Lecture 5:

More Likelihood & Bayes' Theorem

- Joint Analysis
- Constraints
- Extended Likelihood
- Asimov Data Sets
- Binned Histogram PDFs & Error Propagation
- Bayes Theorem

Joint Analysis of Multiple Data Sets

$$-2\log L = \sum_{i=1}^{n} -2\log\left[P\left(x_i \mid H(\mathbf{q})\right)\right]$$

$$= \sum_{i=1}^{k} -2\log\left[P\left(x_i \mid H(\mathbf{q})\right)\right] + \sum_{i=k+1}^{n} -2\log\left[P\left(x_i \mid H(\mathbf{q})\right)\right]$$

Likelihood for one set of data under $\mathrm{H}(\mathbf{q})$.

Likelihood for the same hypothesis, but a different set of data. Could even be from a different experiment and assessed in a completely different way, so long as it is eventually turned into a probability.

Can jointly analyse multiple data sets from multiple experiments to determine the best overall parameter estimations by adding together their likelihoods over the same parameter space

It's always good to show your likelihood space as part of the presentation of results both as an overall summary of the relevant information content of your data and to allow for such joint analyses

Applying Constraints

Assume that several parameters in your model have uncertainties that are subject to external constraints, for example, from a series of calibrations. These can generally also be included as part of the likelihood. For instance, if there are *m* parameters subject to Gaussian constraints, then:

$$L = \left(\prod_{i=1}^{n} P(x_i \,|\, H(\mathbf{q})) \right) \left(\prod_{j=1}^{m} \frac{1}{\sqrt{2\pi}\sigma_j} \; e^{-\frac{(a_j - \mu_j)^2}{2\sigma_j^2}} \right)$$

$$\log L = \sum_{i=1}^{n} \log \left[P\left(x_i \,|\, H(\mathbf{q})\right) \right] - \sum_{j=1}^{n} \frac{(a_j - \mu_j)^2}{2\sigma_j^2} - \sum_{j=1}^{n} \sqrt{2\pi}\sigma_j$$

Can ignore this term, since this is a constant and we're only concerned with derivatives and differences of the likelihood

"Extended" Likelihood

The number of events, n, in a data set is often the result of Poisson fluctuations about the expected mean number of events. If the expected mean is itself a parameter of interest (*e.g.* the "true" flux of signal and/or background events), the associated Poisson fluctuation can then be included in the likelihood as follows:

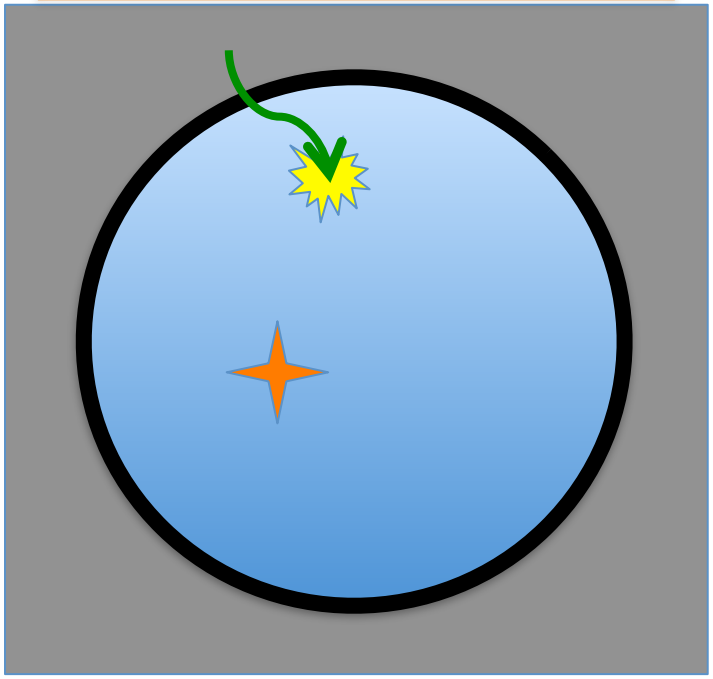
$$L = \left(\frac{\mu^n e^{-\mu}}{n!} \right) \prod_{i=1}^{n} P(x_i \,|\, H(\mathbf{q}))$$

$$\log L = n \log \mu - \mu - log(n!) + \sum_{i=1}^{n} \log \left[P\left(x_i \,|\, H(\mathbf{q})\right) \right]$$

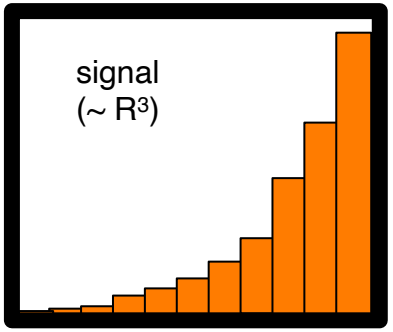
Can ignore this term, since this is a constant and we're only concerned with derivatives and differences of the likelihood

Example of a 2-component model of signal and background:

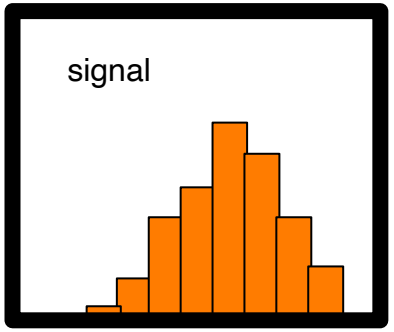
An Experiment Searching for Rare Interactions



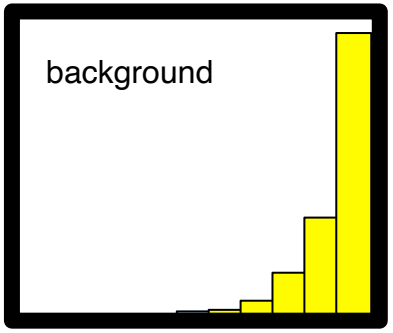
Simulation and/or Calibration Data



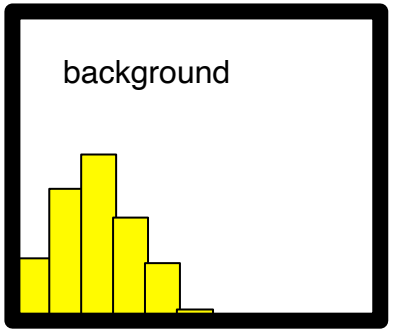
"Radius"



