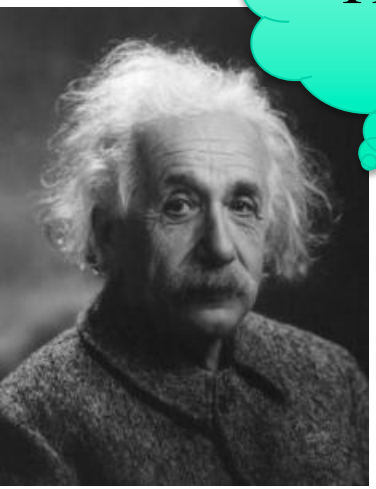
Introduction to Probability and Data Analysis

A. Caldwell
Max Planck Institute for Physics

1. The fundamentals
2. Frequentist and Bayesian approaches
3. Probability of the data and p-values
4. Example of Bayesian analysis – double beta decay with GERDA

Max-Planck-Institut für Physik
(Werner-Heisenberg-Institut)

MAX-PLANCK-GESELLSCHAFT



Theory

$\vec{y}$ Are the theoretical observables

$\vec{\lambda}$ Are the parameters of the theory

M Is the model or theory

$g(\vec{y}|\vec{\lambda}, M)$

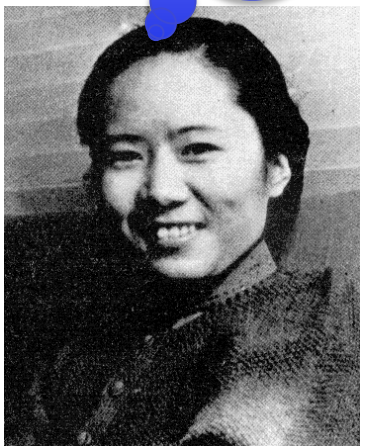
Modeling of experiment

$f(\vec{x}|\vec{\lambda}, M)$

compare

$\vec{D}$

Experiment



$\vec{x}$ Is a possible data outcome

Notation

$$g(\vec{y}|\vec{\lambda}, M)$$

is understood as a probability density. I.e., the probability that

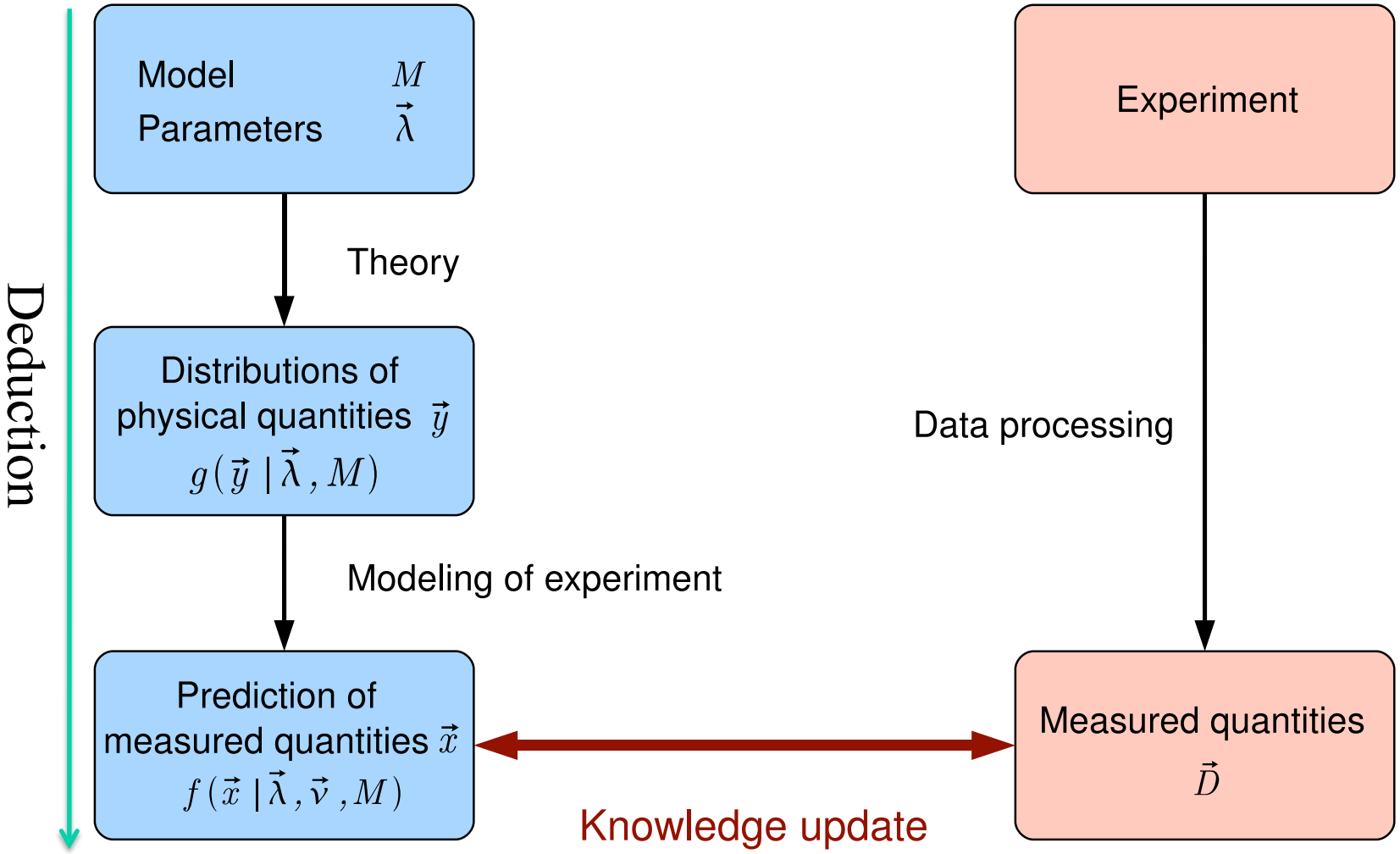
$$\vec{y} \text{ is in the interval } \vec{y} \rightarrow \vec{y} + d\vec{y}$$

given the model *M* and the parameter values specified by $\vec{\lambda}$.

e.g., if we are considering the decay of a unstable particle, we would have

$g(t|\tau) \propto e^{-t/\tau}$ Probability density of decay occurring at time *t* for single particle, assuming constant probability per unit time.

How we learn


Model M
Parameters $\vec{\lambda}$
Theory
Distributions of physical quantities $\vec{y}$
$g(\vec{y}\,|\,\vec{\lambda}, M)$
Modeling of experiment
Prediction of measured quantities $\vec{x}$
$f(\vec{x}\,|\,\vec{\lambda}, \vec{\nu}, M)$
Deduction
Experiment
Data processing
Measured quantities
$\vec{D}$
Knowledge update

How we Learn

We learn by comparing measured data with distributions for predicted results assuming a theory, parameters, and a modeling of the experimental process.

What we typically want to know:

• Is the theory reasonable ? I.e., is the observed data a likely result from this theory (+ experiment).

• If we have more than one potential explanation, then we want to be able to quantify which theory is more likely to be correct given the observations

• Assuming we have a reasonable theory, we want to estimate the most probable values of the parameters, and their uncertainties. This includes setting limits (>< some value at XX% probability).

Logical Basis

Model building and making predictions from models follows deductive reasoning:

Given A➔B (major premise)
Given B➔C (major premise)
Then, given A you can conclude that C is true

etc.

Everything is clear, we can make frequency distributions of possible outcomes within the model, etc. This is math, so it is correct …

Logical Basis

However, in physics what we want to know is the validity of the model given the data. i.e., logic of the form:

Given A➔C
Measure C, what can we say about A ?

Well, maybe A_1➔C, A_2➔C, …

We now need inductive logic. We can never say anything absolutely conclusive about A unless we can guarantee a complete set of alternatives A_i and only one of them can give outcome C. This does not happen in science, so we can never say we found the true model.

