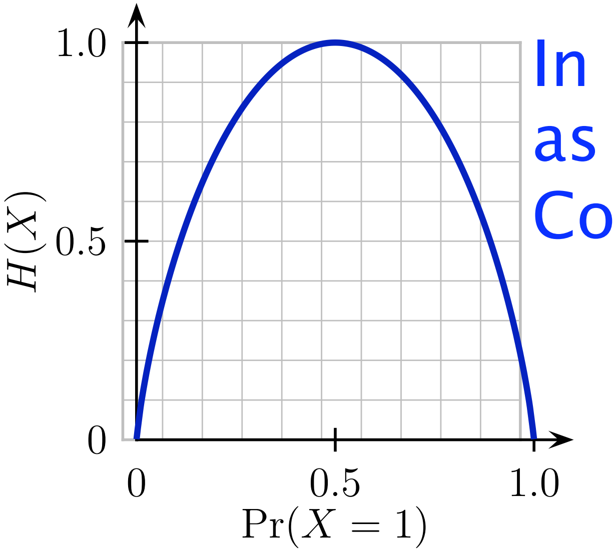
- 16 bits? (**bit** is unit when log is base 2)
- it depends on probabilities of 0,1

In 1870's Boltzman and Gibbs defined entropy:

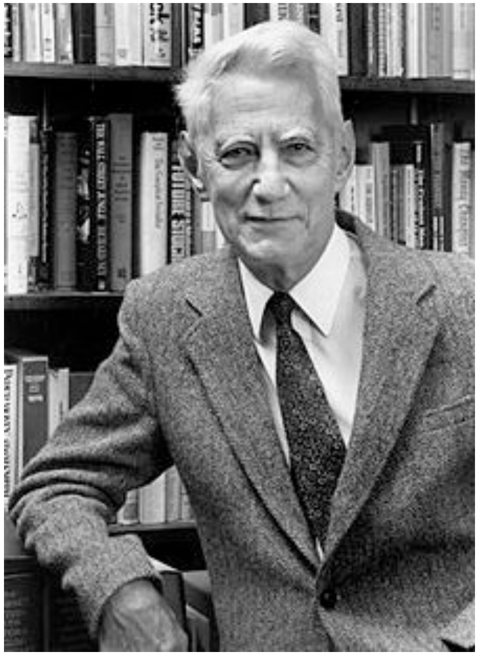
$$S = -k_B \sum_i p_i \ln p_i$$

In 1948, Calude Shannon publishes 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 Theory



How much information in this message?

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

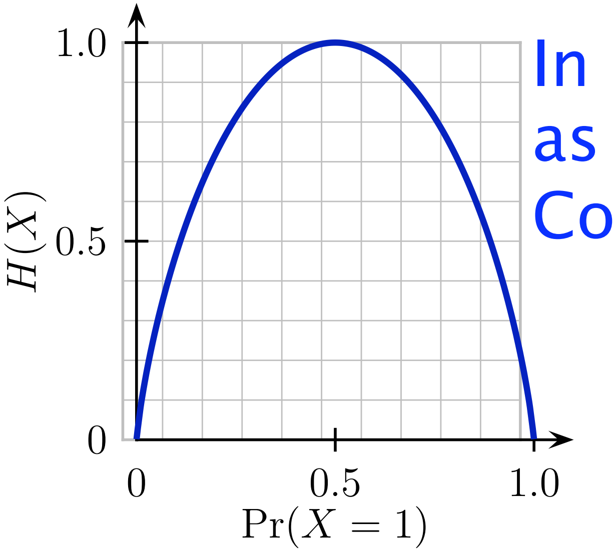
- 16 bits? (**bit** is unit when log is base 2)
- it depends on probabilities of 0,1

In 1870's Boltzman and Gibbs defined entropy:

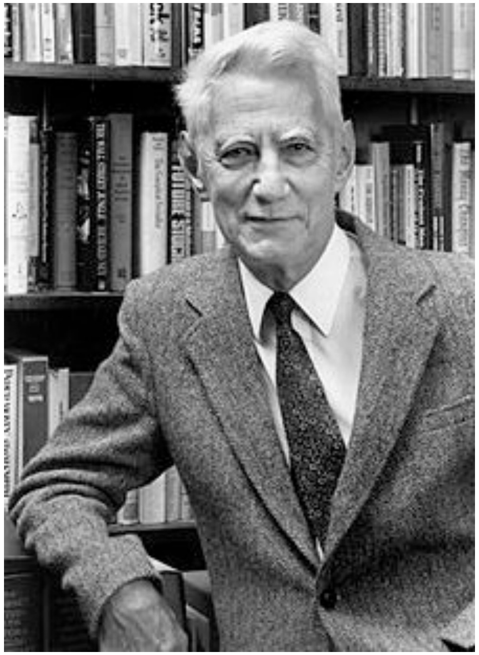
$$S = -k_B \sum_i p_i \ln p_i$$

In 1948, Calude Shannon publishes 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 Theory



How much information in this message?

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

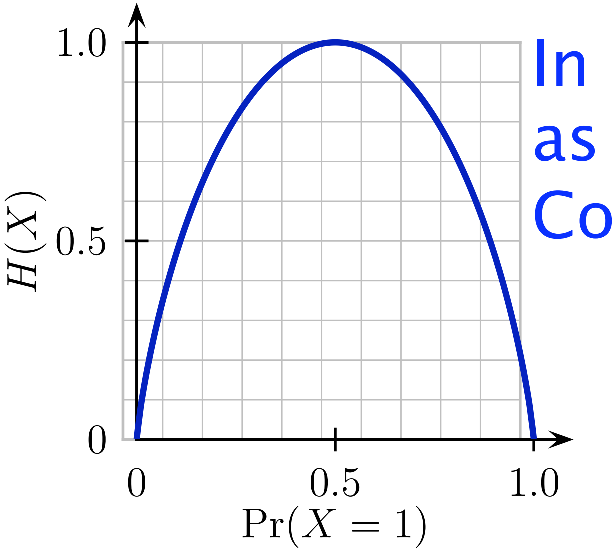
- 16 bits? (**bit** is unit when log is base 2)
- it depends on probabilities of 0,1

In 1870's Boltzman and Gibbs defined entropy:

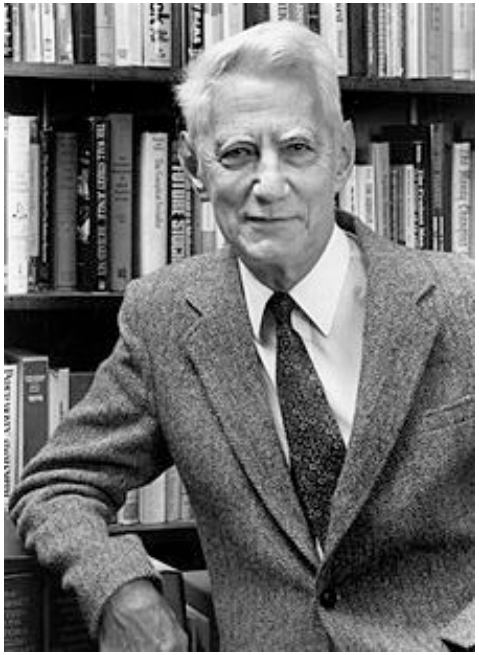
$$S = -k_B \sum_i p_i \ln p_i$$

In 1948, Calude Shannon publishes 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 Theory



How much information in this message?

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

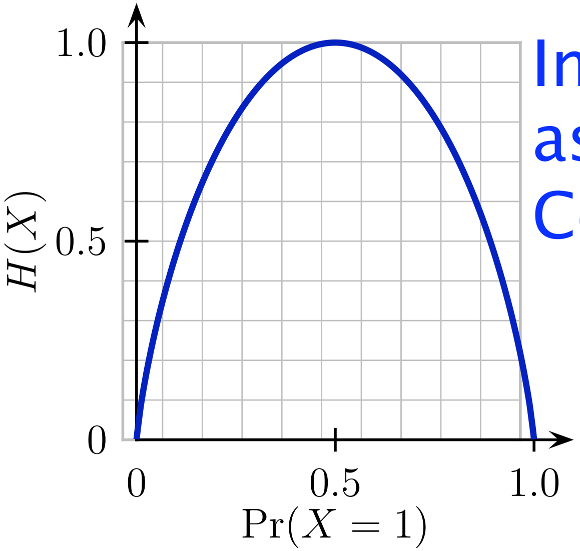
- 16 bits? (**bit** is unit when log is base 2)
- it depends on probabilities of 0,1

In 1870's Boltzman and Gibbs defined entropy:

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

In 1948, Calude Shannon publishes 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

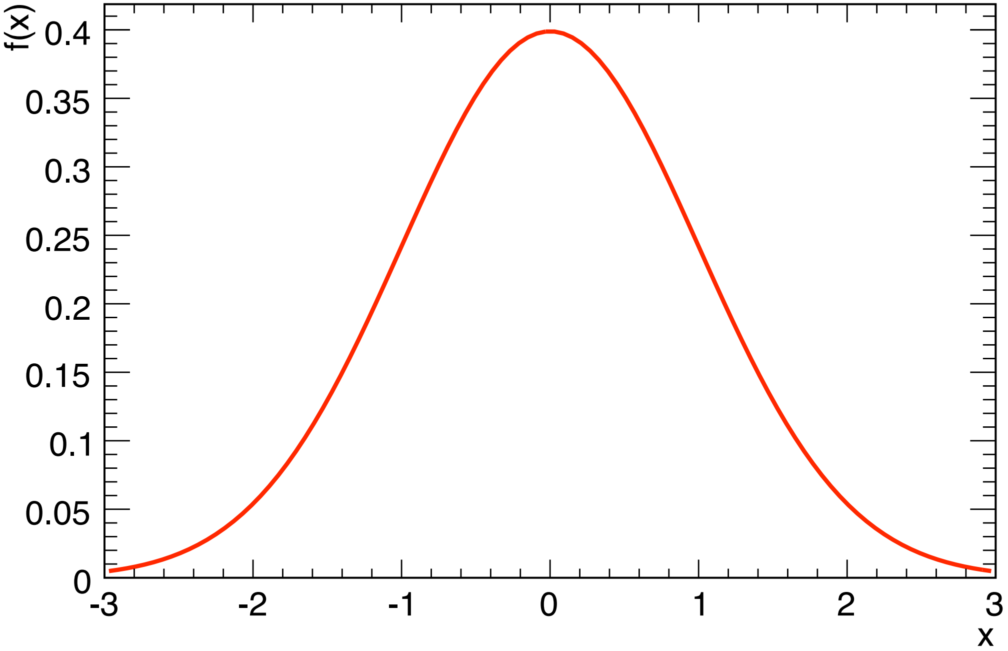
Equivalent of second axiom...

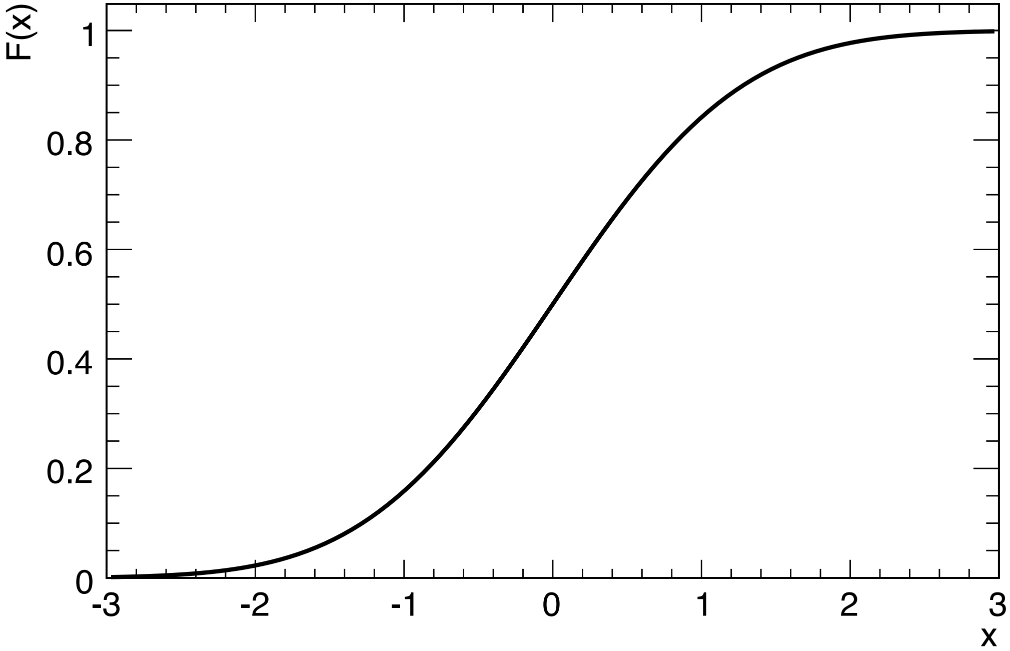
$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Cumulative Density Functions

Often useful to use a cumulative distribution:

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

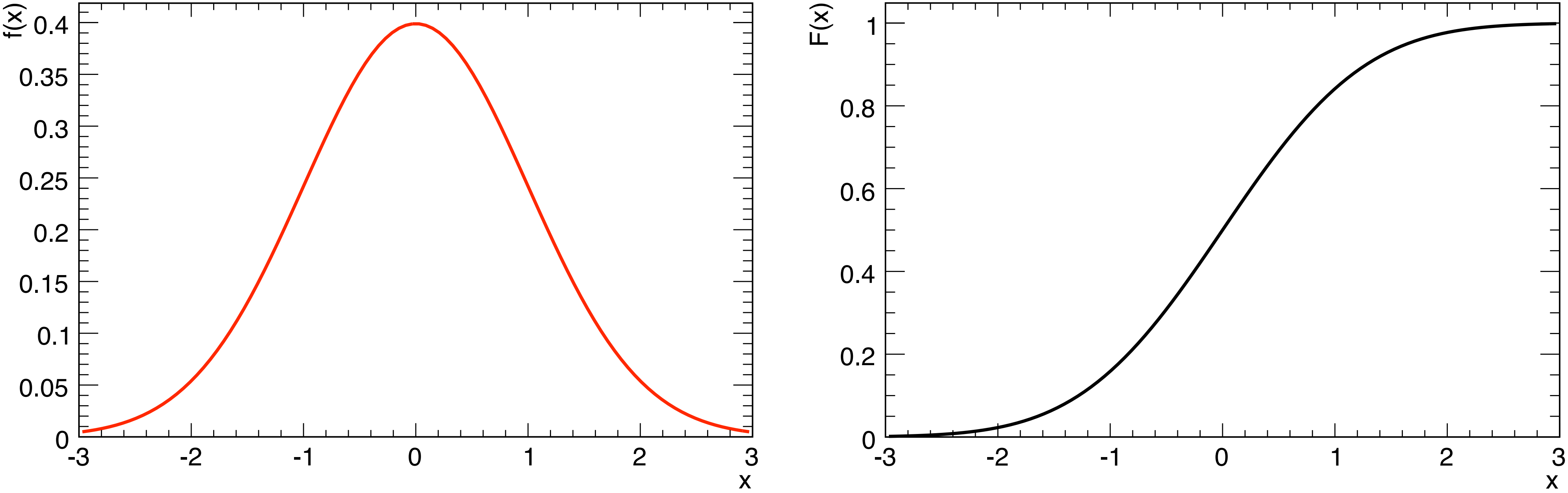




Cumulative Density Functions

Often useful to use a cumulative distribution:

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



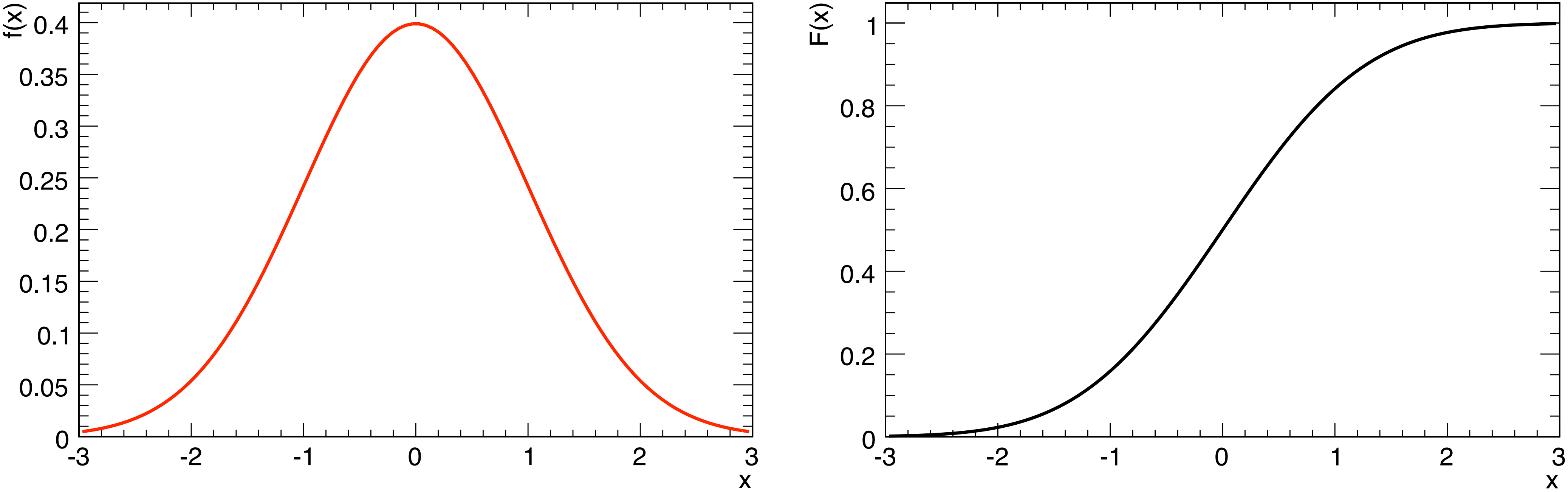
- alternatively, define density as partial of cumulative:

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

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:

