Statistics for Particle Physics

Who am I?

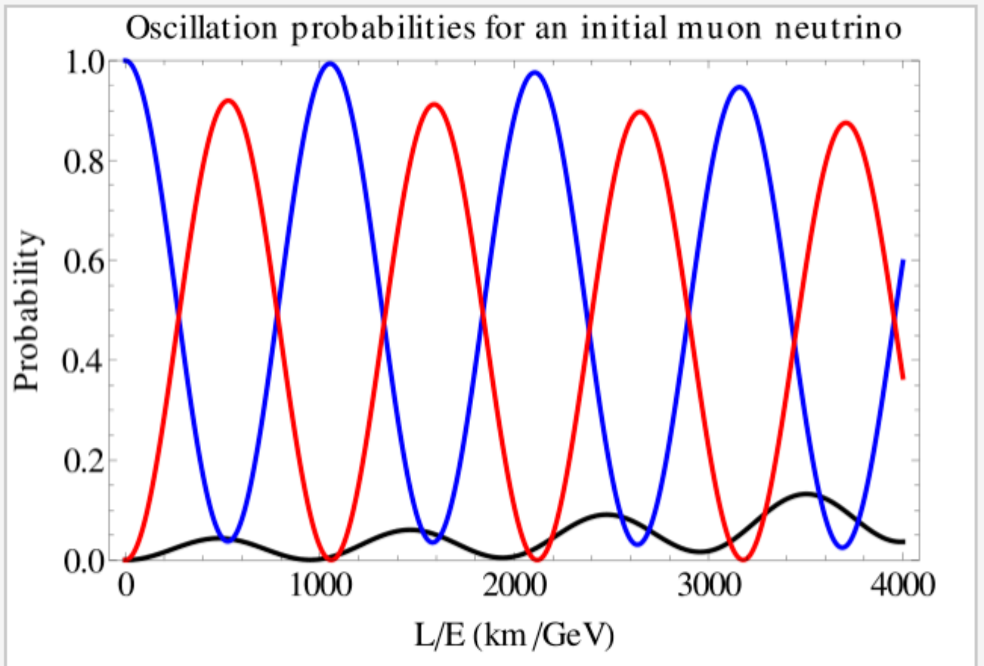
- I work on T2K, DUNE, and LZ
- Interested in analysis challenges of statistics
- Really love onigiri and onsen
- email: asher.kaboth@rhul.ac.uk

Outline

- Hour 1
  - Review of Probability and Basic Terms
  - Frequentist vs Bayesian Statistics
  - Point Estimates
- Hour 2
  - Hypothesis Testing
  - Limit Setting
  - Multivariate Techniques

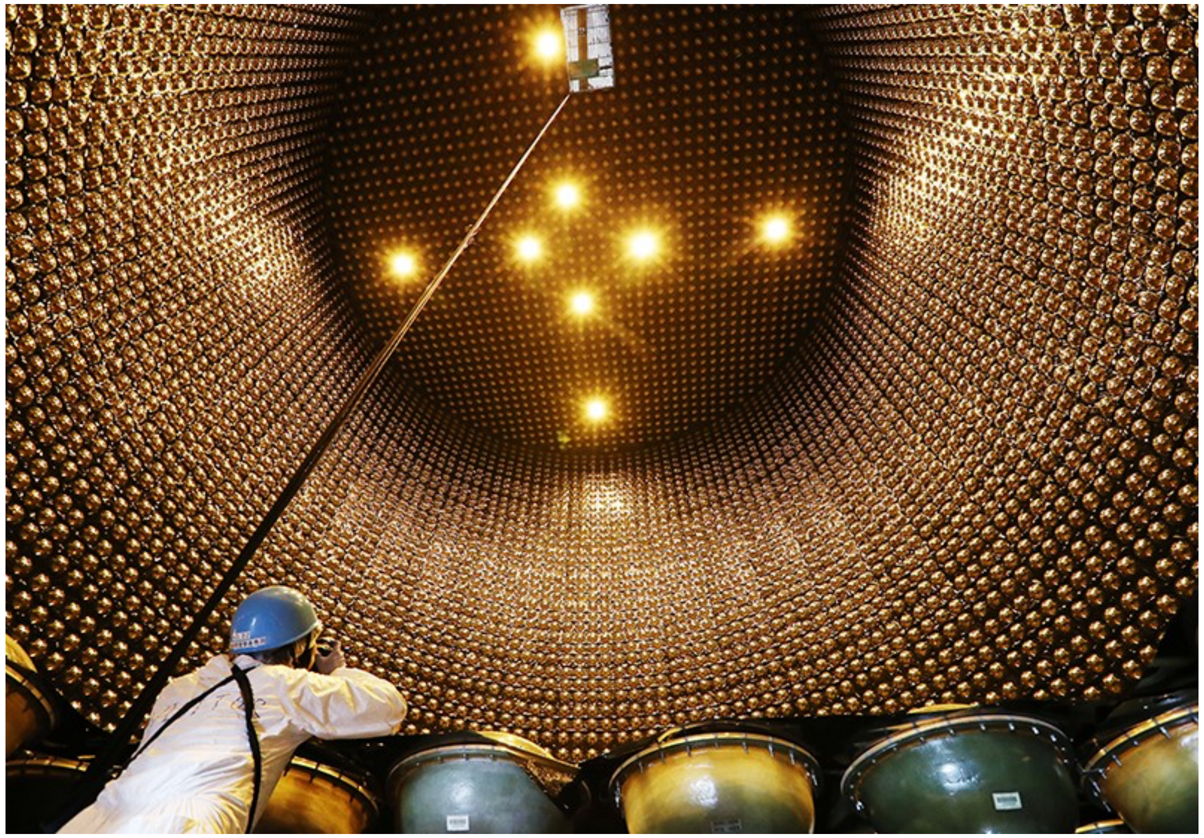
What are we doing here?

We have a nice theory



Muon neutrino oscillations, short range

and a nice experiment



What do they tell us about our natural world?

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*

Tools

- ROOT is the most popular plotting tool in particle physics
- RooStats neatly packages many of the things we'll talk about today
- Many experiments and analyzers are shifting to Python-based analysis

Probability

Frequentist conception: A is the outcome of a repeatable experiment

$$P(A) = \lim_{n \to \infty} \frac{\text{times outcome is } A}{n}$$

Subjective conception/degree of belief: you would make a fair bet on outcome A

Both conceptions obey the Kolmogorov axioms

For all $A \subset S, P(A) \geq 0$

$P(S) = 1$

If $A \cap B = \emptyset, P(A \cup B) = P(A) + P(B)$

$P(\overline{A}) = 1 - P(A)$

$P(A \cup \overline{A}) = 1$

$P(\emptyset) = 0$

if $A \subset B$, then $P(A) \leq P(B)$

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

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

Bayes's Theorem

equal

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(B \cap A)}{P(A)}$$

Therefore:

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

$$P(D) = 0.001$$
$$P(\bar{D}) = 0.999$$
$$P(+|D) = 0.98$$
$$P(-|D) = 0.02$$
$$P(+|\bar{D}) = 0.03$$
$$P(-|\bar{D}) = 0.97$$

Suppose there is a disease and a test with these probabilities. What is $P(D|+)$?

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|\bar{D})P(\bar{D})} = 0.032$$

Frequentist Statistics

- Frequentist statistics is concerned with outcomes of repeated observations (real or hypothetical)
- Probabilities such as P(CP violation exists) are 0 or 1, but we don’t know which
- The preferred theories (models, hypotheses, ...) are those for which our observations would be considered ‘usual’.

Bayesian Statistics

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

- Bayesian statistics uses subjective probabilities for hypotheses
- No prescriptions for priors—informed by knowledge, subjective judgement, and computational feasibility

Probability Density Functions

A random variable is a numerical characteristic assigned to an element of the sample space; can be discrete or continuous.

Suppose outcome of experiment is continuous value x

$$P(x \in [x, x+dx]) = \boxed{f(x)}dx \qquad \int_{-\infty}^{\infty} f(x)dx = 1$$

Probability Density Function

If the variable is discrete

$$P(x_i) = \boxed{p_i} \qquad \sum_i p_i = 1$$

Probability Mass Function

More on PDFs

Joint PDF

$$f(x_1, x_2, ..., x_n) = f(\vec{x})$$

Marginalized PDF

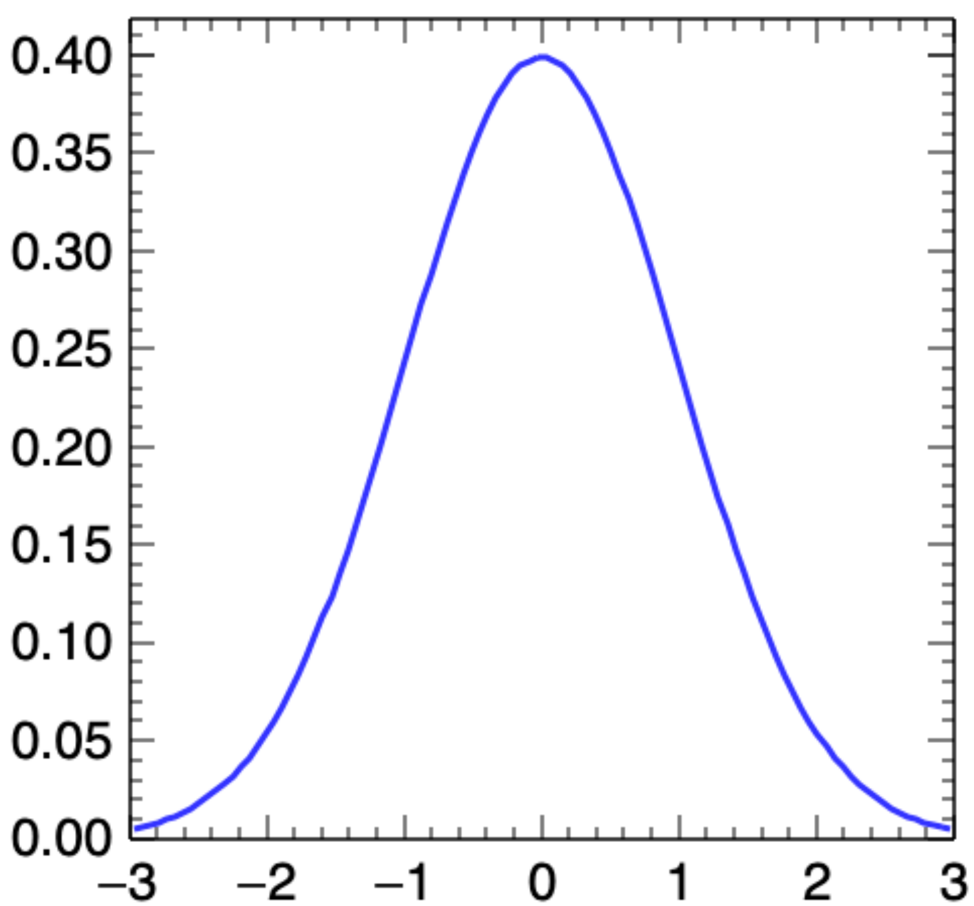
$$f_1(x_1) = \int f(x_1, x_2, ..., x_n) dx_2 dx_3 ... dx_n$$

Conditional PDF

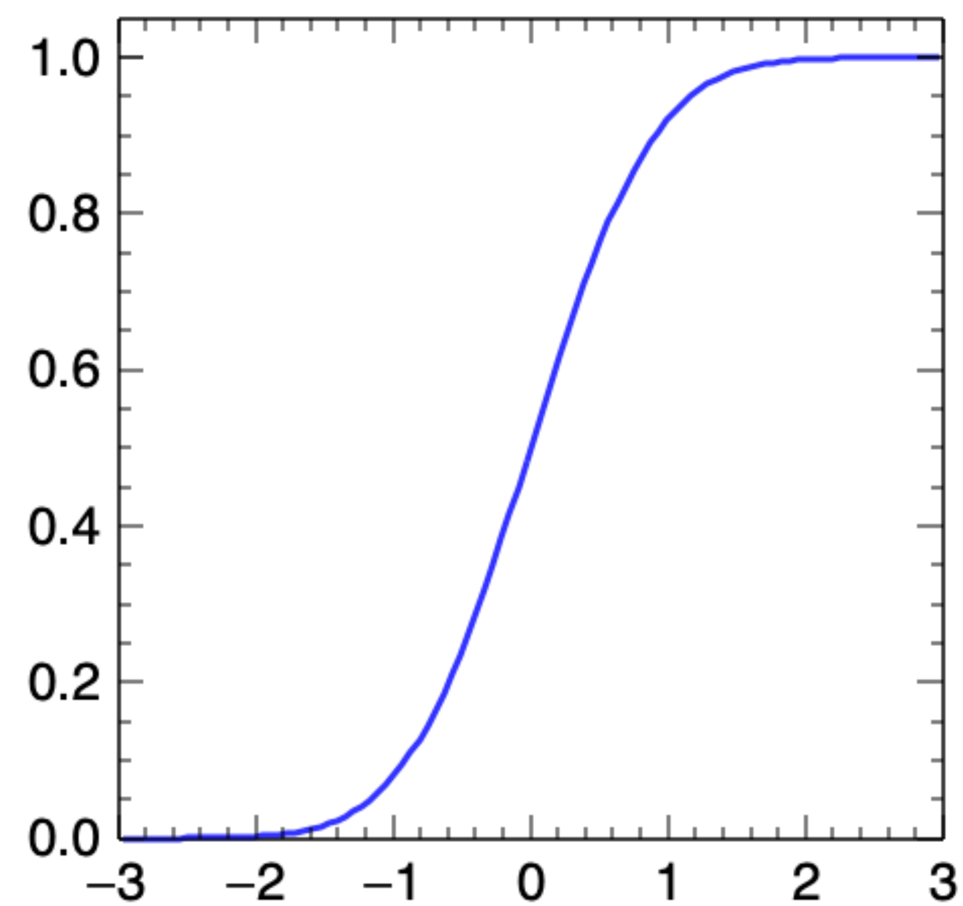
$$g(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}$$

