Bayes Theorem (Recap)

Adrian Bevan

email: a.j.bevan@qmul.ac.uk

Queen Mary University of London

Overview

- Bayes theorem is given by

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

The probability you want to compute: The probability of the hypothesis B given the data A. This is sometimes called the posterior probability

Overview

- Bayes theorem is given by

The probability of the data A, given the hypothesis B. This you can compute given a theory or model

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

The probability you want to compute: The probability of the hypothesis B given the data A. This is sometimes called the posterior probability

Overview

- Bayes theorem is given by

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

The probability of the data A, given the hypothesis B. This you can compute given a theory or model

The prior probability: a level of belief in the feasibility of a hypothesis.

The probability you want to compute: The probability of the hypothesis B given the data A. This is sometimes called the posterior probability

Overview

- Bayes theorem is given by

The probability of the data A, given the hypothesis B. This you can compute given a theory or model

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

The prior probability: a level of belief in the feasibility of a hypothesis.

The probability you want to compute: The probability of the hypothesis B given the data A. This is sometimes called the posterior probability

The probability of the data A given all possible hypotheses.

Overview

- Bayes theorem is given by

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

The probability of the data A, given the hypothesis B. This you can compute given a theory or model

The prior probability: a level of belief in the feasibility of a hypothesis.

The probability you want to compute: The probability of the hypothesis B given the data A. This is sometimes called the posterior probability

The probability of the data A given all possible hypotheses.

$$P(A) = \sum_j P(A|B_j)P(B_j)$$

Example

- What is the probability of being dealt an ace from a deck of 52 cards?
  - Two hypothetical outcomes: you are dealt an ace (B_0), and you are not dealt an ace (B_1).
  - Assume that the deck is unbiased and properly shuffled, so that there are 4 aces out of 52 randomly distributed in the pack of cards.

$$P(B_0) = 4/52$$
$$P(B_1) = 48/52$$

Is this choice of prior reasonable?

  - Compute the probability of being dealt an ace, given that you have an unbiased

$$P(A|B_0) = 4/52$$
$$P(A|B_1) = 48/52$$

  - Now we can compute the posterior probability

$$P(A) = \sum_j P(A|B_j)P(B_j)$$

$$\begin{aligned} P(A) &= (4/52) \times (4/52) + (48/52) \times (48/52) \\ &= (16 + 2304)/2704 \\ &= 0.8580 (4\,\text{s.f.}) \end{aligned}$$

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

$$\begin{aligned} P(A|B_0) &= \frac{(4/52) \times (4/52)}{0.8580} \\ &= 0.007 (4\,\text{s.f.}) \end{aligned}$$

This result does not agree with a frequentist interpretation of the data.

An alternative calculation: assume uniform priors

$$P(A) = \sum_j P(A|B_j)P(B_j)$$

$$\begin{aligned} P(A) &= (4/52) \times (1/2) + (48/52) \times (1/2) \\ &= (4+48)/104 \\ &= 0.5 \end{aligned}$$

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

$$\begin{aligned} P(A|B_0) &= \frac{(4/52) \times (1/2)}{0.5} \\ &= 0.077(3\,\text{d.p.}) \end{aligned}$$

This result agrees with a frequentist interpretation of the data.

Weighted Average (Recap)

Adrian Bevan

email: a.j.bevan@qmul.ac.uk

Queen Mary University of London

Formula for two uncorrelated variables

$$\begin{aligned}\overline{x} \pm \sigma_x &= \frac{x_1/\sigma_1^2 + x_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2} \pm \left(1/\sigma_1^2 + 1/\sigma_2^2\right)^{-1/2}, \\ &= \frac{\sigma_2^2 x_1 + \sigma_1^2 x_2}{\sigma_1^2 + \sigma_2^2} \pm \left(\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right)^{1/2}.\end{aligned}$$

- For example, lets just consider the total precision on the measurement of some observable X. If there are two ways to measure X, and these yield $\sigma_1 = 5$ and $\sigma_2 = 3$, then the total error on the average is

$$\begin{aligned}\left(\frac{\sigma_1^2 \times \sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right)^{1/2} &= \left(\frac{25 \times 9}{25 + 9}\right)^{1/2} \\ &= \sqrt{6.62} \simeq 2.57\end{aligned}$$

- If one has a set of uncorrelated observables, then it is straightforward to compute the average in the same way foreach observable in the set.