"Energy"



"Radius"



"Energy"

Reconstructed energy and position could be correlated (e.g. higher energy events could be easier to reconstruct accurately). So, form 2-D histograms to preserve these correlations and normalise these to one to produce PDFs for each type of event class:

$$P(\tilde{E}_i, \tilde{R}_i | S)$$

$$P(\tilde{E}_i, \tilde{R}_i | B)$$

Consider a hypothesis, H, in which a certain fraction of the data is signal and remaining fraction is background:

$$P(\tilde{E}_i, \tilde{R}_i \,|\, H) = P(\tilde{E}_i, \tilde{R}_i \,|\, S)\left(\frac{\mu_S}{\mu_S + \mu_B}\right) + P(\tilde{E}_i, \tilde{R}_i \,|\, B)\left(\frac{\mu_B}{\mu_S + \mu_B}\right)$$

where $\mu_{total} = \mu_S + \mu_B$

extended likelihood part

$$\log L = n \log(\mu_S + \mu_B) - (\mu_S + \mu_B)$$

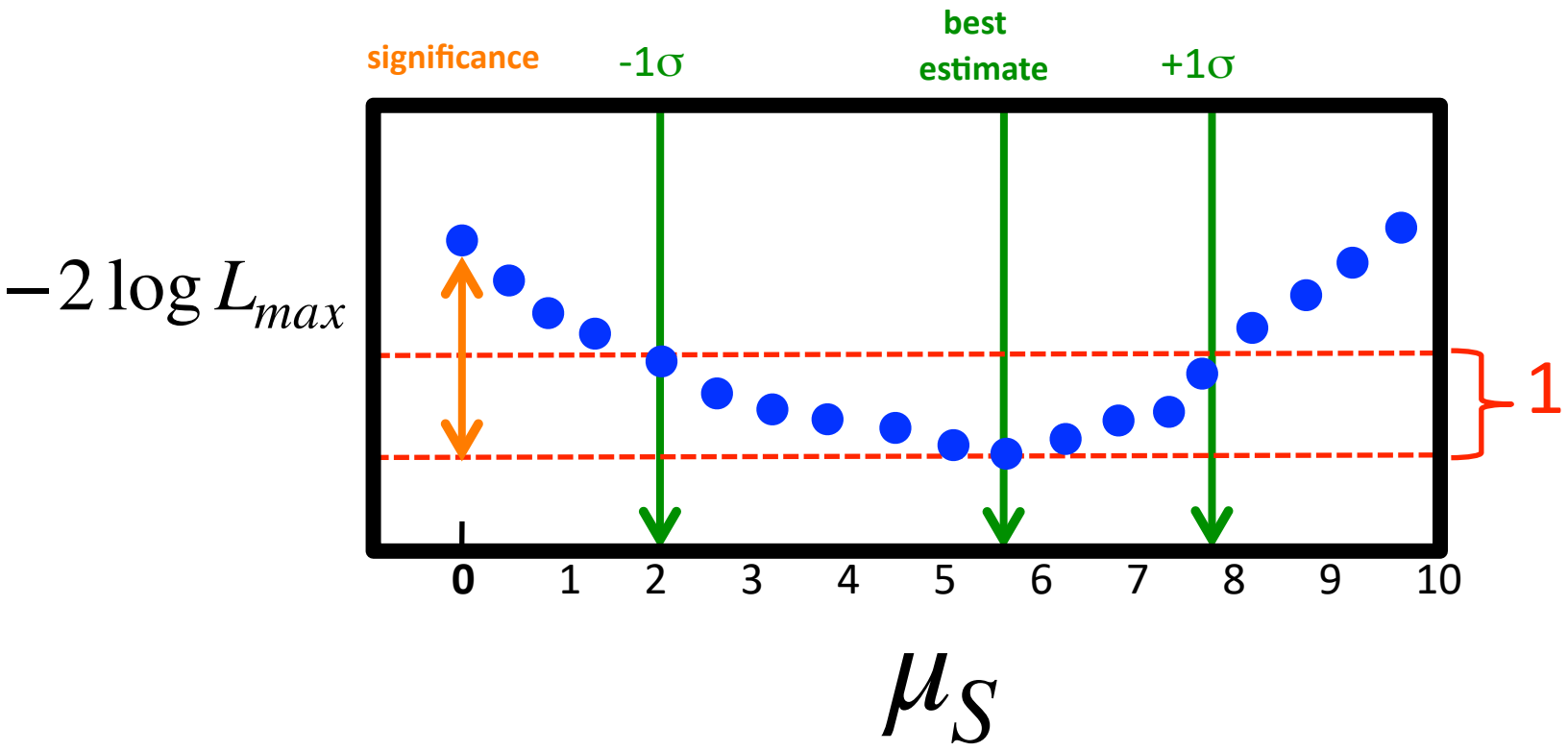
$$+ \sum_{i=1}^{n} \log \left[P(\tilde{E}_i, \tilde{R}_i \,|\, S)\left(\frac{\mu_S}{\mu_S + \mu_B}\right) + P(\tilde{E}_i, \tilde{R}_i \,|\, B)\left(\frac{\mu_B}{\mu_S + \mu_B}\right) \right]$$

→ Maximise $log\ L$ (or minimise $-2log\ L$) over μ_S and μ_B in addition to any other parameters of the model