Cumulative Distribution Function

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



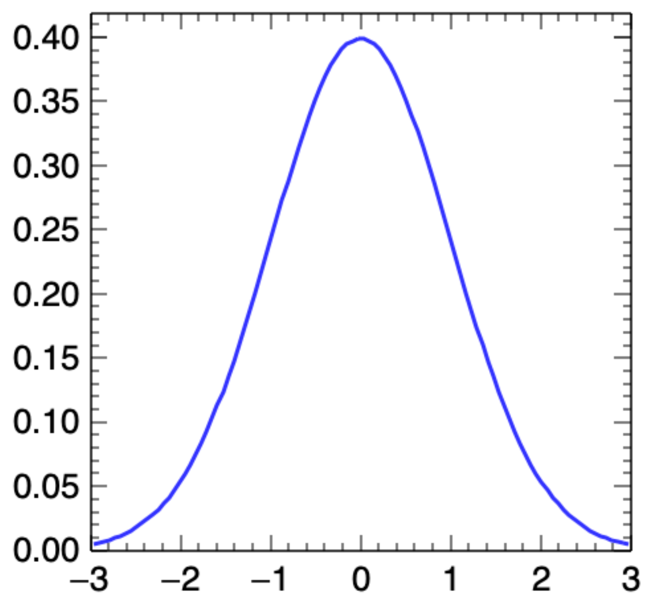
pdf



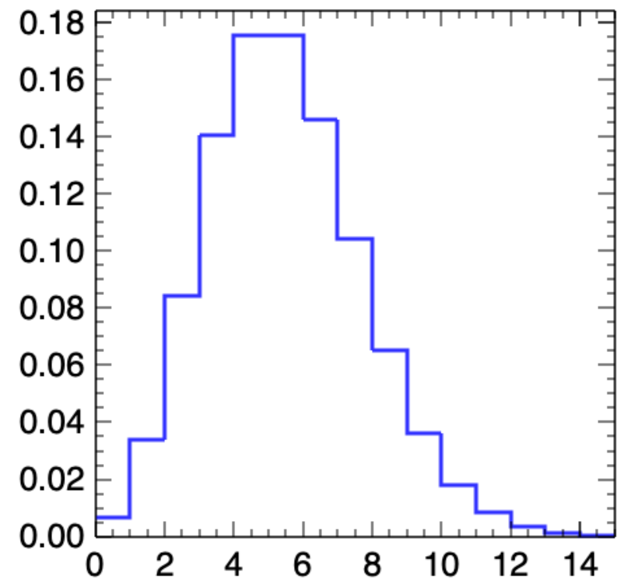
cdf

Common PDFs/PMFs

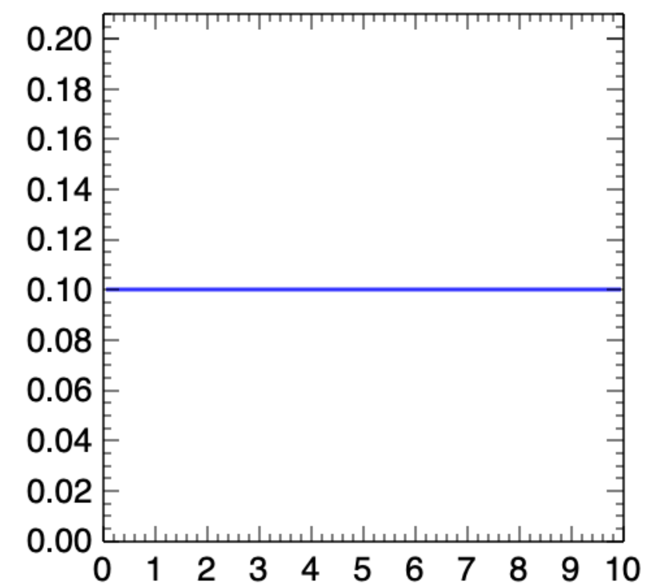
Gaussian



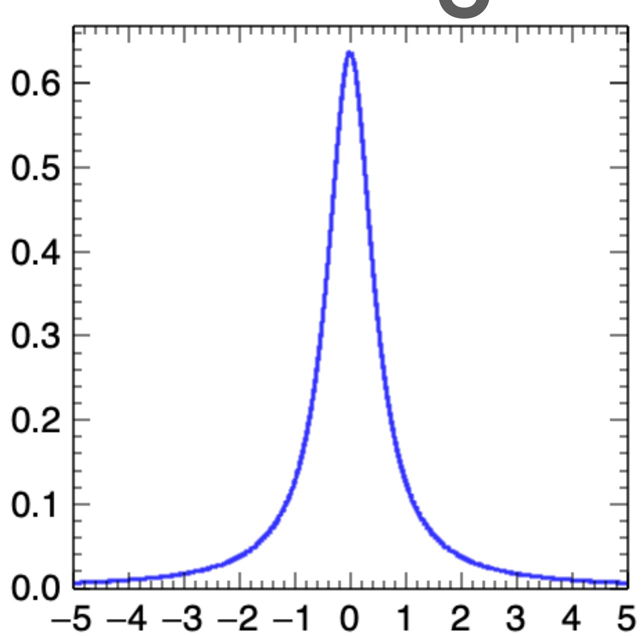
Poisson



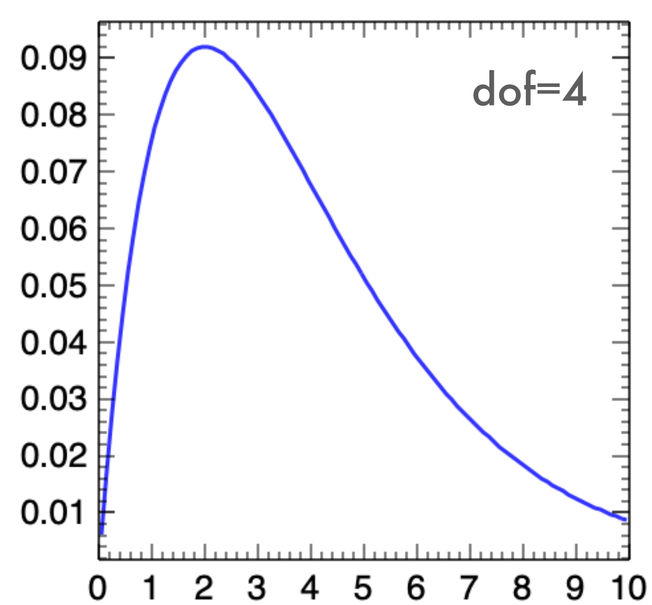
Uniform



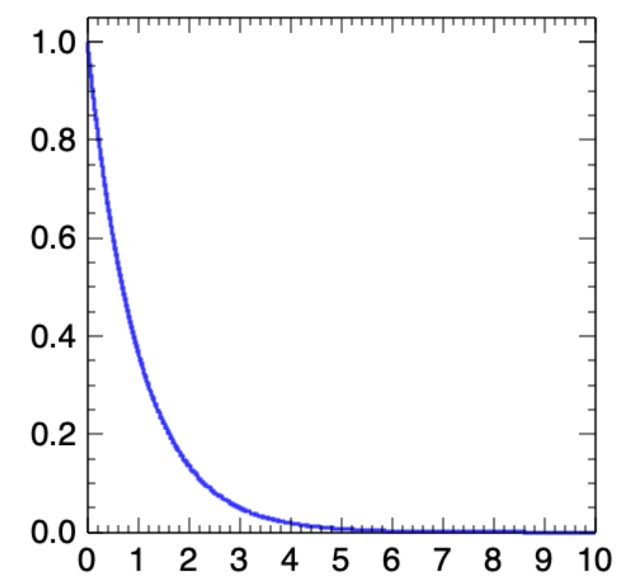
Breit-Wigner



χ^2



Exponential



Means, Covariance, and Correlation

The expectation value, or mean, of a PDF is

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

The variance is

$$V[x] = E[x^2] - (E[x])^2 = \sigma^2$$

Means, Covariance, and Correlation

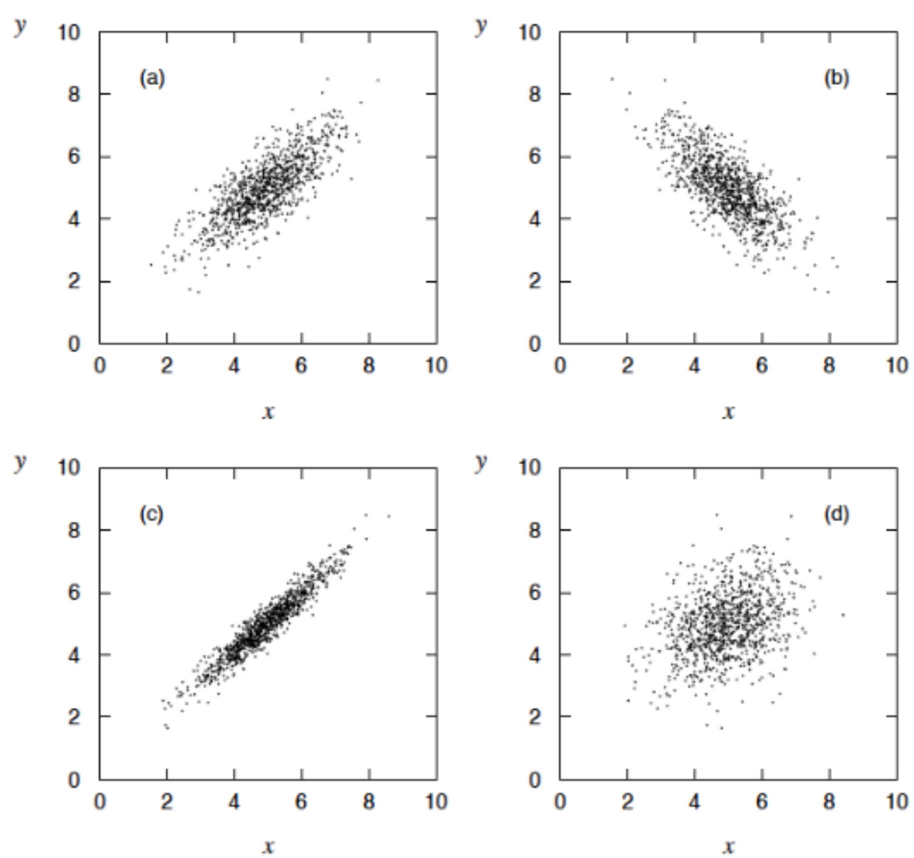
The covariance of two variables in a joint pdf is:

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

The related correlation is: $$\rho_{xy} = \frac{\mathrm{cov}[x,y]}{\sigma_x \sigma_y}$$

$\rho = 0.75$ $\rho = -0.75$

$\rho = 0.95$ $\rho = 0.25$



Comparing Data to Theory

Concept

- We want to know the probability that some set of data comes from some model—the probability of data given a model
- This is called the likelihood
- The model can depend on some vector of parameters, **θ**
- Often use the negative log of the likelihood, as this can be easier to compute, and has some useful properties

$$\mathcal{L}(D|\mathcal{M}(\vec{\theta}))$$

Comparing Data to Theory

Histogram

- Often, we bin data into histograms
- Usually (but not always!) we can assume that the number of events in a bin is poisson distributed

$$\lambda_i = \int_{b_i}^{b_{i+1}} f(x)dx$$

$$\mathcal{L} = \prod_i \frac{\lambda_i^{n_i} e^{-\lambda_i}}{n_i!}$$

$$-\ln \mathcal{L} = \sum_i \lambda_i - n_i \ln \lambda_i + \ln n_i!$$

Comparing Data to Theory

Unbinned

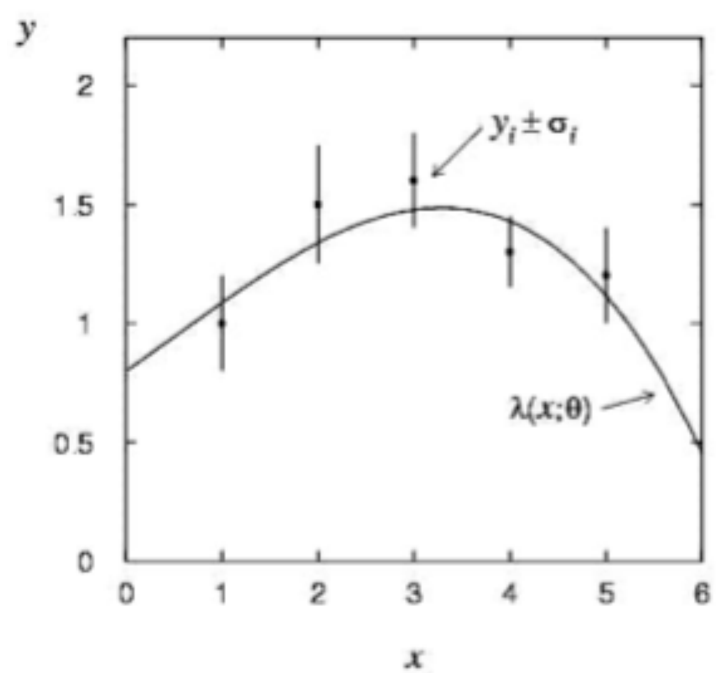
- Sometimes, we can get more information out of our experiment without binning
- If the model predicts a total number of events, we have to include an extended poisson term

Extended Term

$$\mathcal{L} = \prod_{j}^{N} f(x_j) \boxed{\times \frac{\Lambda^N e^{-\Lambda}}{N!}}$$

Comparing Data to Theory

Method of Least Squares

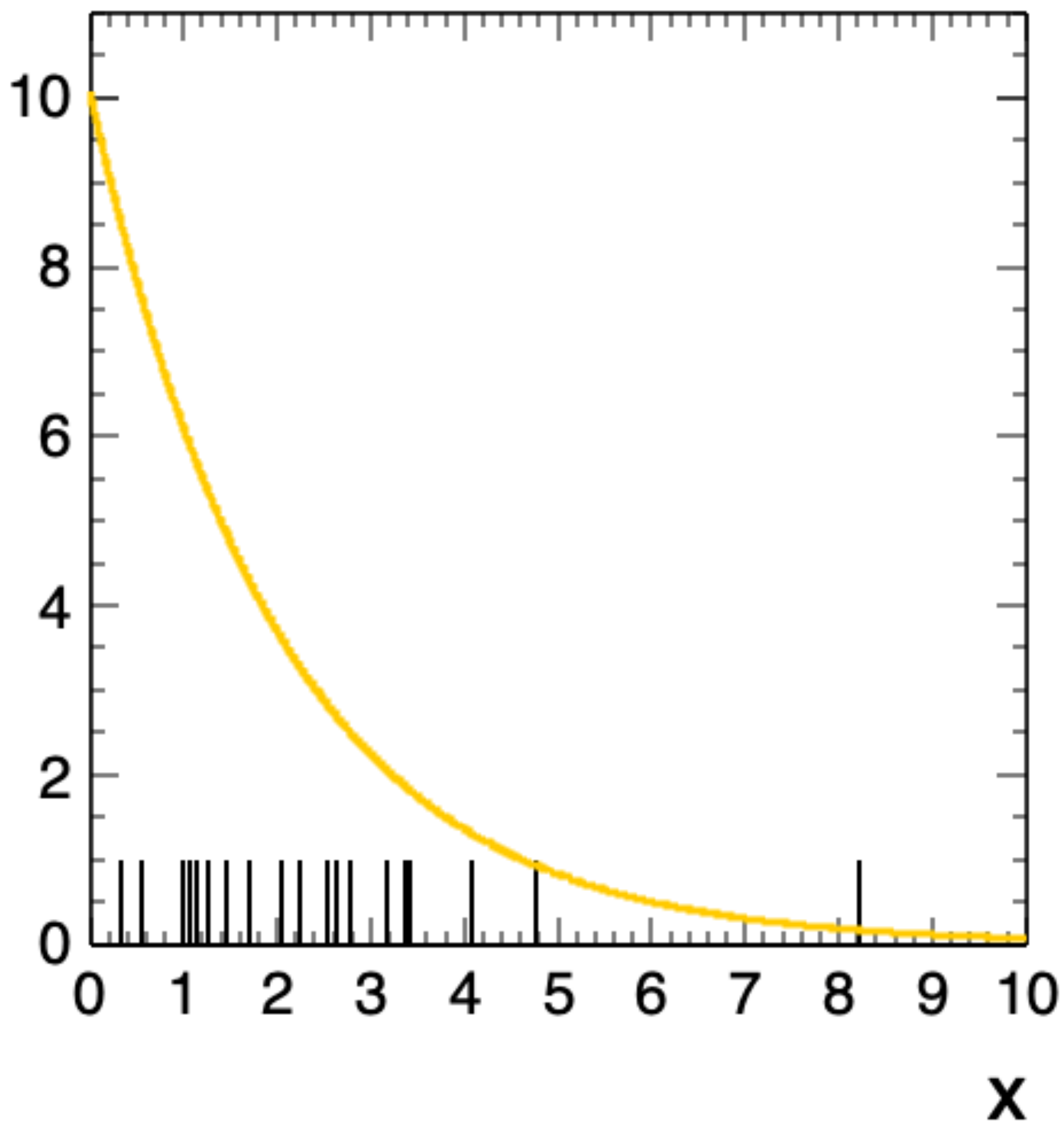


$$L(\theta) = \prod_{i=1}^{N} f(y_i;\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(y_i - \lambda(x_i;\theta))^2}{2\sigma_i^2}\right]$$

$$\ln L(\theta) = -\frac{1}{2}\sum_{i=1}^{N} \frac{(y_i - \lambda(x_i;\theta))^2}{\sigma_i^2}$$

- Sometimes we have data points with associated errors, which we assume are Gaussian
- In this case, we use the method of least squares
- This may look familiar to you as "the" χ^2