- i.e. instead of an n dimensional problem, you have n lots of a one dimensional problem as in the previous example.

- Unfortunately if you have a set of correlated observables from different measurements, then this is no longer the case and the problem becomes a little more complicated.

- A more complicated problem arises when the observables are correlated... In this case the covariance matrix between a set of M measured observables will play a role.

inverse of the covariance matrix for the j[th] measurement

$$\overline{x} = \left[\sum_{j=1}^{M} V_j^{-1}\right]^{-1} \cdot \left[\sum_{j=1}^{M} V_j^{-1} x_j\right],$$

$$V = \left[\sum_{j=1}^{M} V_j^{-1}\right]^{-1}.$$

The j[th] measurement vector (the n correlated observables)

- The first factor is common both to the covariance matrix for the average, and the observable values for the average.

Example: Measuring time-dependent CP asymmetries in B meson decays

- Real example using results from two HEP experiments

Experiment	S	C	ρ_{SC}	cov_{SC}
BABAR (Aubert *et al.*, 2007)	-0.170 ± 0.207	$+0.010 \pm 0.162$	0.035	-0.0012
Belle (Somov *et al.*, 2007)	$+0.190 \pm 0.310$	-0.160 ± 0.225	0.100	0.0070

M = 2
n = 2

Remember if you only have a correlation, you need to compute the covariance in order to compute Vj.

Example: Measuring time-dependent CP asymmetries in B meson decays

- Real example using results from two HEP experiments

Experiment	S	C	ρ_{SC}	cov_{SC}
BABAR (Aubert *et al.*, 2007)	-0.170 ± 0.207	$+0.010 \pm 0.162$	0.035	-0.0012
Belle (Somov *et al.*, 2007)	$+0.190 \pm 0.310$	-0.160 ± 0.225	0.100	0.0070

$$x_1 = \begin{pmatrix} -0.170 \\ +0.010 \end{pmatrix}, \qquad V_1 = \begin{pmatrix} 0.0430 & -0.0012 \\ -0.0012 & 0.0261 \end{pmatrix},$$

$$x_2 = \begin{pmatrix} +0.19 \\ -0.16 \end{pmatrix}, \qquad V_2 = \begin{pmatrix} 0.0964 & 0.0070 \\ 0.0070 & 0.0505 \end{pmatrix},$$

Example: Measuring time-dependent CP asymmetries in B meson decays

- Real example using results from two HEP experiments

Experiment	S	C	ρ_{SC}	cov_{SC}
BABAR (Aubert *et al.*, 2007)	-0.170 ± 0.207	$+0.010 \pm 0.162$	0.035	-0.0012
Belle (Somov *et al.*, 2007)	$+0.190 \pm 0.310$	-0.160 ± 0.225	0.100	0.0070

$$x_1 = \begin{pmatrix} -0.170 \\ +0.010 \end{pmatrix}, \qquad V_1 = \begin{pmatrix} 0.0430 & -0.0012 \\ -0.0012 & 0.0261 \end{pmatrix},$$

$$x_2 = \begin{pmatrix} +0.19 \\ -0.16 \end{pmatrix}, \qquad V_2 = \begin{pmatrix} 0.0964 & 0.0070 \\ 0.0070 & 0.0505 \end{pmatrix},$$

Example: Measuring time-dependent CP asymmetries in B meson decays

- Real example using results from two HEP experiments

Experiment	S	C	ρ_{SC}	cov_{SC}
BABAR (Aubert *et al.*, 2007)	-0.170 ± 0.207	$+0.010 \pm 0.162$	0.035	-0.0012
Belle (Somov *et al.*, 2007)	$+0.190 \pm 0.310$	-0.160 ± 0.225	0.100	0.0070