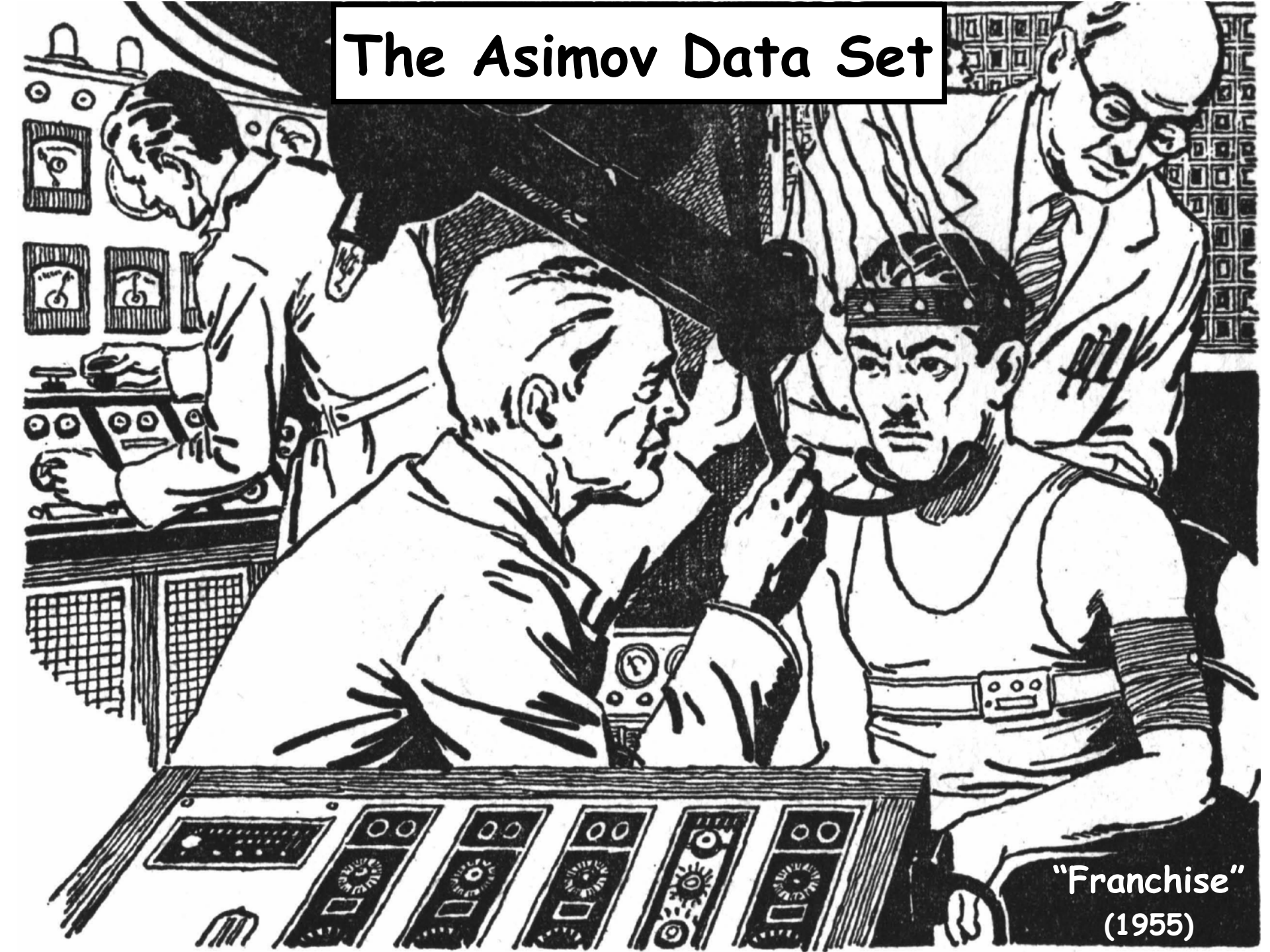
We're particularly interested in the value and significance of the signal, so look at the projection where the likelihood is maximised over all other free model parameters as μ_S is varied:

"nuisance parameters"

"Profile Likelihood"



The Asimov Data Set



"Franchise"
(1955)

Assume that we have a set of multi-dimensional PDFs defined by an arbitrary number of bins (N_{bins}) that can be combined under a particular model ($\boldsymbol{m}$, defined by $\boldsymbol{q}$ parameters) to yield a predicted mean number of observed counts ($\mathbf{n}$) in each bin ($\boldsymbol{i}$) from a given data set. The likelihood can then be expressed as:

$$\mathscr{L} = \prod_{i=1}^{N_{bins}} \left[\frac{m_i(\mathbf{q})^{n_i} e^{-m_i(\mathbf{q})}}{n_i!} \right]$$

$$\log \mathscr{L} = \sum_{i=1}^{N_{bins}} \left[n_i \log m_i(\mathbf{q}) - m_i(\mathbf{q}) - \log n_i! \right]$$

The log-likelihood ratio with respect to some nominal model, $\boldsymbol{m(q_0)}$, is then given by:

$$\log \frac{\mathscr{L}}{\mathscr{L}_0} \equiv \log \mathscr{L}_R$$

$$= \sum_{i=1}^{N_{bins}} \left[n_i \log m_i(\mathbf{q}) - m_i(\mathbf{q}) - n_i \log m_i(\mathbf{q_0}) + m_i(\mathbf{q_0}) \right]$$

$\boldsymbol{q_0}$ might, for example, represent the null hypothesis or could just be the point where the likelihood is maximum

Say we're interested in what to expect on average for the log-likelihood ratio as a function of 'test' parameter values:

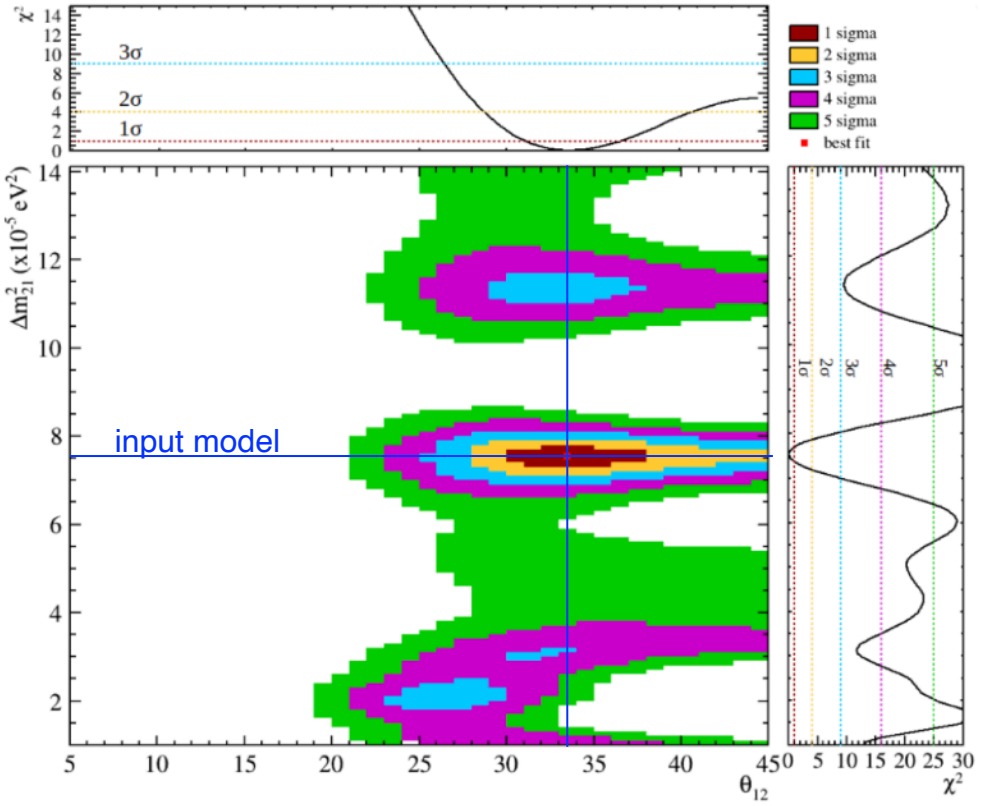
$$\left\langle \log \mathscr{L}_R \right\rangle = \left\langle \sum_{i=1}^{N_{bins}} \left[n_i \log m_i(\mathbf{q}) - m_i(\mathbf{q}) - n_i \log m_i(\mathbf{q_0}) + m_i(\mathbf{q_0}) \right] \right\rangle$$

$$= \sum_{i=1}^{N_{bins}} \left\langle \left[n_i \log m_i(\mathbf{q}) - m_i(\mathbf{q}) - n_i \log m_i(\mathbf{q_0}) + m_i(\mathbf{q_0}) \right] \right\rangle$$

$$= \sum_{i=1}^{N_{bins}} \left[\left\langle n_i \right\rangle \log m_i(\mathbf{q}) - m_i(\mathbf{q}) - \left\langle n_i \right\rangle \log m_i(\mathbf{q_0}) + m_i(\mathbf{q_0}) \right]$$

So we just need to substitute in "perfect, un-fluctuated" expectation values for a representative data set. This could, for example, be taken from scaling the PDF model for some particular set of parameters to the size of a typical data set.

Can be used to find the expected sensitivity for discovering a particular phenomenon, or the expected power to discriminate between different model, or the expected accuracy in constraining model parameters.



Expected ability to constrain oscillation parameters after 5 years of reactor anti-neutrino data from SNO+

(thanks to Iwan Morton-Blake)

Perfect data ought to give perfect results if you're doing things right!

Incredibly useful! Also an excellent way to check if your code is doing the right thing and understanding basic characteristics without having to run the full analysis chain thousands of times!

Likelihood Exercise

1. Generate a data set consisting of 100 random numbers between 0 and 10 (representing a uniform background over some arbitrary energy range) and 15 "signal events," centred on the value 6 and characterised by a Gaussian distribution with a standard deviation of 1 (representing the energy resolution).

2. Construct a likelihood function for a mixed signal and background hypothesis and then write a routine to maximise the likelihood (or minimise -2ln(likelihood)) to find the best overall position of the signal and the number of signal events. Using an Asimov data set, construct a contour plot of -2ln(maximum likelihood) as a function of the two fit parameters. Also plot the profile likelihood for each parameter separately.

3. Make the same plots for several fluctuated data sets and compare.

4. Estimate the uncertainty in each parameter based on Wilks' Theorem using both the Asimov and fluctuated data sets above. Verify this by generating 1000 fluctuated data sets and maximising the likelihoods to find how often the fit values fall within 1σ and 2σ of the true values for each parameter.

5. Separately make a scatter plot of fit signal position vs fit number of signal events along with the 1σ and 2σ contours based on Wilks' Theorem, now assuming 2 degrees of freedom. Verify the fraction of events within each contour.

6. Draw 1σ and 2σ Bayesian contours on the scatter plot of step 4, assuming priors that are uniform in position and event rate.

7. **(Bonus question)** Repeat the generation and fitting in step 3 for true signal values of 5, 10, 20, 30 and plot the average significance of signal detection, in terms of standard deviations based on Wilks' Theorem, as a function of the signal strength. Similarly, repeat with the number of signal fixed to 15 again, but with the number of backgrounds taken as 50, 200, 400, 1000. Again, plot the average significance as a function background number.

Binned Histogram PDFs

A common approach to producing PDFs involves binning data generated by simulation or calibration runs and then normalising the resulting histogram by the number of entries.

It is important to ensure that sufficient statistics are used in the production of these histograms so as to accurately characterise the true PDFs, preserve important correlations and avoid sampled bins that appear to have ‘zero probability’ due to fluctuations, which will zero out the entire likelihood calculation (and cause infinities in the log).