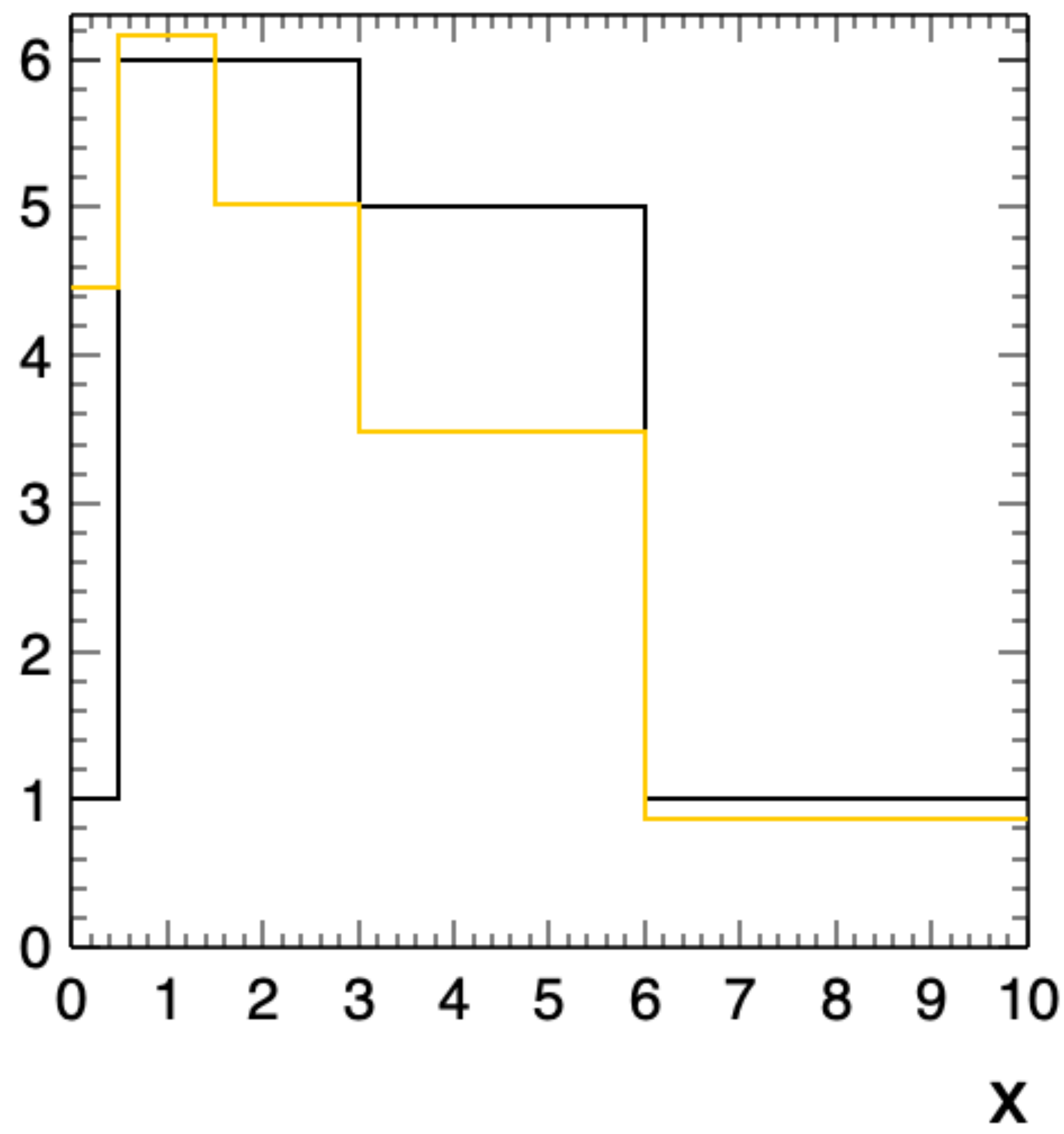
Example



- Model is a decaying exponential that predicts 20 events with a decay constant of -0.5 ($e^{-0.5x}$)
- One example possible data set from this model, N=17, -lnL=0.703

Example

- The same data, but in 5 unequal bins
- Orange shows the model prediction, black shows the data
- -lnL=2.09876



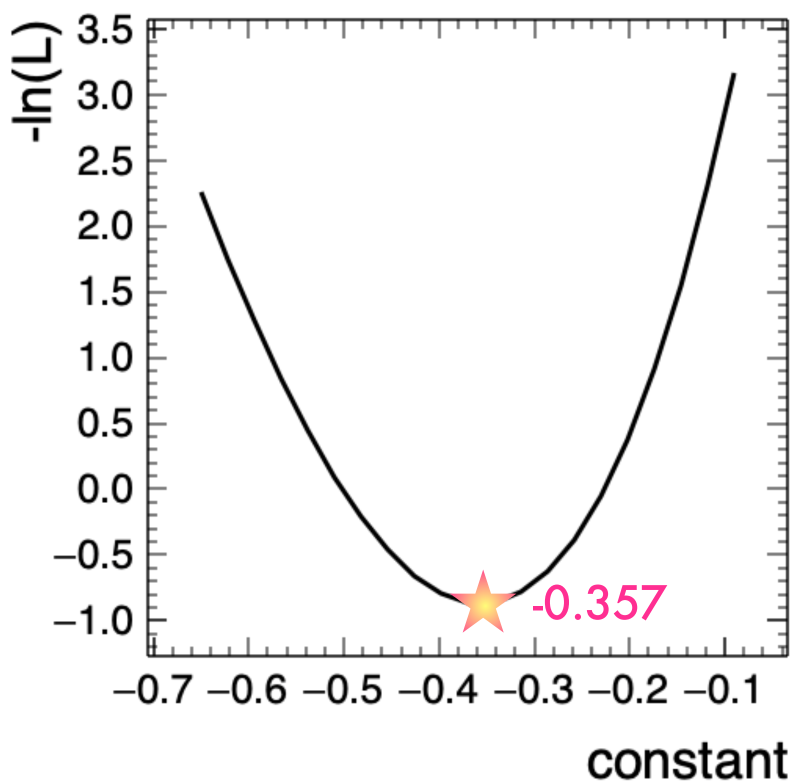
Parameter Estimation

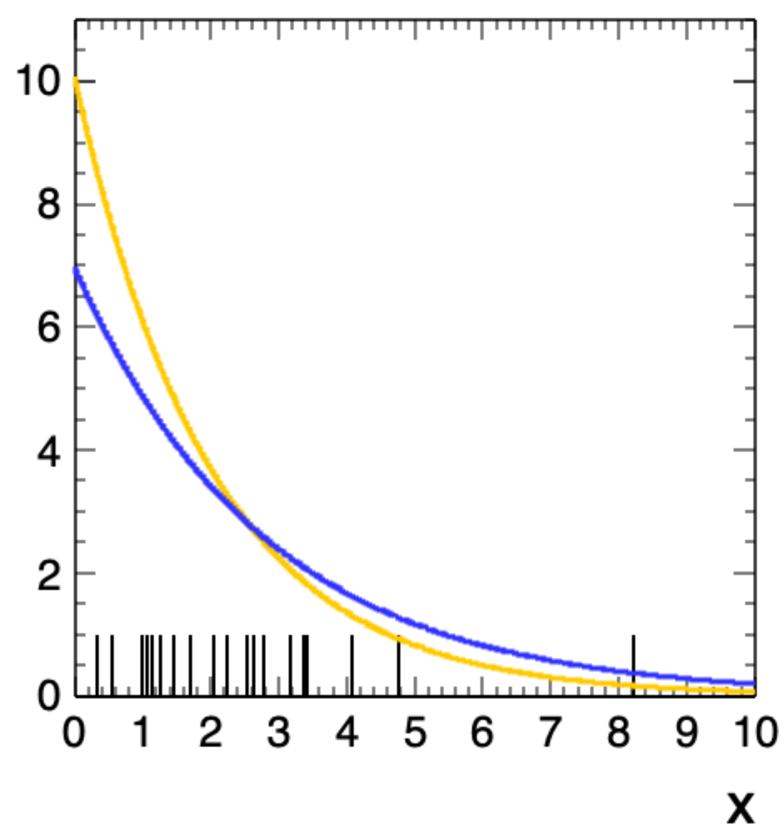
- We have a model, *M*, with some parameters **θ**
- We would like to estimate what the value of these parameters are
- We also what to know what the range of possible values is

Parameter Estimation

- Our ideal estimate would be *unbiased* and have a *small variance*
- Generally these goals are in tension
- The usual (frequentist) tool is a Maximum Likelihood Estimator
- If we're close to the true value of a parameter, then we have a high probability to get the data we observe

Parameter Estimation





- In practice, typically the -lnL is *minimized*, rather than L *maximized*
- The usual tool for this is MINUIT or another gradient descent algorithm
- We denote the values of the model parameters at the minimum as $\hat{\vec{\theta}}$
- The exponential has an analytic solution—the mean is -1/constant
- In this case we get the exact answer!

Parameter Estimation

We'd also like to estimate the uncertainty on our parameters

$$-\ln(\mathcal{L}) = -\ln(\mathcal{L}(\hat{\theta})) - \cancelto{0}{\frac{\partial \mathcal{L}}{\partial \theta}\bigg|_{\theta=\hat{\theta}}} (\theta - \hat{\theta}) - \frac{1}{2!}\ \frac{\partial^2 \mathcal{L}}{\partial \theta^2}\bigg|_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2$$

Expand lnL around the minimum

$$-\ln(\mathcal{L}) = -\ln(\mathcal{L}_{\min}) + \frac{\theta - \hat{\theta})^2}{2\hat{\sigma^2}_{\hat{\theta}}}$$

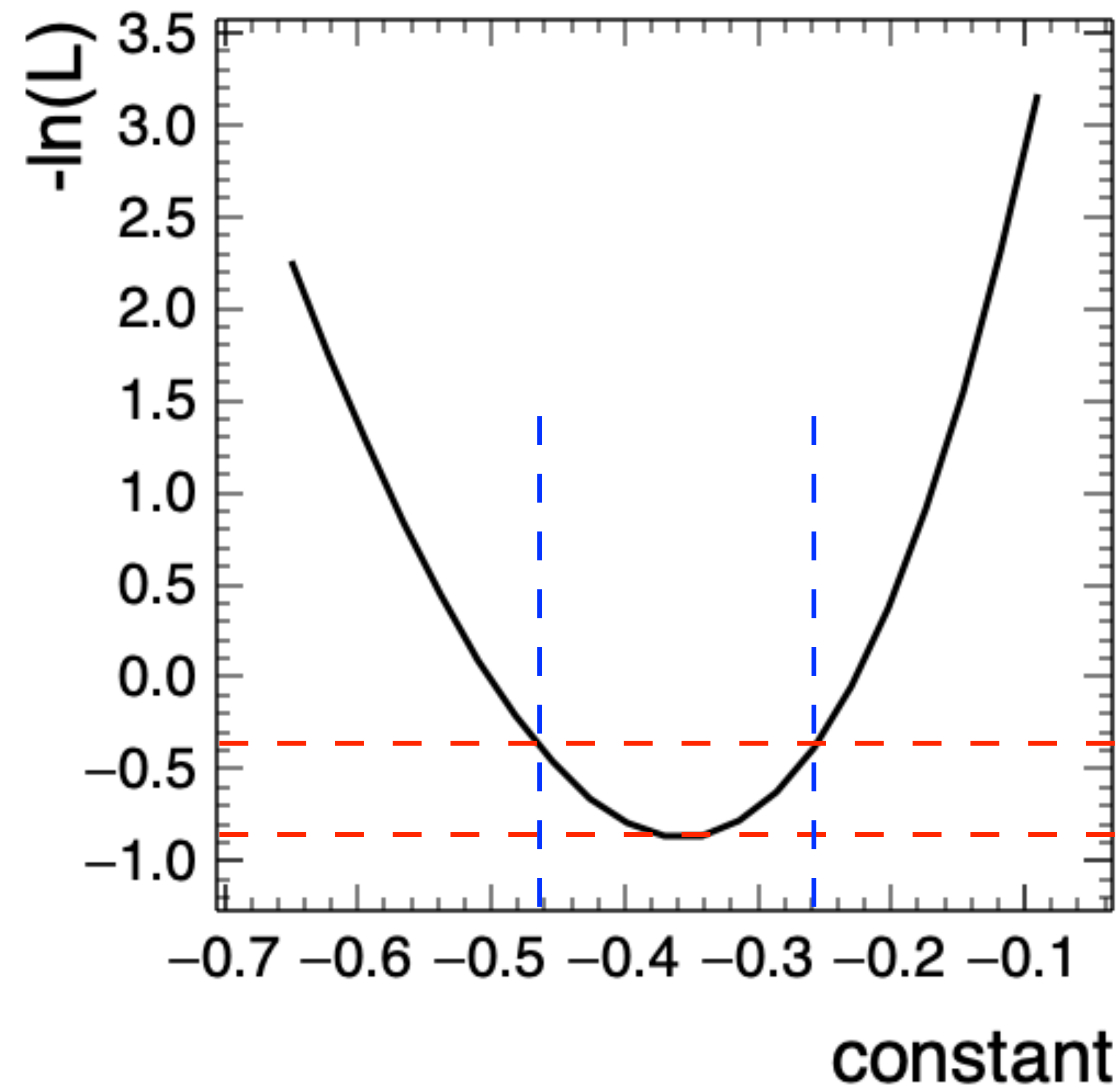
Using a result from information theory (information inequality)

$$-\ln(\mathcal{L}(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) = -\ln(\mathcal{L}_{\min}) + \frac{1}{2}$$

change θ away from θ until -ln L increases by 1/2

Parameter Estimation

- Can see the result from the previous slide graphically
- Remember that this is a confidence interval—if we repeated this experiment many times, 68% of the time, the true value would fall in our calculated interval

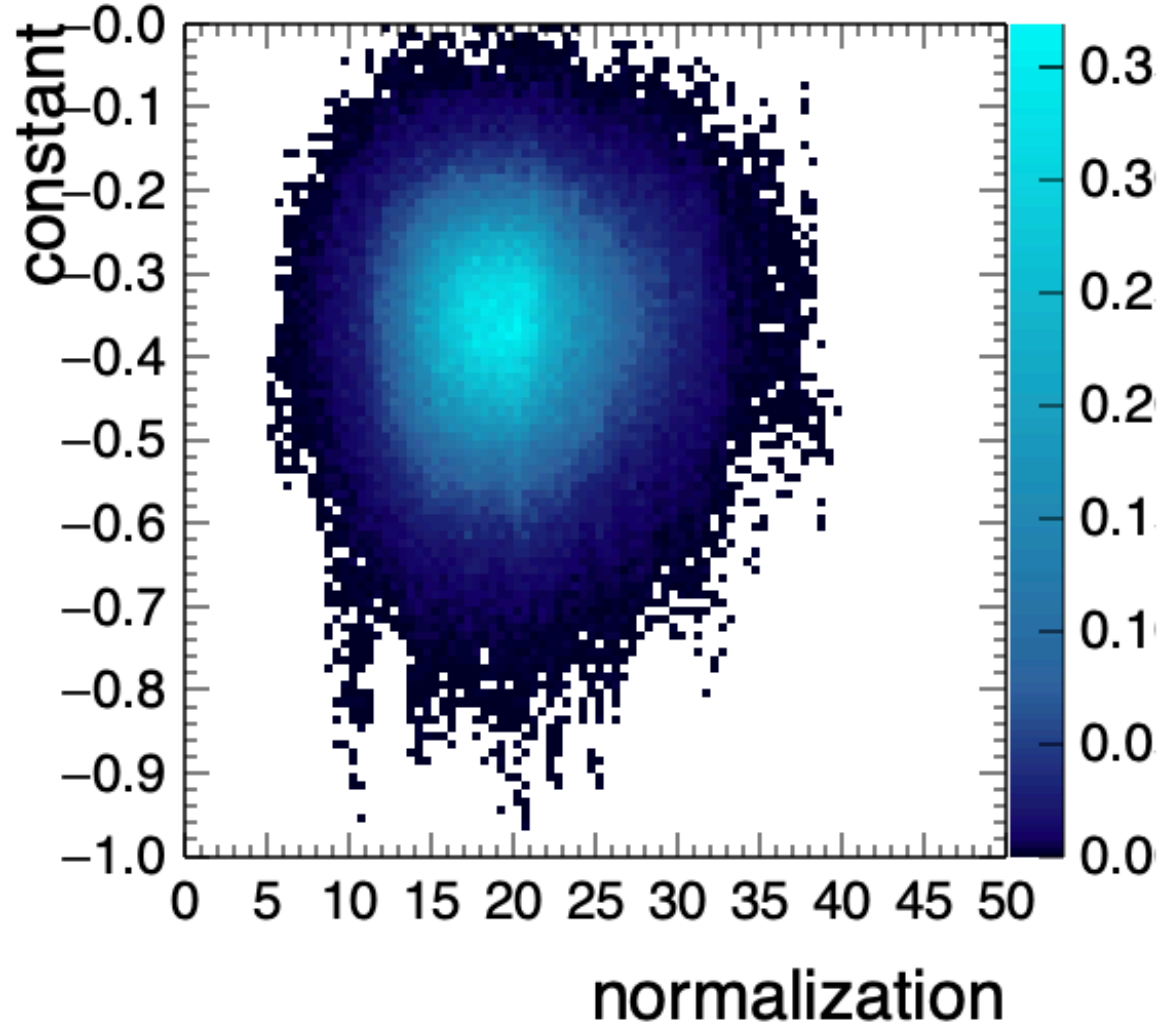


Bayesian Parameter Estimation

- Less interested in a point estimate of a parameter and more interested in the whole posterior
- Need to account for prior in the analysis
- Usually use some numeric tool to build up the posterior