σ_i^2 σ_i^2

$$x_1 = \begin{pmatrix} -0.170 \\ +0.010 \end{pmatrix}, \qquad V_1 = \begin{pmatrix} 0.0430 & -0.0012 \\ -0.0012 & 0.0261 \end{pmatrix},$$

$$x_2 = \begin{pmatrix} +0.19 \\ -0.16 \end{pmatrix}, \qquad V_2 = \begin{pmatrix} 0.0964 & 0.0070 \\ 0.0070 & 0.0505 \end{pmatrix},$$

Example: Measuring time-dependent CP asymmetries in B meson decays

- Real example using results from two HEP experiments

Experiment	S	C	ρ_{SC}	cov_{SC}
BABAR (Aubert *et al.*, 2007)	-0.170 ± 0.207	$+0.010 \pm 0.162$	0.035	-0.0012
Belle (Somov *et al.*, 2007)	$+0.190 \pm 0.310$	-0.160 ± 0.225	0.100	0.0070

$$x_1 = \begin{pmatrix} -0.170 \\ +0.010 \end{pmatrix}, \qquad V_1 = \begin{pmatrix} 0.0430 & -0.0012 \\ -0.0012 & 0.0261 \end{pmatrix},$$

$$x_2 = \begin{pmatrix} +0.19 \\ -0.16 \end{pmatrix}, \qquad V_2 = \begin{pmatrix} 0.0964 & 0.0070 \\ 0.0070 & 0.0505 \end{pmatrix},$$

- Having constructed x_j and V_j, one can compute the average ...

- Starting with the error matrix:

$$V = \left[\sum_{j=1}^{M} V_j^{-1}\right]^{-1}.$$

- Where

$$V_1^{-1} + V_2^{-1} = \begin{pmatrix} 33.749 & -0.402 \\ -0.402 & 58.363 \end{pmatrix},$$

- So

$$V = \begin{pmatrix} 0.0296 & 0.0002 \\ 0.0002 & 0.0171 \end{pmatrix},$$

Results:
Covariance = 0.0002
Variances are read off of the diagonal

- Similarly for the average value we can now compute S and C (the n observables of interest in our example) via:

$$\overline{x} = \left[\sum_{j=1}^{M} V_j^{-1}\right]^{-1} \cdot \left[\sum_{j=1}^{M} V_j^{-1} x_j\right],$$

- Thus the average of the two measurements is:

$$\begin{pmatrix} S \\ C \end{pmatrix} = \begin{pmatrix} -0.05 \pm 0.17 \\ -0.06 \pm 0.13 \end{pmatrix}$$

- with a covariance of 0.0002 between S and C.