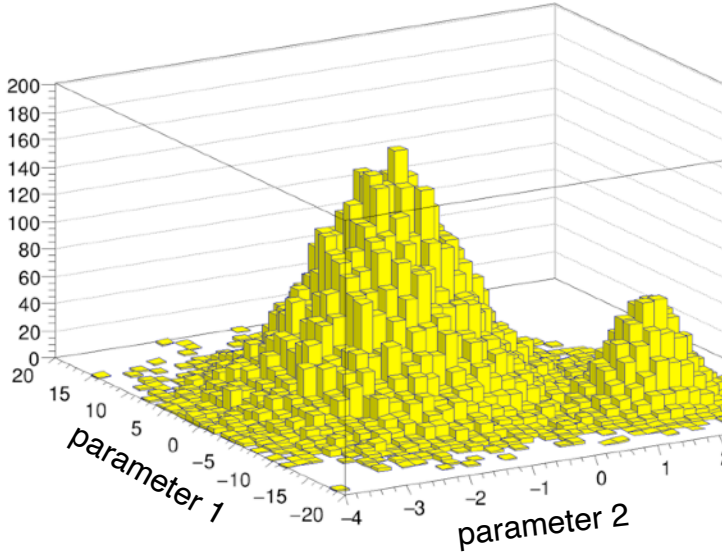
This can be particularly tricky for multi-dimensional PDFs with a large phase space…

Useful Rules of Thumb

- Focus on the main parameters and most important correlations;
- Minimise the dimensionality by choose parameters and parameter combinations that are as independent as possible to reduce sharp correlation features and allow factorisation;
- Choose the largest binning that still provides adequate resolution for important features;
- Use overflow and underflow bins at the distribution edges to avoid empty bins;
- PDF statistics should be at least an order of magnitude larger than the data set to which it will be applied.

Accounting for Statistical Uncertainties in PDFs

Let's say you create a set of PDFs for some parameters by running lots of simulations, binning the resulting distributions of parameter values and then normalising the areas of each histogram to one.



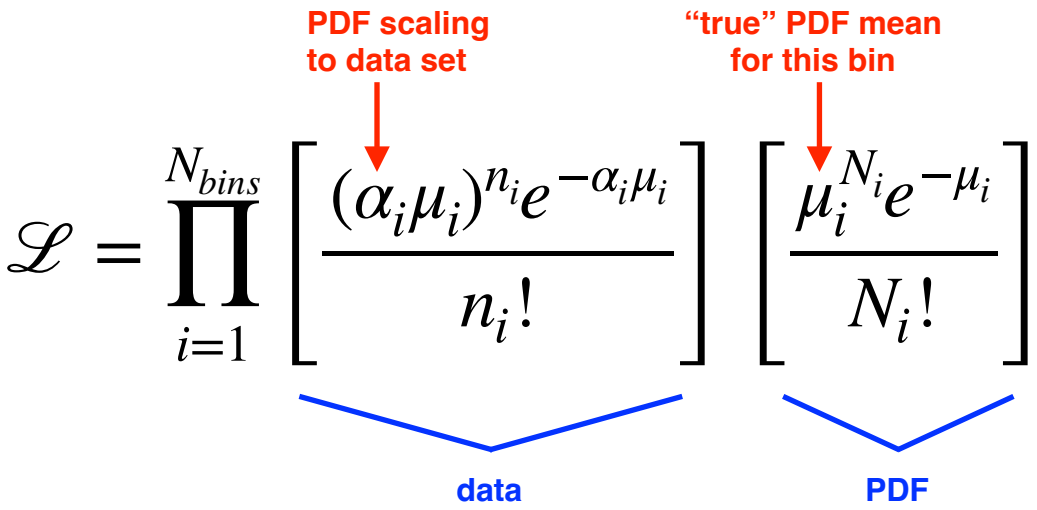
How do you deal with the statistical uncertainties in the constructed PDFs?

Could smooth PDFs, but can be tricky in multiple dimensions and has the potential to produce artefacts

$$P(Data \mid Model) \rightarrow \cancel{P(Data \mid PDF\ Histograms)}$$

$$\searrow$$

$$P(Data, PDF\ Histograms \mid Model)$$



Want to maximise overall likelihood, so maximise here over the set of "true" PDF means

Complicated by model correlations between bins and multi-component models

First analysed by Barlow and Beeston (Comp. Phys. Comm. 77, 219, 1993)

A much more practical approximation was given by Conway (PHYSTAT 2011, arXiv:1103.0354), which is what we'll follow here.

For the ith bin in the data and PDF histogram, the contribution to the extended likelihood is:

$$-\ln \mathscr{L}_i = -n_i \ln \mu_i + \mu_i$$

where n_i = number observed and μ_i = model prediction based on the PDFs

Make Two Simplifying Assumptions:

1) Take model systematics to be uncorrelated between bins to allow bin-by-bin error propagation (conservative);
2) Assume the uncertainty in μ due to statistical fluctuations in the contributing PDFs can be approximated by a single Gaussian scaling.

We can then drop the bin subscript for simplicity incorporate the Gaussian uncertainty scaling into the likelihood for that bin as follows:

$$-\ln \mathscr{L} = -n \ln \beta\mu + \beta\mu + \frac{(\beta - 1)^2}{2\sigma^2}$$

We want to maximise the likelihood (minimise $-\ln\mathscr{L}$), which can be explicitly done bin-by-bin in the parameter β by differentiation:

$$\beta^2 + (\mu\sigma^2 - 1)\beta - n\sigma^2 = 0$$

Solve for β in each bin and calculate the likelihood…

What is σ for the bin?

Assume we are using k PDFs to model the total number of events predicted in this bin:

$$\mu = \sum_{j=1}^{k} \mu_j$$

where

$$\mu_j = f_j \left(\frac{m_j}{N_j} \right)$$

simulated events for this PDF that fall in this bin

PDF normalisation

total # simulated events for this PDF

$$\sigma_j \equiv \frac{\Delta \mu_j}{\mu_j} \simeq \frac{\Delta m_j}{m_j} \simeq \frac{1}{\sqrt{m_j}} = \sqrt{\frac{f_j}{\mu_j N_j}}$$

So just need to remember this

$$\sigma^2 = \sum_{j=1}^{k} \sigma_j^2$$

Still issues for $m_j \sim 0$, (hard to get around)

Nuisance Parameters in Binned PDFs

Nuisance parameters can affect the PDFs definitions in two ways:

1) Though unknown model parameters that define the PDF shape
2) Through systematic uncertainties in the modelled parameters themselves (such as calibration uncertainties in the energy scale etc.)