Parameter Estimation

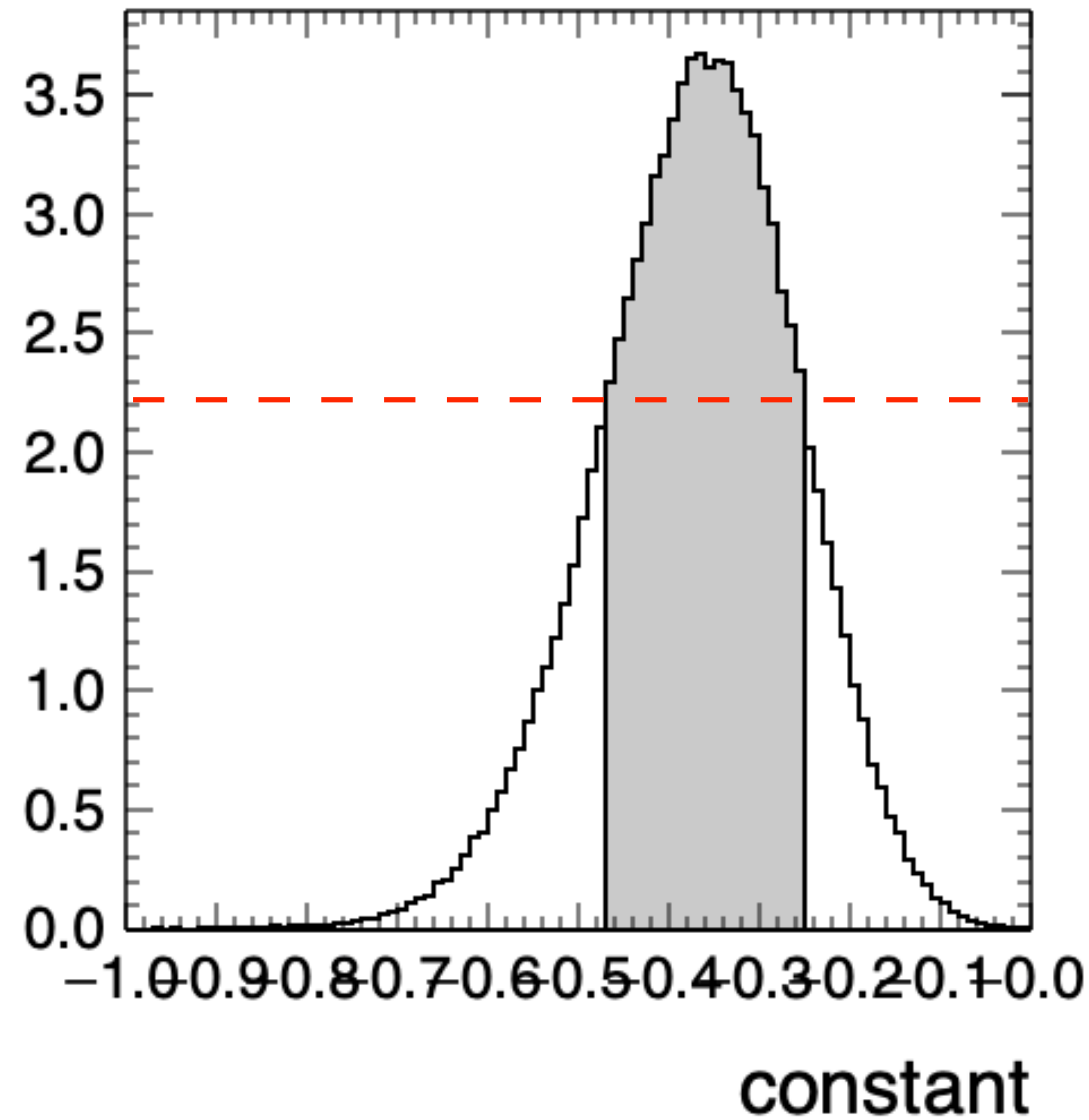
- Used a Markov Chain Monte Carlo to calculate the posterior*—essentially numerically integrating the posterior
- Uniform prior on normalization between 0 and 50, uniform prior on constant between 0 and -1



*More on this if we have time

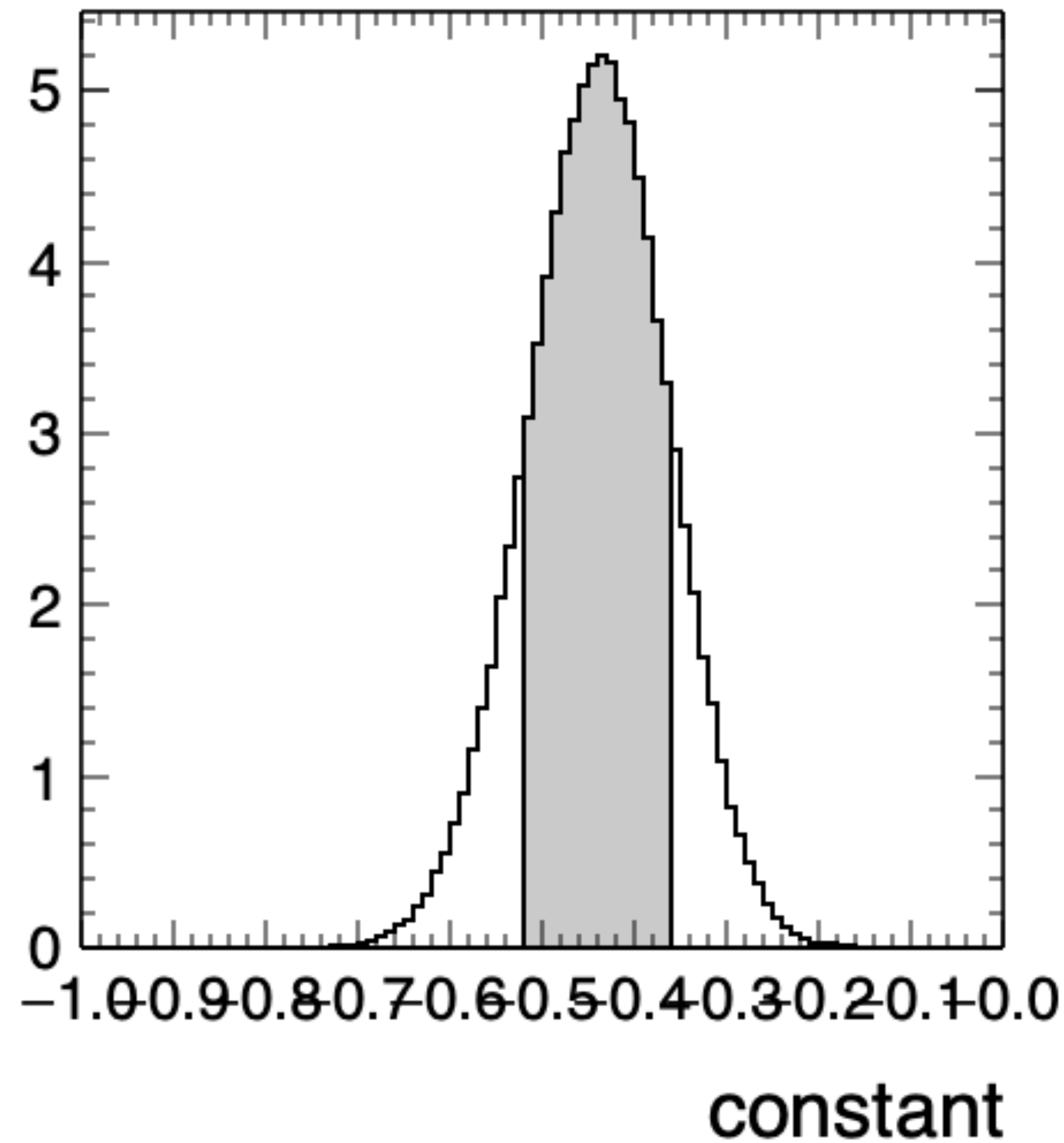
Parameter Estimation

- Can select ANY 68% of the probability—so what should we select?
- Have chosen a Highest Posterior Density Interval—the probability of any value inside the interval is higher than the probability of any value outside the inteval, and it contains 68% of the probability



Parameter Estimation

- What if a theorist told us: “I’m sure the value of the constant is 0.5±0.1”
- We can use that as a Gaussian prior and compare our answer to the previous result



Nuisance Parameters

- So far have only been interested in a parameter of interest (our decay constant)
- What if there are other parameters (detector, model) that we don't care about, but have some knowledge of?

Nuisance Parameters

- In the frequentist method, we add 'constraint terms' to the likelihood
- In the Bayesian framework, we just have a bunch more priors!

Nuisance Parameters

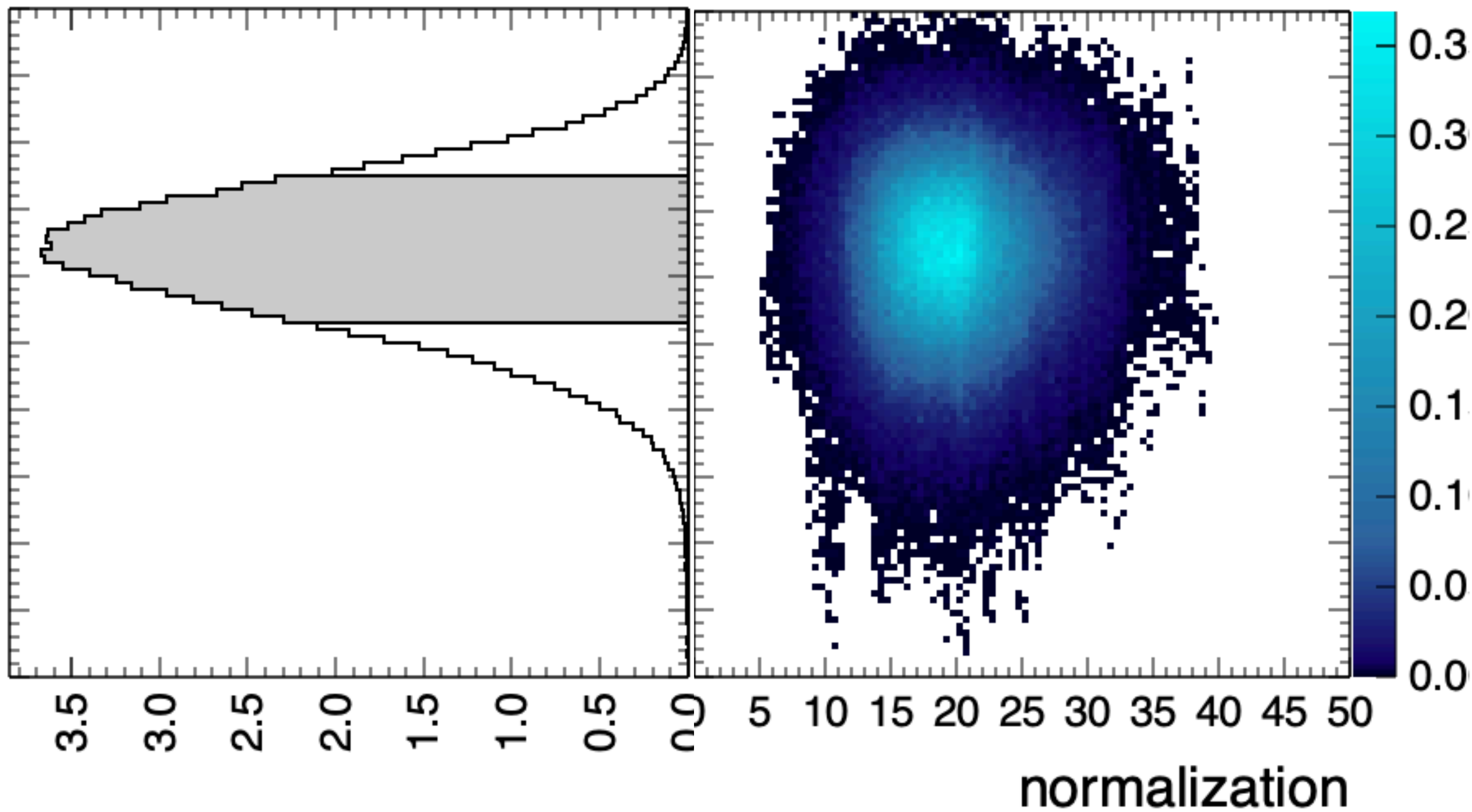
Profiling

- When we minimize a likelihood, we can just add our nuisance parameters to the list of things to minimize
- Find a global minimum across all parameters
- Look at the variation of the parameter of interest at the best estimate of the nuisance parameters

$$\mathcal{L}_p \equiv \mathcal{L}(f, \hat{\theta})$$

Nuisance Parameters

Marginalizing

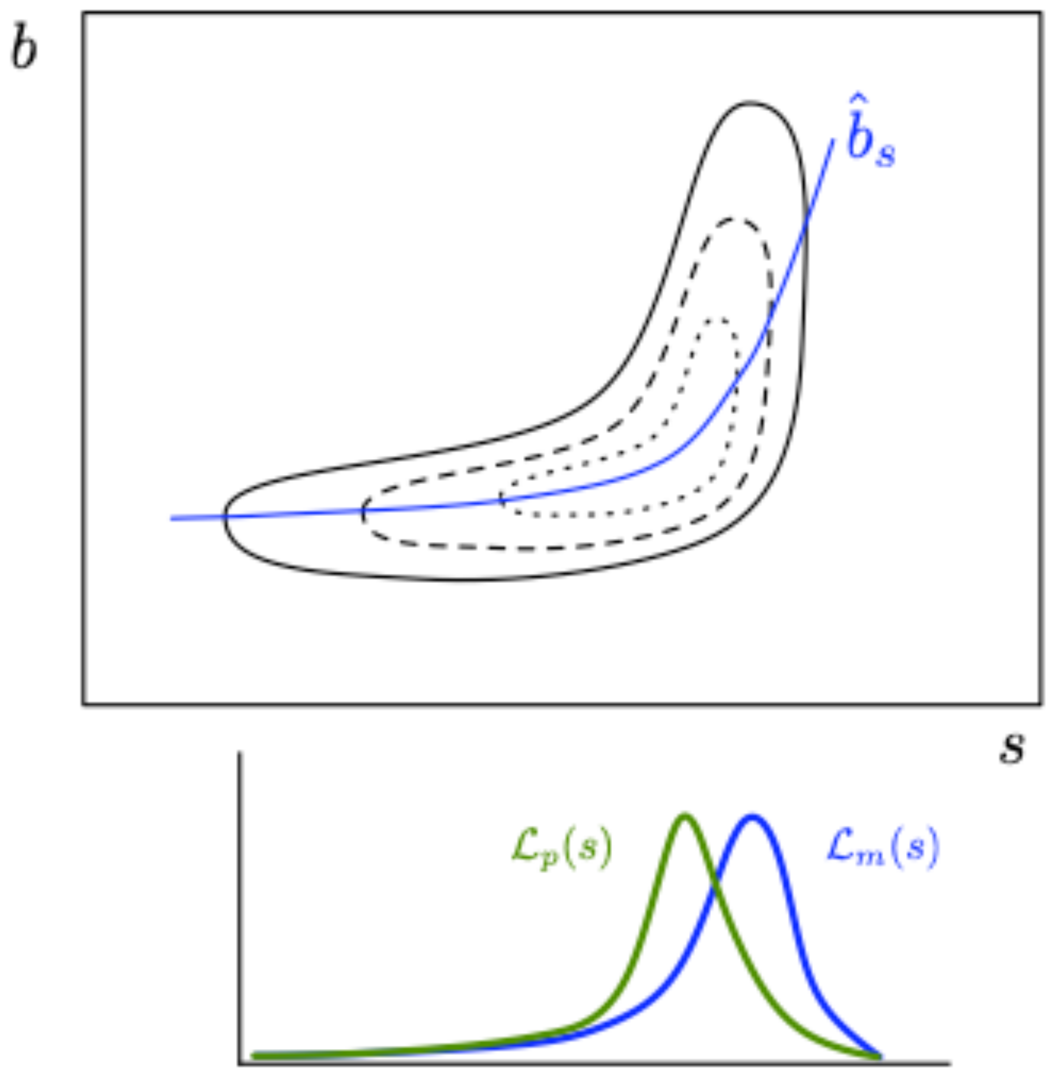


- When we calculate a posterior, we can include all our nuisance parameters
- Typically, when looking at a parameter of interest, we integrate (or marginalize) over the other parameters

So. Does it matter?

Yes!