Poisson limits

Adrian Bevan

email: a.j.bevan@qmul.ac.uk

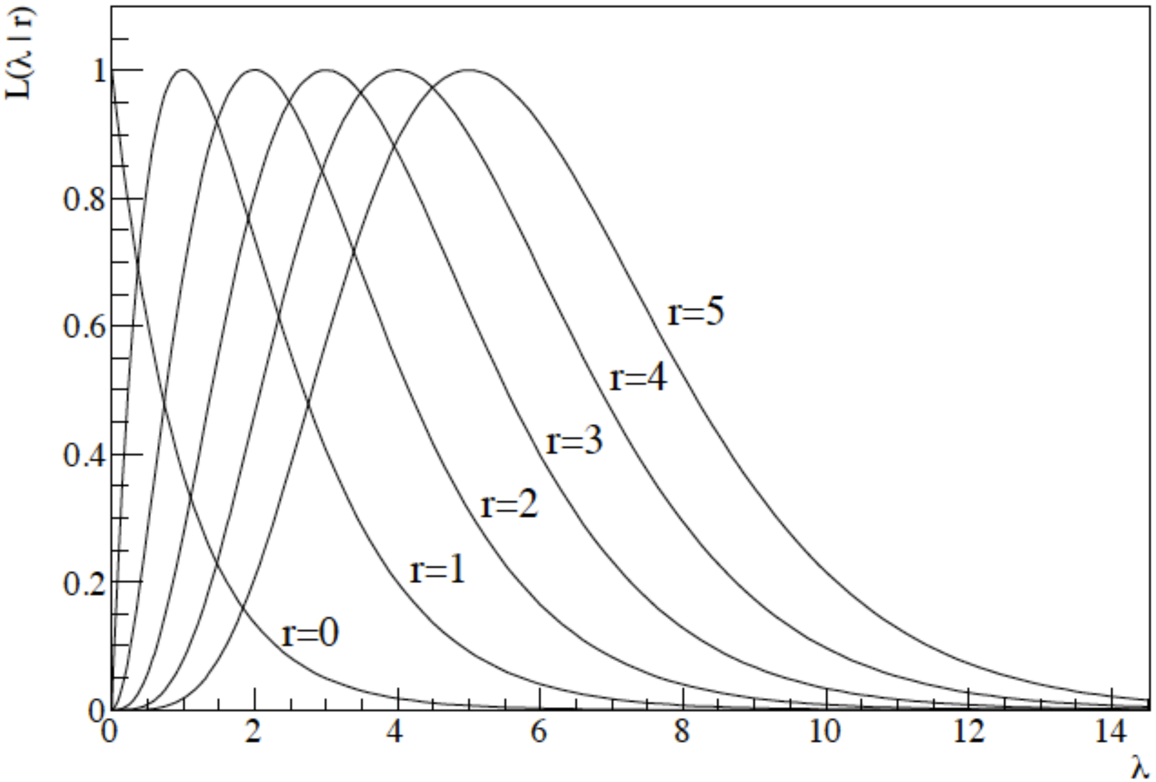
Queen Mary
University of London

- For a given number of observed events resulting from the study of a rare process one wants to compute a limit on the true value of some underlying theory parameters (e.g. the mean occurrence of the rare process).
- This is a Poisson problem, where:

$$f(r, \lambda) = \frac{\lambda^r e^{-\lambda}}{r!},$$

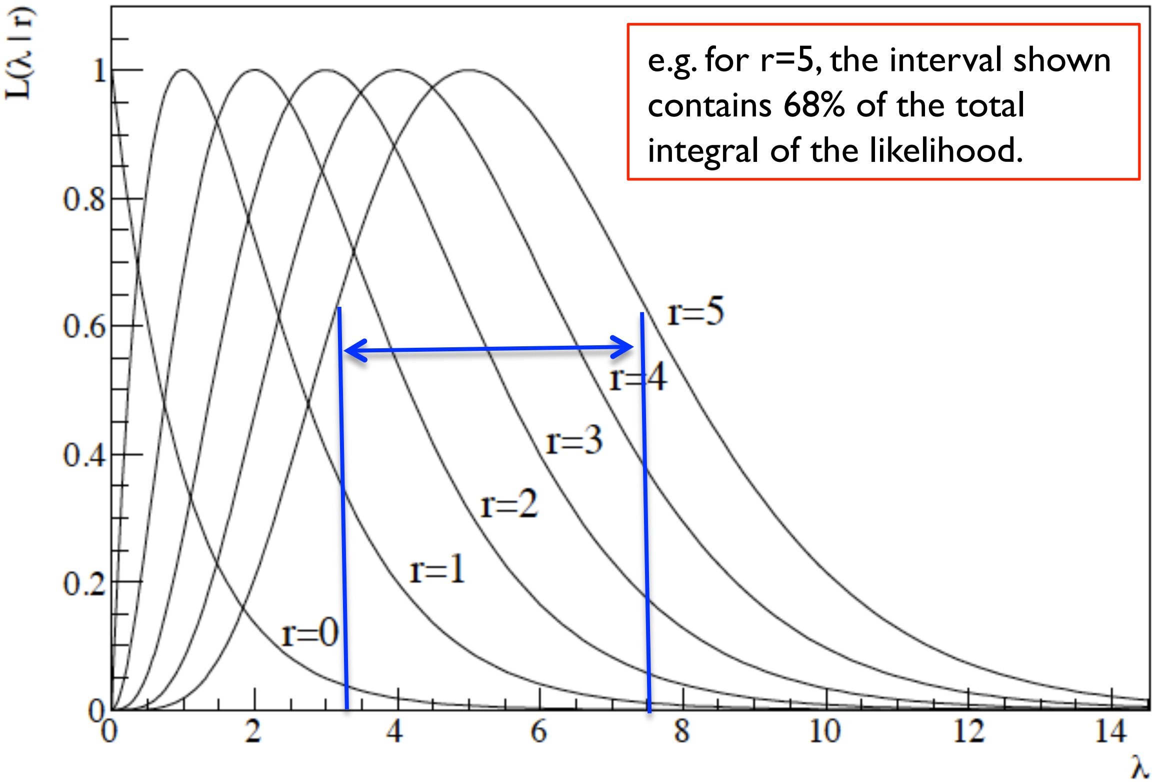
λ = mean/variance of the underlying distribution
r = observed number of events