Logical basis

Instead of truth, we consider knowledge

Knowledge = **justified ~~true~~ belief**

Justification comes from the data. Make predictions from your model, and see if they are correct. If yes, your belief increases …

Starting point: prior knowledge or maybe plain belief

Do the experiment

Data analysis gives updated knowledge. Experimental results in line with model predictions give justification for believing our model.

Elements of Data Analysis

Probability of the data (Likelihood) $P(D|\lambda, M) = \mathcal{L}(\lambda|D)$

e.g., Poisson process

$$P(n|\lambda) = \frac{e^{-\lambda}\lambda^n}{n!} \quad \lambda \text{ fixed}$$

$$\mathcal{L}(\lambda|n) = \frac{e^{-\lambda}\lambda^n}{n!} \quad n \text{ fixed}$$

All that is used in frequentist or classical approach

In a Bayesian analysis, also need the prior probability $P_0(\lambda|M)$

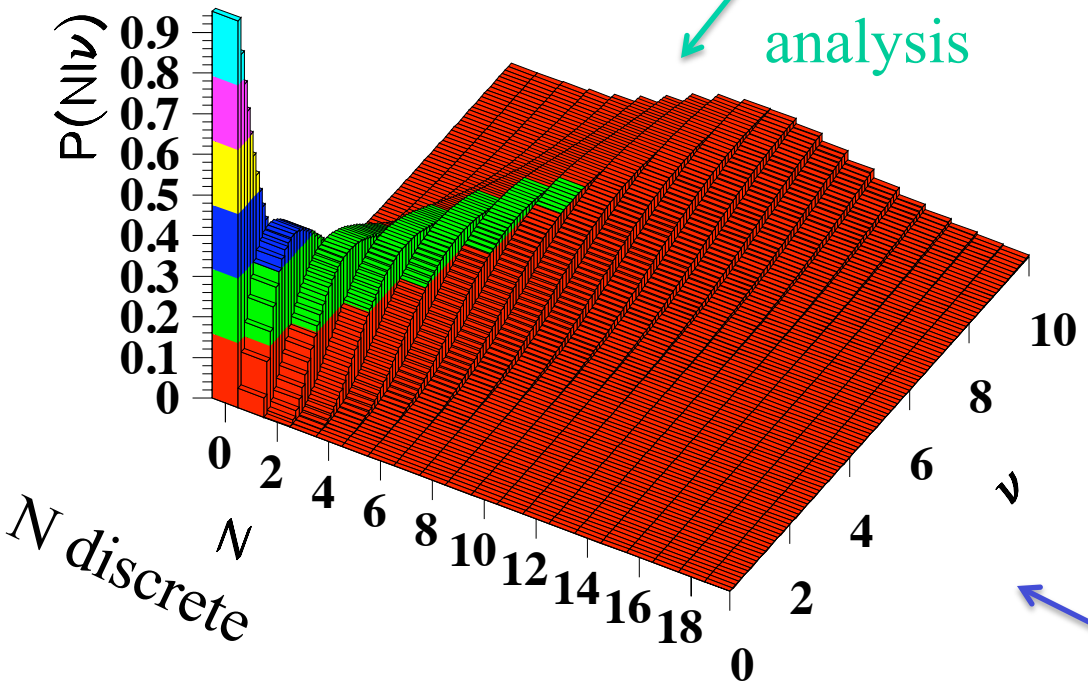
Then use $$P(\lambda|M, D) = \frac{P(D|\lambda, M)P_0(\lambda|M)}{P(D|M)}$$

$$Z = P(D|M) = \int d\lambda P(D|\lambda, M)P_0(\lambda|M)$$

"evidence" often not needed

Poisson Example

$$P(N|\nu) = \frac{e^{-\nu}\nu^N}{N!}$$ likelihood when viewed as function of ν for fixed N.



Likelihood used in likelihood analysis

Probability of the data used in confidence level setting

Bayesians and Frequentists

Frequentists make statements of the kind:

'Assuming the model is correct, this result will occur in XX% of the experiments'

The model is **assumed true**, and estimators for the true parameters in the model are produced from the data.

In the 'classical' approach, this is then converted to 'assuming the model, the bounds [a,b] will contain the true value in XX% of experiments performed' (confidence levels). Does not imply that the true value is in the range [a,b] with probability XX !

The decision on whether to then believe the model/parameters is left to the individual (subjective). *The inductive part of the reasoning is left out of the analysis.*

Bayesians and Frequentists

Bayesians make statements of the kind:

'the degree-of-belief in model A is XX (between 0,1)'

Given the new data, the degree-of-belief is updated using the frequencies of possible outcomes in the context of the models (full set)

Credible regions are then defined: with XX% credibility, the parameter is in the interval [a,b]. **Note – very different from a CL.**

The inductive part of the reasoning is built in to the analysis, and the connection between prior beliefs and posterior beliefs is made clear.

Subjective, but the subjective element is made explicit.

Bayesians and Frequentists

In both approaches, work with models and frequencies of outcomes within the model.

Many elements are the same: calculating the frequencies of possible outcomes given the model AND the experimental conditions; picking the most sensitive variables to test the theory, …

There is no right and wrong approach, but you have to understand what you get out of each type of analysis. E.g., don't confuse confidence levels with probabilities, p-values with support for a model, …

Why isn't everyone a Bayesian ?

There is the worry about stating prior beliefs – doesn't this make the result subjective ?

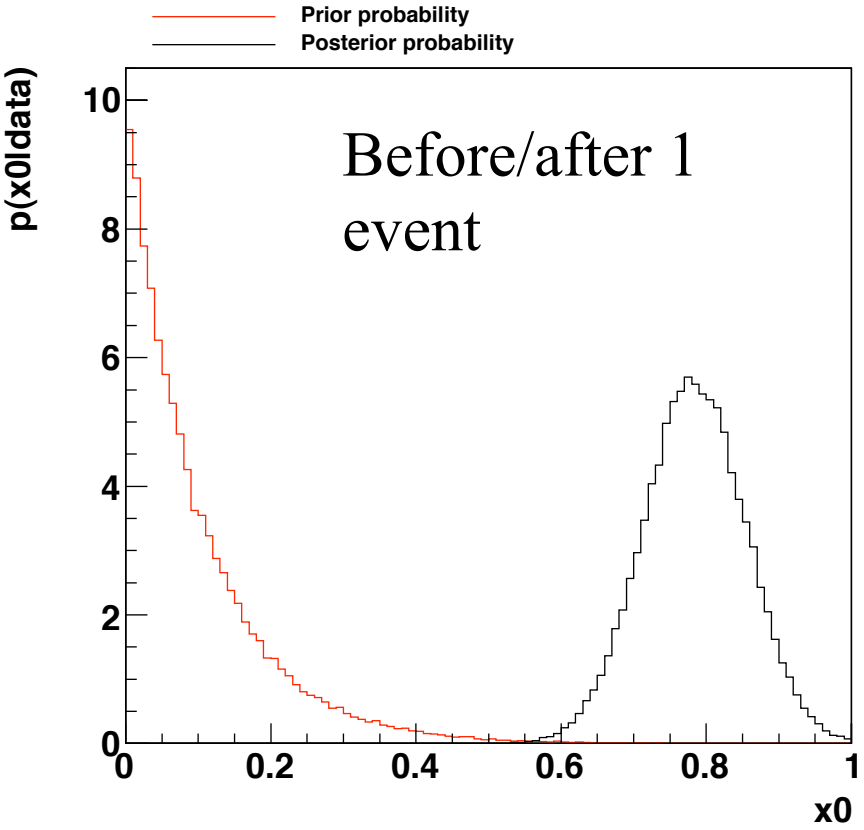
Answer – all statements concerning the 'truth' or 'correctness' of a model are subjective. In the Bayesian approach, the connection between prior belief and posterior belief is made explicit.