- If everything is a gaussian, then there is no difference
- Oddly shaped distributions can cause significant difference
- Can also use hybrid techniques that marginalize over some parameters and profile over others



BREAK TIME

Hypotheses

- A hypothesis H specifies the probability for the data, i.e., the outcome of the observation, here symbolically: x.
  - x could be uni-/multivariate, continuous or discrete.
  - x could represent e.g. observation of a single particle, a single event, or an entire "experiment".
- Possible values of x form the sample space S (or "data space").
- Simple (or "point") hypothesis: $f(x|H)$ completely specified.
- Composite hypothesis: H contains unspecified parameter(s).
- The probability for x given H is also called the likelihood of the hypothesis, written $L(x|H)$.

Defining your hypotheses carefully is probably the most critical part of your statistical exercise

Definition of a Test

- Consider e.g. a simple hypothesis H_0 and alternative H_1.
- A test of H_0 is defined by specifying a critical region W of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

$$P(x \in W | H_0) \leq \alpha$$

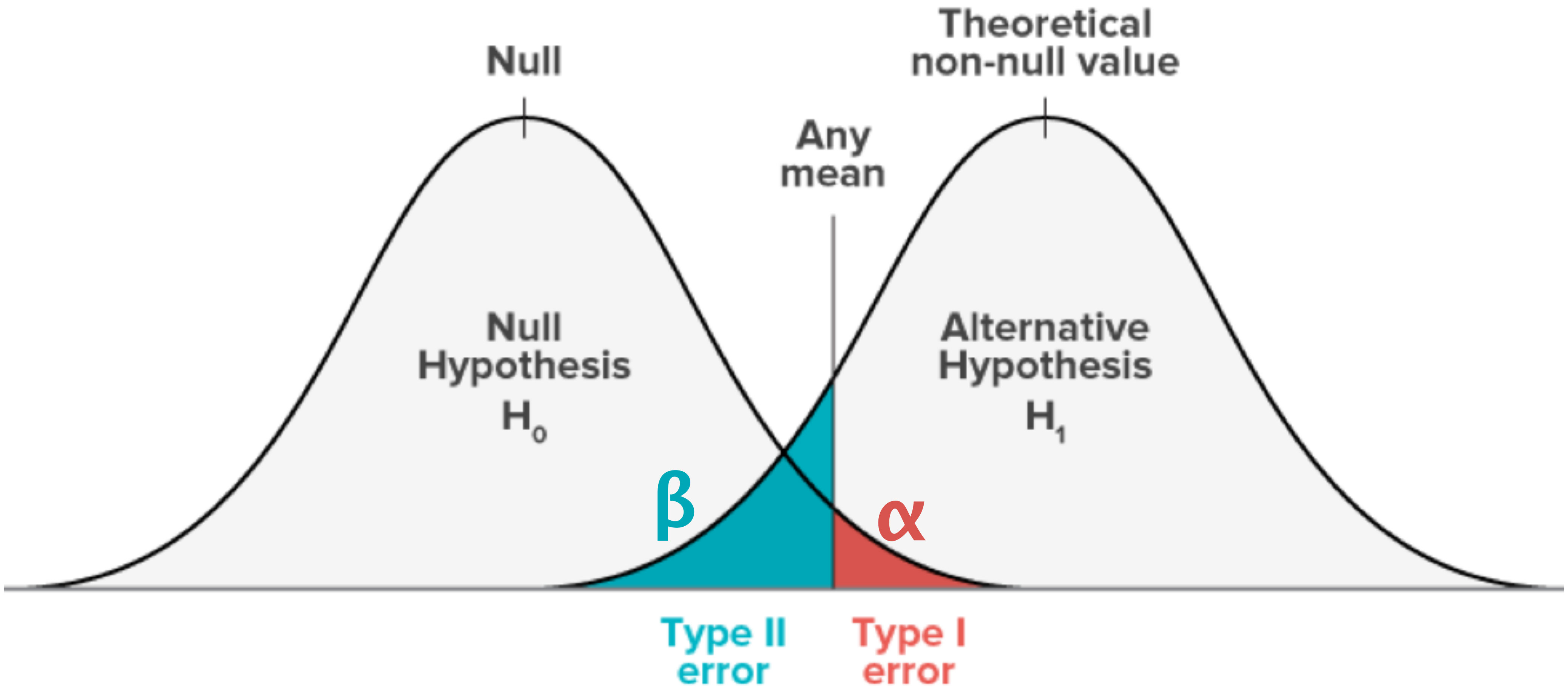
If x is observed in the critical region, reject H0.

- α is called the size or significance level of the test.
- Critical region also called "rejection" region

Definition of a Test

- There are an infinite number of possible critical regions that give the same significance level α.
- So the choice of the critical region for a test of H_0 needs to take into account the alternative hypothesis H_1.
- Roughly speaking, place the critical region where there is a low probability to be found if H_0 is true, but high if H_1 is true

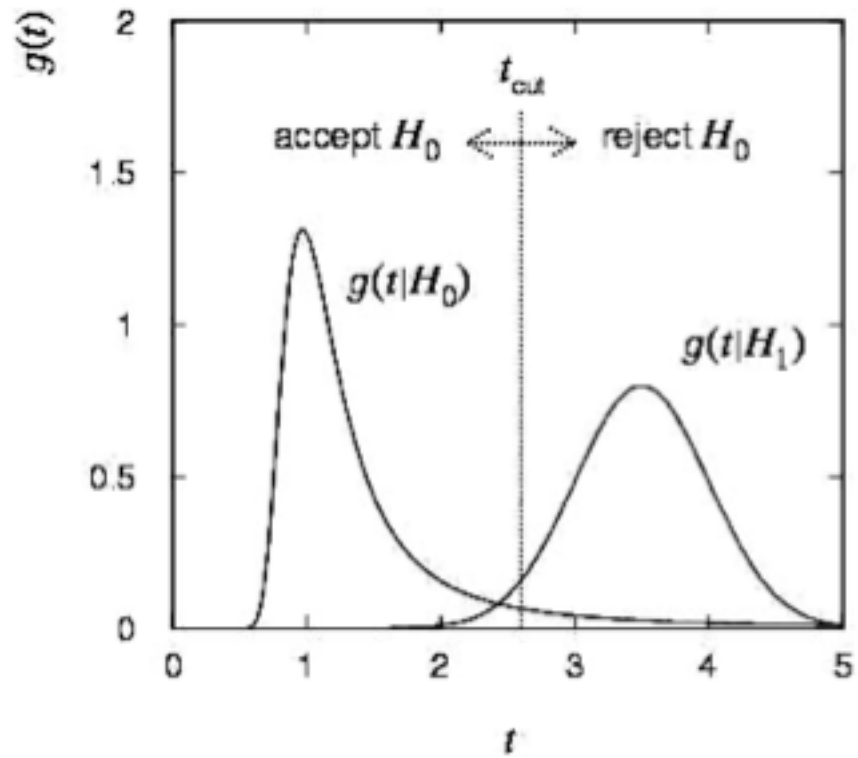
Type-1 and Type-2 Errors



- Type 1 error: Reject H_0 when it is true
- Type 2 error: Fail to reject H_0 when H_1 is true, occurs with probability β
- The *power* of a test is defined as 1-β
- Generally you can pick 2 of 3 of α, β, and the amount of data in your experiment

Test Statistics

- In general, we'll have lots of information about events from our detector
- We want to distill this down to a 1D problem
- The variable we'll choose is called the test statistic
- The Neyman-Pearson Lemma tells us that the highest power for a given significance level is given by t(x)



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

p-values

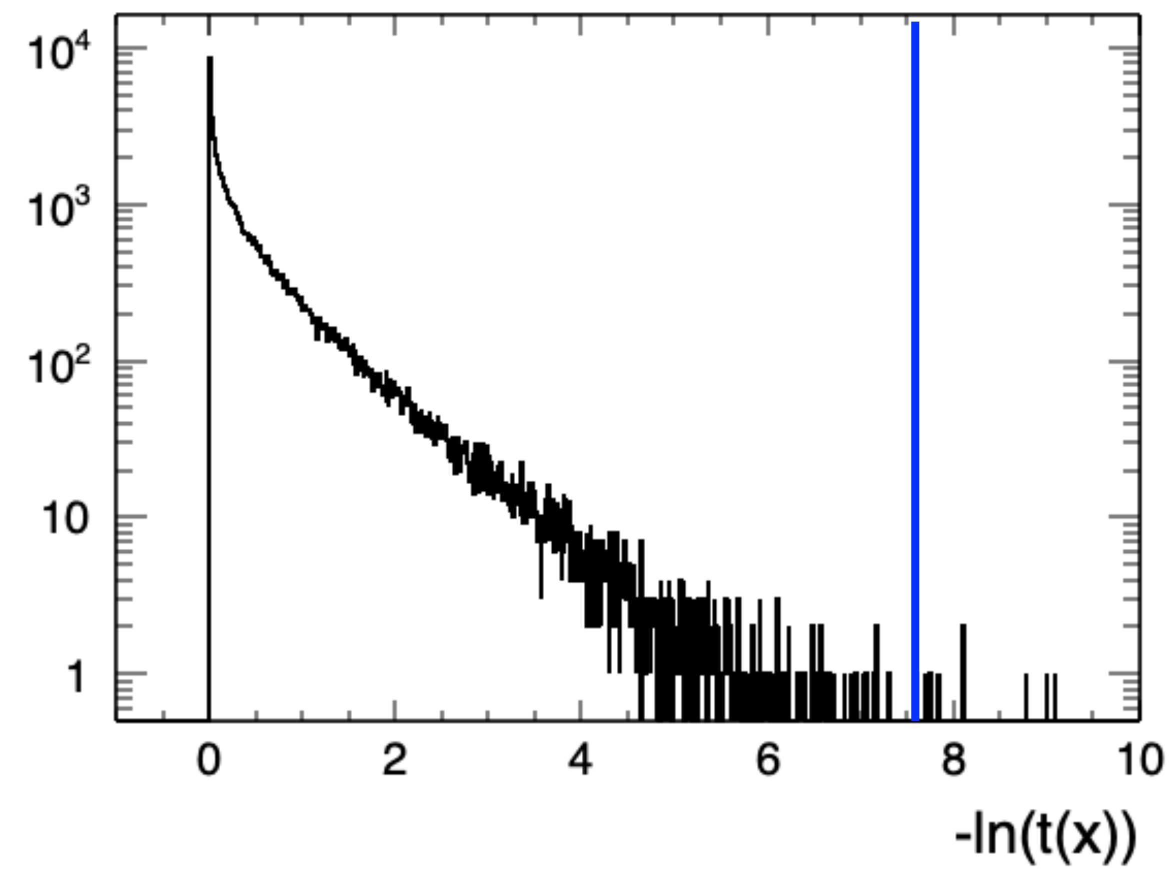
- p = probability, under assumption of H, to observe data with equal or lesser compatibility with H relative to the data we got.
- This is NOT the probability that H is true!!
- Often define significance as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same p-value ("5σ" discovery)

Example

- Use our exponential example:
  - H_0: the data comes from a uniform distribution (i.e., the exponential constant is 0)
  - H_1: the data comes from an exponential distribution
- Generate 100k example data sets from H_0 and generate

Our example data set from Hour 1

p=8E-5, 3.9σ



Nuisance Parameters

- When we have nuisance parameters, nothing is optimal
- "Near optimal" is the profile likelihood ratio test

$$t(x) = \frac{\mathcal{L}(f, \hat{\hat{\theta}})}{\mathcal{L}(\hat{f}, \hat{\theta})}$$

Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554

New Example

- Consider the case of trying to find some signal on top of some background with only a counting experiment: $n=n_s+n_b$
- n_s and n_b are Poisson random variables with means s and b
- Assume b is known
- If n and b are close, then we won't be able to say we've distinguished s from 0 → set an upper limit

Limit Setting

- In our example—or any physics application—we want to find the value of the signal parameter such that there is a given small probability (say α =0.05) to find as few events as we saw or fewer
- This is hypothesis testing 'in reverse': H_0: s=some value; H_1: s=0
- We adjust s until we can't reject H_0 at the given level any more

$$\alpha = \sum_{k}^{n} \frac{(s+b)^k e^{s+b}}{k!}$$

Tests and Confidence Intervals

- Carry out a test of size α for all values of hypothesized θ. The values that are not rejected constitute a confidence region (or interval) for θ at confidence level CL = $1 - \alpha$.
- The confidence interval will by construction contain the true value of θ with probability of at least $1 - \alpha$. The interval will cover the true value of θ with probability $\geq 1 - \alpha$.
- Usually use a p-value of θ to define critical region of test as having $p_\theta \leq \alpha$.
- The parameter values in the confidence region/interval have p-values of at least α.
- To find boundary of region/interval, set $p\theta = \alpha$ and solve for θ.

Limit Setting

Suppose n=0 and b=0

$$0.05 = e^{-s}$$
$$s_{upp} = -\ln(0.05) = 2.996$$

Suppose n=0 and b=3.1

$$0.05 = e^{-s+b}$$
$$s_{upp} = -\ln(0.05) - b = -0.1$$

What Happened?!?

Physicist:
We already knew $s \geq 0$ before we started; can't use negative upper limit to report result of expensive experiment!

Statistician:
The interval is designed to cover the true value only 95% of the time – this was clearly not one of those times.

If we were frequentists with infinite budget and time, if we repeated our experiment many times, the mean upper limit is ~5

Nuisance Parameters

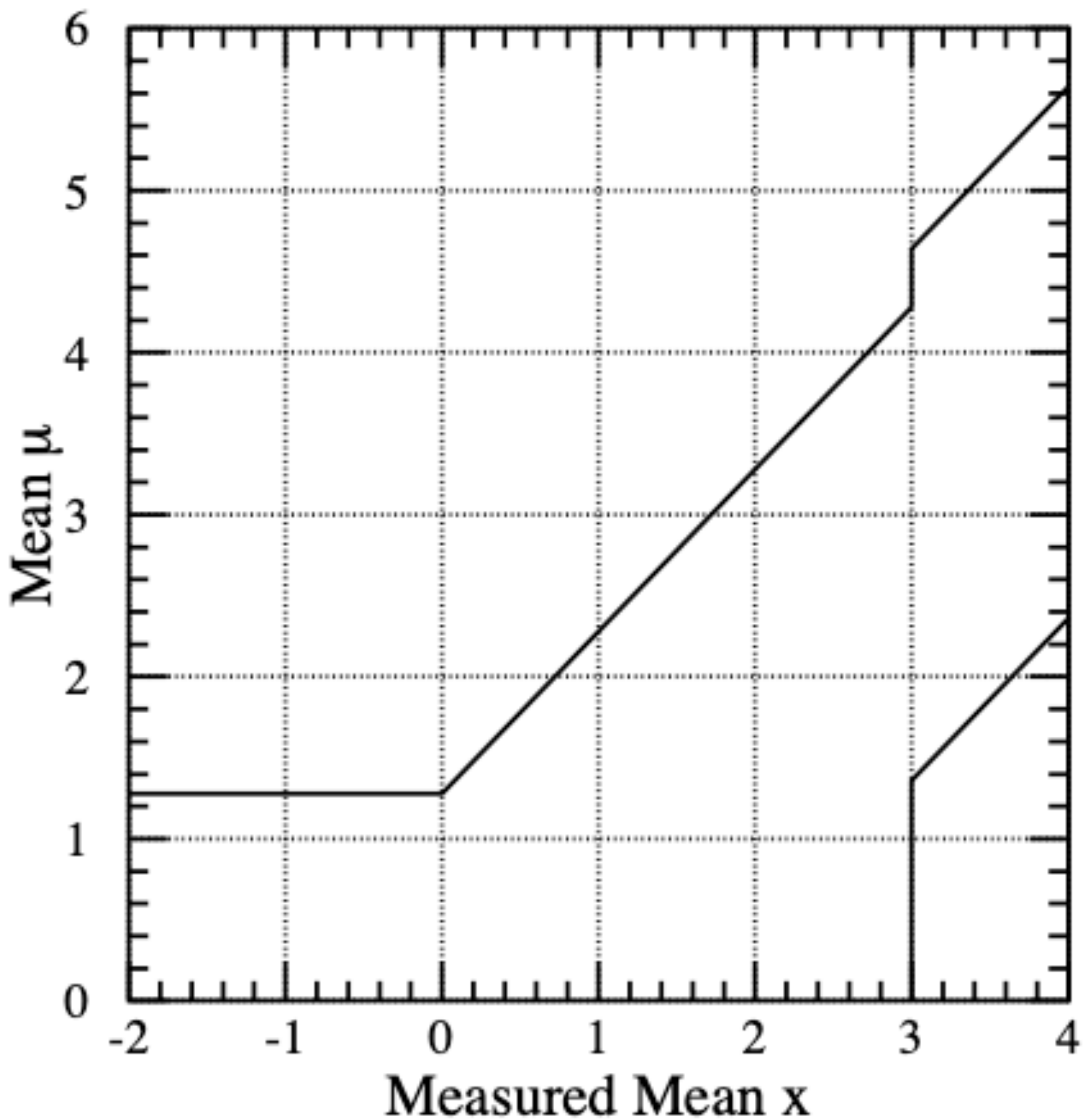
$$\mathcal{L}(s,b) = \frac{(s+b)^n e^{-(s+b)}}{n!} \frac{(\tau\beta)^m e^{-\tau\beta}}{m!}$$

$$\lambda(s) = \frac{\mathcal{L}(s,\hat{\hat{b}})}{\mathcal{L}(\hat{s},\hat{b})}$$