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

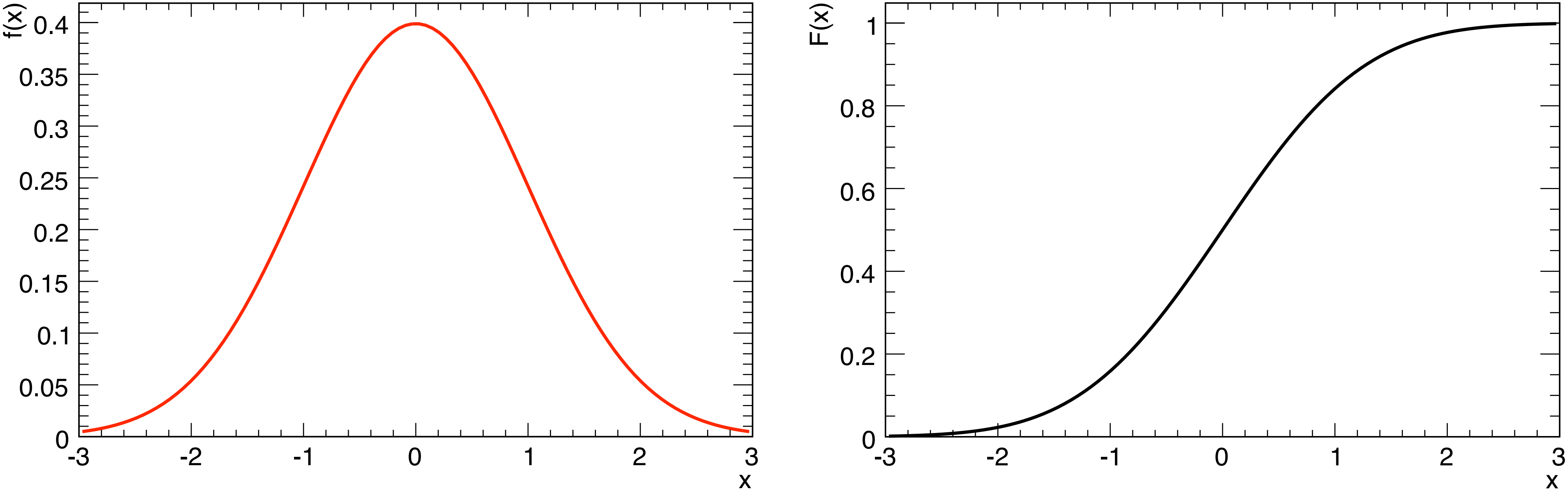
- similar to relationship of total and differential cross section:

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

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:

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

Impact of continuous variables & PDFs

Bayes' theorem basically untouched

$$f_X(x|Y=y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_Y(y|X=x)\,f_X(x)}{f_Y(y)} = \frac{f_Y(y|X=x)\,f_X(x)}{\int_{-\infty}^{\infty} f_Y(y|X=\xi)\,f_X(\xi)\,d\xi}.$$

- need to be careful that marginal PDFs are well behaved

Information theory

- the obvious generalization

$$h[f] = -\int\limits_{-\infty}^{\infty} f(x)\log f(x)\,dx, \quad (*)$$

- is not the continuous limit of the discrete case

$$h[f] = \lim_{\Delta\to 0}\left[H^{\Delta} + \log\Delta\right] = -\int_{-\infty}^{\infty} f(x)\log f(x)\,dx,$$

- it's not invariant to re–parameterizations

Parametric vs. Non-Parametric PDFs

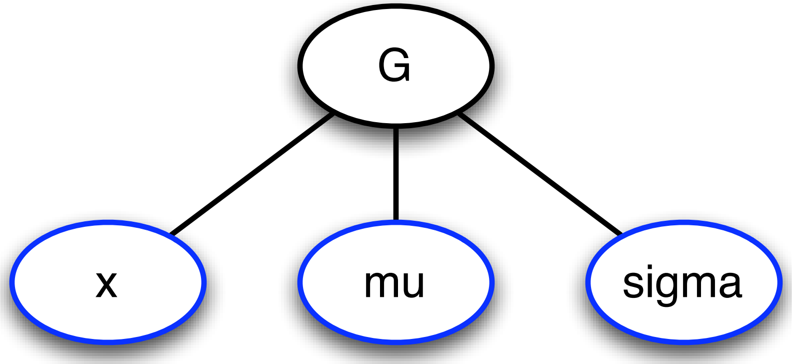
Many familiar pdfs are considered **parametric**

- 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**

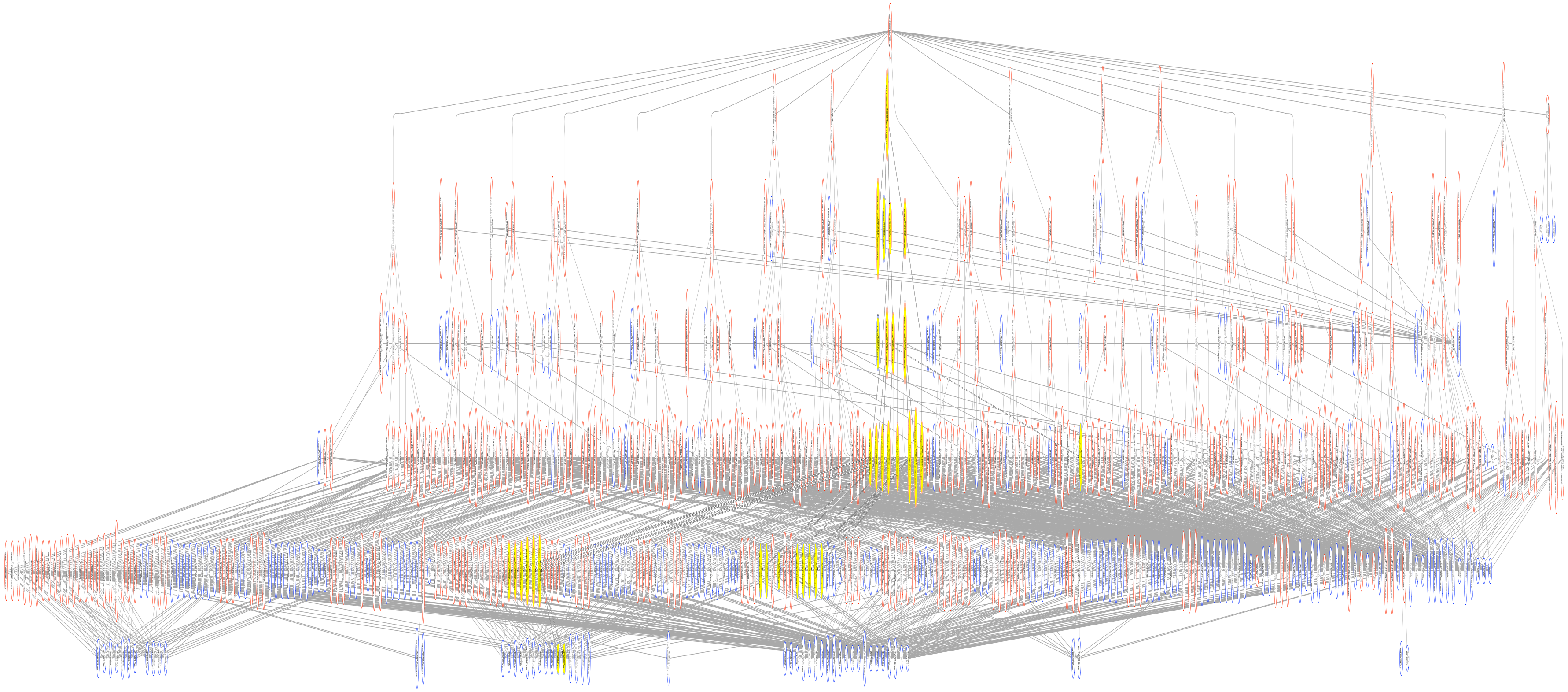
- eg. a Gaussian $G(x|\mu,\sigma)$ is parametrized by (μ,σ)
- defines a family of functions
- allows one to make inference about parameters
- some examples have very complicated parametric pdfs



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

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

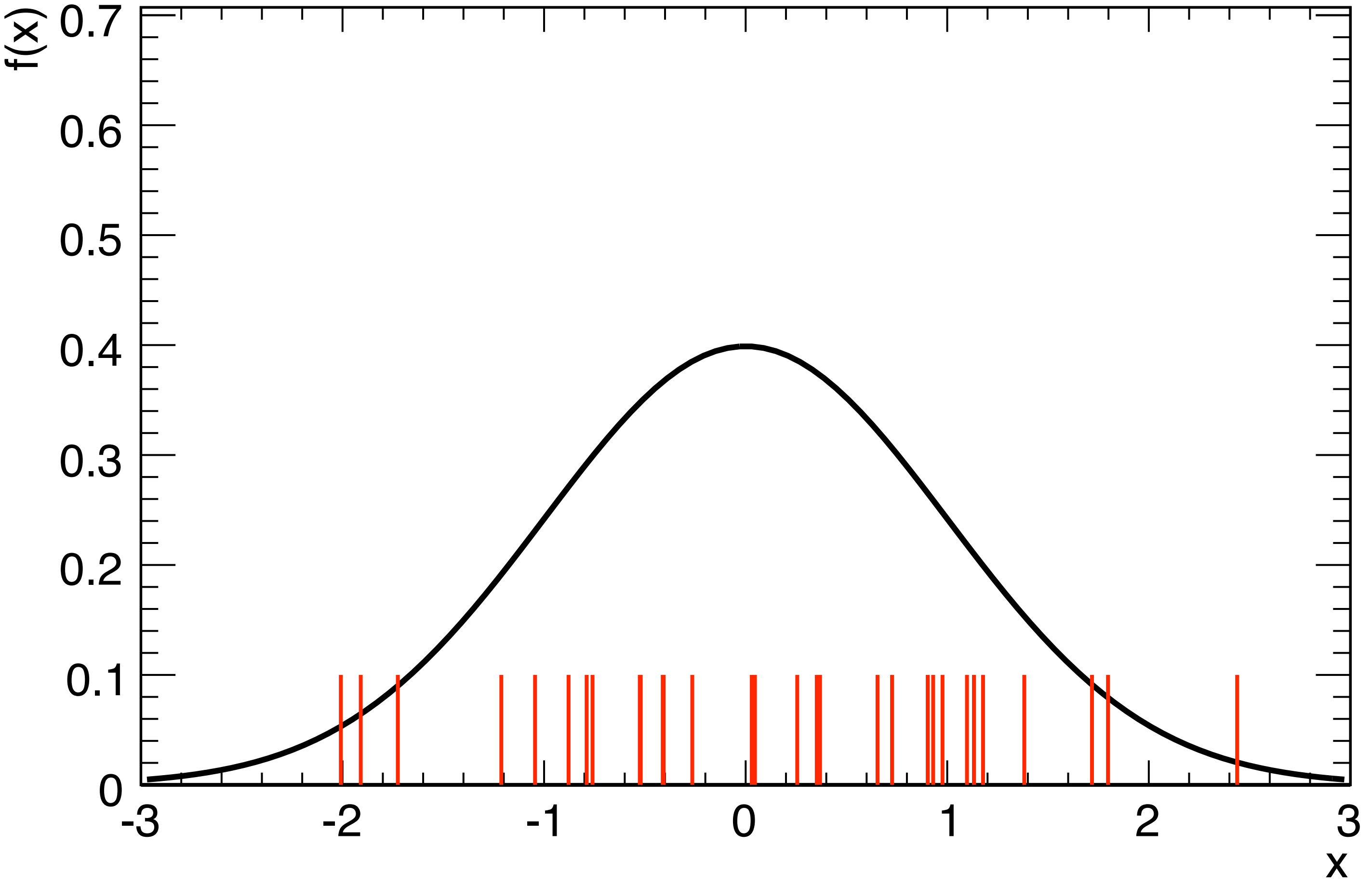
$$f_{emp} = \frac{1}{N} \sum_{i}^{N} \delta(x - x_i)$$

Parametric vs. Non-Parametric PDFs

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

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

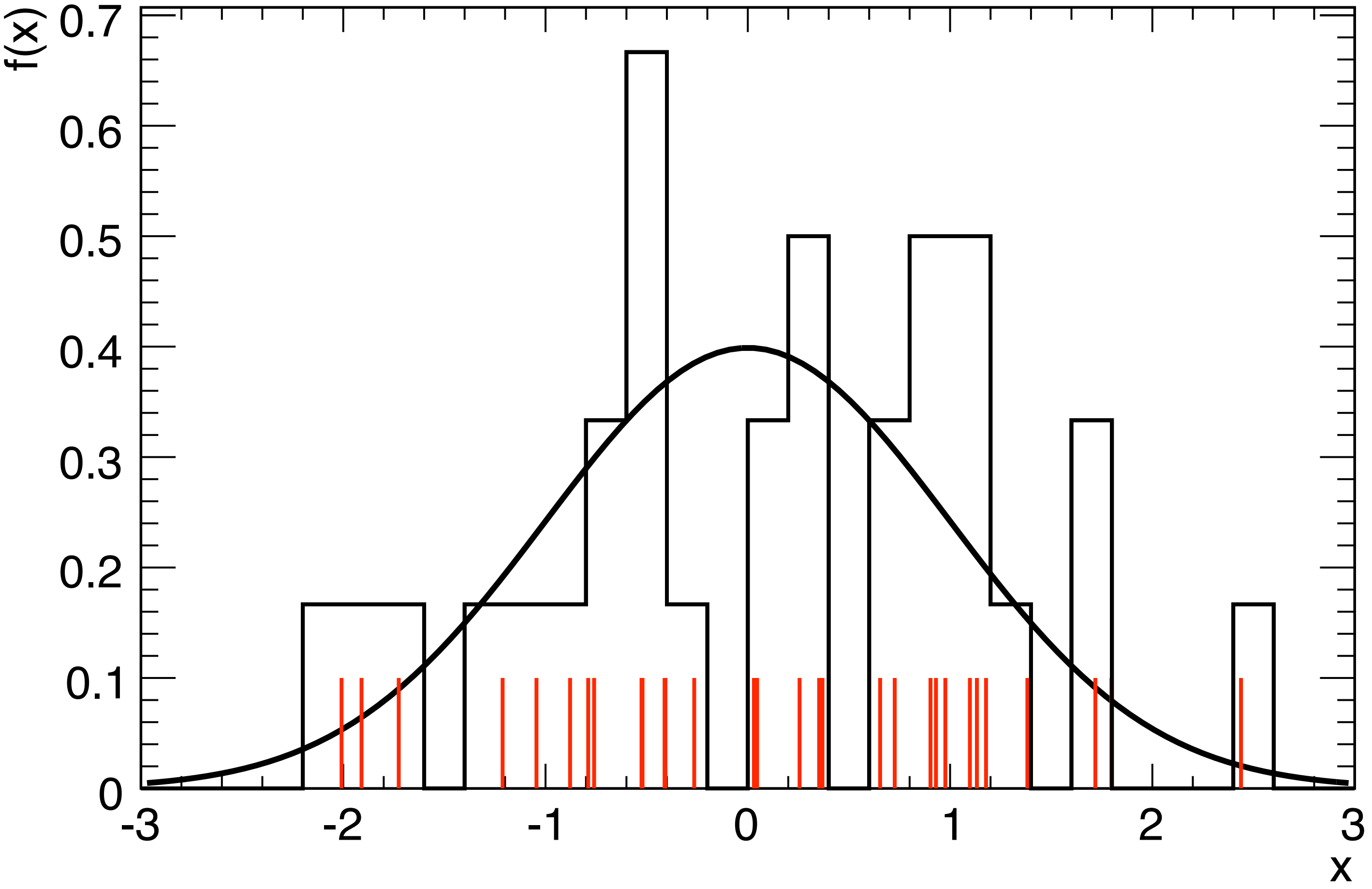
$$f^{w,s}_{hist}(x) = \frac{1}{N}\sum_i h_i^{w,s}$$

Parametric vs. Non-Parametric PDFs

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

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

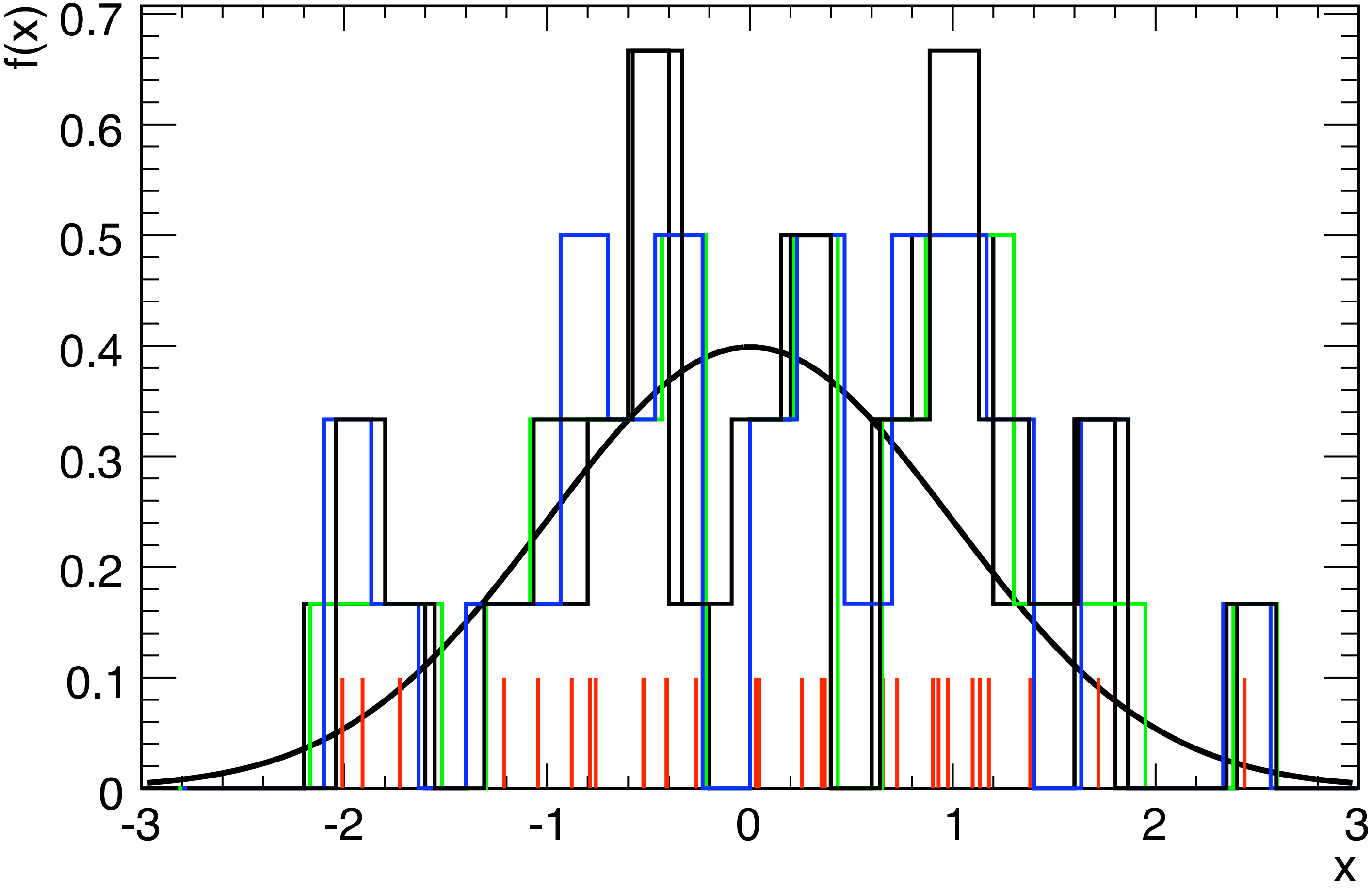
$$f_{hist}^{w,s}(x) = \frac{1}{N}\sum_i h_i^{w,s}$$