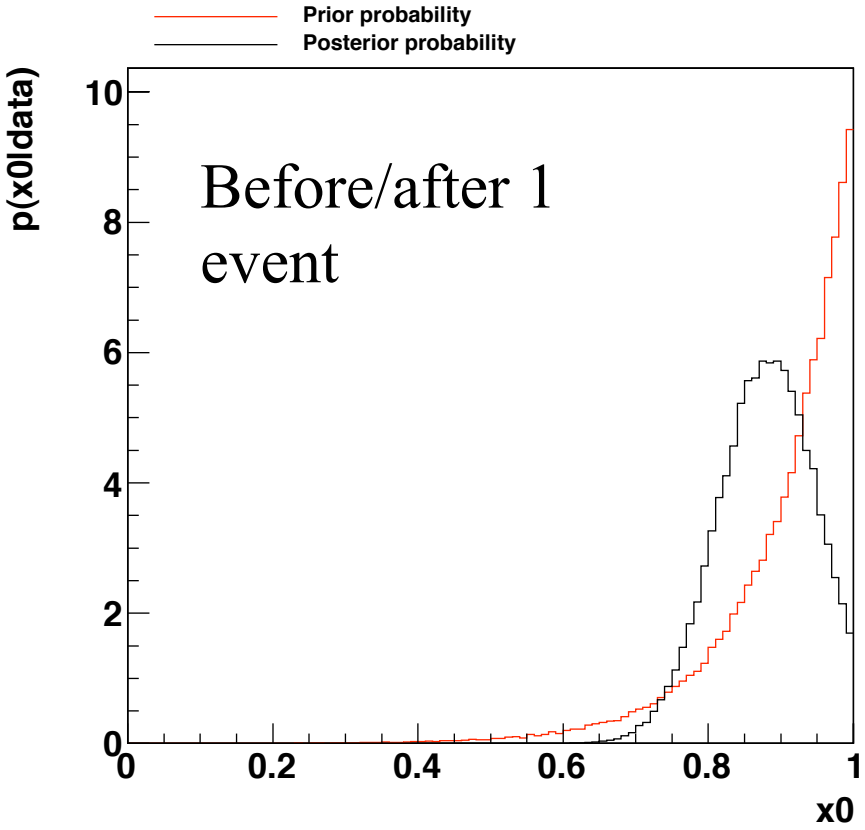
But doesn't the answer depend strongly on your priors ?

Answer – it depends how good your data is. If the data is strong, then the answer depends only weakly on the priors.

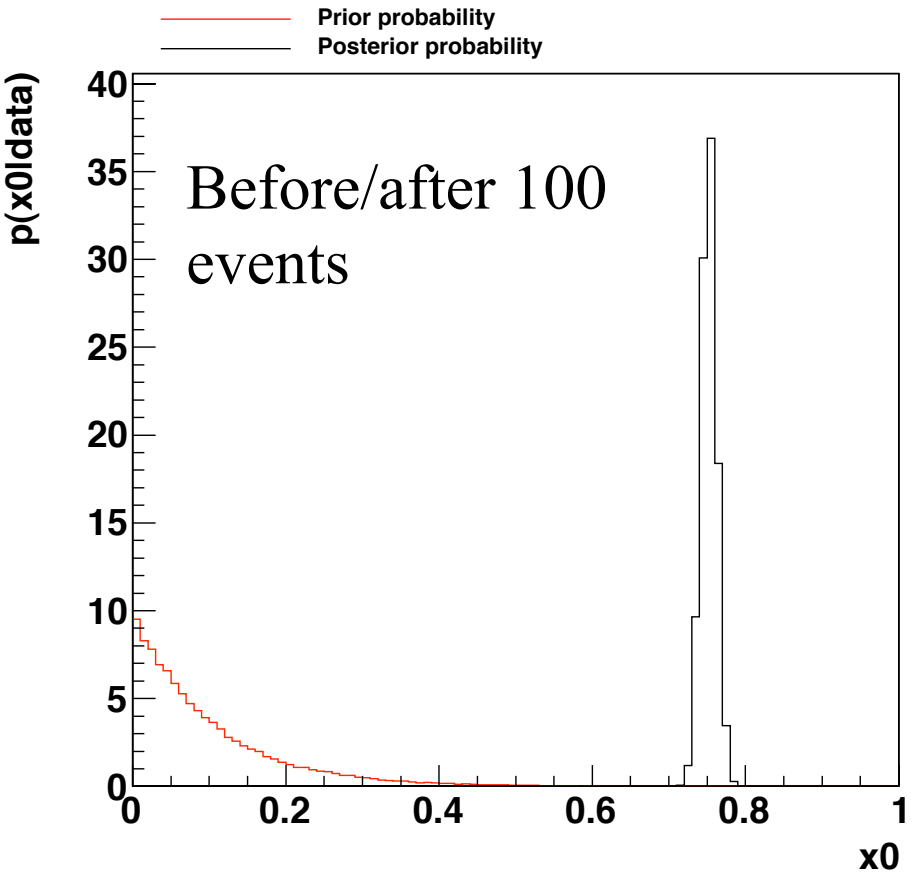
Learning with Gaussian Distributions

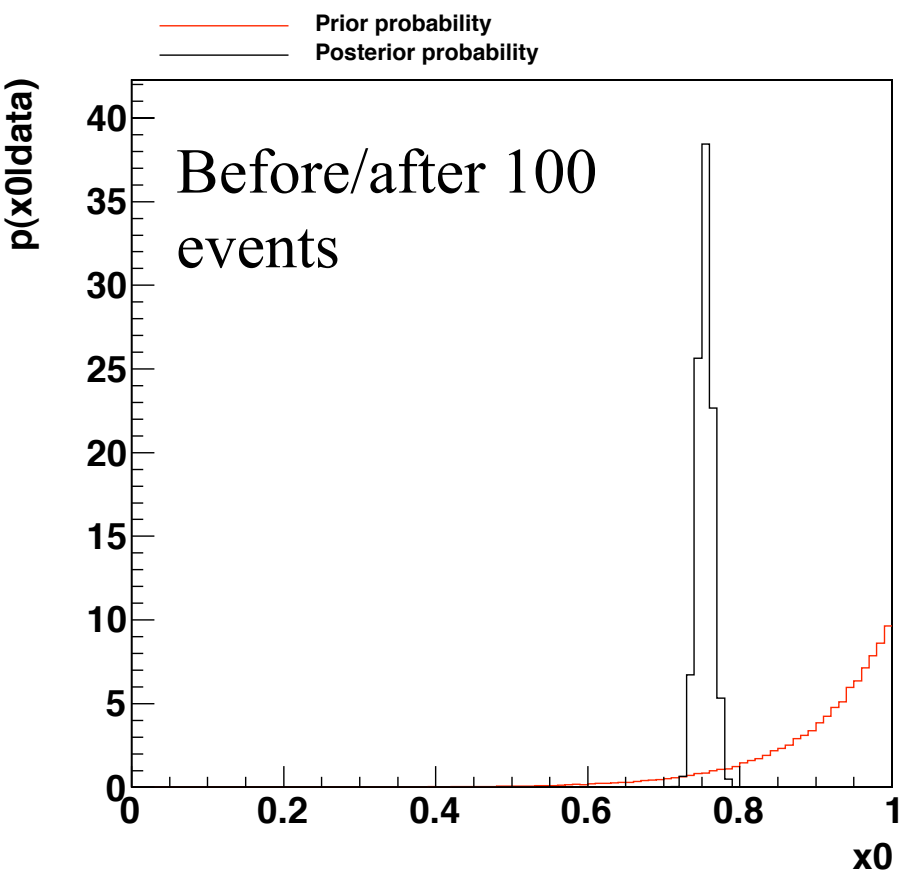
The Bayes' formula has built in the learning feature. Example with Gaussian probability distribution: start with very different priors. Sample from a Gaussian distribution centered at 0.75 (width 0.1).