- Imagine we have some other set of data that can constrain the value of b—a sideband
- It has m events, with m~Poisson(τβ)
- Now we can use our PLR statistic

'Flip-Flopping'

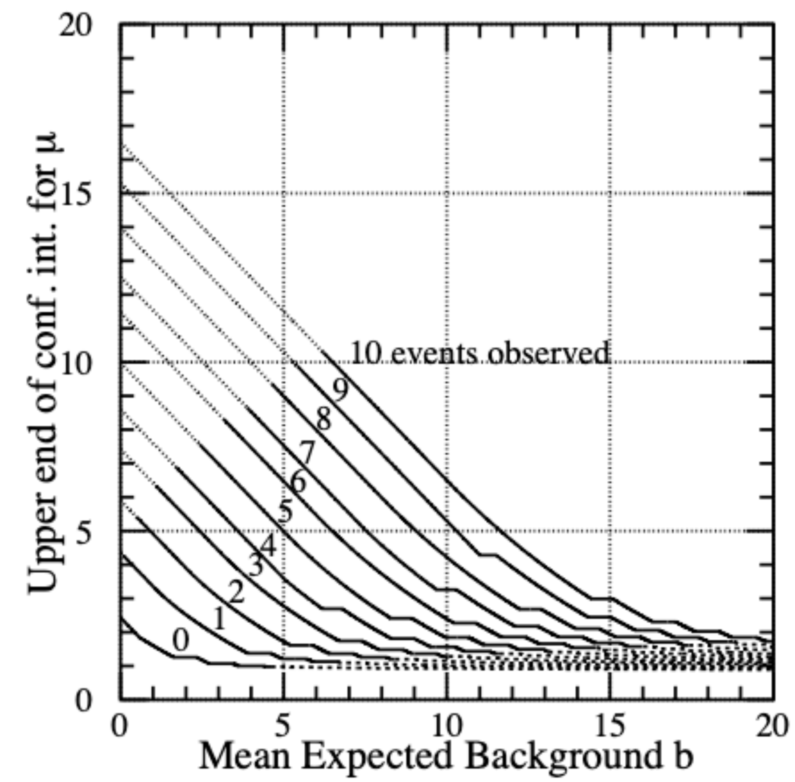
- What if we don't know whether we should set an upper limit or have a two-sided interval?
- "If the result x is less then 3σ, I will state an upper limit from the standard tables. If the result is greater than 3σ, I will state a central confidence interval from the standard tables."

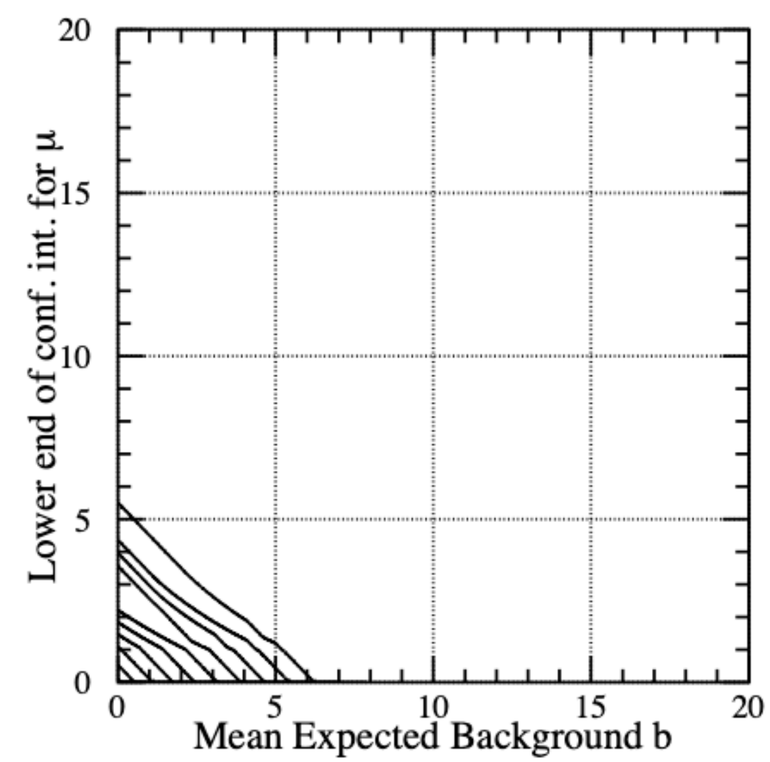


Feldman-Cousins

- The Feldman-Cousins ordering principle describes a way around the flip-flopping problem
- Use our PLR test statistic with a treatment so that our parameter of interest cannot go below zero

$$\tilde{t}_\mu = \begin{cases} -2\ln \frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(0,\hat{\hat{\theta}}(0))} & \hat{\mu} < 0 , \\ -2\ln \frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\hat{\mu},\hat{\theta})} & \hat{\mu} \geq 0 . \end{cases}$$





OK, what if I'm a Bayesian?

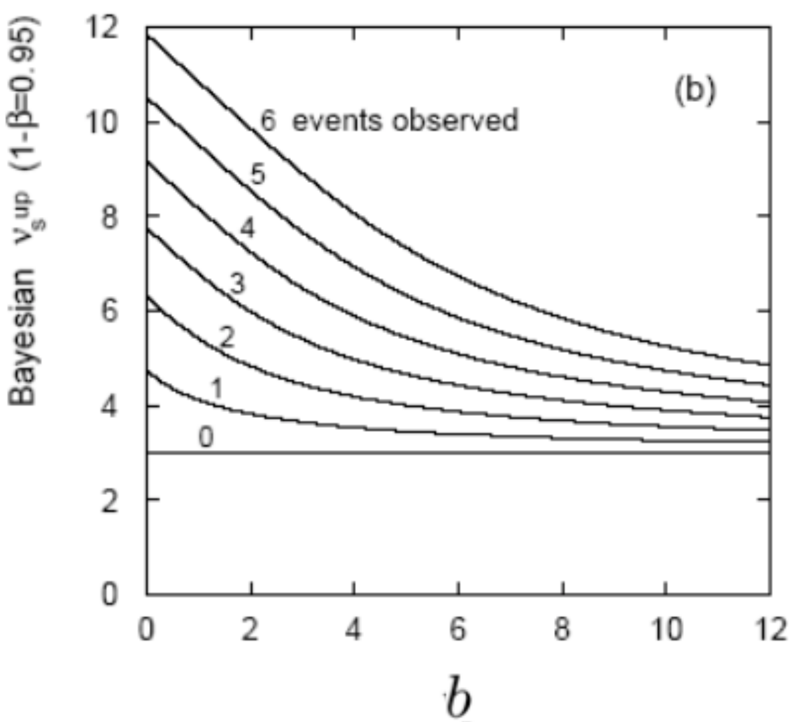
First we need a prior, let's start with

$$\pi(s) \sim \mathrm{Uniform}(0, 100)$$

In our example, when n=0, b=3.1

$$p(s|n) \sim p(n|s)\pi(s) \sim e^{-(s+b)} * 0.01(s \in [0, 100])$$

We find s_{upp} = -2.996 *no matter the value of b*



Bayesian flip-flopping

- Using a HPD interval will naturally produce either a one- or two-sided credible interval
- Or, you can always choose to set an upper limit—even if that's dumb
- **However**, these intervals do not have the conjugate properties of testing that confidence intervals do
- We don't have time today to talk about Bayesian hypothesis testing

Goodness-of-Fit

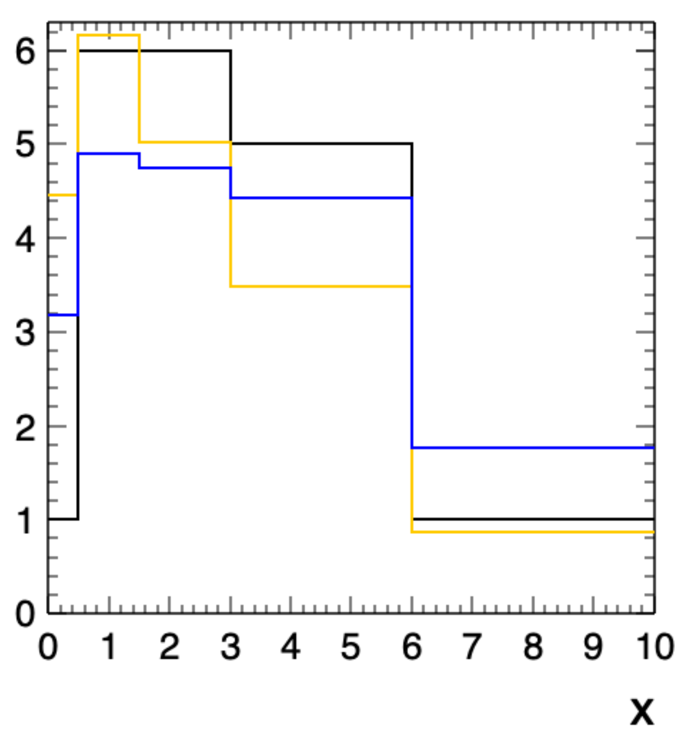
- Sometimes we want to know: "does my model with optimized parameters represent the data well?"
- In this case, H_0 is the 'saturated model', that exactly matches the data, and H_1 is the model we used to fit the data
- This is only well defined for binned likelihoods

Poisson Likelihood Ratio

$$t(x) = \prod_i \frac{\lambda_i^{n_i} e^{-\lambda_i}}{n_i!} \prod_i \frac{n_i!}{n_i^{n_i} e^{-n_i}}$$

$$t(x) = \prod_i \left(\frac{\lambda_i}{n_i}\right)^{n_i} e^{n_i - \lambda_i}$$

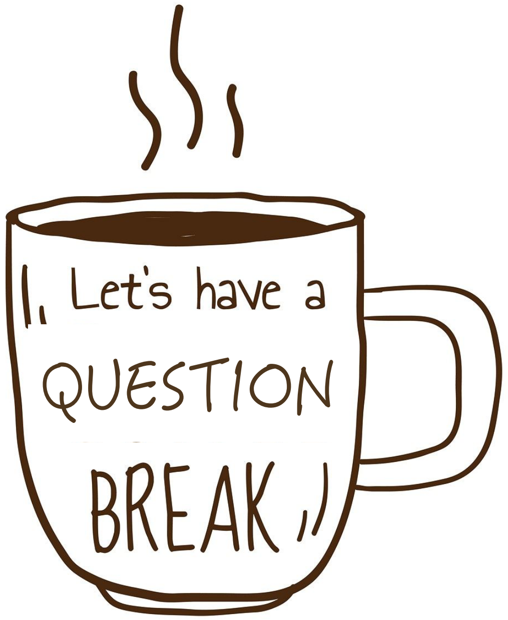
$$-\ln(t(x)) = \sum_i \lambda_i - n_i + n_i \ln\left(\frac{n_i}{\lambda_i}\right)$$



-ln(t(x))=1.526
p=0.466

- Note that this can be used any other place you'd use a likelihood!
- This will be distributed as a χ^2 with dof as the number of bins minus the number of free parameters -1

BREAK TIME



Multivariate Techniques

- Generally we refer to multivariate techniques as a way of going from many dimensions of information to one dimension. This includes:
  - Analytic techniques
  - Machine Learning
- We've seen one multivariate technique already— likelihood ratios!
- I'm mostly going to talk about this in the light of a classification problem, but there's active, ongoing research in applying these methods to MC generation, fitting, limit setting and more

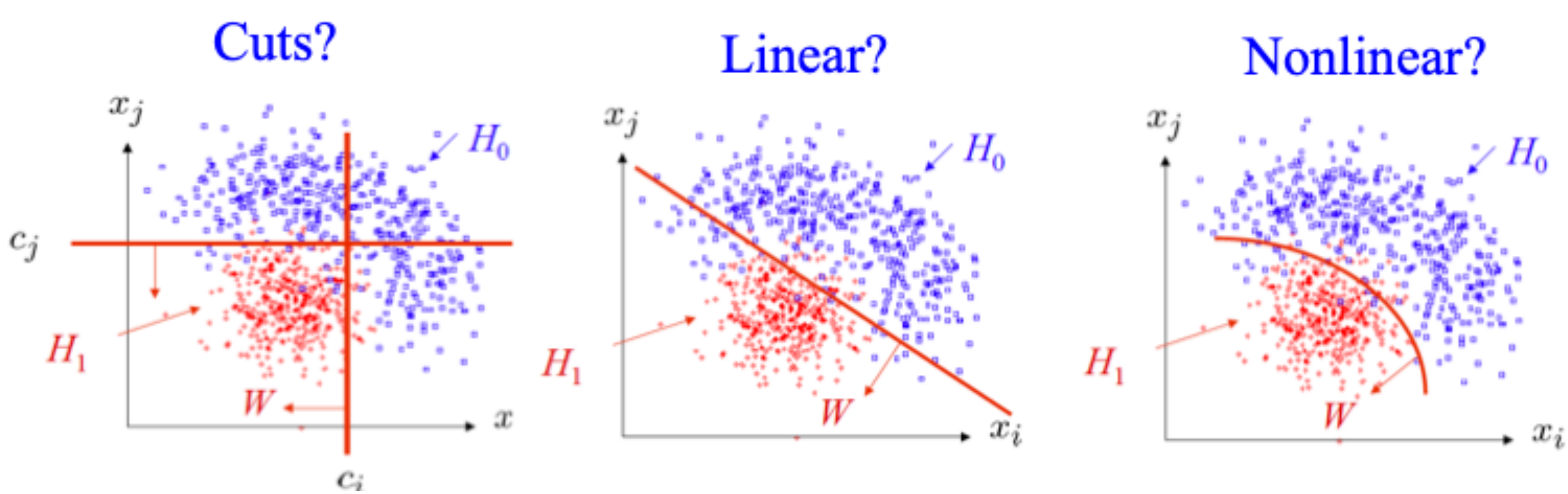
With thanks to G. Cowan for many of the examples

Tools

- ROOT has a number of multivariate tools available in TMVA
- Python packages Scikit-learn and TensorFlow are the standards

Classification

If we had good knowledge of our PDFs, this would be easy! But what if we don't?