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

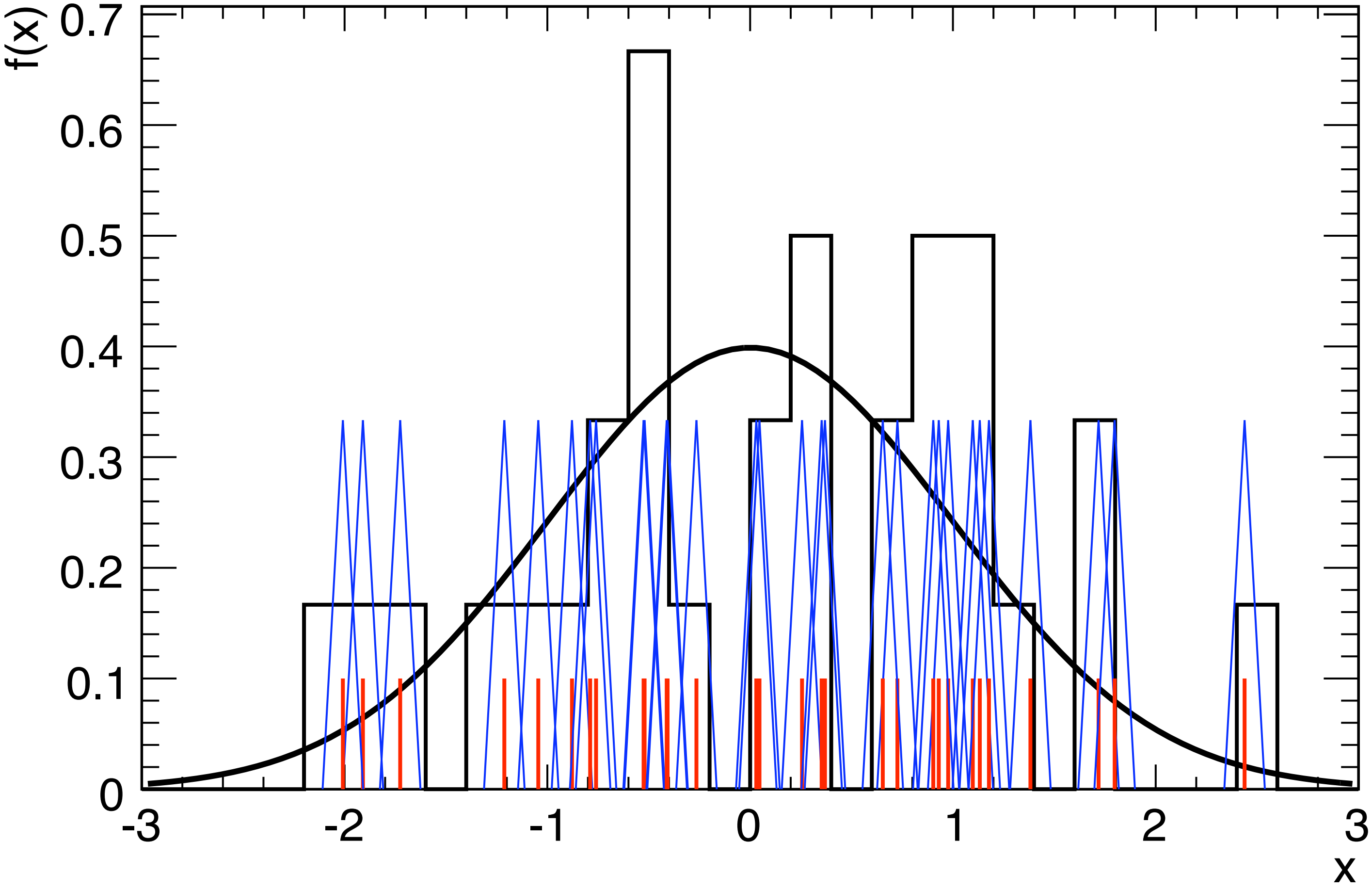
$$f^w_{ASH}(x) = \frac{1}{N}\sum_i^N K^w(x - x_i)$$

Parametric vs. Non-Parametric PDFs

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

"Average Shifted Histogram" minimizes effect of binning

$$f^w_{ASH}(x) = \frac{1}{N}\sum_i^N K^w(x - x_i)$$

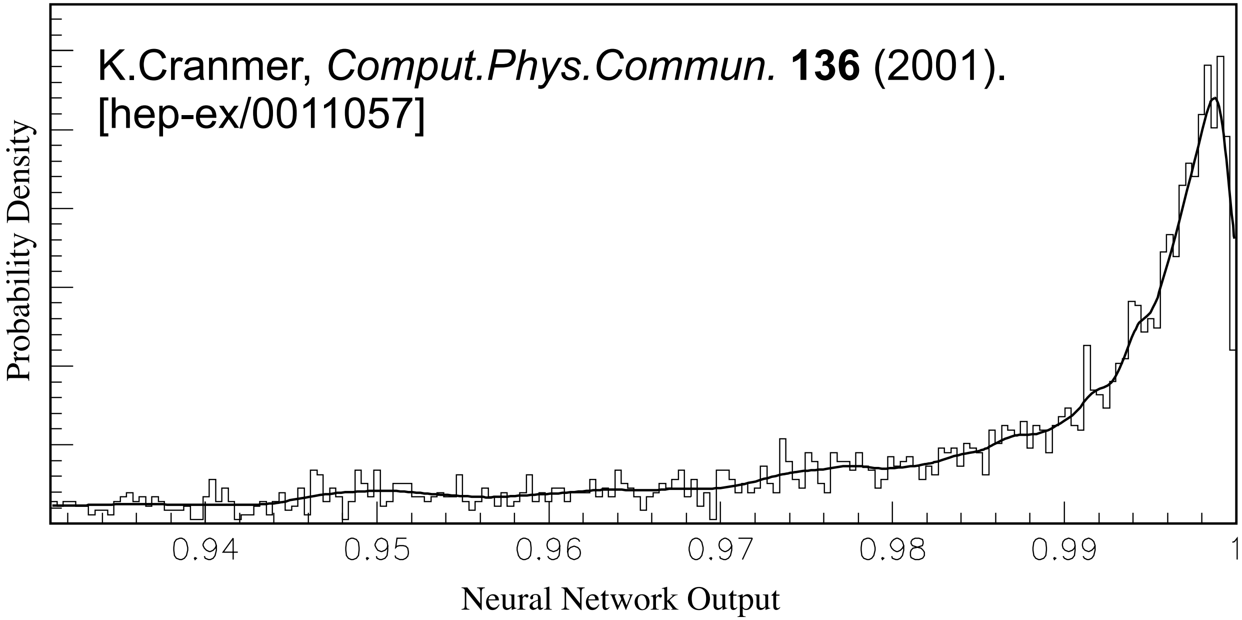


Kernel Estimation

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

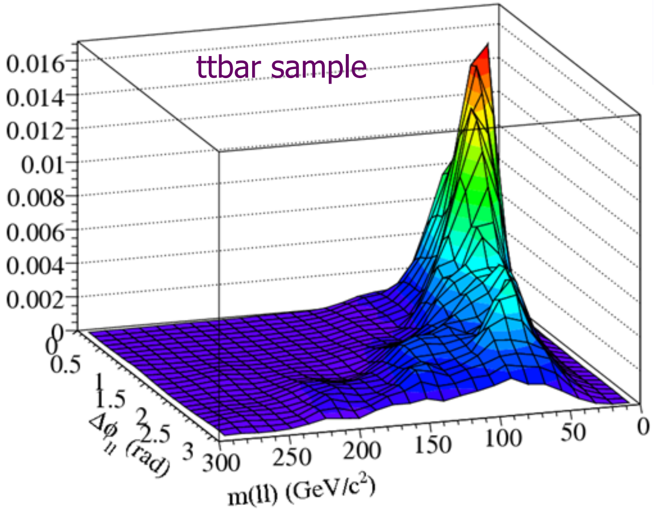
Max Baak has coded N-dim KEYS pdf described in Comput.Phys.Commun. **136** (2001) in RooFit.

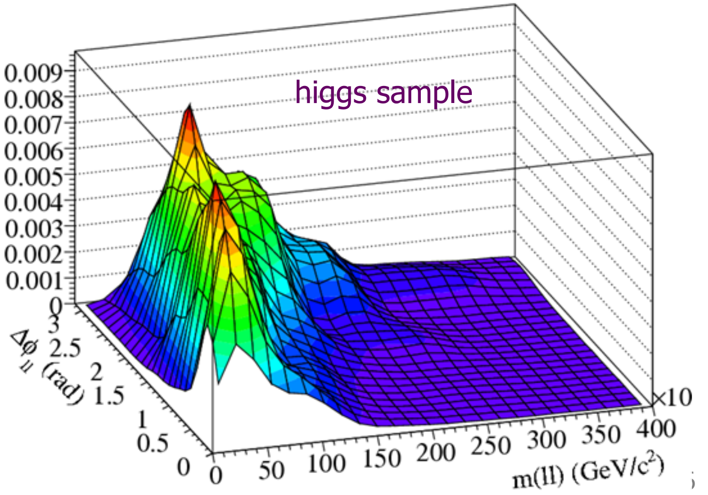
These pdfs have been used as the basis for a multivariate discrimination technique called "PDE"

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

Correlations

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





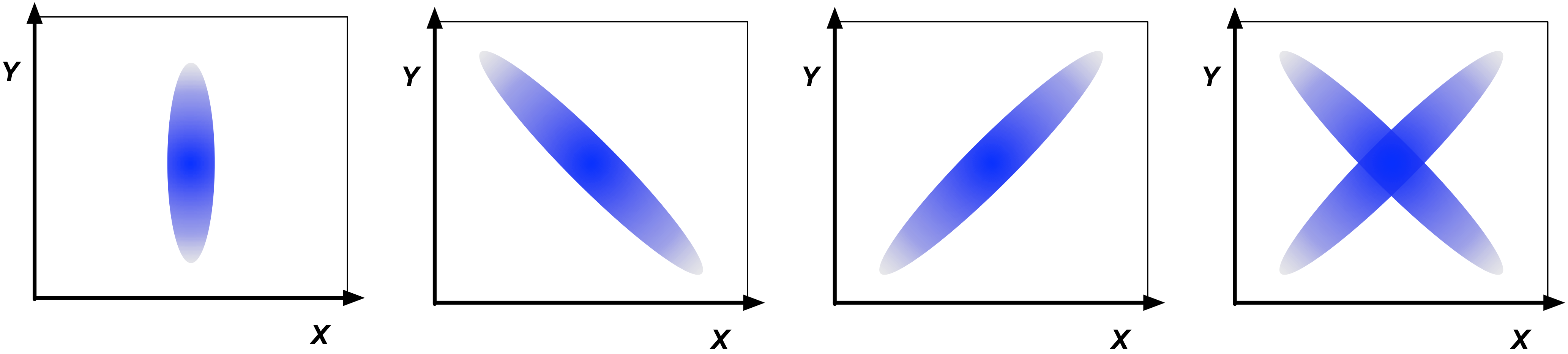
Max Baak

Correlation / Covariance

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

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

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



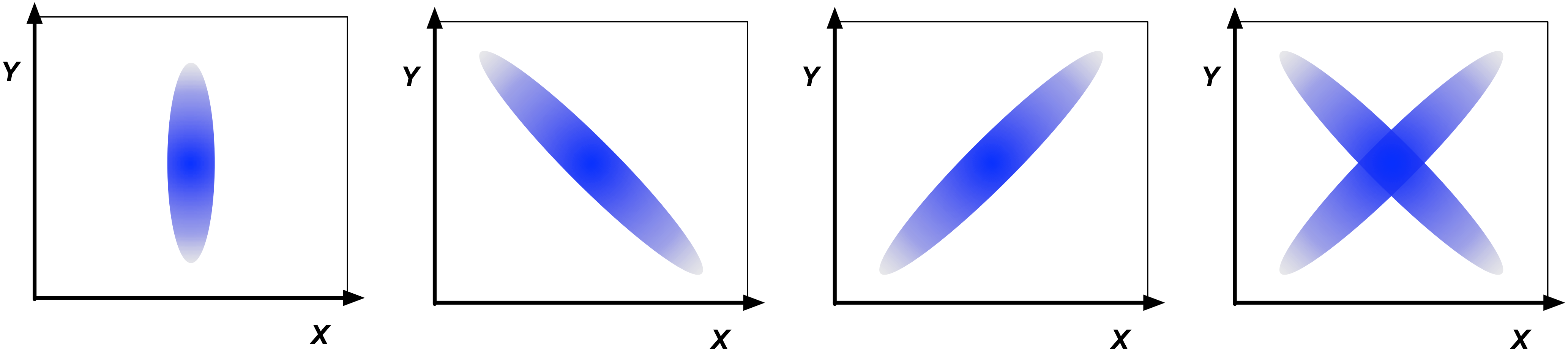
Correlation / Covariance

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

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

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



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

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

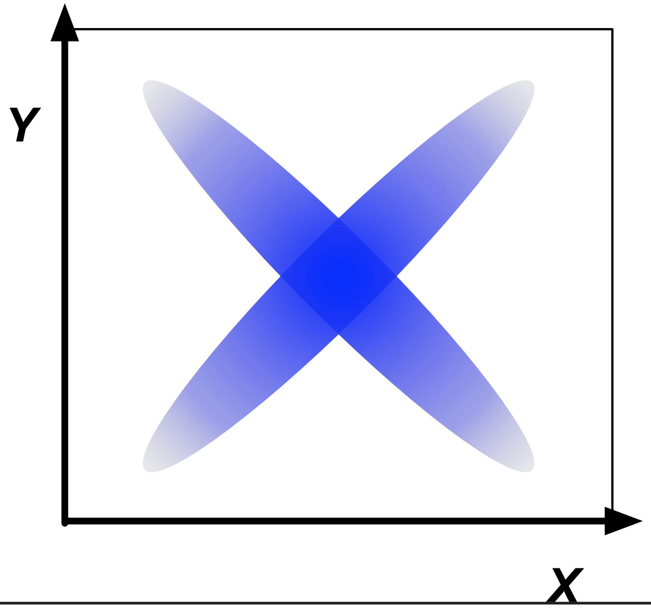
Mutual Information

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



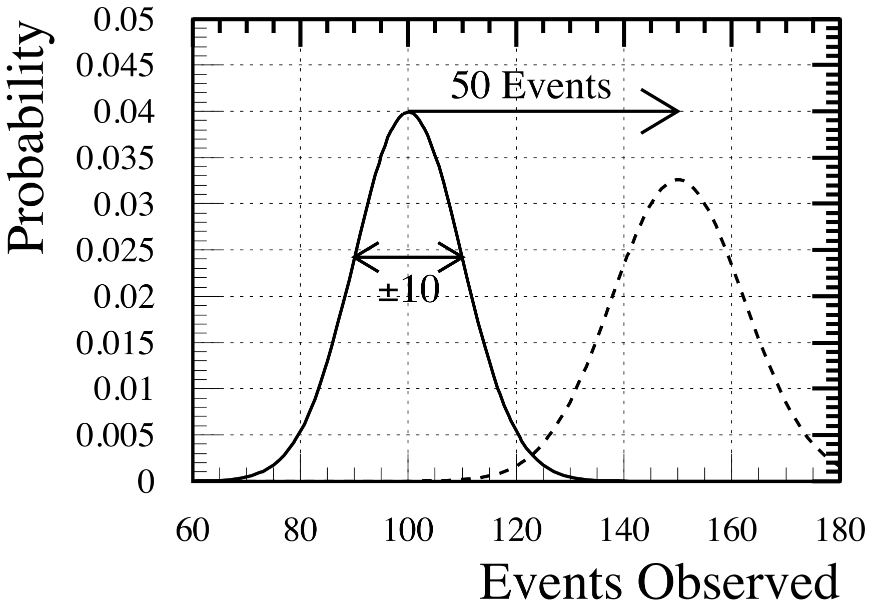
Mutual Information doesn't seem to be used much within HEP, but it seems quite useful

Hypothesis Testing

Hypothesis testing

One of the most common uses of statistics in particle physics is Hypothesis Testing

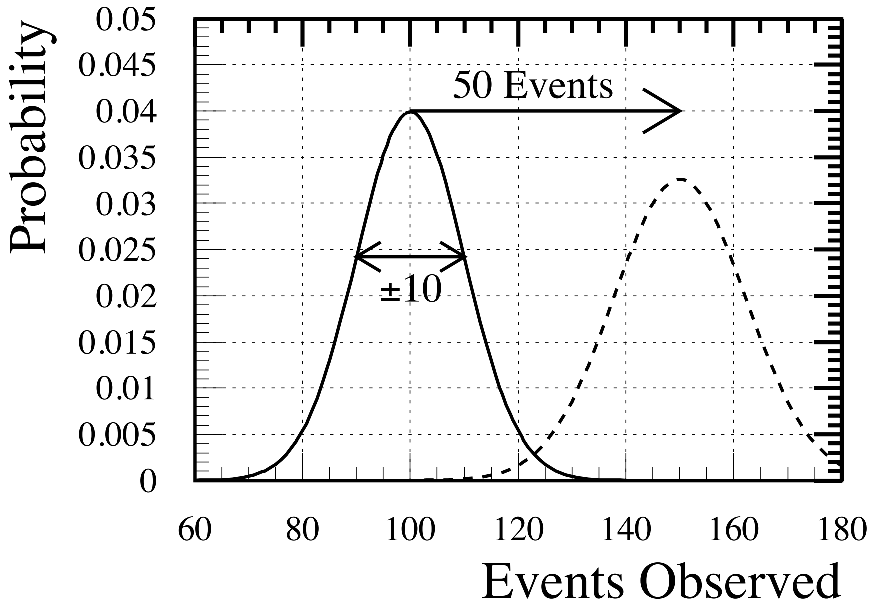
- assume one has pdf for data under two hypotheses:
  - Null-Hypothesis, H_0: eg. background-only
  - Alternate-Hypothesis H_1: eg. signal-plus-background



Hypothesis testing

One of the most common uses of statistics in particle physics is Hypothesis Testing