Learning with Gaussian Distributions





Moral: prior not important if you have informative data.

Why isn't everyone a Bayesian ?

My suspicion: it is because most people do not understand the frequentist approach. Frequentist statements and Bayesian statements are thought to be about the same logical concept, and the frequentist statement does not require a prior, so …

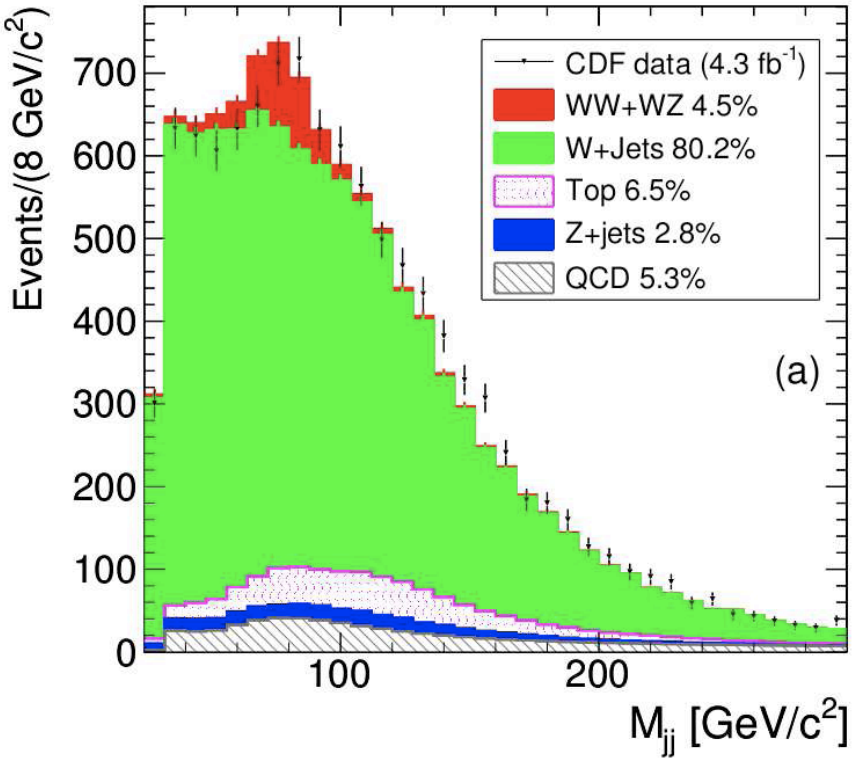
A. L. Read, *Presentation of search results: the CL_S technique,* J. Phys. G: Nucl. Part. Phys. **28** (2002) 2693-2704.

nearly all physicists tend to misinterpret frequentist results as statements about the theory given the data.

Frequentist statements are not statements about the model – only about the data in the context of the model.

Why isn't everyone a Bayesian ?

G. D'Agostini, Probably a discovery: Bad mathematics means rough scientific communication, arXiv:1112.3620v2 [physics.data-an]



Quoting a Discovery article:
It is what is known as a ``three-sigma event," and this refers to the statistical certainty of a given result. In this case, this result has a 99.7 percent chance of being correct (and a 0.3 percent chance of being wrong)."

$$1 - P(D|H_0) = P(H_1|D)$$

This is nonsense !

The Higgs announcement

Gemeinsame Presseerklärung des
Komitee für Elementarteilchenphysik KET
Forschungsschwerpunkt ATLAS (BMBF-FSP 101 ATLAS)
Forschungsschwerpunkt CMS (BMBF-FSP 102 CMS)
Deutsches Elektronen-Synchrotron DESY
Max-Planck-Institut für Physik
Helmholtz-Allianz „Physik an der Teraskala“

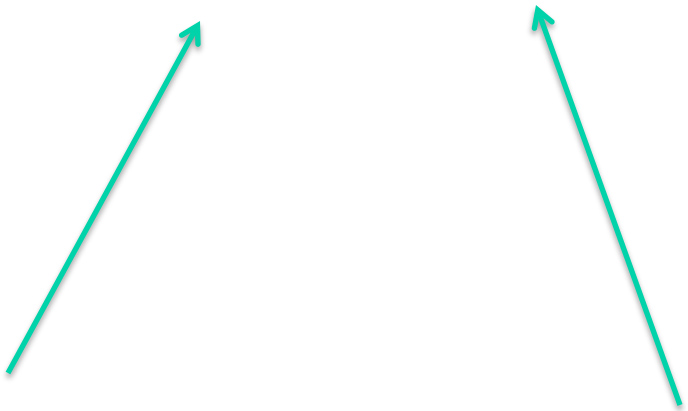
Der Nachweis eines neuen Teilchens wird in der Teilchenphysik klassischerweise auf zwei Stufen gestellt: Die Messungen, die die Wissenschaftler an ihren Experimenten durchführen, beruhen auf Statistik. Sie geben daher zu jedem ihrer Ergebnisse die Sicherheit als so genannte Signifikanz an. Die Einheit, die sie dafür verwenden ist sigma, dargestellt durch den griechischen Buchstaben σ. Die erste Stufe eines Teilchenfunds („evidence“) ist erreicht, wenn sich das Signal des Teilchens mit einer Deutlichkeit zeigt, dass die Physiker mit 99,75 Prozent Sicherheit von seiner Echtheit ausgehen. Dies entspricht einer Signifikanz von 3σ. Von einer „Entdeckung“ und damit der zweiten Stufe sprechen die Forscher bei einer Signifikanz von 5σ, das entspricht einer Fehlerwahrscheinlichkeit von 0,000057%.

Translation - Probability of error is 0,000057%

Error on what ????? That the Higgs is found - not correct

What happened

equated $1 - P(D|H_0) = P(H_1|D)$



Probability of observing the data or something more extreme given the background only hypothesis

Probability that the Higgs exists

This is logical nonsense …

Modeling and Data

We imagine flipping a coin, or rolling dice, or picking a lottery number. The initial conditions are not known, so we assume symmetry and say every outcome is equally likely (Laplace):

-coin, heads or tails each have equal chance, probability of each is ½

-- rolling dice - each number on each die equally likely: each pair (6x6) equally likely. E.g., (1,1) (3,4)…

- i.e., we make a model for the physical process. The model contains assumptions (e.g., each outcome equally likely).

- given the model, we can make predictions and compare these to the data.
- from the comparison, we decide if the model is reasonable.

Probability of the data

The expected distribution (density) of the data assuming a model M and parameters $\vec{\lambda}$ is written as $P(\vec{x}|\vec{\lambda}, M)$ where $\vec{x}$ is a possible realization of the data. There is usually **no unique definition** of the 'probability of the data.' Different choices incorporate different information.

Imagine we flip a coin 10 times, and get the following result:

T H T H H T H T T H

We now repeat the process with a different coin and get

T T T T T T T T T T

Which outcome has higher probability ?

Take a model where H, T are equally likely. Then,

outcome 1 $prob = (1/2)^{10}$

And

outcome 2 $prob = (1/2)^{10}$

Something seem wrong with this result ? This is because (in our head) we evaluate many probabilities at once. The result above is the probability for any sequence of ten flips of a fair coin. Given a fair coin, we could also calculate the chance of getting n times H:

$$\left(\begin{array}{c} 10 \\ n \end{array} \right) \left(\frac{1}{2} \right)^{10}$$

And we find the following result:

n	p
0	$1 \cdot 2^{-10}$
1	$10 \cdot 2^{-10}$
2	$45 \cdot 2^{-10}$
3	$120 \cdot 2^{-10}$
4	$210 \cdot 2^{-10}$
5	$252 \cdot 2^{-10}$
6	$210 \cdot 2^{-10}$
7	$120 \cdot 2^{-10}$
8	$45 \cdot 2^{-10}$
9	$10 \cdot 2^{-10}$
10	$1 \cdot 2^{-10}$

There are many more ways to get 5 H than 0, so this is why the first result somehow looks more probable, even if each sequence has exactly the same probability in the model.