In principle, each exploration of different possible nuisance parameter values would involve re-making each PDF from scratch before applying the likelihood calculation, which can be a real pain in the neck and expensive in terms of computation time!

There are a couple useful tricks for dealing with this…

1. Re-interpret the binning ("shift the data")

Here, you basically re-interpret the PDF binning as representing modified data values.

For example, say there are systematic uncertainties in the scale and offset for reconstructed energies, $\hat{E}_i$, due to limitations in calibrations. We can take the binned PDF to represent the 'corrected' energy estimator:

$$\hat{E}_i^* = \alpha \hat{E}_i + \beta$$

where α and β are nuisance parameters varied in the fit and applied to each individual data point, but without the need to recompute the PDFs themselves.

This works well for some cases but, for example, is less straightforward for resolution systematics or for model parameter uncertainties that affect bin-to-bin correlations

2. Transform the PDFs

A more general approach* is to treat the different bins of a PDF as a representing a vector that can then be transformed to a new PDF using a modification matrix:

$$b'_i = \sum_{j=0}^{Nbins} M_{ij} b_j$$

where b'_i is the modified bin content for the new PDF. M is the matrix of modification weights based on the nature of the nuisance parameters.

This process is fast for 3 reasons:

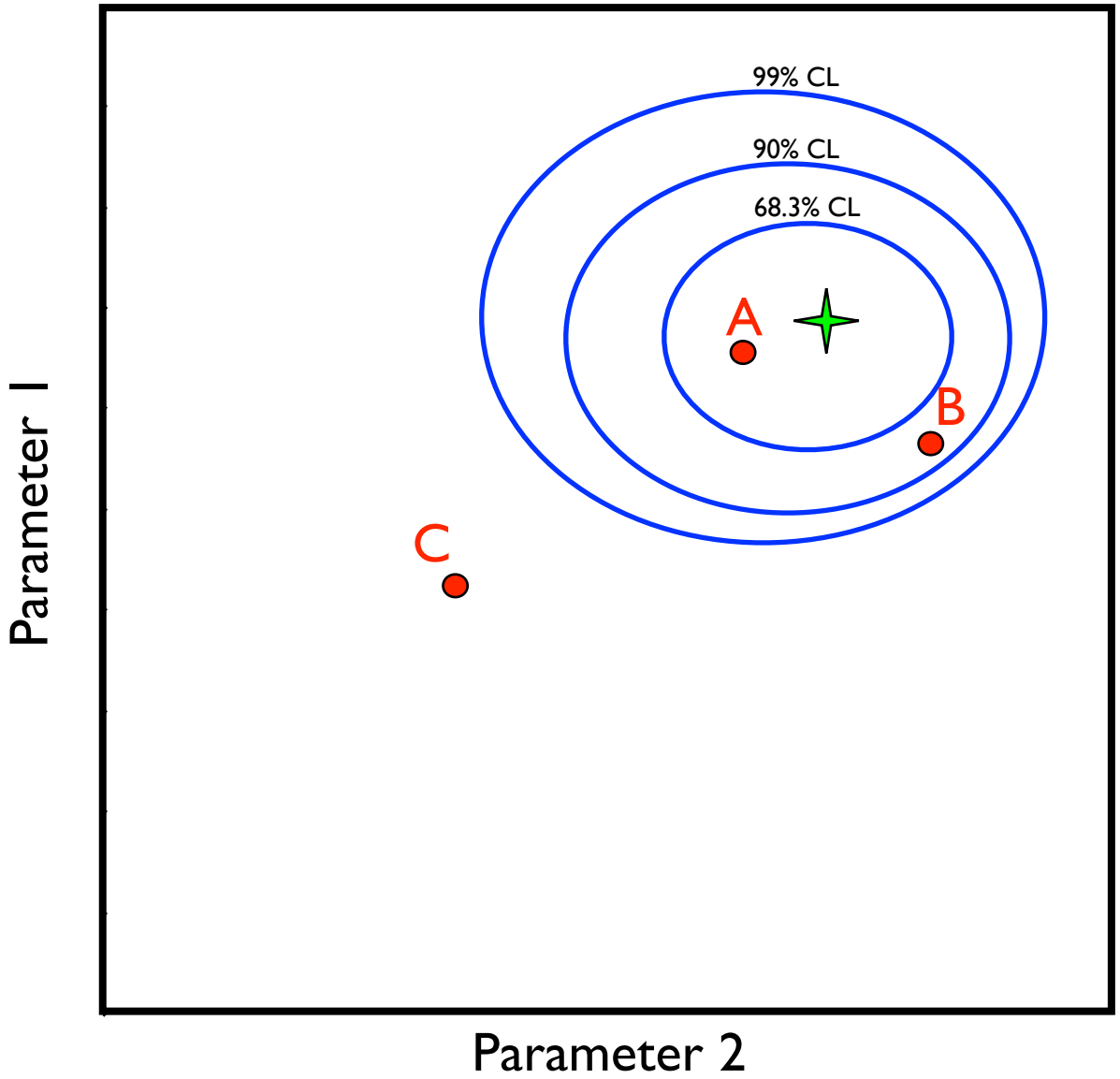
1) A single matrix can be used to represent all systematic distortions in a single step;
2) Typically, # bins being manipulated << # of events used to build the PDFs
3) Highly optimised code exists for matrix operations like this for both CPUs and GPUs

Care must be taken near distribution edges, since modifications can move events into and out of the nominal fiducial fitting region. This can be handled using adjustable buffer regions that extend beyond the edges and keeping track of normalisations.