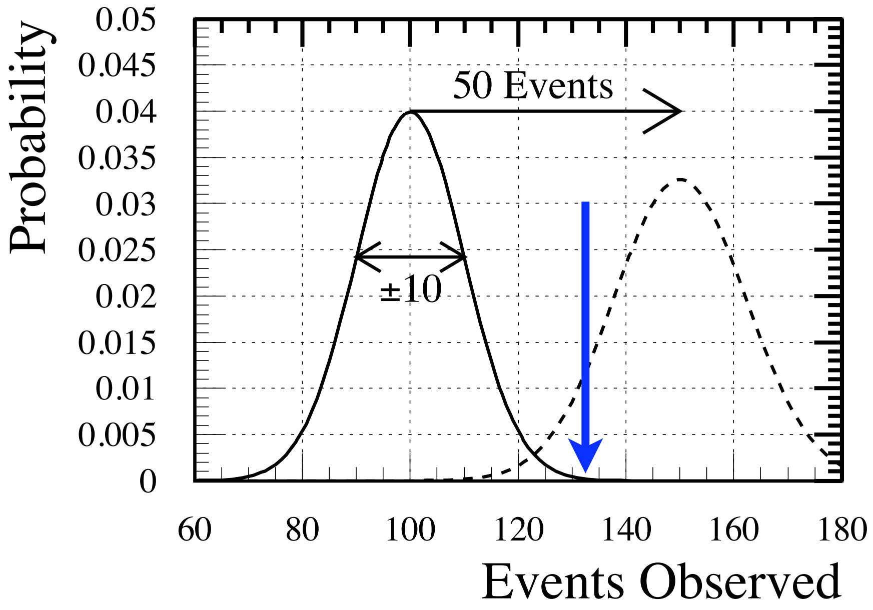
- assume one has pdf for data under two hypotheses:
  - Null-Hypothesis, H_0: eg. background-only
  - Alternate-Hypothesis H_1: eg. signal-plus-background
- one makes a measurement and then needs to decide whether to **reject** or **accept** H_0



Hypothesis testing

One of the most common uses of statistics in particle physics is Hypothesis Testing

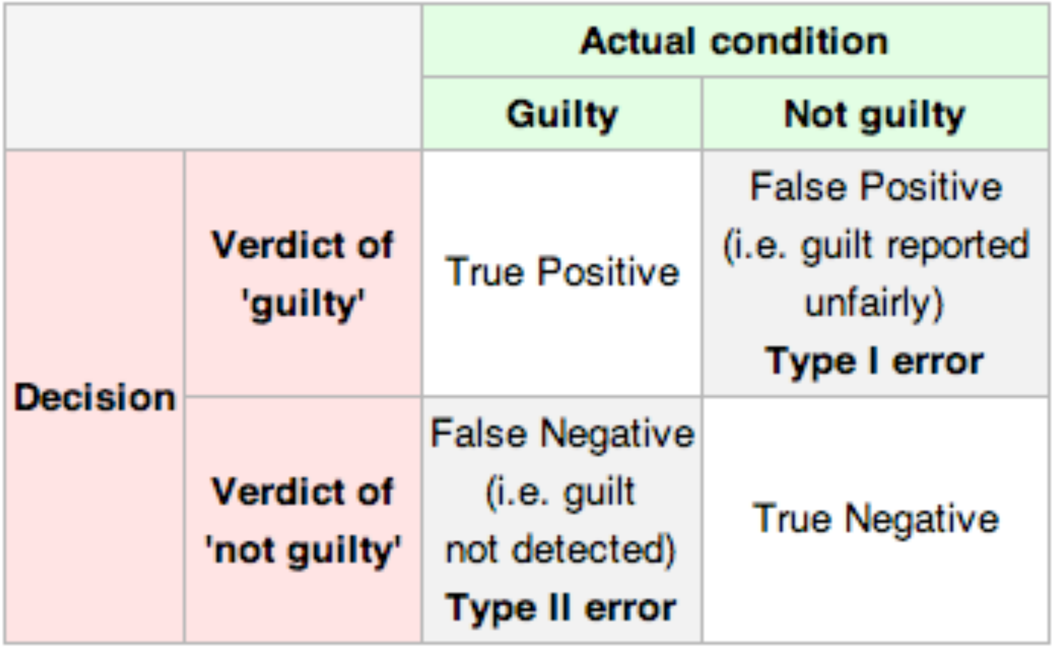
- assume one has pdf for data under two hypotheses:
  - Null-Hypothesis, H_0: eg. background-only
  - Alternate-Hypothesis H_1: eg. signal-plus-background
- one makes a measurement and then needs to decide whether to **reject** or **accept** H_0



Hypothesis testing

Before we can make much progress with statistics, we need to decide what it is that we want to do.

- first let us define a few terms:
  - Type I error: reject H_0 when it is true
  - Type II error: accept H_0 when H_1 is true
    - basically the same as “reject H_1 when H_1 is true”

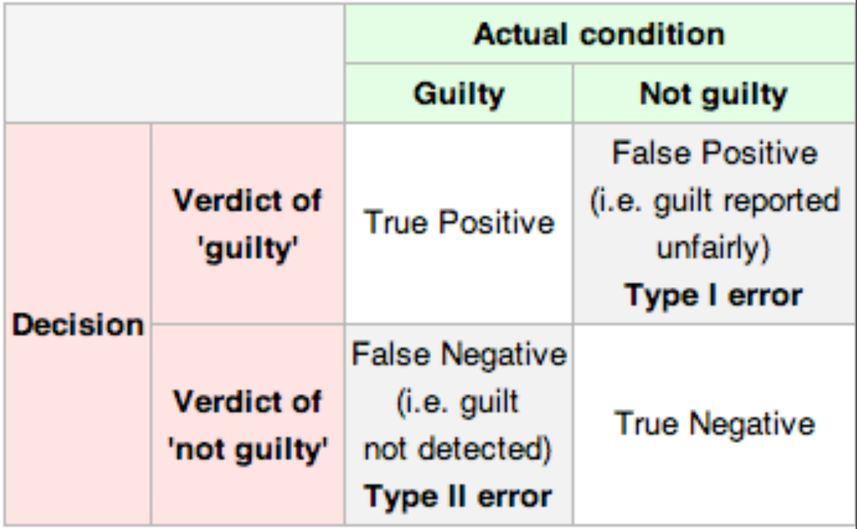


		Actual condition	
		Guilty	**Not guilty**
Decision	**Verdict of 'guilty'**	True Positive	False Positive (i.e. guilt reported unfairly) **Type I error**
	Verdict of 'not guilty'	False Negative (i.e. guilt not detected) **Type II error**	True Negative

Hypothesis testing

Before we can make much progress with statistics, we need to decide what it is that we want to do.

- first let us define a few terms:
  - Rate of Type I error α
  - Rate of Type II β
  - Power = $1 - \beta$



		Actual condition	
		Guilty	Not guilty
Decision	Verdict of 'guilty'	True Positive	False Positive (i.e. guilt reported unfairly) **Type I error**
	Verdict of 'not guilty'	False Negative (i.e. guilt not detected) **Type II error**	True Negative

Treat the two hypotheses asymmetrically

- the Null is special.
  - Fix rate of Type I error, call it “the size of the test”

Now one can state “a well-defined goal”

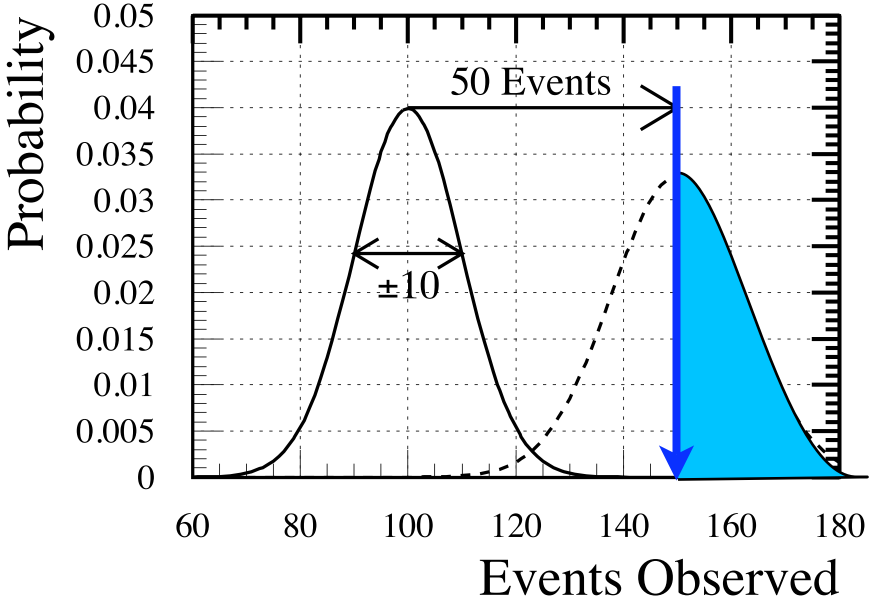
- Maximize power for a fixed rate of Type I error

Hypothesis testing

The idea of a “5σ“ discovery criteria for particle physics is really a conventional way to specify the size of the test

- usually 5σ corresponds to $\alpha = 2.87 \cdot 10^{-7}$
  - eg. a very small chance we reject the standard model

In the simple case of number counting it is obvious what region is sensitive to the presence of a new signal



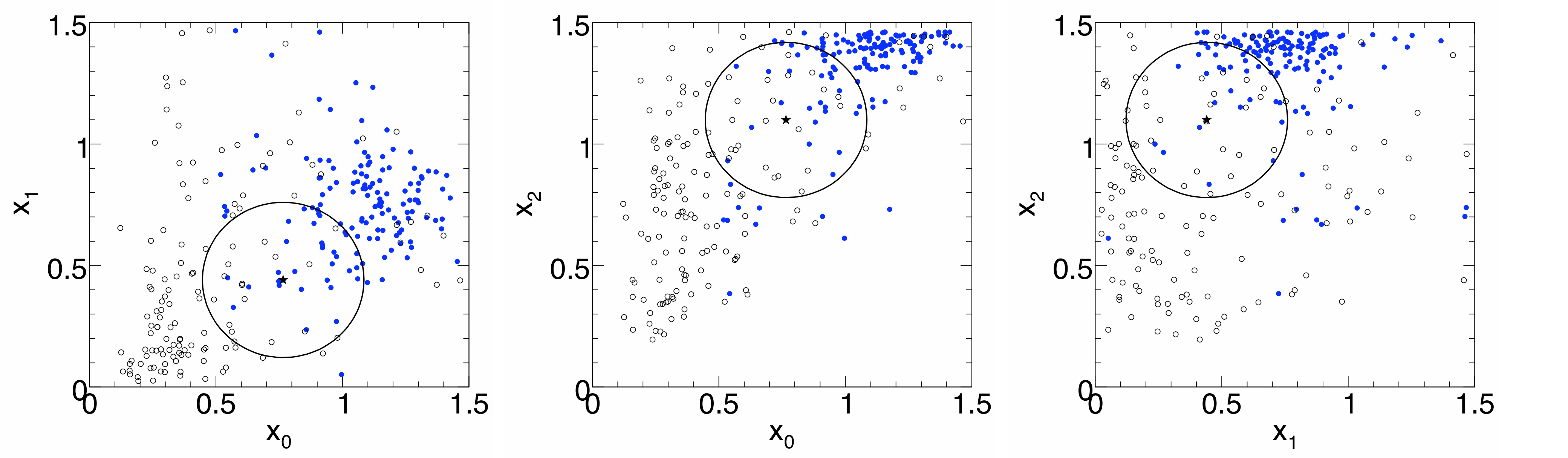
Hypothesis testing

The idea of a “5σ“ discovery criteria for particle physics is really a conventional way to specify the size of the test

- usually 5σ corresponds to $\alpha = 2.87 \cdot 10^{-7}$
  - eg. a very small chance we reject the standard model

In the simple case of number counting it is obvious what region is sensitive to the presence of a new signal

- but in higher dimensions it is not so easy



The Neyman-Pearson Lemma

In 1928-1938 Neyman & Pearson developed a theory in which one must consider competing Hypotheses:

- the Null Hypothesis H_0 (background only)
- the Alternate Hypothesis H_1 (signal-plus-background)

Given some probability that we wrongly reject the Null Hypothesis

$$\alpha = P(x \notin W | H_0)$$

Find the region W such that we minimize the probability of wrongly accepting the H_0 (when H_1 is true)

$$\beta = P(x \in W | H_1)$$

Note, if data falls in W then we accept H_0

The Neyman-Pearson Lemma

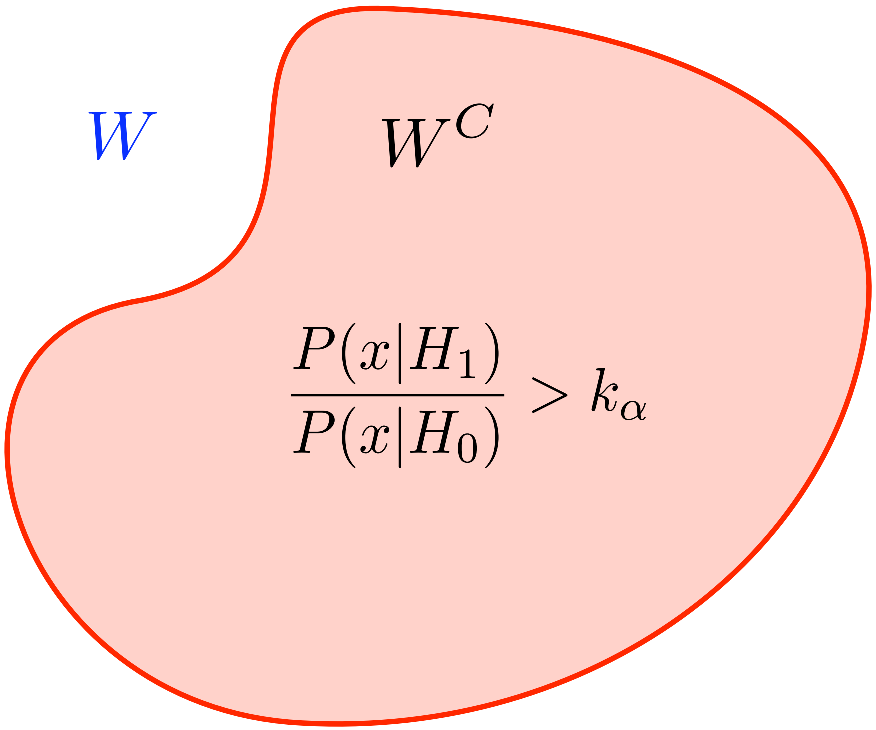
The region W that minimizes the probability of wrongly accepting H_0 is just a contour of the Likelihood Ratio

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

Any other region of the same size will have less power

The likelihood ratio is an example of a Test Statistic, eg. a real-valued function that summarizes the data in a way relevant to the hypotheses that are being tested

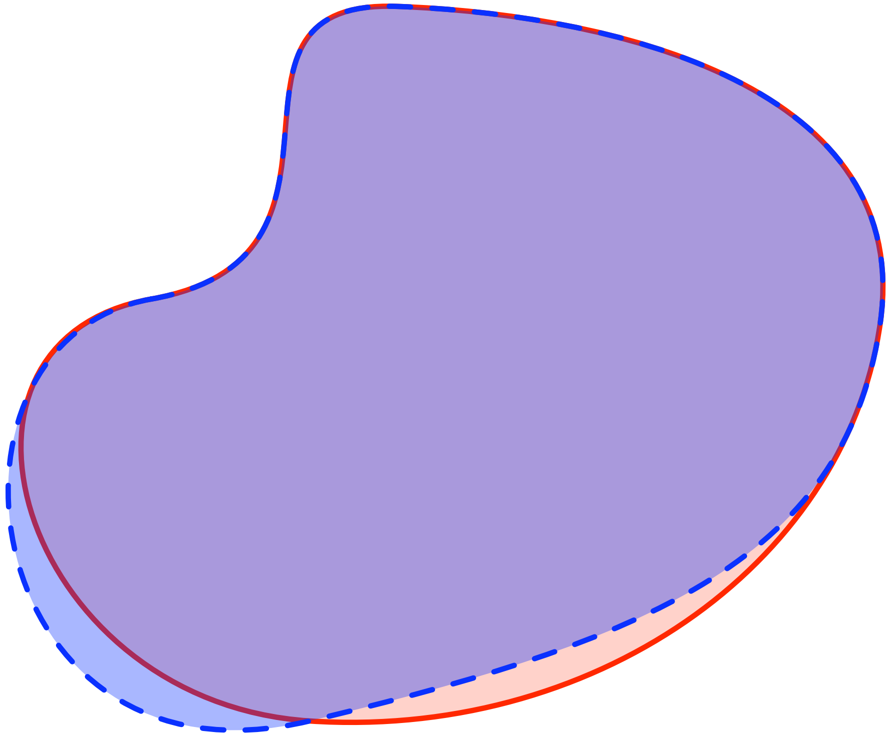
A short proof of Neyman-Pearson



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

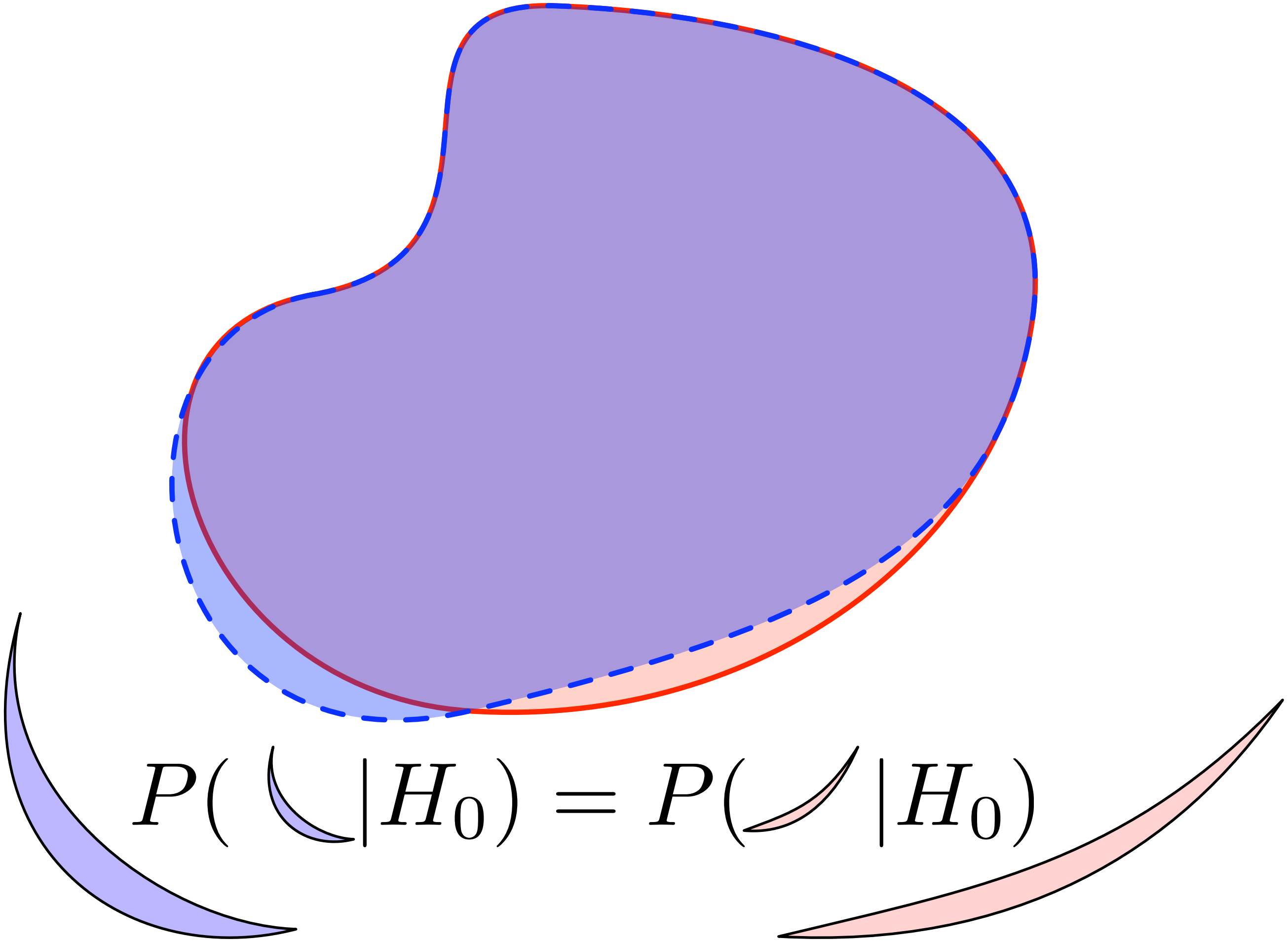
Consider the contour of the likelihood ratio that has size a given size (eg. probability under H_0 is $1-\alpha$)