- For a given observation, we can compute a likelihood as a function of λ, e.g.



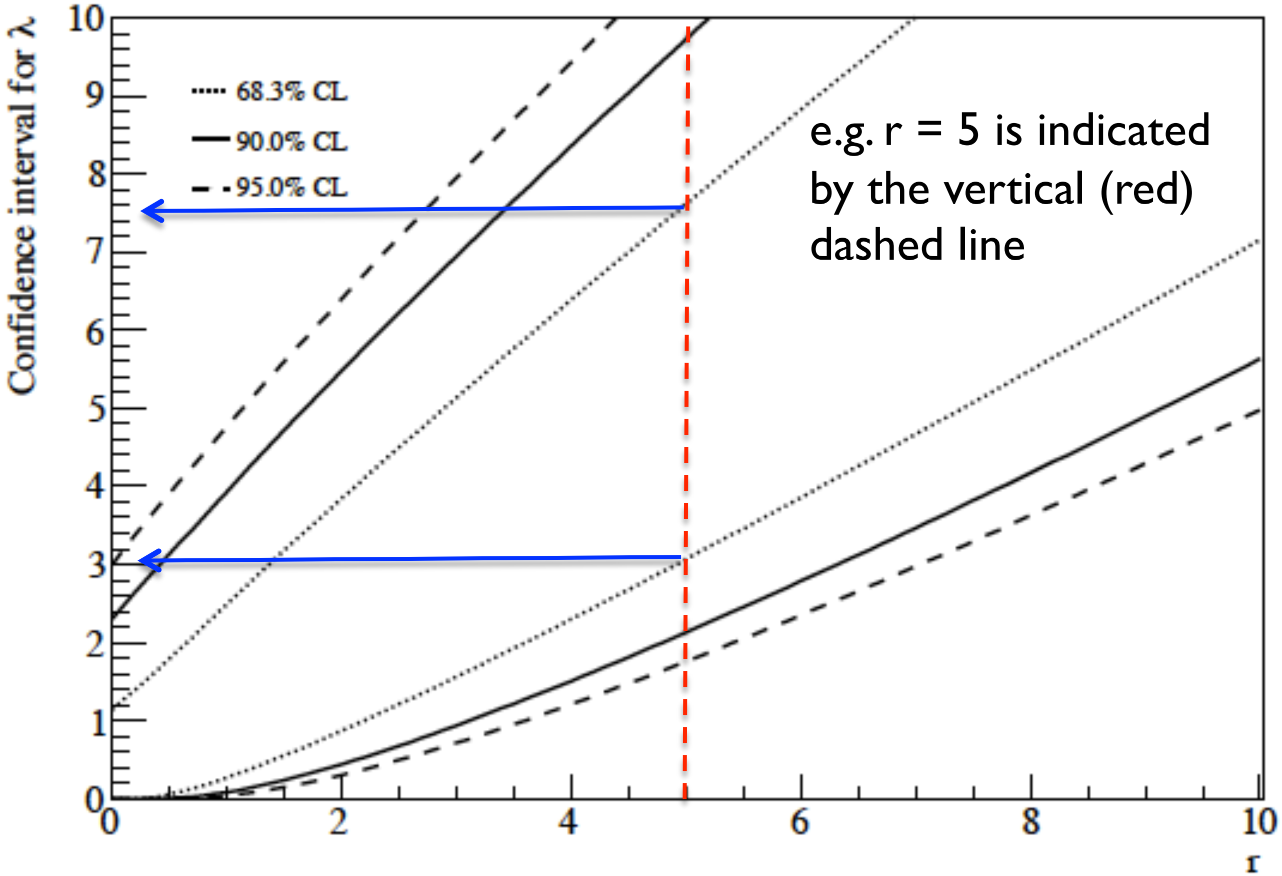
From each likelihood distribution one can construct a one or two sided interval by integrating L(λ ,r) to obtain the desired coverage.