* Jack Dunger, Springer Theses (Oxford). Springer International Publishing, Cham, 2019, 10.1007/978-3-030-31616-7

Bayesians vs Frequentists

Consider a single experiment in which 2 parameters are measured (✦) and compared with predictions from 3 different theoretical models (A, B, C)



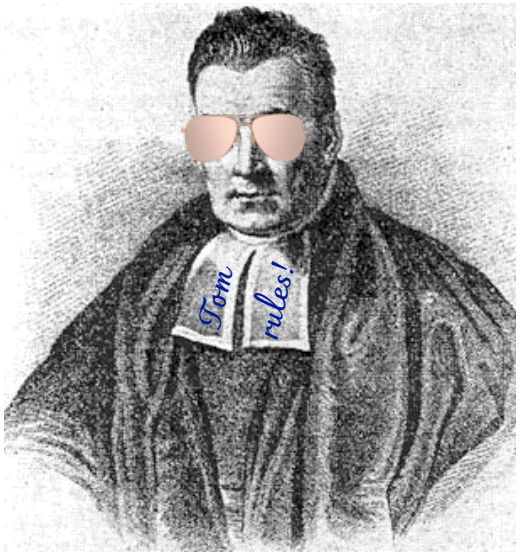
Different Definitions of Probability in relation to models:

Bayesian:

Degree of belief. Given a single measurement, ascribe "betting odds" to the phase space of possible models. Requires an assumed context for the comparison of these models (prior). There is no relevance to the "statistical coverage of a confidence interval," because there is only one measurement (which is not repeated over and over again).

Frequentist:

Frequency of occurrence given a hypothetical ensemble of 'identical' experiments. Individual measurements are not used to assess the validity of a model. There is no such thing as a "probability" for a model parameter to lie within derived bounds - either it does or it doesn't. However, if everyone played the same game, the correct model would be bounded a known fraction of the time.



Bayes' Theorem

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

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

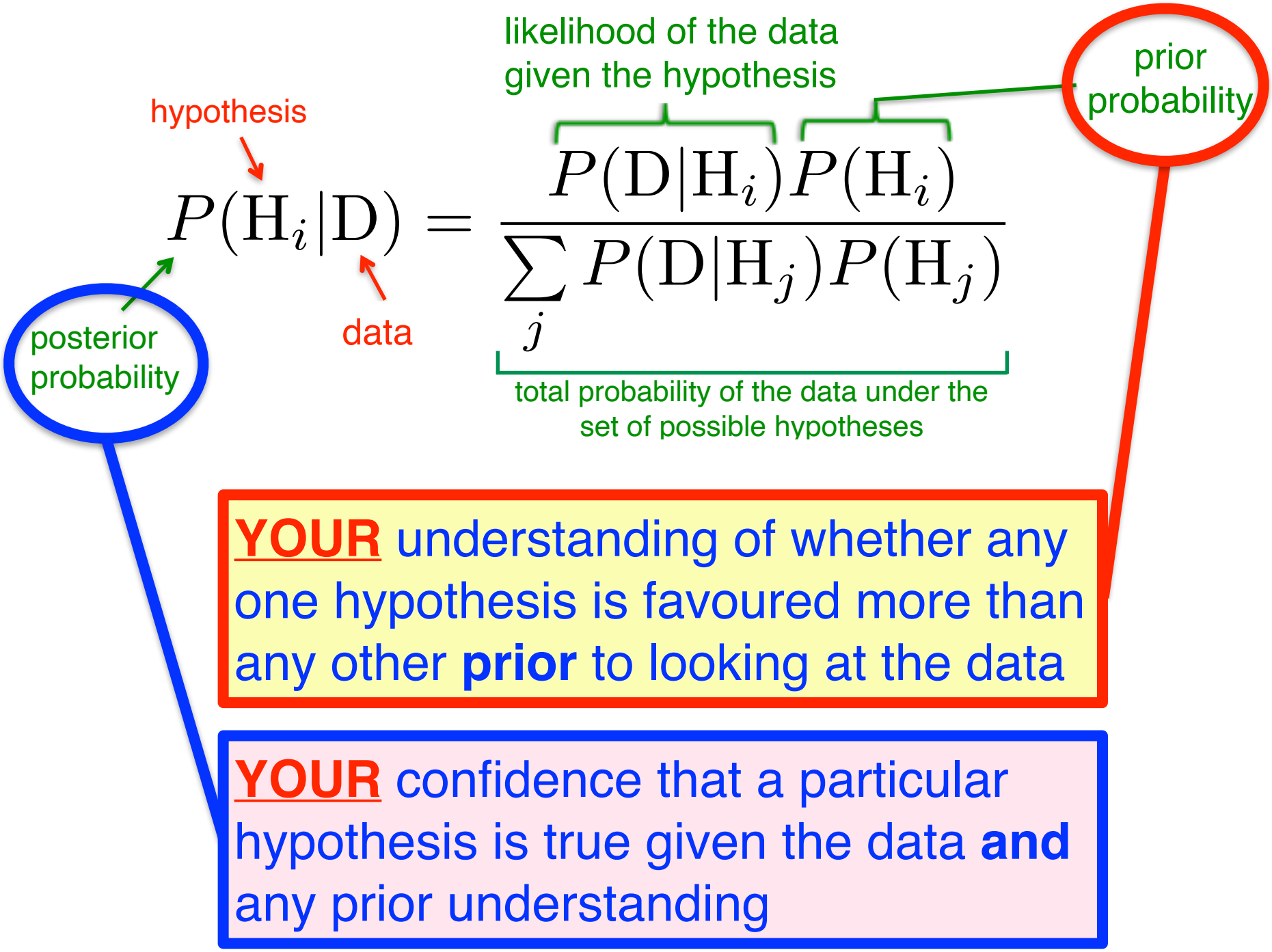
note:
$P(A\,|\,B) \neq P(B\,|\,A)$

If there are multiple versions of A to choose from, then

(*e.g.* different hypotheses)

$$P(B) = \sum_j P(B\,|\,A_j)P(A_j)$$

$$P(A_i\,|\,B) = \frac{P(B\,|\,A_i)P(A_i)}{\sum_j P(B\,|\,A_j)P(A_j)}$$


likelihood of the data
given the hypothesis
prior
probability
hypothesis
$P(\mathrm{H}_i|\mathrm{D}) = \frac{P(\mathrm{D}|\mathrm{H}_i)P(\mathrm{H}_i)}{\sum_j P(\mathrm{D}|\mathrm{H}_j)P(\mathrm{H}_j)}$
posterior
probability
data
total probability of the data under the
set of possible hypotheses
YOUR understanding of whether any one hypothesis is favoured more than any other prior to looking at the data
YOUR confidence that a particular hypothesis is true given the data and any prior understanding