A short proof of Neyman-Pearson



Now consider a variation on the contour that has the same size

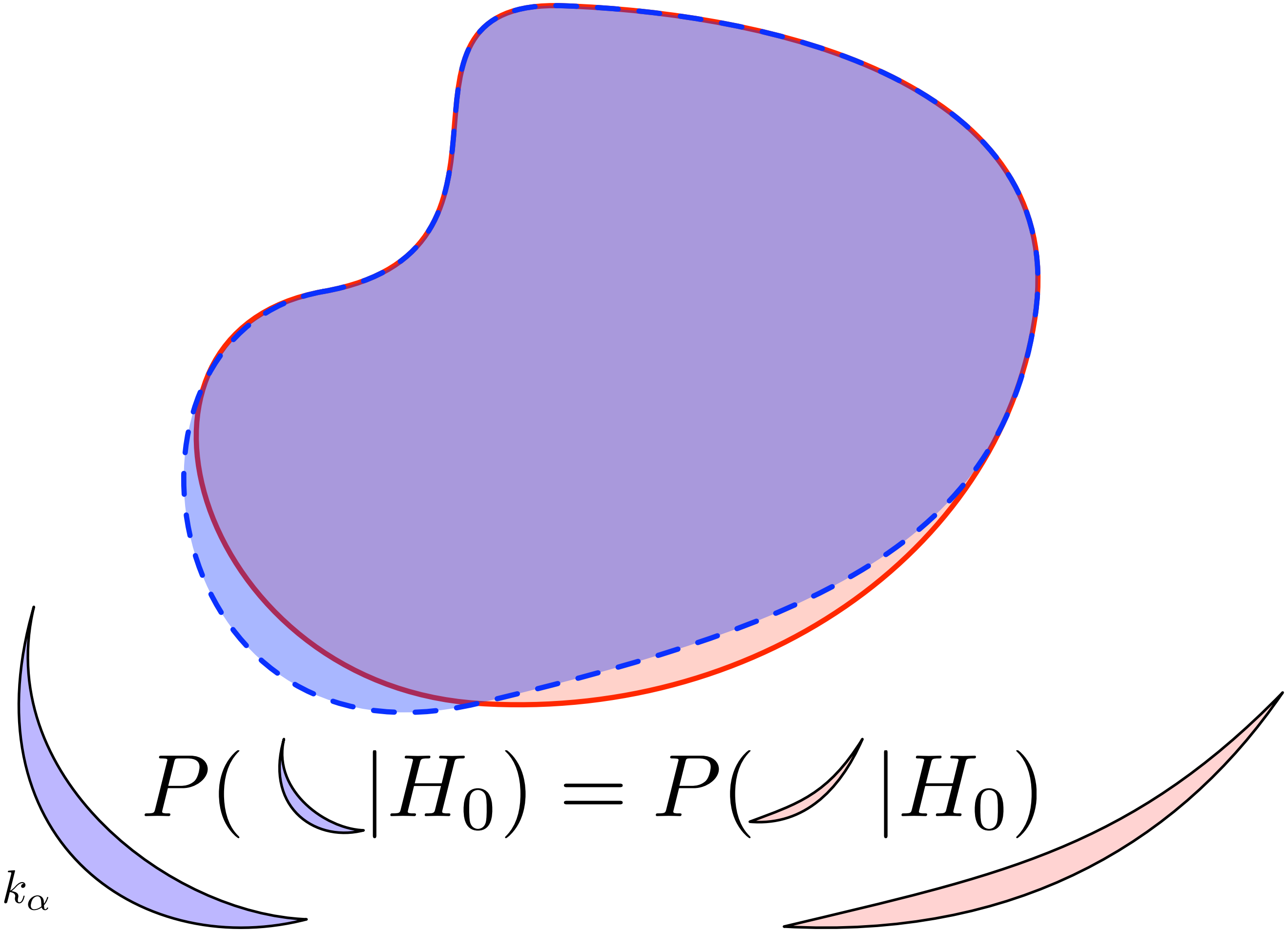
A short proof of Neyman-Pearson



$$P(\quad|H_0) = P(\quad|H_0)$$

Now consider a variation on the contour that has the same size (eg. same probability under H$_0$)

A short proof of Neyman-Pearson



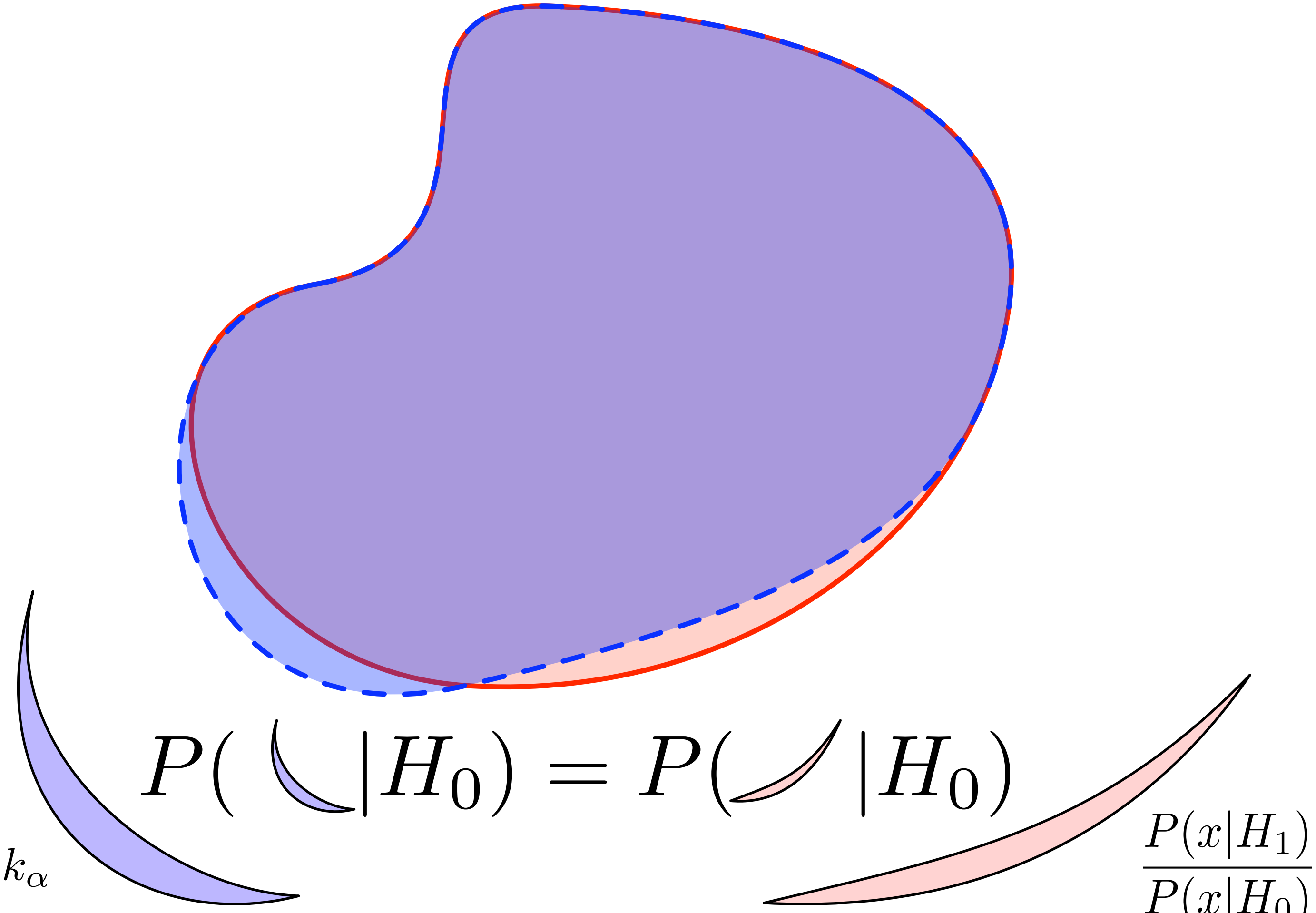
$$P(\text{◟}|H_0) = P(\text{◞}|H_0)$$

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

$$P(\text{◟}|H_1) < P(\text{◟}|H_0)k_\alpha$$

Because the new area is outside the contour of the likelihood ratio, we have an inequality

A short proof of Neyman-Pearson



$$P(\text{◟}|H_0) = P(\text{◞}|H_0)$$

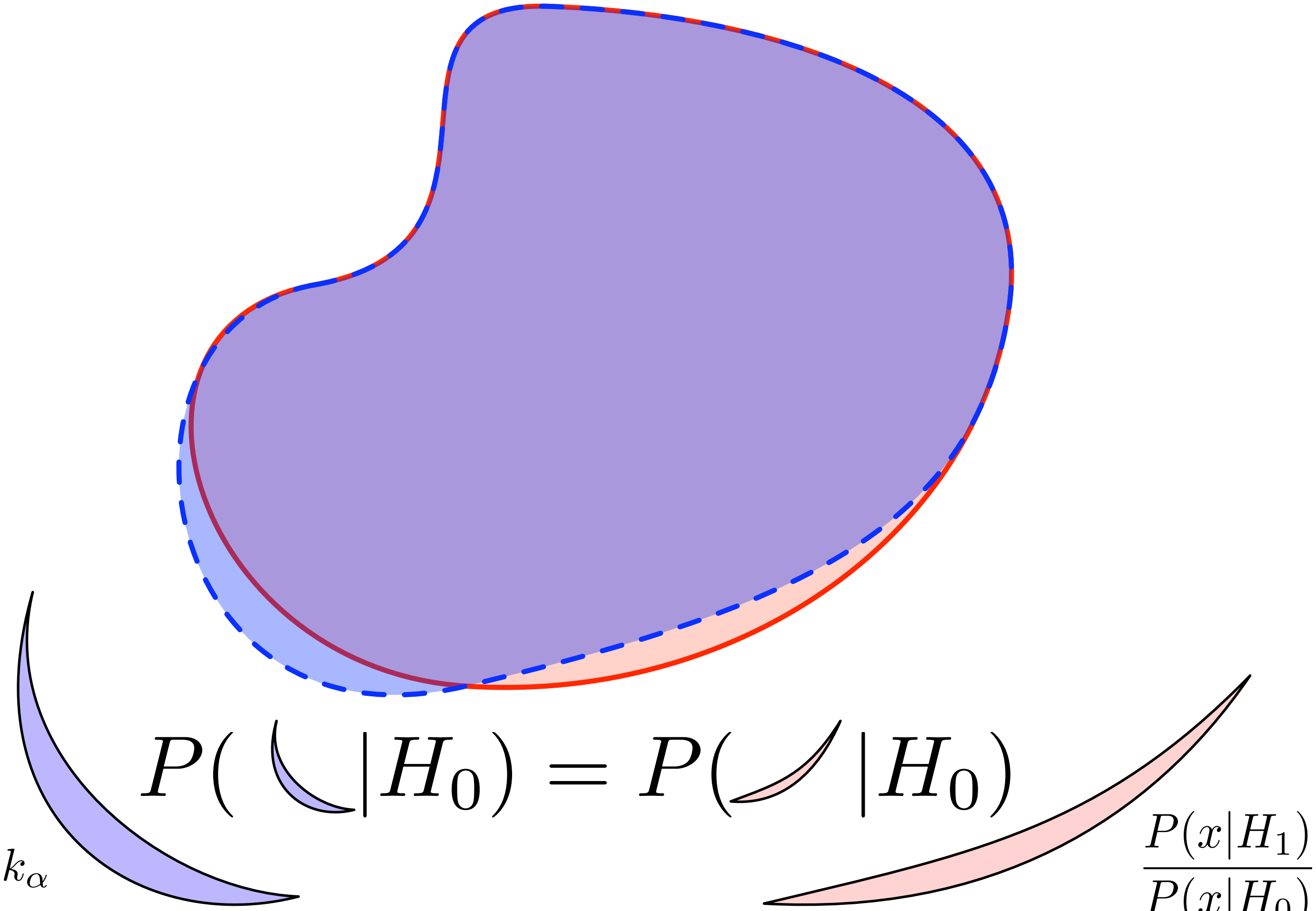
$$\frac{P(x|H_1)}{P(x|H_0)} < k_\alpha \qquad\qquad \frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

$$P(\text{◟}|H_1) < P(\text{◟}|H_0)k_\alpha \qquad\qquad P(\text{◞}|H_1) > P(\text{◞}|H_0)k_\alpha$$

And for the region we lost, we also have an inequality

Together they give...

A short proof of Neyman-Pearson



$$P(\text{◟}|H_0) = P(\text{◞}|H_0)$$

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

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

$$P(\text{◟}|H_1) < P(\text{◟}|H_0)k_\alpha$$

$$P(\text{◞}|H_1) > P(\text{◞}|H_0)k_\alpha$$

$$\boxed{P(\text{◟}|H_1) < P(\text{◞}|H_1)}$$

The new region region has less power.

Decision Theory

One of the deficiencies of the Neyman–Pearson approach is that one must specify the size of the test α

- But where does α come from?
  - is it purely conventional or is there a reason?

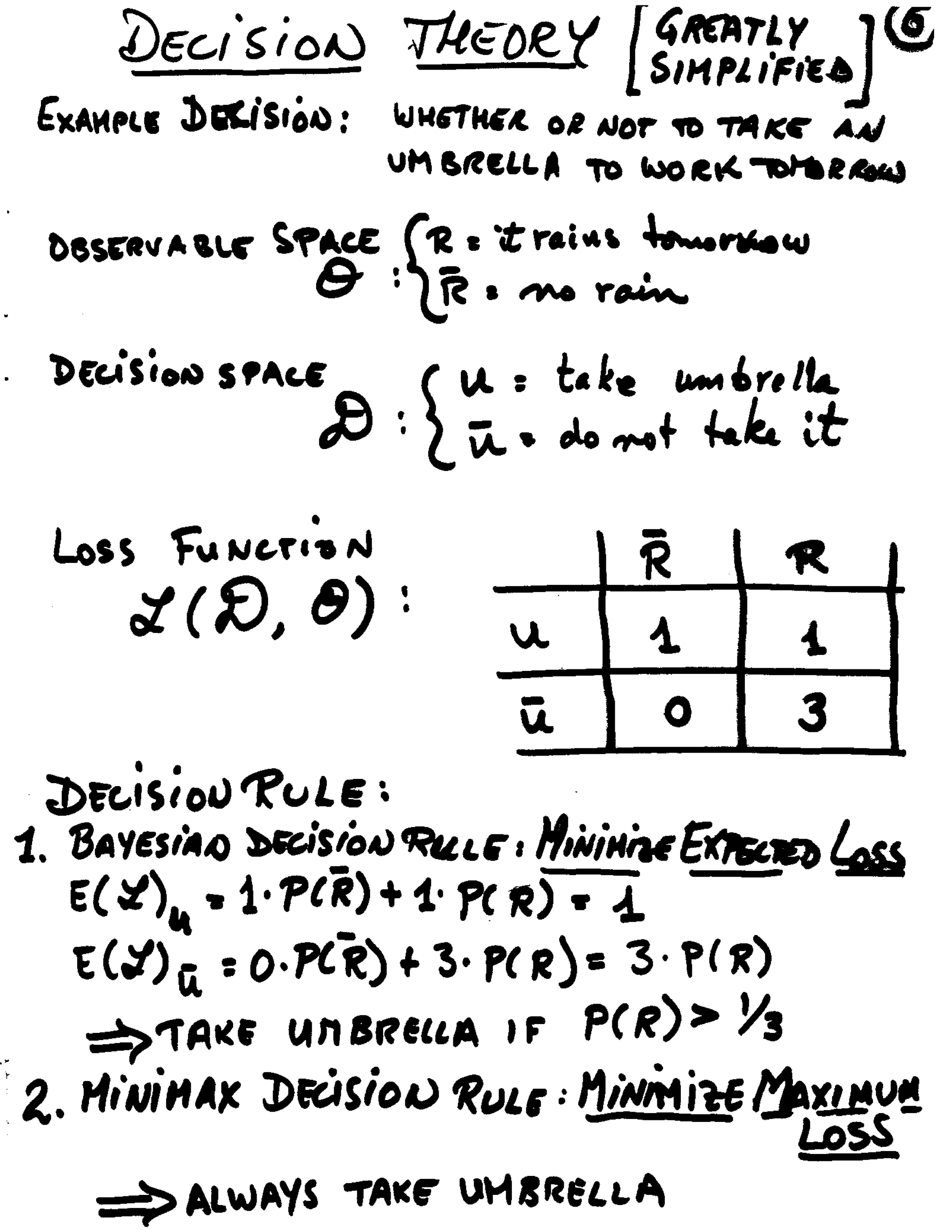
Much of statistics (and economics, etc.) is devoted to making **decisions**.