- For a given observation, we can compute a likelihood as a function of λ, e.g.



From each likelihood distribution one can construct a one or two sided interval by integrating L(λ ,r) to obtain the desired coverage.

We can build a 2D region from a family of likelihood curves for different values of r.



Physically there are discrete observed numbers of events for a background free process. If however background plays a role, then the problem becomes more complicated, and non-integer values may be of relevance..

Multi-dimensional constraint

An example of constraining a model using data.

Adrian Bevan

email: a.j.bevan@qmul.ac.uk

Queen Mary University of London

Motivation

- The Standard Model of particle physics describes all phenomenon we know at a sub-atomic level.
- The model is incomplete:
  - Universal matter-antimatter asymmetry unknown
  - Nature of neutrinos unknown
  - What is Dark Matter
  - What is Dark Energy [related to a higher order GUT]
  - etc.

- Many particle physicists want to test the Standard Model precisely for two (related) reasons:
  - (i) understand the model better
  - (ii) see where it breaks...

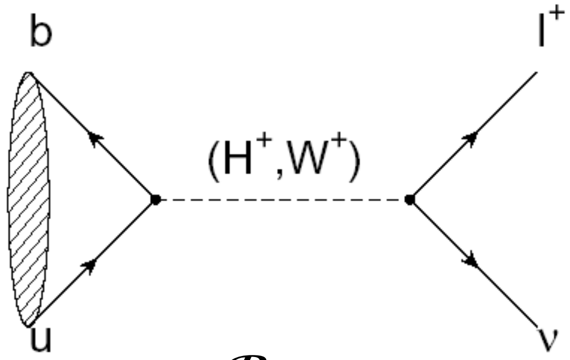
The problem

- The decay $B^{\pm} \to \tau^{\pm}\nu$ has been measured, and can be compared with theoretical expectations.
- Measurement:

$$\mathcal{B}(B^{\pm} \to \tau^{\pm}\nu) = (1.15 \pm 0.23) \times 10^{-4}$$

- Standard Model expectation:

$$\mathcal{B}(B^{\pm} \to \tau^{\pm}\nu)_{SM} = (1.01 \pm 0.29) \times 10^{-4}$$



$$r_H = \frac{\mathcal{B}_{SM+NP}}{\mathcal{B}_{SM}}$$

For a simple extension of the Standard Model, called the type II 2 Higgs Doublet Model we know that r_H depends on the mass of a charged Higgs and another parameter, β .

$$r_H = \left(1 - \frac{m_B^2}{m_H^2}\tan^2\beta\right)^2$$

What can we learn about m_H and tanβ for this model?