Maybe the model is wrong and one coin is not fair ? How would we test this ?

The message: there are usually many ways to define the probability for your data. Which is better, or whether to use several, depends on what you are trying to do.

E.g., have measured times in exponential decay. Can define the probability density as

$$P(\vec{t}|\tau) = \prod_{i=1}^{N} \frac{1}{\tau} e^{-t_i/\tau}$$

Or you can count events in a time interval and compare to expectations

$$P(\vec{t}|\tau) = \prod_{j=1}^{M} \frac{e^{-\nu_j} \nu_j^{n_j}}{n_j!}$$

ν_j = expected events in bin j

n_j = observed events in bin j

Probability of the data

Which probability (ies) of the data you want to use depend on what you are trying to extract.

E.g., unbinned likelihood can give the best information on the parameter value, but

Unbinned likelihood can contain no information on goodness-of-fit. Each new problem has to be analyzed in detail.

p-values and Goodness-of-fit

Requirement:

- Assume a model *M* with parameters $\vec{\lambda}$

Test statistic:

- Any scalar function of data $T(D)$
- Interpret: large $T(D) =$ discrepancy between *M* and *D*

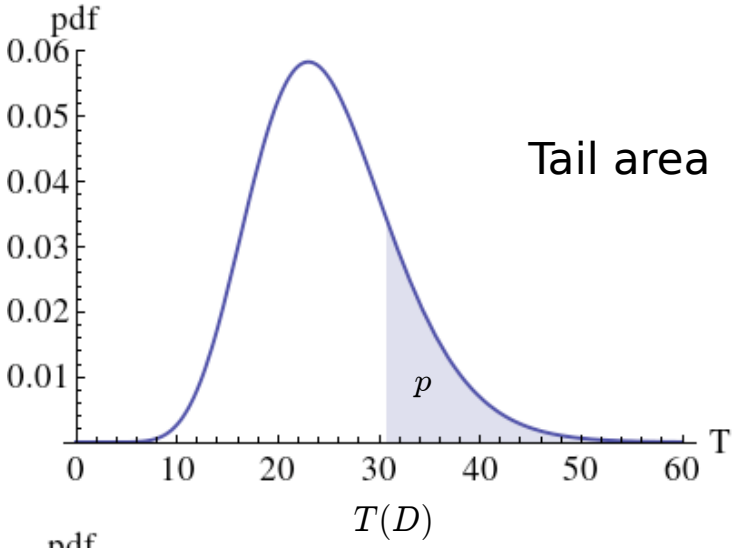
Example:

- Probability of the data $$P(D|\vec{\lambda}) \propto \prod \exp\left\{-\frac{\left(y_i - f(x_i|\vec{\lambda})\right)^2}{2\sigma_i^2}\right\} = \exp\left\{-\frac{\chi^2}{2}\right\}$$
- Familiar choice $$T(D) = \chi^2(D)$$
- Extension: discrepancy variable $T(D|\vec{\lambda})$. Fitting procedure important!

p-values and Goodness-of-fit

- Definition:

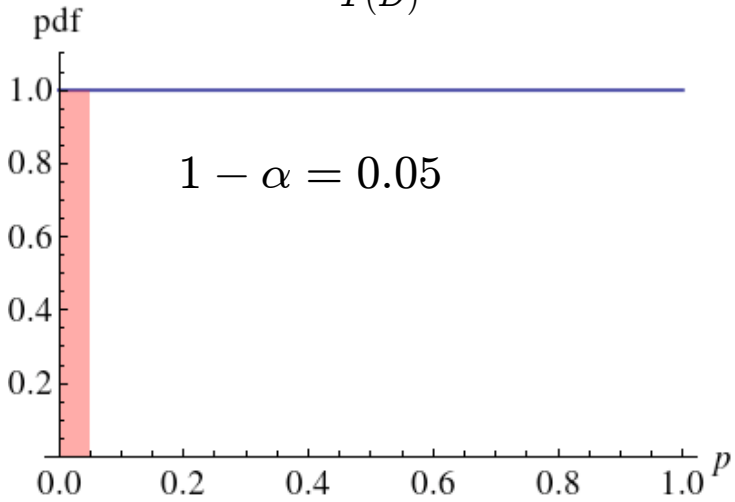
$$p \equiv P(T > T(D)|M)$$



- Assuming *M* and before data is taken:
  p uniform in [0,1]

- Confidence level α:

$$p < 1 - \alpha \Rightarrow \text{ reject model}$$



Why do we reject the small p-values if all are equally likely ?

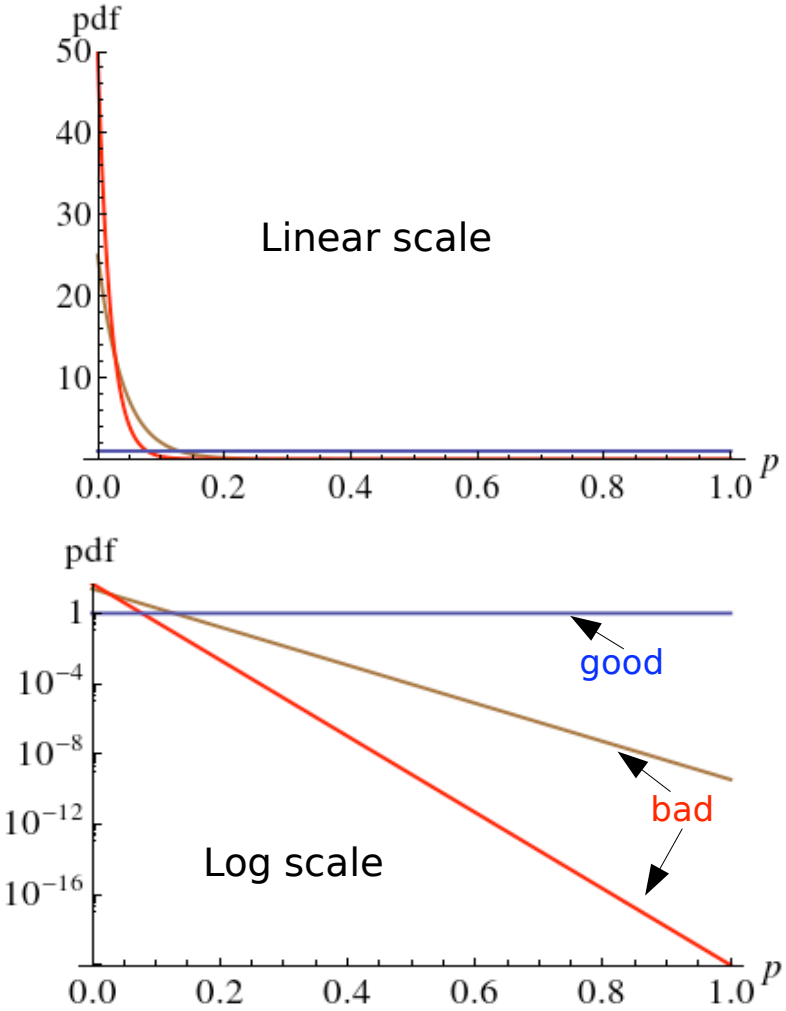
Comment on reasoning behind p-values

- Need prior knowledge about alternatives
- Good model: flat p-value

$$P(p|M_0) = 1$$

- Bad model: peak at p=0, sharply falling

$$P(p|M_i) \approx c_i e^{-c_i p} \ , \quad c_i \gg 1$$



Reasoning behind p-values

- Similar prior for all models $P(M_i) \approx P(M_j)$
- Bayes Theorem: $P(M_0|p) \approx \frac{P(p|M_0)}{\sum_{i=0}^{K} P(p|M_i)}$

Small p

$$P(M_0|p \approx 0) \approx \frac{1}{1 + \sum_{i=1}^{K} c_i} \ll 1$$

Large p

$$P(M_0|p \approx 1) \approx 1$$

Bayes Theorem gives justification to p-values

Pitfalls of p-values

p-values depend critically on how you have chosen the test statistic (or discrepancy variable). The same data set can have hugely varying p-values resulting from different choices of the test quantity.