Relative probability ratio between two different hypothesis (or variants of some hypothesis) given the same observed data:

$$\frac{P(H_i|D)}{P(H_k|D)} = \frac{P(D|H_i)}{P(D|H_k)} \frac{P(H_i)}{P(H_k)}$$

likelihood ratio "odds" ratio

1. Informative:

Permits known, physical constraints to be imposed (*e.g.* energies and masses must be greater than zero; the position of observed events must be inside the detector, *etc.*) and allows known attributes of the physical system to be taken into account (*e.g.* energies are being sampled from some particular spectrum; the relative probabilities for different event classes are drawn from some given distribution, *etc.*).

The probabilities of different hypotheses are the same in what metric?

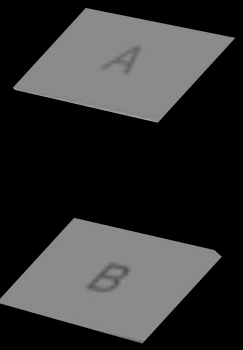
All values of **A** are equally likely $\neq$ All values of $\mathbf{A}^2$ are equally likely

2. Non-Informative:

(A Case of Too Much History!)

When there is no clear *a priori* preference, you must still choose a context to be used for comparing models.

Your brain inherently makes Bayesian inferences:

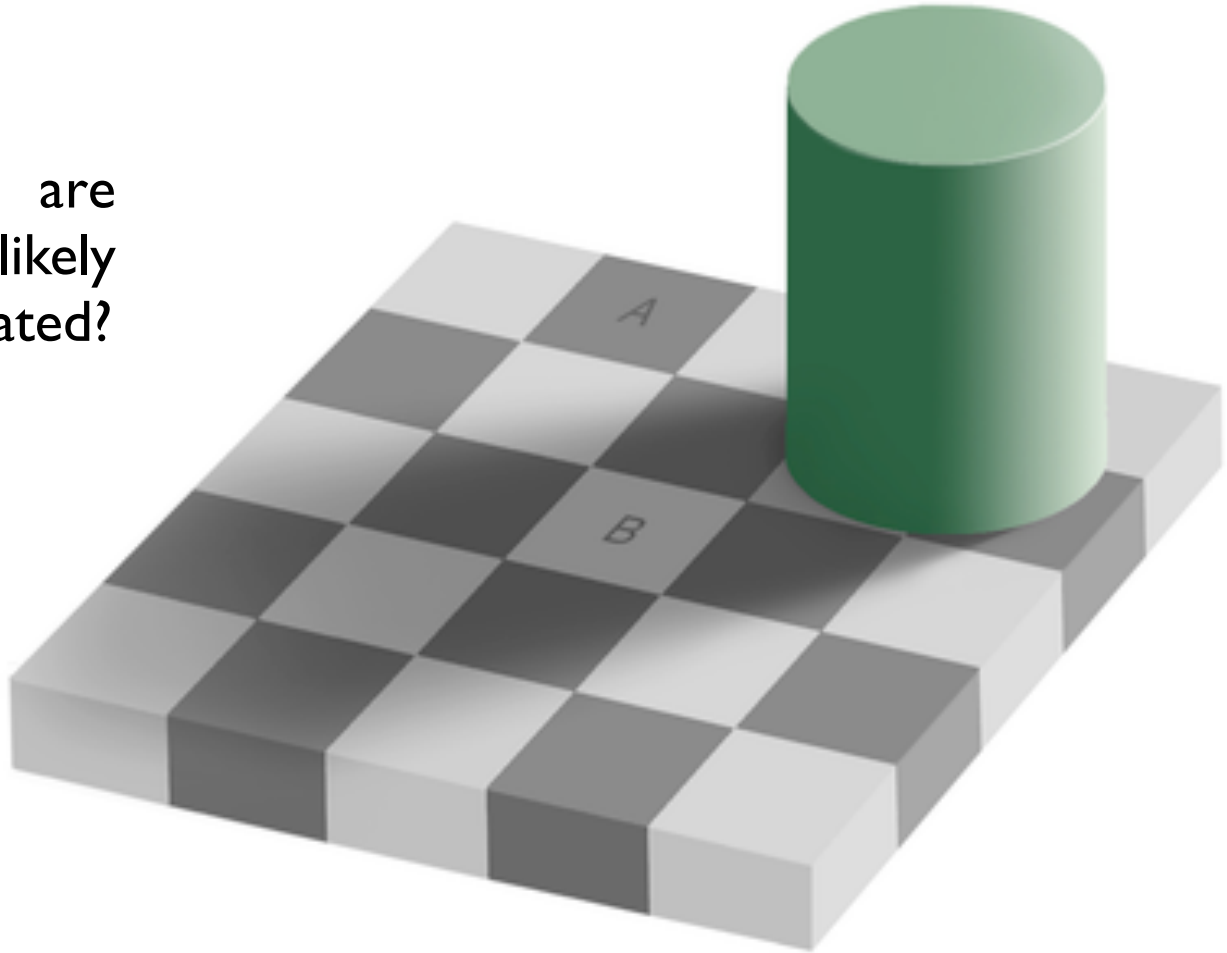


Your brain inherently makes Bayesian inferences:

Context is necessary to relate data to model parameters

(visual observation) (optical properties of surface)

Prior: How are the squares likely being illuminated?



The <u>model</u> is of central importance to enable predictions