- need to consider **Utility** of different outcomes

In the context of decision and utility theory there can be a justification, but this is rarely done in particle physics

Decision Theory

From Fred James lectures



DECISION THEORY [GREATLY SIMPLIFIED] 6

EXAMPLE DECISION: WHETHER OR NOT TO TAKE AN UMBRELLA TO WORK TOMORROW

OBSERVABLE SPACE Θ : $\begin{cases} R = \text{it rains tomorrow} \\ \bar{R} = \text{no rain} \end{cases}$

DECISION SPACE $\mathcal{D}$: $\begin{cases} u = \text{take umbrella} \\ \bar{u} = \text{do not take it} \end{cases}$

LOSS FUNCTION $\mathcal{L}(\mathcal{D}, \Theta)$:

	$\bar{R}$	R
u	1	1
$\bar{u}$	0	3

DECISION RULE:

1. BAYESIAN DECISION RULE: MINIMIZE EXPECTED LOSS

$$E(\mathcal{L})_u = 1 \cdot P(\bar{R}) + 1 \cdot P(R) = 1$$

$$E(\mathcal{L})_{\bar{u}} = 0 \cdot P(\bar{R}) + 3 \cdot P(R) = 3 \cdot P(R)$$

$\Rightarrow$ TAKE UMBRELLA IF $P(R) > 1/3$

2. MINIMAX DECISION RULE: MINIMIZE MAXIMUM LOSS

$\Rightarrow$ ALWAYS TAKE UMBRELLA

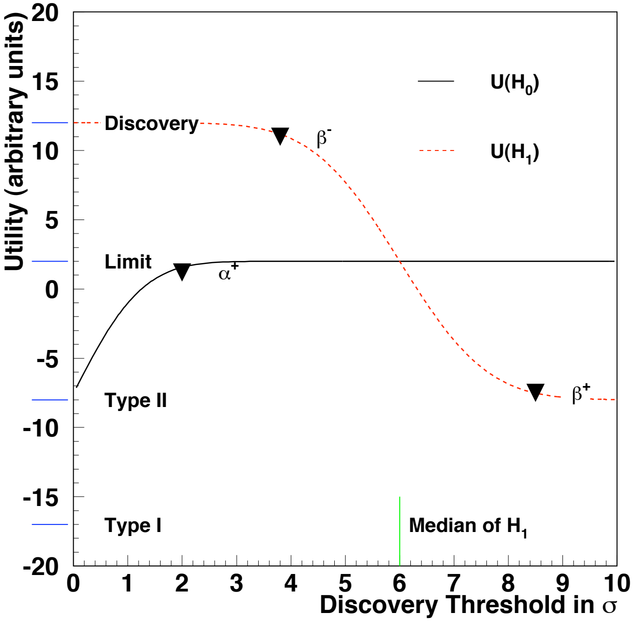
One take on “Why 5σ?”: Utility Theory

(taken from my thesis)

Instead of arguing about convention, derive threshold from utility theory:

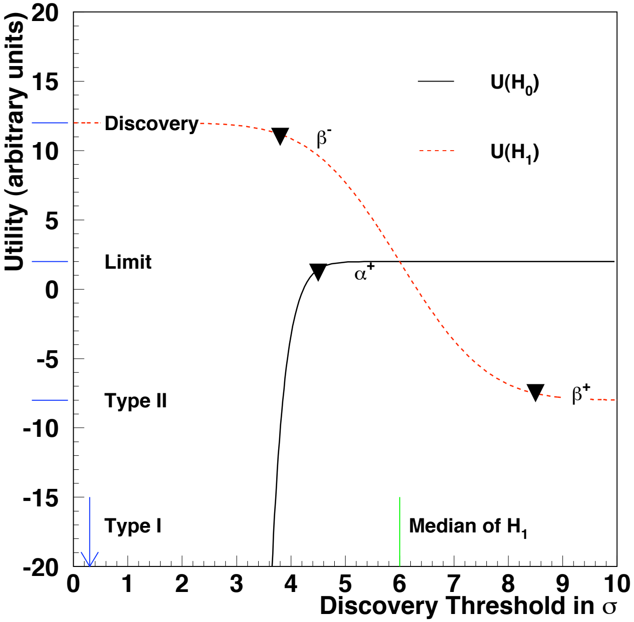
- assumptions of Game Theory not appropriate
- let size of the test for discovery be α and for limit setting be α’

$$U(H_0) = (1-\alpha')\cdot U(\text{Type I}) + \beta'\cdot U(\text{Limit}) + (\beta-\beta')\cdot U(\text{No Result})$$
$$U(H_1) = (1-\beta)\cdot U(\text{Discovery}) + \beta'\cdot U(\text{Type II}) + (\beta-\beta')\cdot U(\text{No Result}).$$



With a prior on H₀/H₁ one can use a richer decision theory. But in a frequentist way, one obtains:

$$\alpha^+ = \epsilon\left[1-\frac{U(\text{Type I})}{U(\text{Limit})}\right]^{-1}$$



Ideally, the field would establish these utilities instead of working with the purely conventional 5σ requirement. Since that is not the case, it is reasonable to ask “what is this ratio of utilities which justifies a 5σ discovery threshold?” If we take $\epsilon = 1\%$ and $\alpha = 10^{-7}$, then $|U(\text{Type I})/U(\text{Limit})| > 10^5$. Perhaps this ratio is reasonable, perhaps not, but it is the ratio under which we operate today.

Figure 2.5: Utility as a function of the discovery threshold for a channel w

Simple vs. Compound Hypotheses

The Neyman–Pearson lemma is **the answer** for simple hypothesis testing

- a hypothesis is **simple** if it has no free parameters and is totally fixed $f(x|H_0)$ vs. $f(x|H_1)$

What about cases when there are free parameters?

- eg. the mass of the Higgs boson $f(x|H_0)$ vs. $f(x|H_1, m_H)$

A test is called **similar** if it has size α for all values of the parameters

A test is called **Uniformly Most Powerful** if it maximizes the power for all values of the parameter

Uniformly Most Powerful tests don't exist in general

The Likelihood Function

When a hypothesis is composite typically there is a pdf that can be parametrized $f(\vec{x}|\theta)$

- for a fixed θ it defines a pdf for the random variable x
- for a given measurement of x one can consider $f(\vec{x}|\theta)$ as a function of θ called the **Likelihood function**
- Note, this is not Bayesian, because it still only uses P(data | theory) and
  - **the Likelihood function is not a pdf!**

Sometimes θ has many components, generally divided into:

- parameters of interest: eg. masses, cross-sections, etc.
- nuisance parameters: eg. parameters that affect the shape but are not of direct interest (eg. energy scale)
  - more tomorrow when I discuss systematics

A common likelihood function

Consider an experiment with multiple channels indexed by i

Each channel has n_i events indexed by j

- with s_i signal and b_i background expected

Each event has discriminating variables x_{ij} (possibly N-dim)

- with $f_s(x_{ij})$ and $f_b(x_{ij})$ describing signal & bkg components
- and assume signal and background don't interfere quantum mechanically, so that the probabilities just add

Then one can write the following pdf / likelihood function

$$L(x_{ij}|s_i, b_i) = \prod_i^{N_{chan}} Pois(n_i|s_i + b_i) \prod_j^{n_i} \frac{s_i f_s(x_{ij}) + b_i f_b(x_{ij})}{s_i + b_i}$$

A common likelihood function

Consider an experiment with multiple channels indexed by i

Each channel has n_i events indexed by j

- with s_i signal and b_i background expected

Each event has discriminating variables x_{ij} (possibly N-dim)

- with $f_s(x_{ij})$ and $f_b(x_{ij})$describing signal & bkg components
- and assume signal and background don't interfere quantum mechanically, so that the probabilities just add

Then one can write the following pdf / likelihood function

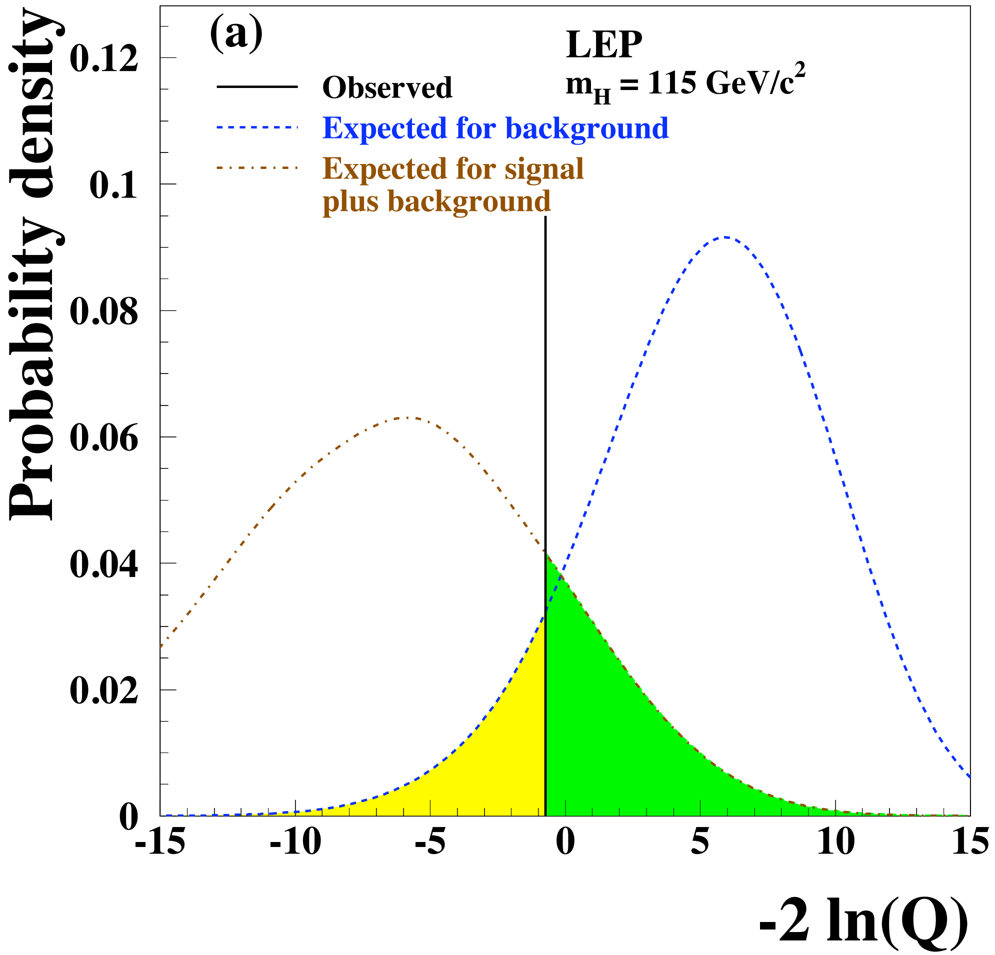
$$L(x_{ij}|s_i, b_i, \nu_i) = \prod_i^{N_{chan}} Pois(n_i|s_i + b_i) \prod_j^{n_i} \frac{s_i f_s(x_{ij};\nu_i) + b_i f_b(x_{ij};\nu_i)}{s_i + b_i}$$

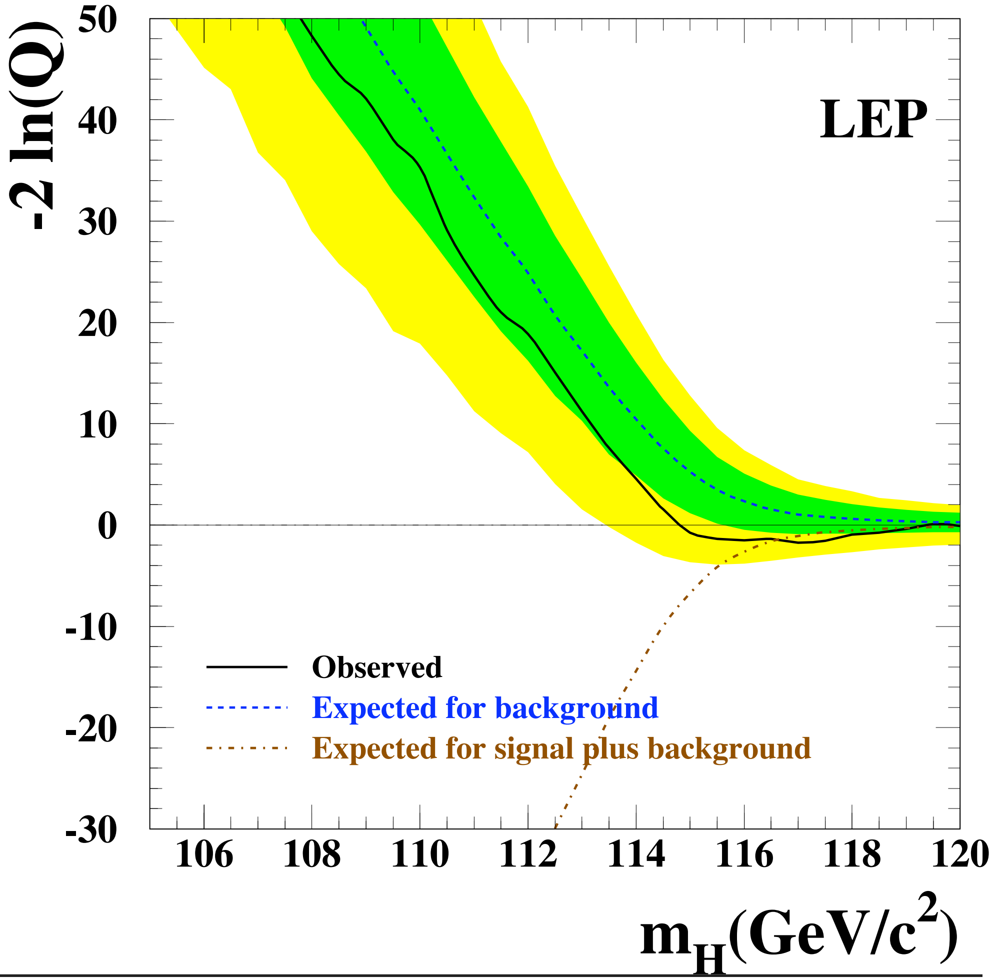
An explicit likelihood ratio

In that case:

$$Q = \frac{L(x|H_1)}{L(x|H_0)} = \frac{\prod_i^{N_{chan}} Pois(n_i|s_i+b_i) \prod_j^{n_i} \frac{s_i f_s(x_{ij}) + b_i f_b(x_{ij})}{s_i+b_i}}{\prod_i^{N_{chan}} Pois(n_i|b_i) \prod_j^{n_i} f_b(x_{ij})}$$

$$q = \ln Q = -s_{tot} \sum_i^{N_{chan}} \sum_j^{n_i} \ln\left(1 + \frac{s_i f_s(x_{ij})}{b_i f_b(x_{ij})}\right)$$



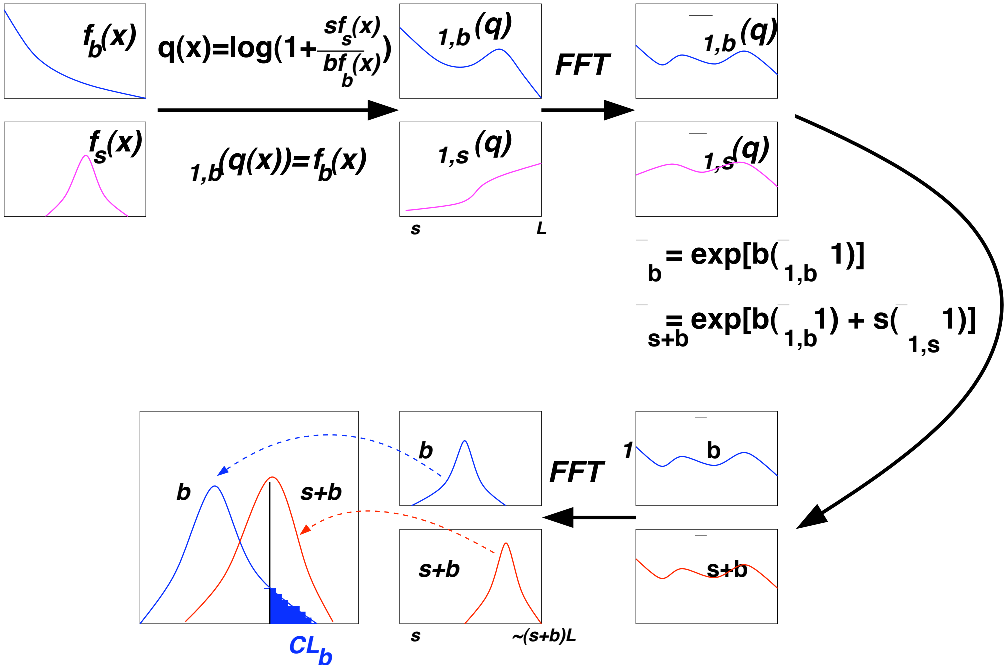


Distribution of the test statistic

LEP Higgs Working group developed formalism to combine channels and take advantage of discriminating variables in the likelihood ratio.

$$Q = \frac{L(x|H_1)}{L(x|H_0)} = \frac{\prod_i^{N_{chan}} Pois(n_i|s_i+b_i) \prod_j^{n_i} \frac{s_i f_s(x_{ij}) + b_i f_b(x_{ij})}{s_i+b_i}}{\prod_i^{N_{chan}} Pois(n_i|b_i) \prod_j^{n_i} f_b(x_{ij})}$$

$$q = \ln Q = -s_{tot} \sum_i^{N_{chan}} \sum_j^{n_i} \ln \left(1 + \frac{s_i f_s(x_{ij})}{b_i f_b(x_{ij})} \right)$$

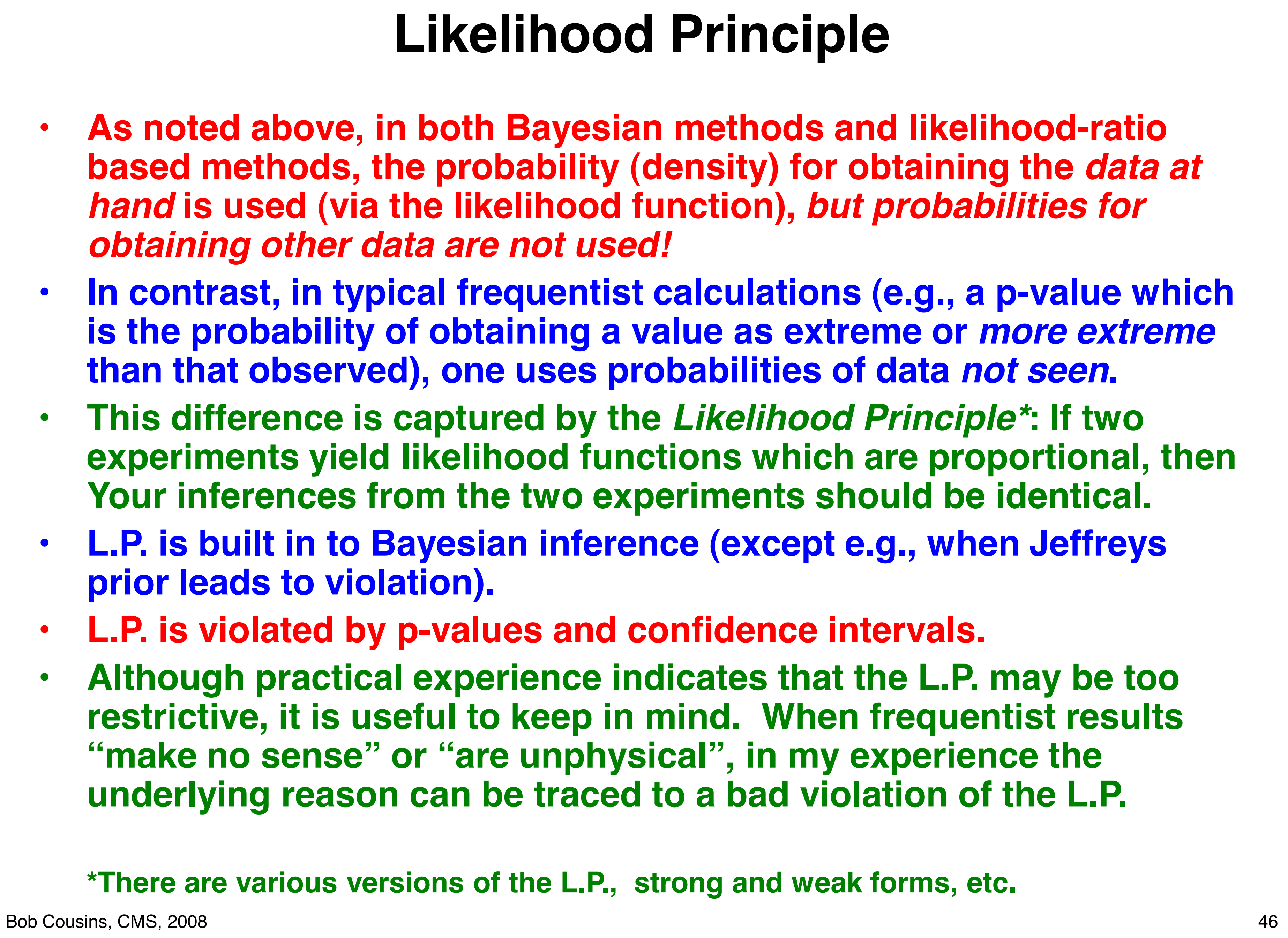


Hu and Nielsen's CLFFT used Fourier Transform and exponentiation trick to transform the log-likelihood ratio distribution for one event to the distribution for an experiment

Cousins-Highland was used for systematic error on background rate.