E.g., consider a model where we assume an exponential decay law. As mentioned earlier, we can define the following probabilities of the data:

Unbinned likelihood

$$P(\vec{t}|\tau) = \prod_{i=1}^{N} \frac{1}{\tau} e^{-t_i/\tau}$$

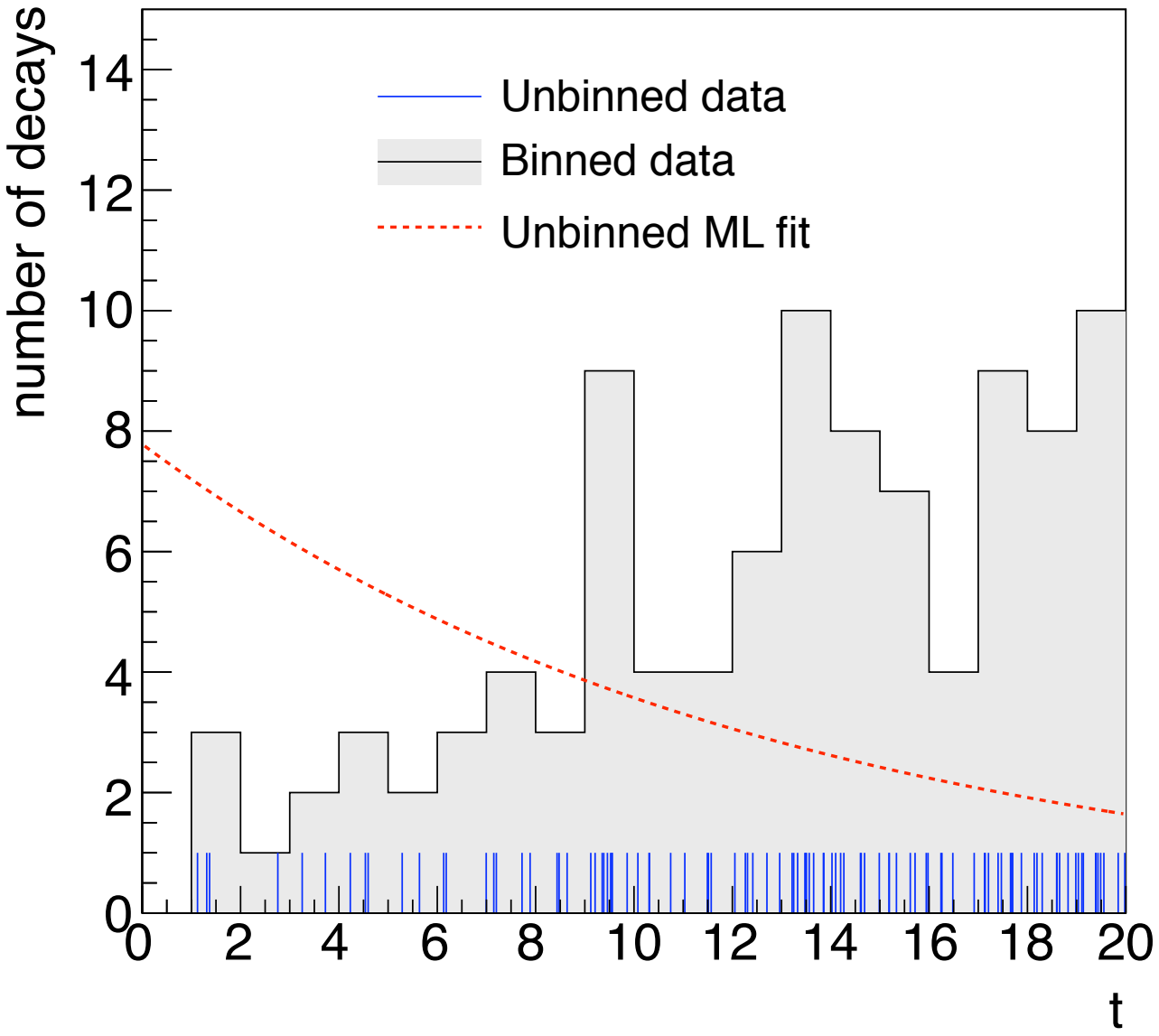
Binned Poisson distribution

$$P(\vec{t}|\tau) = \prod_{j=1}^{M} \frac{e^{-\nu_j} \nu_j^{n_j}}{n_j!}$$

$\nu_j =$ expected events in bin j

$n_j =$ observed events in bin j

pitfalls



Assumed model is exponential. Data actually from linearly increasing function.

pitfalls

We take the best fit probability as our test statistic. For the unbinned fit, it is easy to show that

$$\tau^* = \frac{1}{N}\sum_{i=1}^{N} t_i$$

$$p = \int_{\sum t_i' > \xi} \mathrm{d}t_1' \int \mathrm{d}t_2' \ldots (\tau^*)^{-N} e^{-\sum t_i'/\tau^*} = 1 - P(N, N)$$

Regularized incomplete gamma function

$$P(s,x) = \frac{\gamma(s,x)}{\Gamma(s)} = \frac{\int_0^x t^{s-1}e^{-t}\mathrm{d}t}{\int_0^\infty t^{s-1}e^{-t}\mathrm{d}t}$$

Doesn't depend on the data ! In fact, for large *N*, $p \approx 0.5$

pitfalls

The p-value from the maximum likelihood is about 0.5 !

The p-value from the binned fit is 0

What happened ? The maximum likelihood quantity does not know anything about the distribution of the events, and the result only depends on

$$\tau^* = \frac{1}{N}\sum_{i=1}^{N} t_i$$

and the p-value only depends on N !

Lesson: make sure your test statistic is sensitive to what you want to test !

Bayesian Parameter Estimation

The posterior pdf gives the full probability distribution for all parameters, including all correlations – no approximations. If interested in subset of parameters, then marginalize. E.g., for one parameter:

$$P(\lambda_i|\vec{D},M) = \int P(\vec{\lambda}|\vec{D},M)d\vec{\lambda}_{\mathrm{J}\neq i}$$

Can calculate what you need from the posterior pdf. E.g.,

Mode $\max_{\lambda_i} \{P(\lambda_i|D,M)\}$ + probability intervals, …

Mean of λ_i $< \lambda_i >= \int P(\lambda_i|\vec{D},M)\lambda_i d\lambda_i$

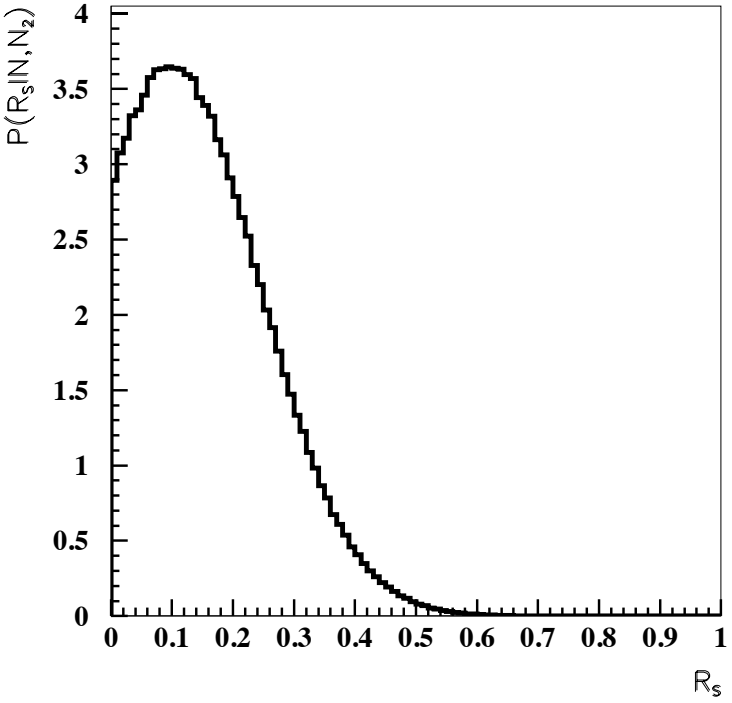
Median $\int_{\lambda_{min}}^{\lambda_{med}} P(\lambda_i|\vec{D},M)d\lambda_i = 0.5$