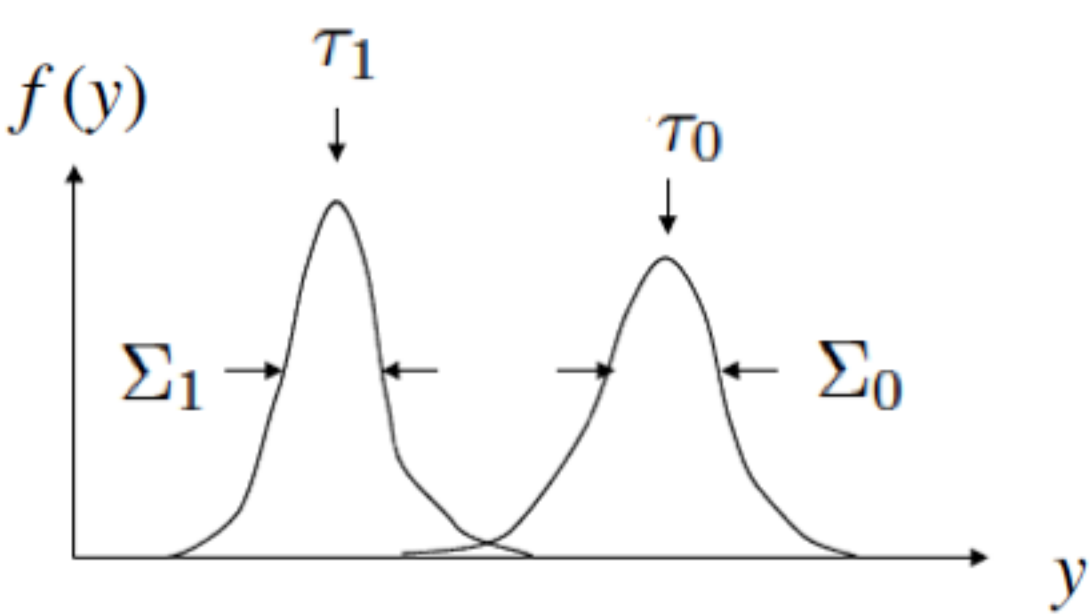
General Terms

- Purity: fraction of signal events of selected events
- Efficiency: fraction of all signal events which are in the selection
- Training sample: MC used to optimize the discriminator
- Testing sample: MC used after optimization to test discrimination

Fisher (or Linear) Discriminant

$$y(\vec{x}) = \sum_{i=1}^{n} w_i x_i = \vec{w}^T \vec{x}$$

Choose w_i for maximum separation and minimum width

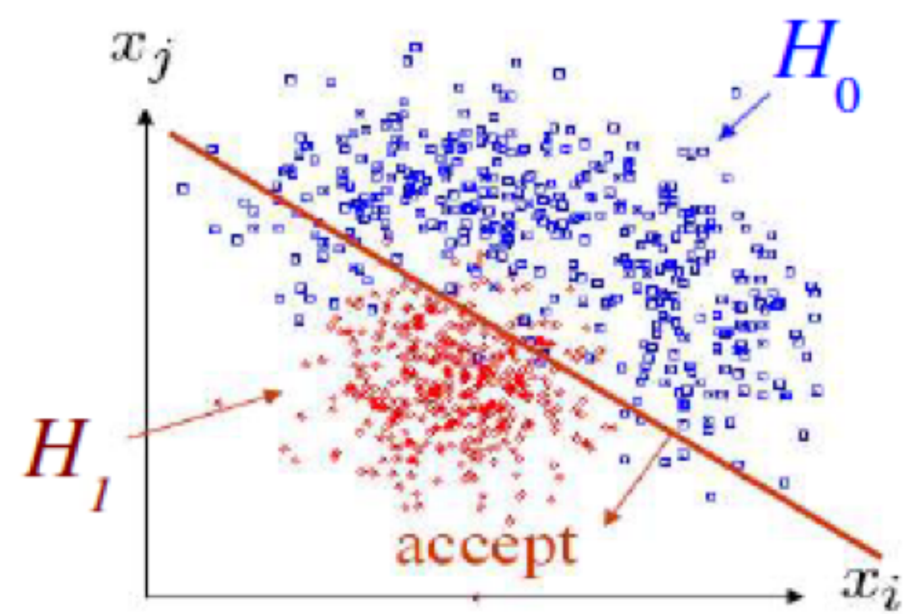


maximize

$$J(\mathbf{w}) = \frac{(\tau_1 - \tau_0)^2}{\Sigma_0^2 + \Sigma_1^2}$$

$$y(\vec{x}) = \vec{w}^T \vec{x} \qquad \text{with } \vec{w} \propto W^{-1}(\vec{\mu}_0 - \vec{\mu}_1)$$

$$W_{ij} = (V_0 + V_1)_{ij}$$



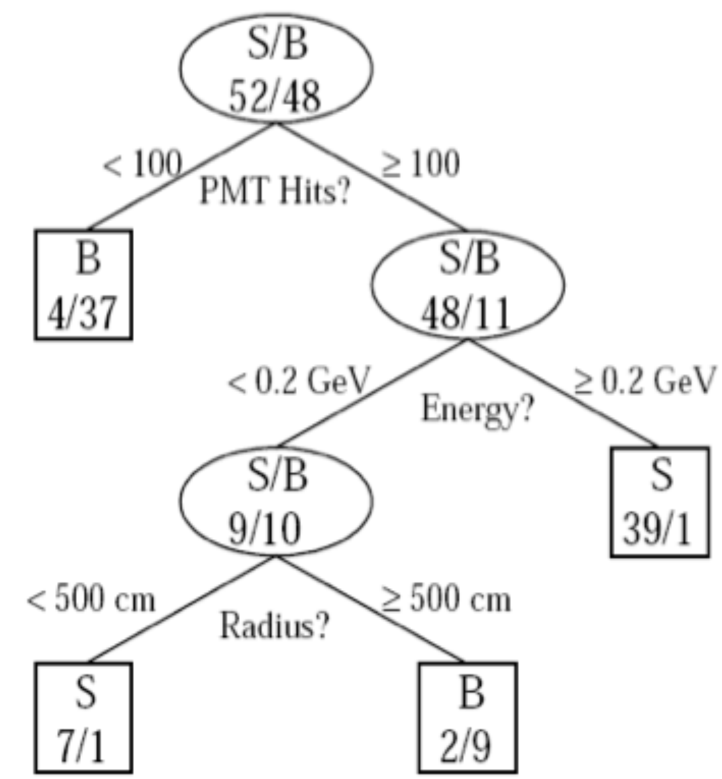
Projecting on an axis transverse to the decision boundary shows maximum separation

Decision Trees

- From the set of input variables, find the single variable that, with a cut, creates the greatest increase in sample purity
- Subsequent nodes classified as Signal or Background
- Iterate until a stop condition is reached

$$P = \frac{\sum_{\text{signal}} w_i}{\sum_{\text{signal}} w_i + \sum_{\text{background}} w_i}$$

$w_i = \text{weight}$



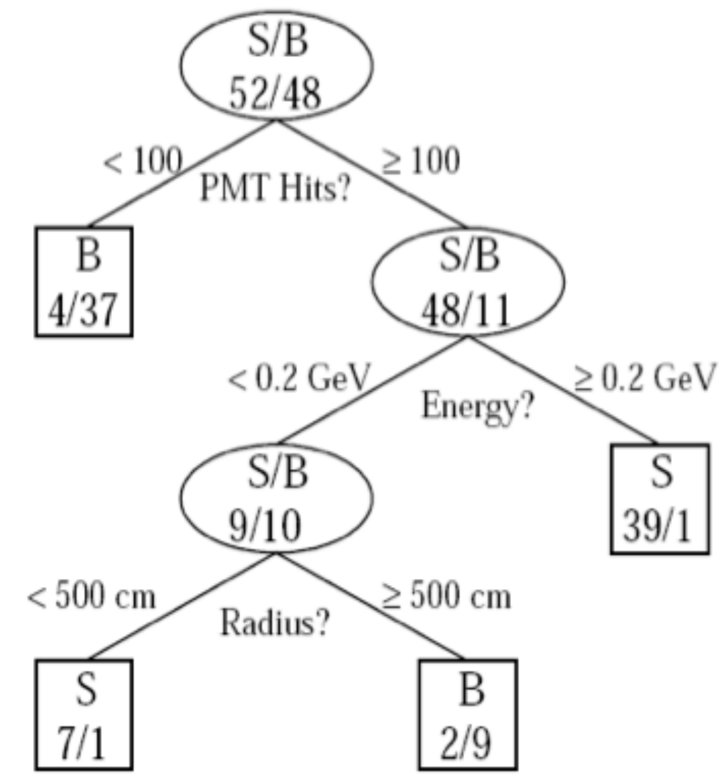
Example by MiniBooNE experiment, B. Roe et al., NIM 543 (2005) 577

Finding the Best Cut

- The level of separation within a node can be quantified by the Gini Coefficient: $G = p(1-p)$
- If a cut separates set A into subsets B and C, maximize $\Delta = W_a G_a - W_b G_b - W_c G_c$, with $W_a = \sum_{i \in a} w_i$

Decision Trees

- Terminal nodes are classified as Signal or Background by majority
- This method tends to react strongly to fluctuations in the training sample
- *Boosting* the tree can smooth out these effects

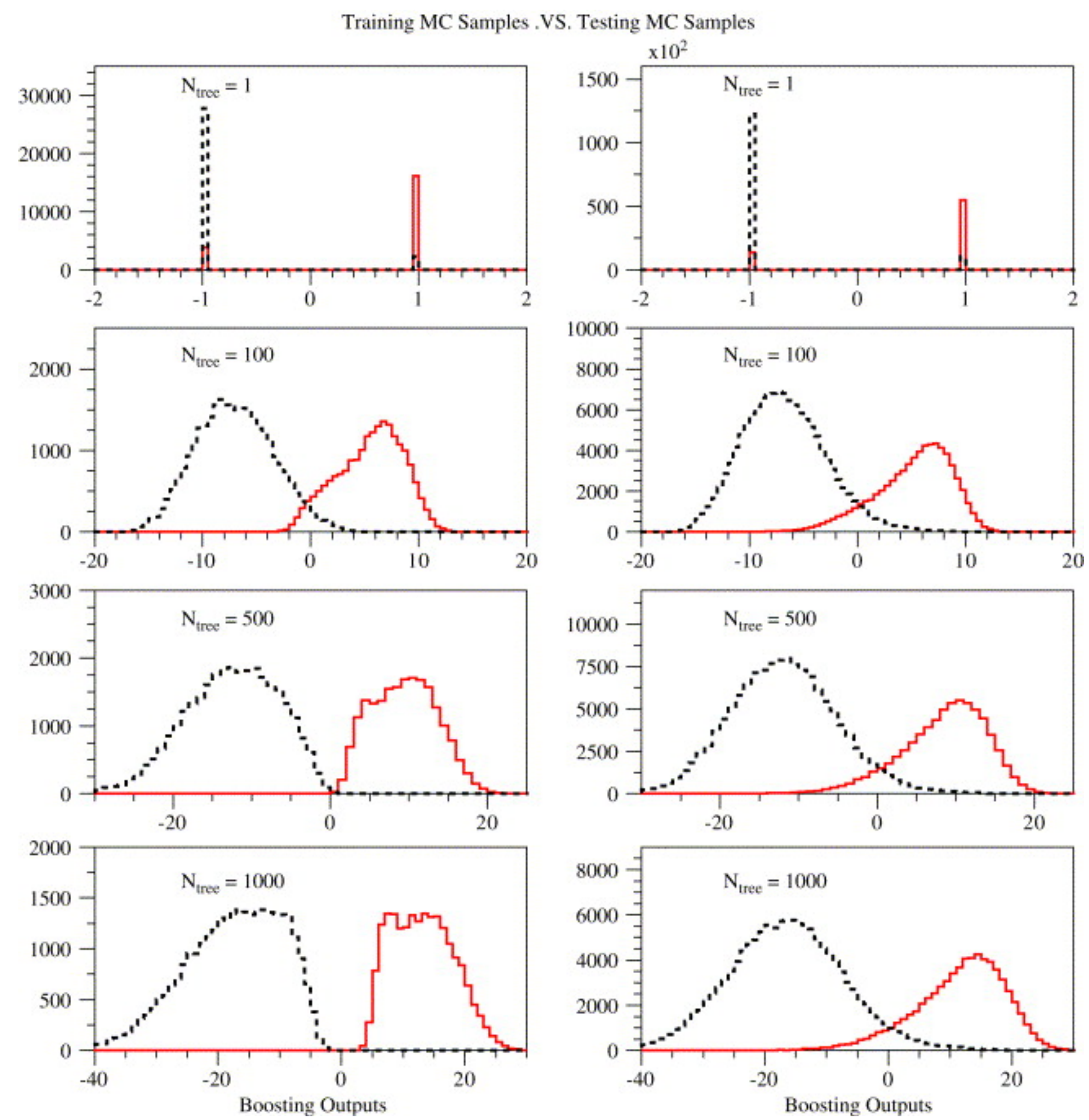


Example by MiniBooNE experiment, B. Roe et al., NIM 543 (2005) 577

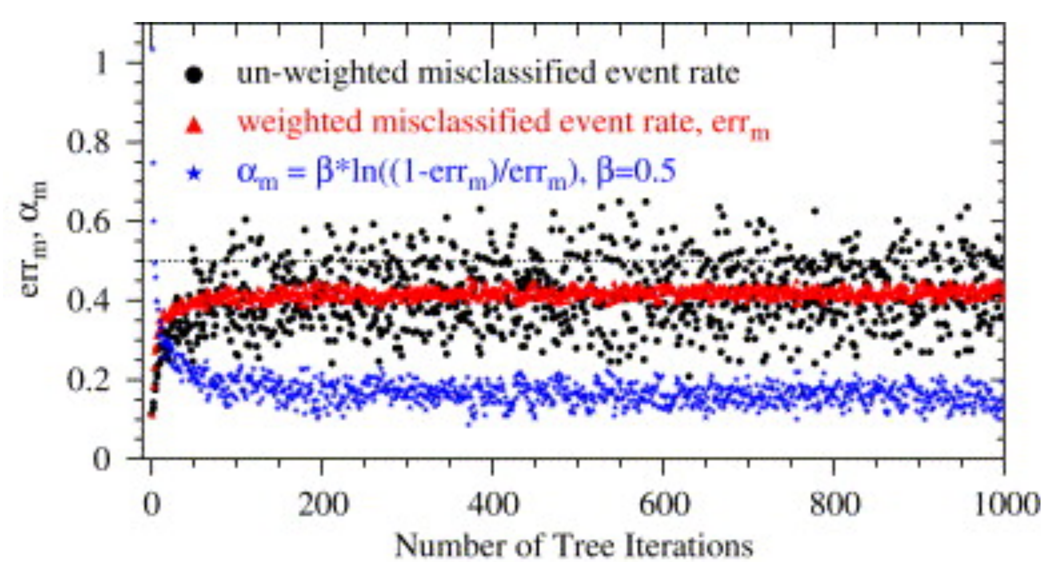
Boosted Decision Trees

- Many kinds of boosting algorithm—not just for decision trees!
- AdaBoost, ε-Boost, LogitBoost, etc
- General principle is to boost the weights of misclassified events in subsequent iterations to improve performance

MiniBooNE Example



MiniBooNE use AdaBoost, and finds stability after a few hundred iterations



Neural Networks

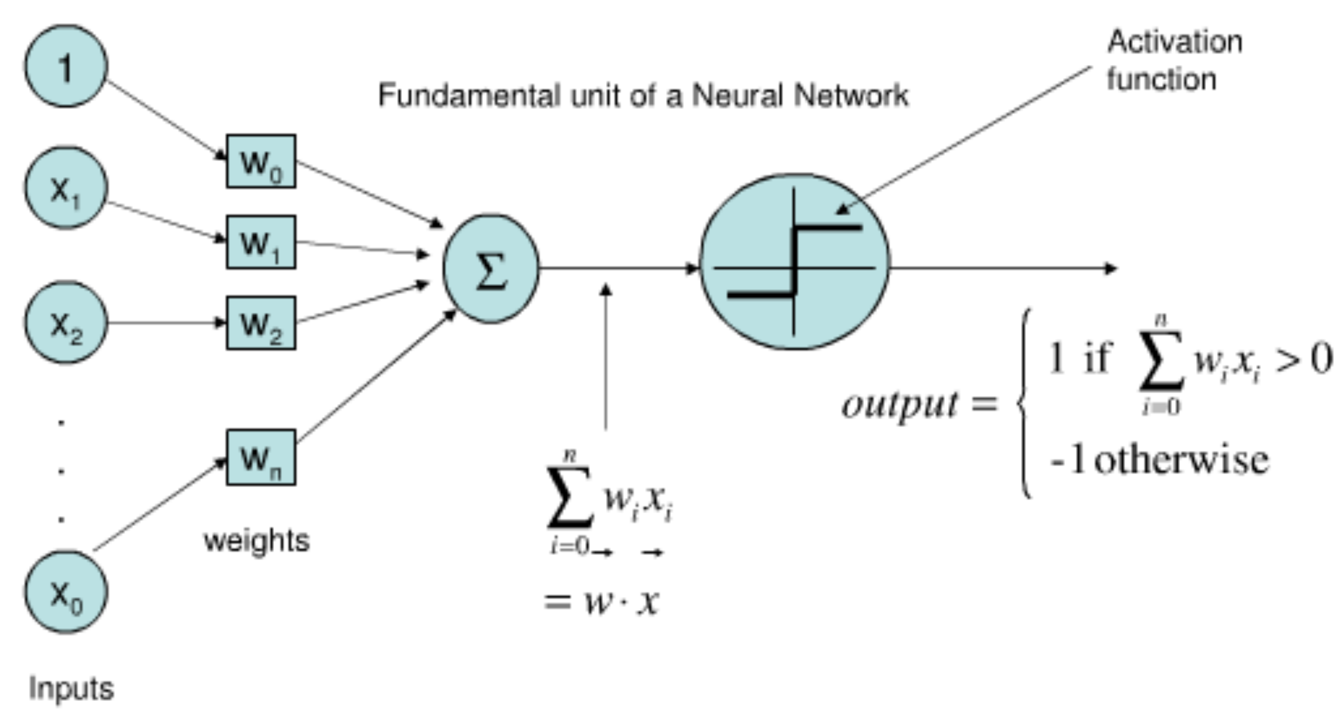
- Neural Networks are an attempt to model neural processes
- They've been around more than 80 years—widely used in ML and AI
- Essentially a way of parameterizing a set of basis functions defining the transformation of a feature space