Getting this to work at the LHC is tricky numerically because we have channels with n_i from 10-10000 events (physics/0312050)

The Likelihood Principle



Likelihood Principle

- As noted above, in both Bayesian methods and likelihood-ratio based methods, the probability (density) for obtaining the *data at hand* is used (via the likelihood function), *but probabilities for obtaining other data are not used!*
- In contrast, in typical frequentist calculations (e.g., a p-value which is the probability of obtaining a value as extreme or *more extreme* than that observed), one uses probabilities of data *not seen*.
- This difference is captured by the *Likelihood Principle**: If two experiments yield likelihood functions which are proportional, then Your inferences from the two experiments should be identical.
- L.P. is built in to Bayesian inference (except e.g., when Jeffreys prior leads to violation).
- L.P. is violated by p-values and confidence intervals.
- Although practical experience indicates that the L.P. may be too restrictive, it is useful to keep in mind. When frequentist results "make no sense" or "are unphysical", in my experience the underlying reason can be traced to a bad violation of the L.P.

*There are various versions of the L.P., strong and weak forms, etc.

Bob Cousins, CMS, 2008 46

Examples of Likelihood Analysis

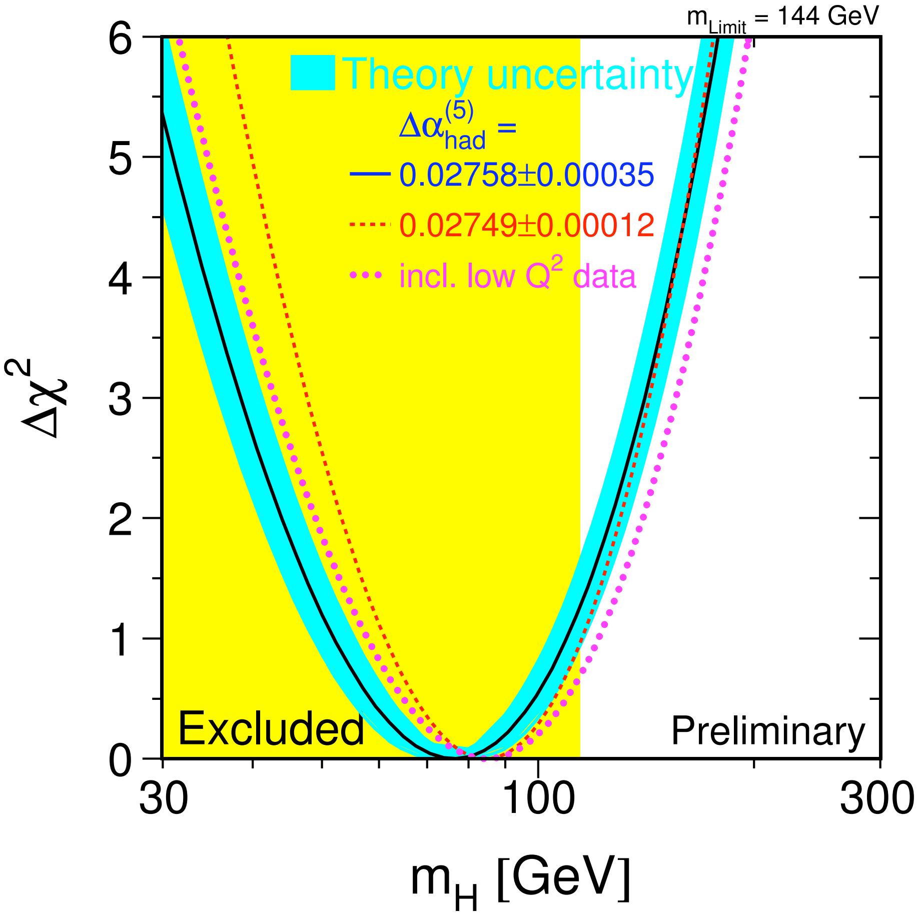
In these examples, a model that relates precision electroweak observables to parameters of the Standard Model was used

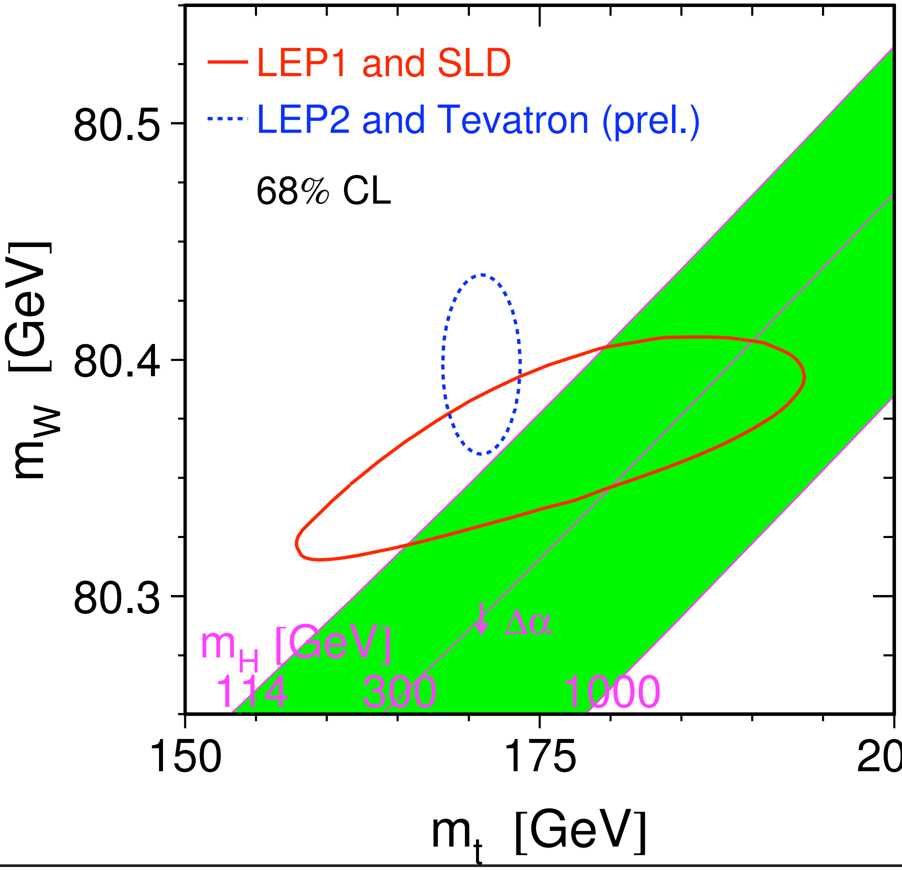
- the inference is based only on the likelihood function
  - there is no prior, so it’s not Bayesian
  - not a classical confidence interval either: discuss tomorrow

Examples of Likelihood Analysis

In these examples, a model that relates precision electroweak observables to parameters of the Standard Model was used

- the inference is based only on the likelihood function
  - there is no prior, so it’s not Bayesian
  - not a classical confidence interval either: discuss tomorrow





A Note on Multivariate Algorithms

Use of Multivariate Methods

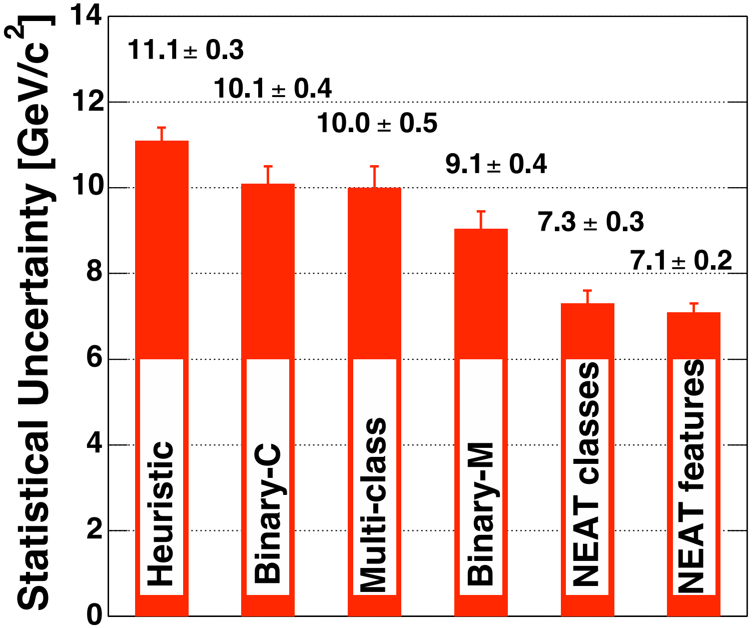
Multivariate methods are now ubiquitous in high-energy physics, the nagging problem is that:

- most multivariate techniques are borrowed from other fields, and they optimize some heuristic that physicists aren't interested in (like a score, or ad hoc training error)
- the difference can be quite large when systematic uncertainties are taken into account

A few recent developments

- Evolutionary techniques
- Matrix Element techniques

Whiteson & Whiteson, hep-ex/0607012



The Neyman-Pearson Lemma

The region W that minimizes the probability of wrongly accepting the H_0 is just a contour of the Likelihood Ratio:

$$\frac{L(x|H_0)}{L(x|H_1)} > k_\alpha$$

This is the goal!

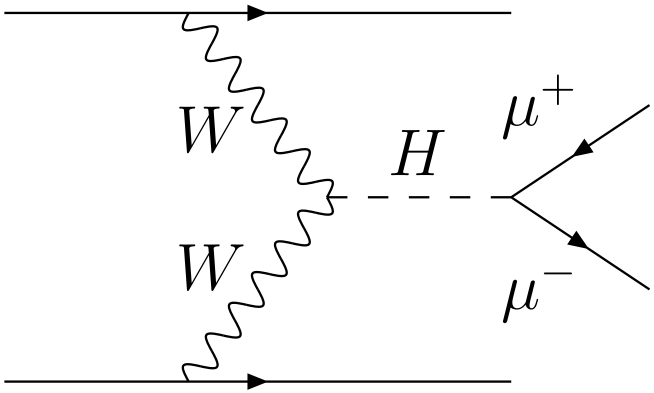
The problem is we don't have access to $L(x|H_0)$ & $L(x|H_1)$

The Neyman-Pearson Lemma

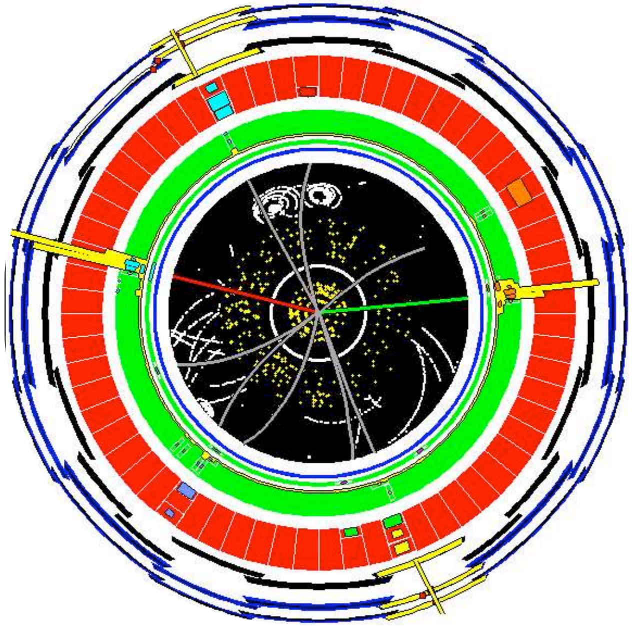
The region W that minimizes the probability of wrongly accepting the H_0 is just a contour of the Likelihood Ratio:

$$\frac{L(x|H_0)}{L(x|H_1)} > k_\alpha$$

$L(x|H_0) =$



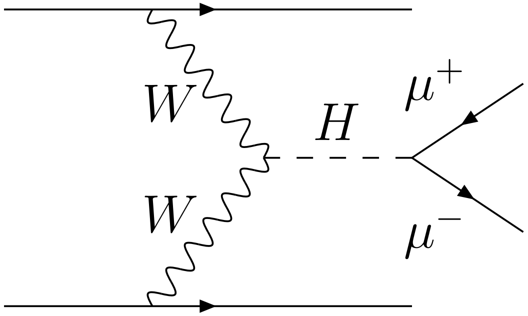
$\oplus$



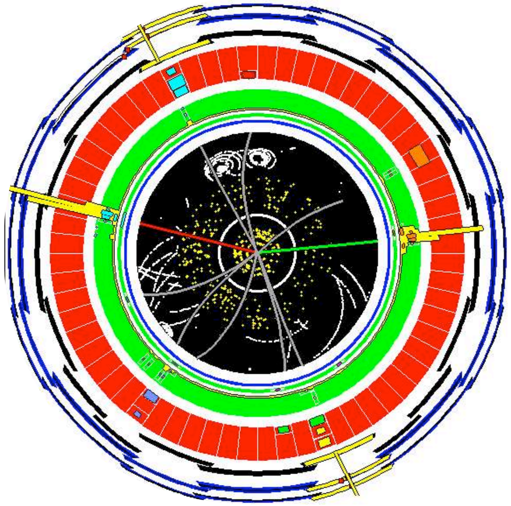
Matrix Element Techniques

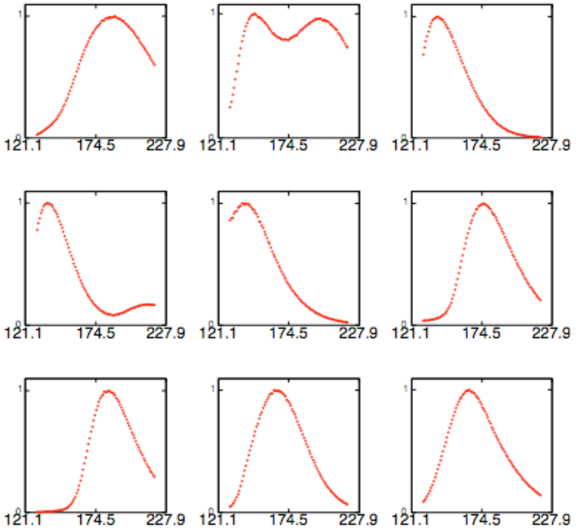
Instead of using generic machine learning algorithms, some members of the Tevatron experiments are starting to attack this convolution numerically

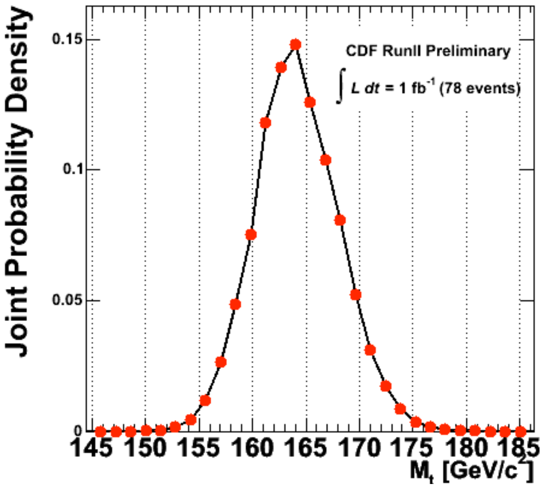
$L(x|H_0) =$



$\oplus$







Matrix Element Techniques

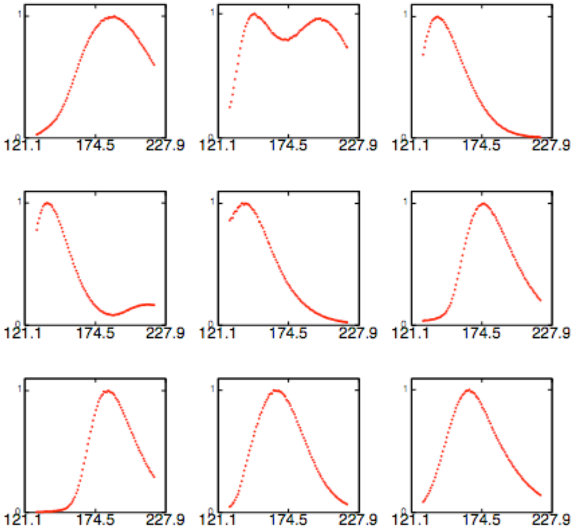
Instead of using generic machine learning algorithms, some members of the Tevatron experiments are starting to attack this convolution numerically

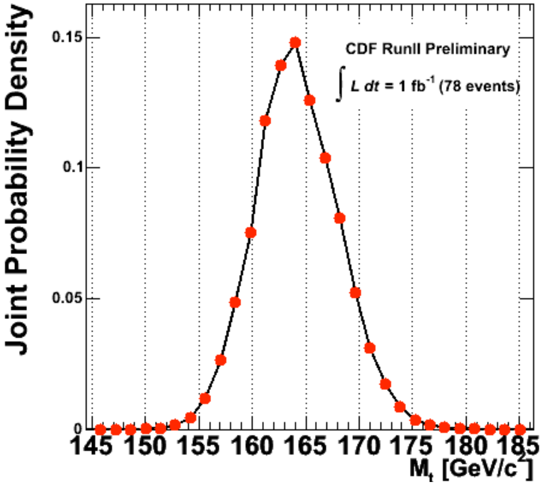
$$P(\mathbf{x}|M_t) = \frac{1}{N}\int d\Phi\; |\mathcal{M}_{t\bar{t}}(p;M_t)|^2 \prod_{jets} f(p_i, j_i) f_{PDF}(q_1) f_{PDF}(q_2)$$

Phase-space Integral

Matrix Element

Transfer Functions





Matrix Element Techniques for Theorists

About 2 years ago, I realized that phenomenologists doing sensitivity studies can use the Neyman-Pearson lemma directly

- directly integrate likelihood ratio
- model detector effects with transfer functions
  - numerically much easier than experimental situation because one generates hypothetical data
- just as one computes a cross-section for a new signal, one can compute a maximum significance (at leading order)

Experimental:
$x \sim$ observable

$$Q(\mathbf{x}) = \frac{L(\mathbf{x}|H_1)}{L(\mathbf{x}|H_0)} = \frac{\mathrm{Pois}(n|s+b)\ \prod_j^n f_{s+b}(x_j)}{\mathrm{Pois}(n|b)\ \prod_j^n f_b(x_j)}$$

$$q(\mathbf{x}) \equiv \ln Q(\mathbf{x}) = -s + \sum_{j=1}^{n} \ln\left(1 + \frac{s f_s(x_j)}{b f_b(x_j)}\right)$$

Theoretical:
$\vec{r} \sim$ phase space

$$\boxed{q(\vec{r}) = -\sigma_{\mathrm{tot},s}\, \mathcal{L}\ + \ln\left(1 + \frac{d\sigma_s(\vec{r})}{d\sigma_b(\vec{r})}\right)}$$

Cranmer, Plehn. EPJ & hep-ph/0605268

Statistical Learning Theory

*When solving a given problem,
try to avoid solving a more general problem as an intermediate step.*

-V.N. Vapnik

Learning Machines

Multivariate Algorithms / Learning Machines
are essentially Black Boxes with some parameters.