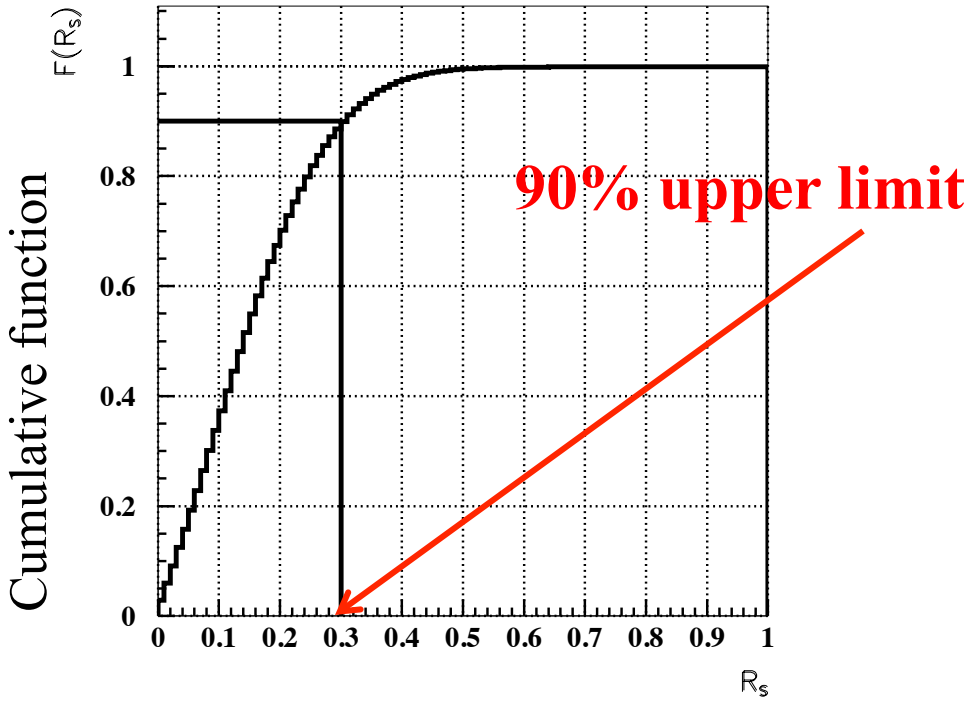
Can also perform uncertainty propagation w/o approximations

Setting Limits

Setting limits is easy – just integrate the posterior pdf. E.g., 90% upper limit:

$$0.9 = \int_{\lambda_{min}}^{\lambda_{upper}} P(\lambda_i|\vec{D}, M)d\lambda_i$$





Or calculate contours in higher dimensional spaces

Example: Double Beta Decay

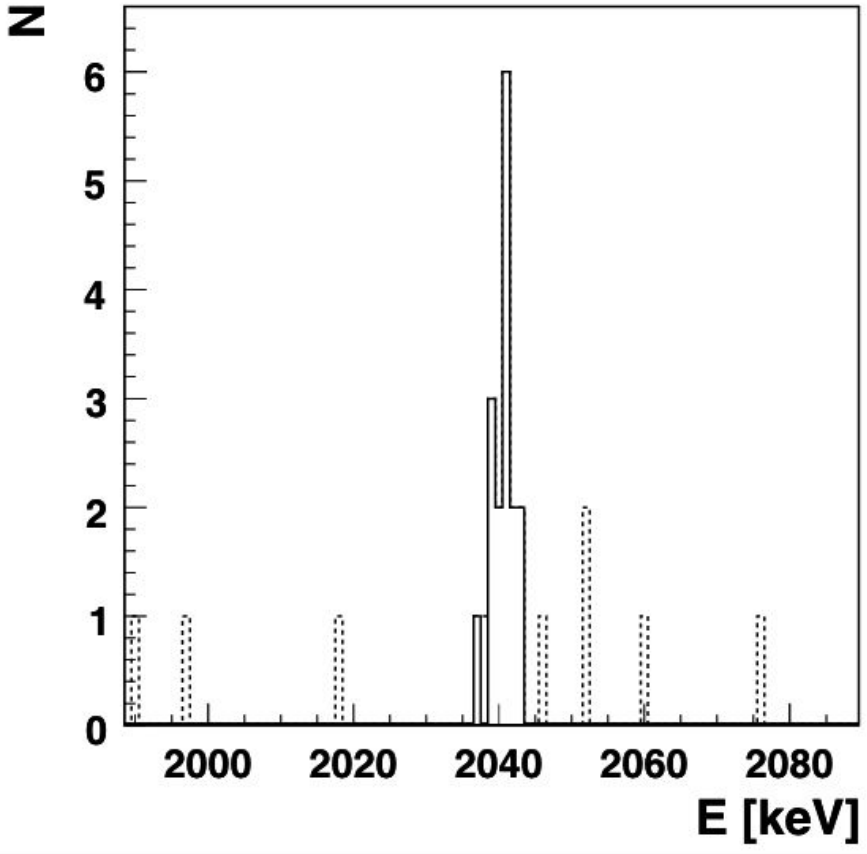
One of the outstanding questions in Particle Physics is whether the neutrino is its own antiparticle (so-called Majorana particle).

The only practical way which has been found to search for the Majorana nature of neutrinos (particle same as antiparticle) is double beta decay (because of the light mass of neutrinos, helicity flip is very unlikely unless the neutrinos have very low energy).

For us, what is interesting is that we are looking for a peak at a well-defined energy in a sparse spectrum.

A. Caldwell, K. Kröninger, Phys. Rev. D 74 (2006) 092003

Discovery or not ?



Analyze energy spectrum and decide if there is evidence for a signal.
Counting experiment – Poisson statistics.

Sparse Spectra

Define the propositions:

H = **The observed spectrum is due to background only**

$\overline{H}$ = **The spectrum is due to background + signal (neutrinoless double beta decay).**

I.e., we assume backgrounds are known up to normalization and some smoothly varying shape, and the only possibility other than known background is signal from neutrinoless double beta decay.

$$p(H \mid spectrum) + p(\overline{H} \mid spectrum) = 1$$

so

$$p(H \mid spectrum) = \frac{p(spectrum \mid H)\, p_0(H)}{p(spectrum \mid H)\, p_0(H) + p(spectrum \mid \overline{H})\, p_0(\overline{H})}$$

$$p(\overline{H} \mid spectrum) = \frac{p(spectrum \mid \overline{H})\, p_0(\overline{H})}{p(spectrum \mid H)\, p_0(H) + p(spectrum \mid \overline{H})\, p_0(\overline{H})}$$

Double Beta Decay Example

p_0(H):

The existing limits are $T_{1/2}>4\ 10^{25}$ yr; a positive claim for a signal exists at the level $T_{1/2}=1.2\ 10^{25}$ yr; my favorite theorist believes strongly that neutrinos are Majorana particles, but he wont tell me the neutrino mass; the theorist at a neighboring university says that he believes strongly in Leptogenesis, and in that context the neutrino is a Majorana particle but it must be very light, such that neutrinoless double beta decay is unobservable,...