Single Layer Perceptron

Define a discriminant: $y(\vec{x}) = h\left(w_0 + \sum w_i x_i\right)$

Typically h is some sigmoid function, called the activation function

This is called the 'single layer perceptron' and, when h is monotonic, equivalent to a linear discriminant



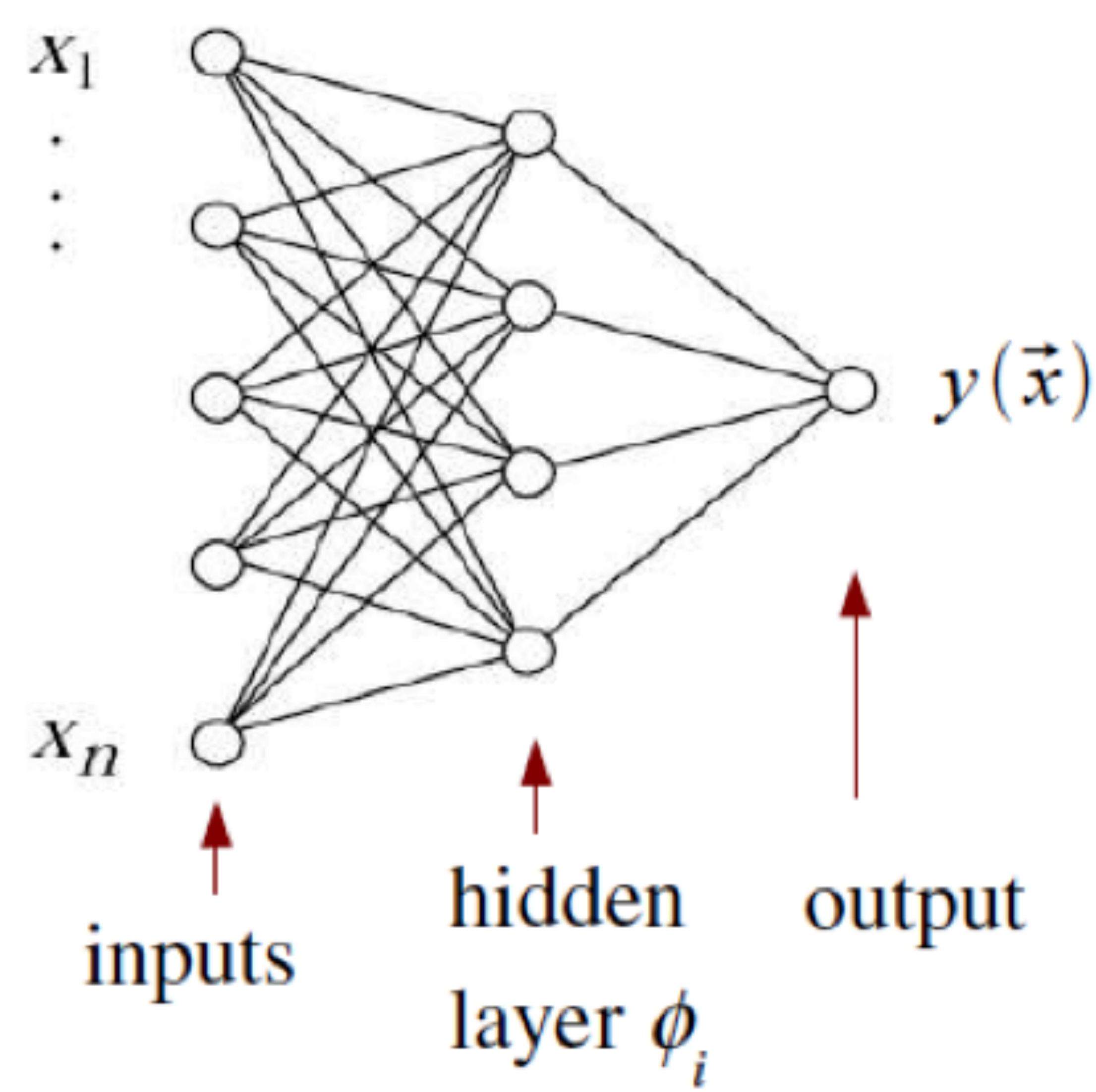
Multilayer Perceptron

Generalize to more than one layer

Superscript for weights indicates layer number

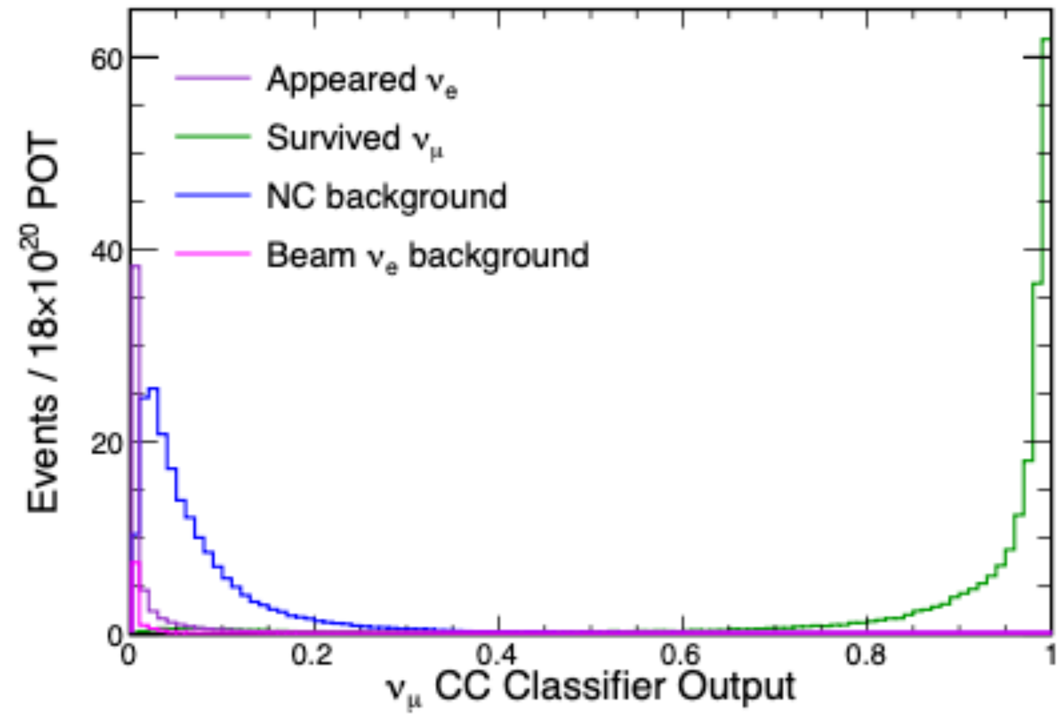
$$\varphi_i(\vec{x}) = h\left(w_{i0}^{(1)} + \sum_{j=1}^{n} w_{ij}^{(1)} x_j\right)$$

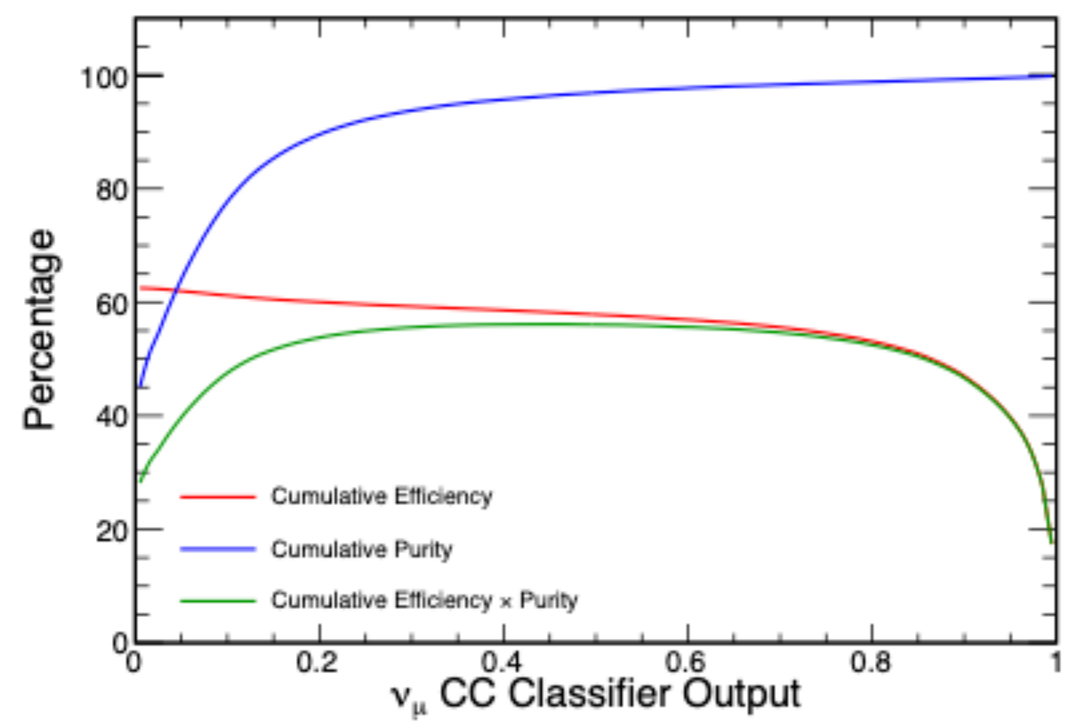
$$y(\vec{x}) = h\left(w_{10}^{(2)} + \sum_{j=1}^{n} w_{1j}^{(2)} \varphi_j(\vec{x})\right)$$



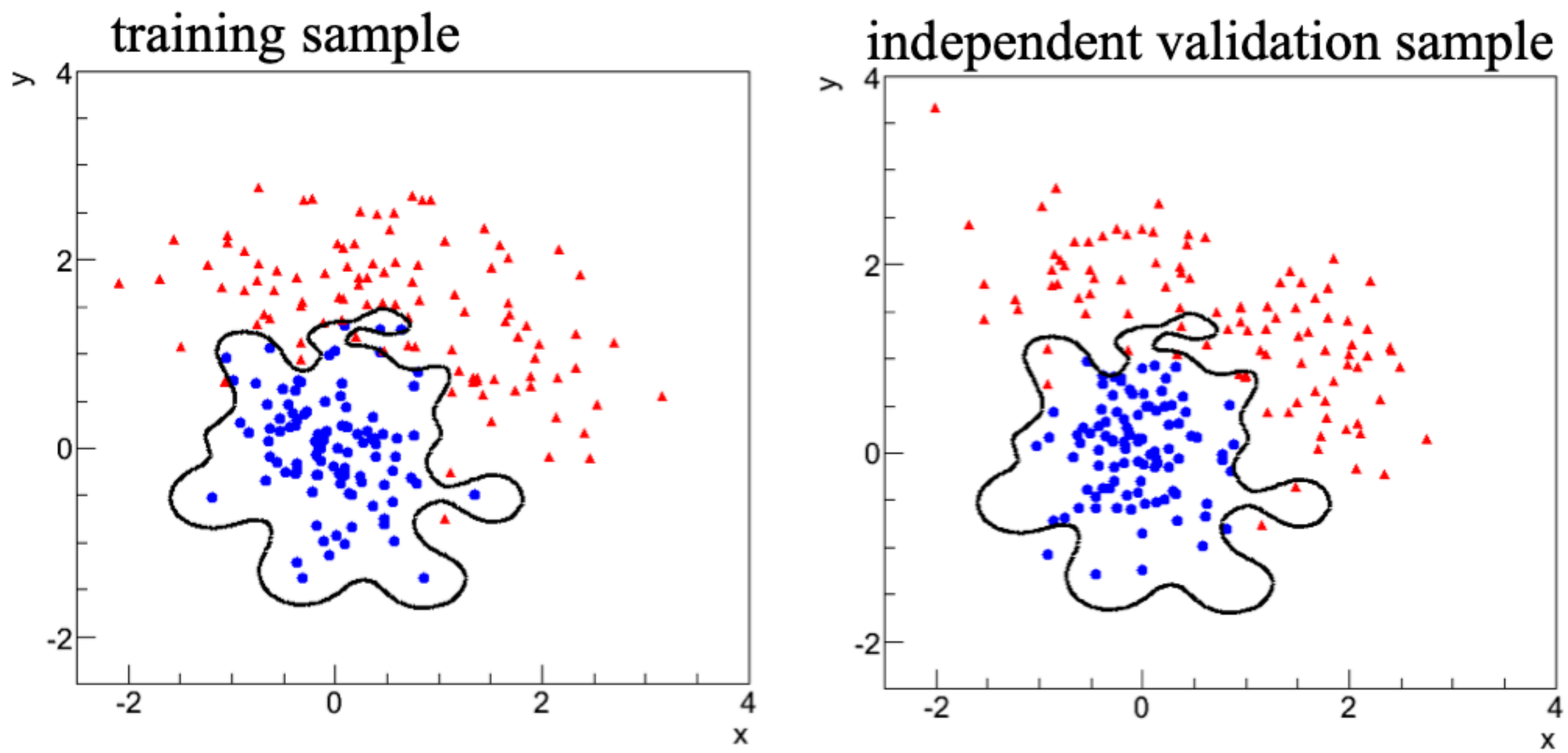
Example: NOvA

- Classifying event types as νe, νμ, or NC
- Uses a convolutional neural network (CNN)
  - CNNs do some dimensionality reduction in hidden layers
  - Reduces computational complexity





Common Pitfalls



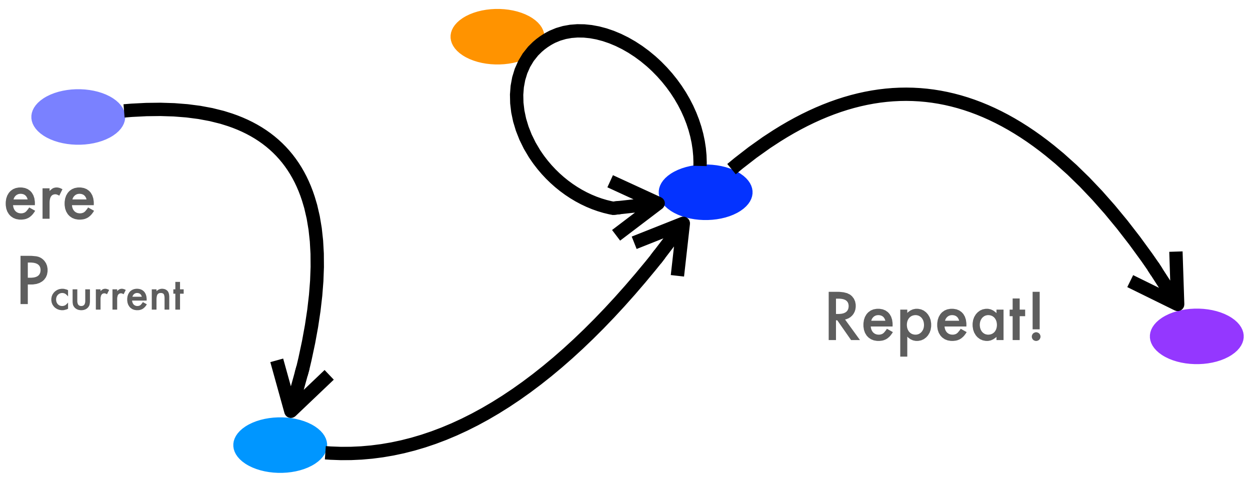
- Overtraining—making your acceptance region too sensitive to your training sample
- Data/MC disagreement—ensuring that you don't have a garbage-in-garbage-out problem

BREAK TIME

Markov Chain Monte Carlo

A Markov Chain maps out the probability density of a likelihood function, P

Don't step here– P too small



Start here
Calculate $P_{current}$

Repeat!

Propose another point
Calculate $P_{proposed}$; if better, step to that point
if not, step with probability $P_{proposed}/P_{current}$

- Use Metropolis-Hastings algorithm with MCMC; doesn't require calculating likelihood derivatives

Estimating Parameters and Uncertainties