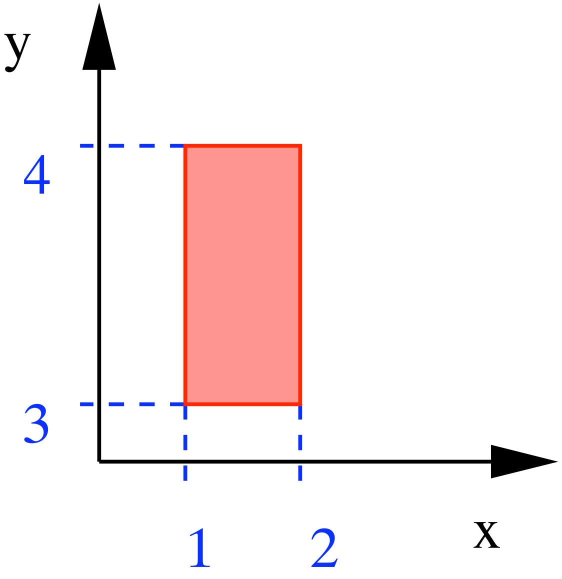
Formally, a learning machine looks like a family of functions from an input space I to an output space O, each specified by some parameters α.

$$f(x \in I; \alpha) = y \in O$$

Training Data is a set of pairs $\{x_i, y_i\}$

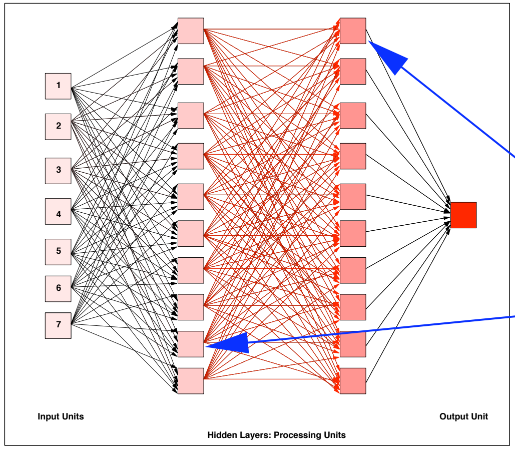
The way in which the function's parameters are determined from training data is associated *learning*.

Examples of Learning Machines



Cuts can be viewed as learning machines

$$f = \begin{cases} 1 & 1 < x < 2 \text{ and } 3 < y < 4 \\ 0 & \text{else} \end{cases}$$



Neural Nets can be viewed as learning machines

weights & biases make up the parameters

Risk

Goal of Learning = minimizing some notion of Risk.

$$R(\alpha) \;=\; \int Q(x,y;\alpha)p(x,y)dxdy$$

- Use different $Q(x,y;\alpha)$ for different problems
- Note: in general we don't know $p(x,y)$.

In practice, we only have the *Empirical Risk*

$$R_{\mathrm{emp}}(\alpha) = \sum_{i=1}^{l} Q(x_i, y_i; \alpha).$$

Risk

Goal of Learning = minimizing some notion of Risk.

$$R(\alpha) \;=\; \int Q(x,y;\alpha)p(x,y)dxdy$$

- Use different $Q(x,y;\alpha)$ for different problems
- Note: in general we don't know $p(x,y)$exactly

In practice, we only have the *Empirical Risk*

$$R_{\mathrm{emp}}(\alpha) = \sum_{i=1}^{l} Q(x_i, y_i; \alpha).$$

Examples of Risk

Goal of Learning = minimizing some notion of Risk.

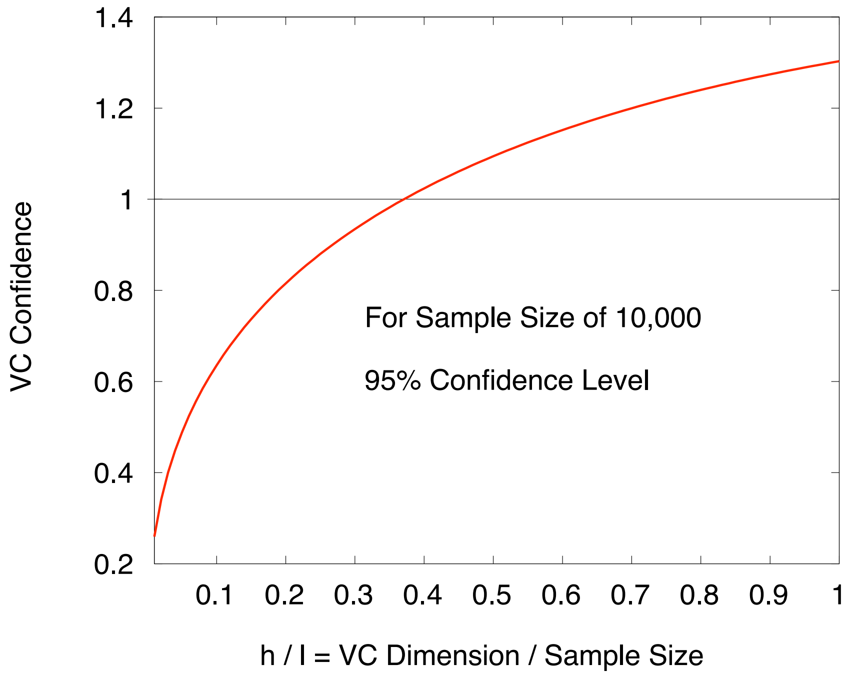
$$R(\alpha) \ = \ \int Q(x,y;\alpha)p(x,y)dxdy$$

Problem	Appropriate $Q(x,y;\alpha)$	Used by
Classification	$\lvert y-f(x;\alpha)\rvert$	Support Vector
Regression	$\lvert y-f(x;\alpha)\rvert^2$	Neural Networks
New Particle Search	$y\Theta(k_\alpha+f(x;\alpha))$	??

Limits on Risk

Surprisingly, there are general bounds on the Risk of a Multivariate Algorithm, given by:

$$\begin{aligned} R(\alpha) &= \int Q(x,y;\alpha)p(x,y)dxdy \\ &\leq R_{\mathrm{emp}}(\alpha) + \sqrt{\left(\frac{h(\log(2l/h) - \log(\eta/4))}{l}\right)} \end{aligned}$$



$1-\eta$ → the confidence the bound holds.

l → the sample size

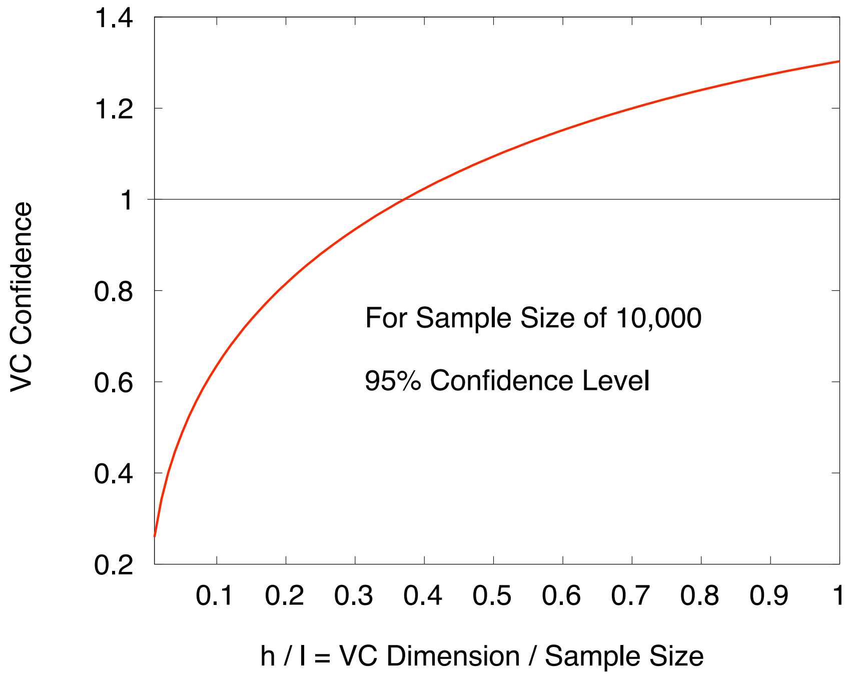
h → the Vapnik Chervonenkis dimension

(holds for $0 < Q < 1$)

Limits on Risk

Surprisingly, there are general bounds on the Risk of a Multivariate Algorithm, given by:

$$\begin{aligned} R(\alpha) &= \int Q(x,y;\alpha)p(x,y)dxdy \\ &\leq R_{\mathrm{emp}}(\alpha) + \sqrt{\left(\frac{h(\log(2l/h) - \log(\eta/4))}{l}\right)} \end{aligned}$$



$1-\eta \rightarrow$ the confidence the bound holds.

$l \rightarrow$ the sample size

$h \rightarrow$ the Vapnik Chervonenkis dimension

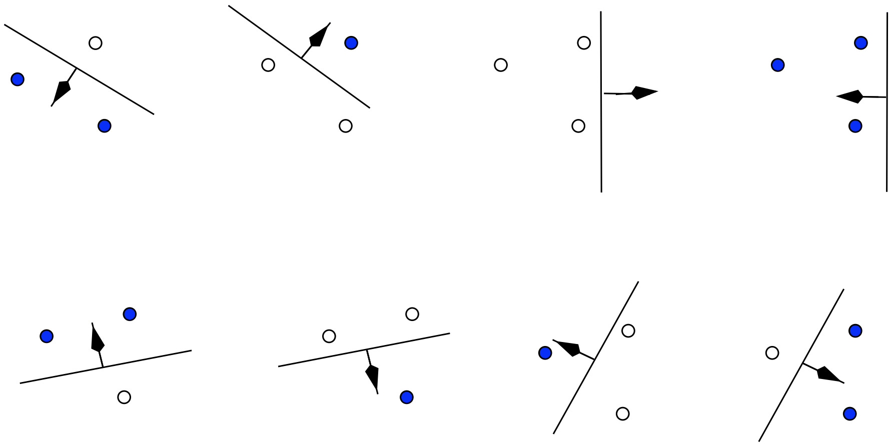
(holds for $0 < Q < 1$)

Support Vector Machines aim to minimize the limit on Risk by balancing R_{emp} and complexity of learning machine characterized by h

VC Dimension

The VC dimension h is equal to the maximal number of points that can be *shattered* by the learning machine $f(x;\alpha)$.

"A set $\{x_i\}$ is shattered by $f(x;\alpha)$" means that for every permutation of classifications $\{x_i, y_i\}$, there is an α such that $f(x_i;\alpha) = y_i$.



Examples:

An oriented line can shatter 3 points in $\mathbb{R}^2$

A Hyperplane can shatter $d+1$ points in $\mathbb{R}^d$

Note: Not every set of h elements must be shattered by $f(x;\alpha)$, but just one.

Importance of VC Dimension

Suggests: (# Training Samples) > (20 x VC dim)

Higher VC dimension → Higher Generalization Capacity → Higher Risk

The Risk bound essentially describes potential for over-training.

Tighter bounds are possible with an independent testing set.

Algorithm	VC Dim	Equivalent # Training Samples
cuts (7-d)	$2d = 14$	≈1,000
Genetic Programming	100	≈7,000
NN (7-10-10-1)	400	≈25,000

Importance of VC Dimension

Suggests: (# Training Samples) > (20 x VC dim)

Higher VC dimension → Higher Generalization Capacity → Higher Risk

The Risk bound essentially describes potential for over-training.

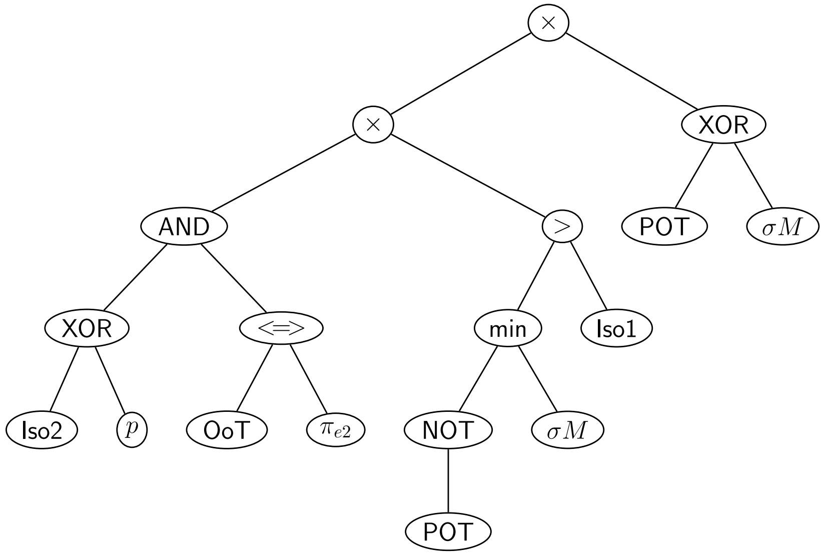
⭐ Tighter bounds are possible with an independent testing set.

Algorithm	VC Dim	Equivalent # Training Samples
cuts (7-d)	$2d = 14$	≈1,000
Genetic Programming	100	≈7,000
NN (7-10-10-1)	400	≈25,000

Because we usually have an independent testing set, the limit on true Risk is often not very useful in practice

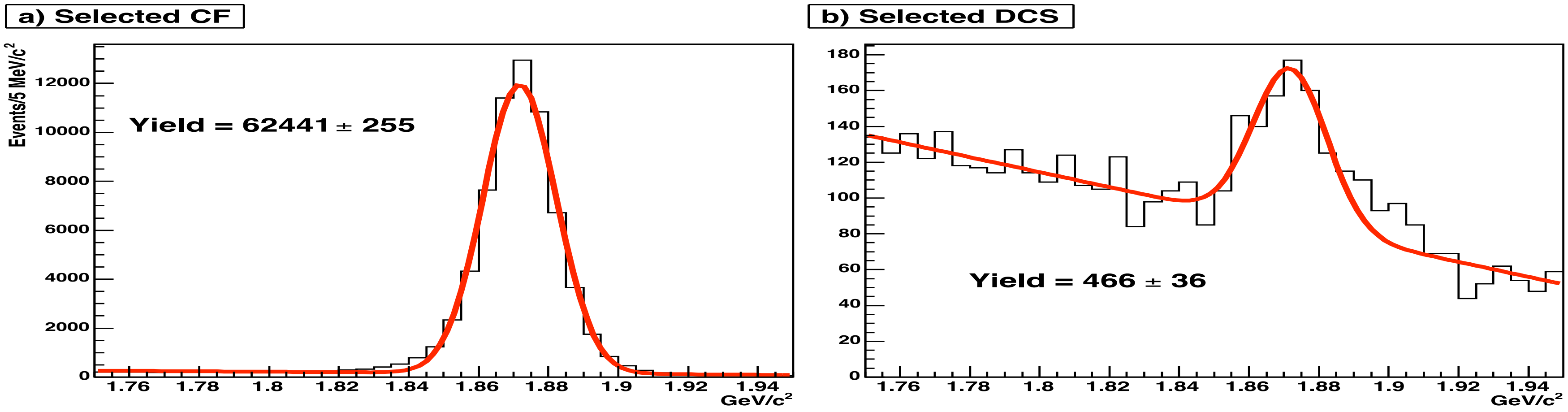
Genetic Programming

R.S. Bowman and I brought a technique called Genetic Programming to HEP. It’s a program that actually writes programs to search for the Higgs! Comput. Phys. Commun [physics/0402030]



The FOCUS collaboration has recently used Genetic Programming to study doubly Cabibbo suppressed decay of $D^+ \to K^+\pi^+\pi^-$ relative to Cabbibo favored $D^+ \to K^-\pi^+\pi^+$

hep-ex/0503007


a) Selected CF

b) Selected DCS

Review

Axioms of Probability

- Frequentist & Subjective Bayesian interpretations

Bayes' Theorem

- Frequentist and Subjective Bayesian examples

Basic Information Theory

- entropy and mutual information

Probability Density Functions

- parametric and non-parametric

Hypothesis Testing and Decision Making

- Type I and II errors; size and power
- the Neyman-Pearson lemma
- simple vs. compound hypotheses

The Likelihood Function & Likelihood Principle

- nuisance parameters

Multivariate Analysis & Statistical Learning Theory

Next time...

The Neyman-Construction

Coverage as a calibration for our statistical device

Systematics, Systematics, Systematics

The strategical challenge of searches for Beyond the Standard Model