DBD example

We know how to perform all calculations:

$$p(spectrum \mid H) = \int p(spectrum \mid B)\, p_0(B)\, dB$$

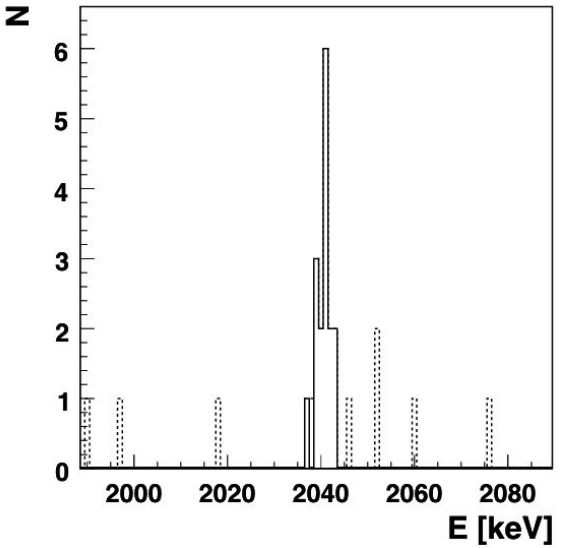
$$p(spectrum \mid \overline{H}) = \int p(spectrum \mid S,B)\, p_0(S)\, p_0(B)\, dB$$

Where B is the expected number of background events and S is the expected number of signal events. These quantities come with their own priors.

n_i = observed number of events in bin i

λ_i = expected number of events in bin i

$$\lambda_i = S \int_{\Delta E_i} f_S(E)\, dE + B \int_{\Delta E_i} f_B(E)\, dE$$



Where f_S and f_B are the normalized signal and background probability densities as functions of energy.

DBD example

then $$p(spectrum \mid B) = \prod_{i=1}^{N} \frac{\lambda_i(0,B)^{n_i}}{n_i!} e^{-\lambda_i(0,B)}$$

$$p(spectrum \mid S,B) = \prod_{i=1}^{N} \frac{\lambda_i(S,B)^{n_i}}{n_i!} e^{-\lambda_i(S,B)}$$

To determine parameter values or set limits, we need

$$p(S,B \mid spectrum) = \frac{p(spectrum \mid S,B)\, p_0(S)\, p_0(B)}{\int p(spectrum \mid S,B)\, p_0(S)\, p_0(B)\, dS dB}$$

and then marginalize

$$p(S \mid spectrum) = \int p(S,B \mid spectrum)\, dB$$

e.g., 90% probability upper limit, S_{90} from solving

$$\int_0^{S_{90}} p(S \mid spectrum)\, dS = 0.90$$

GERDA example

Assumptions for GERDA:

$$p_0(H) = p_0(\overline{H}) = 1/2$$

$$p_0(S) = \frac{1}{S_{\max}} \quad 0 \le S \le S_{\max} \qquad p_0(S) = 0 \text{ otherwise}$$

$$p_0(B) = \frac{e^{-\frac{(B-\mu_B)^2}{2\sigma_B^2}}}{\int_0^\infty e^{-\frac{(B-\mu_B)^2}{2\sigma_B^2}} dB} \quad B \ge 0; \quad p_0(B) = 0 \; B < 0$$

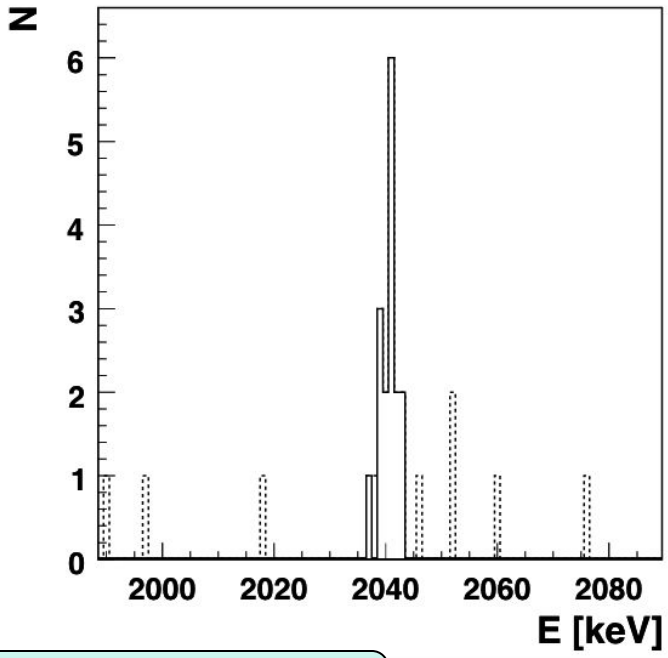
S_{max} was calculated assuming $T_{1/2}$=0.5 10^{25} yr

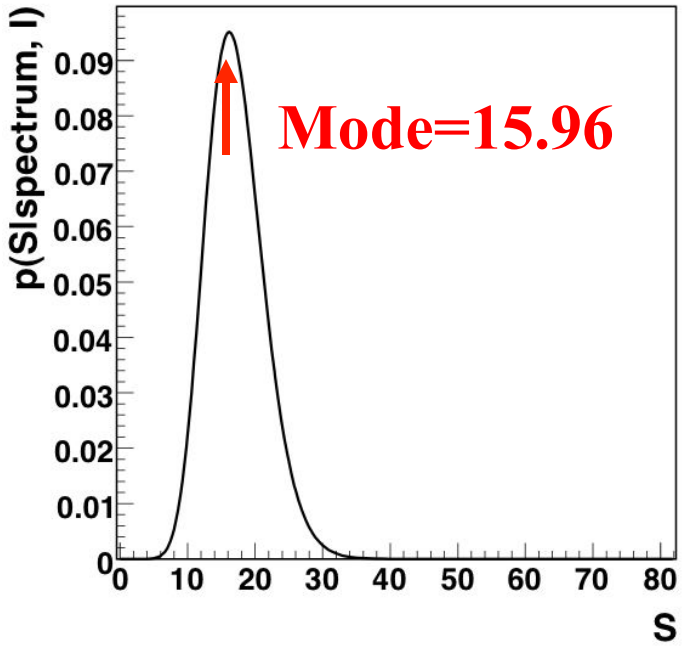
μ_B=B_0, σ_B=B_0/2

100 keV window analyzed. B_0 total background in this window.

Example:

$S_{true}=16$, $B_{true}=9$

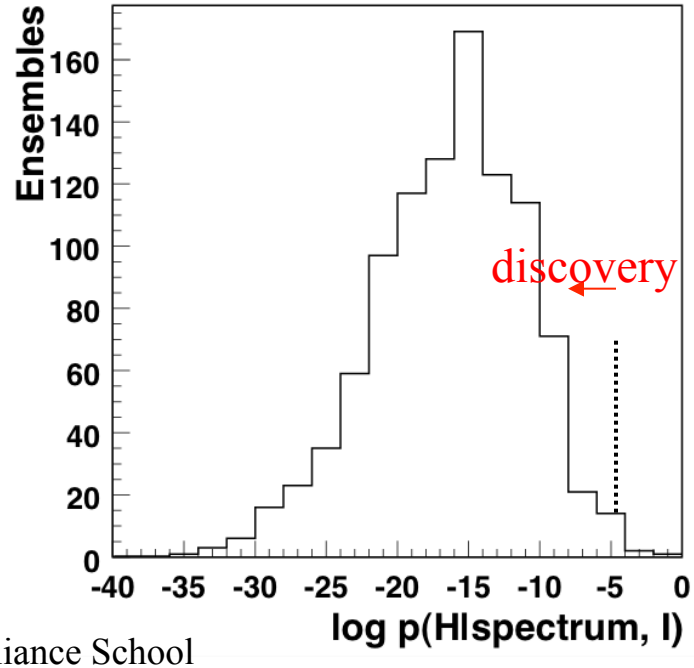




$$p(H \mid spectrum) = 2.2 \cdot 10^{-12}$$

1000 experiments simulated with
$T_{1/2}=2\ 10^{25}$ yr, 10^{-3}/(kg keV yr)
Exposure 100 kg-yr

About 95% chance a discovery could be claimed



The End