- We can compute r_H from our knowledge of the measured and predicted branching fractions:

$$r_H = 1.14 \pm 0.40$$

- How can we use this to constrain m_H and tanβ?

$$r_H = \left(1 - \frac{m_B^2}{m_H^2}\tan^2\beta\right)^2$$

Method 1: χ^2 approach

- Construct a χ^2 in terms of r_H

$$\chi^2 = \left(\frac{r_H - \widehat{r}_H(m_H, \tan\beta)}{\sigma_{r_H}} \right)^2$$

Method 1: χ^2 approach

- Construct a χ^2 in terms of r_H

From SM theory and experimental measurement

$$\chi^2 = \left(\frac{r_H - \widehat{r}_H(m_H, \tan\beta)}{\sigma_{r_H}} \right)^2$$

From SM theory and experimental measurement

Method 1: χ^2 approach

- Construct a χ^2 in terms of r_H

Calculate using

$$r_H = \left(1 - \frac{m_B^2}{m_H^2}\tan^2\beta\right)^2$$

One has to select the parameter values.

From SM theory and experimental measurement

$$\chi^2 = \left(\frac{r_H - \widehat{r}_H(m_H, \tan\beta)}{\sigma_{r_H}}\right)^2$$

From SM theory and experimental measurement

Method 1: χ^2 approach

- For a given value of m_H and $\tan\beta$ you can compute χ^2.

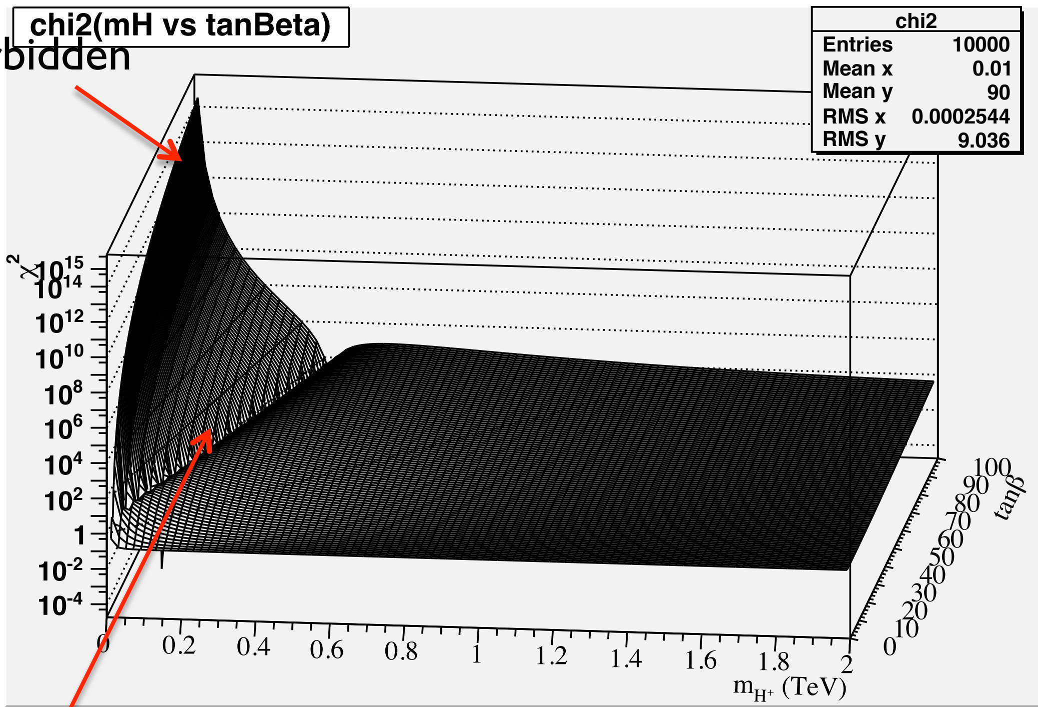
- e.g.

$$
\begin{aligned}
m_H &= 0.2\text{TeV} \\
\tan\beta &= 10 \\
\widehat{r}_H(m_H, \tan\beta) &= 0.93 \\
\chi^2 &= \left(\frac{1.14 - 0.93}{0.4}\right)^2 \\
&= 0.28
\end{aligned}
$$

- So the task at hand is to scan through values of the parameters in order to study the behaviour of constraint on r_H.

Method 1: χ^2 approach

Forbidden



Allowed region
(the valley)

A large χ^2 indicates a region of parameter space that is forbidden.

A small value is allowed.

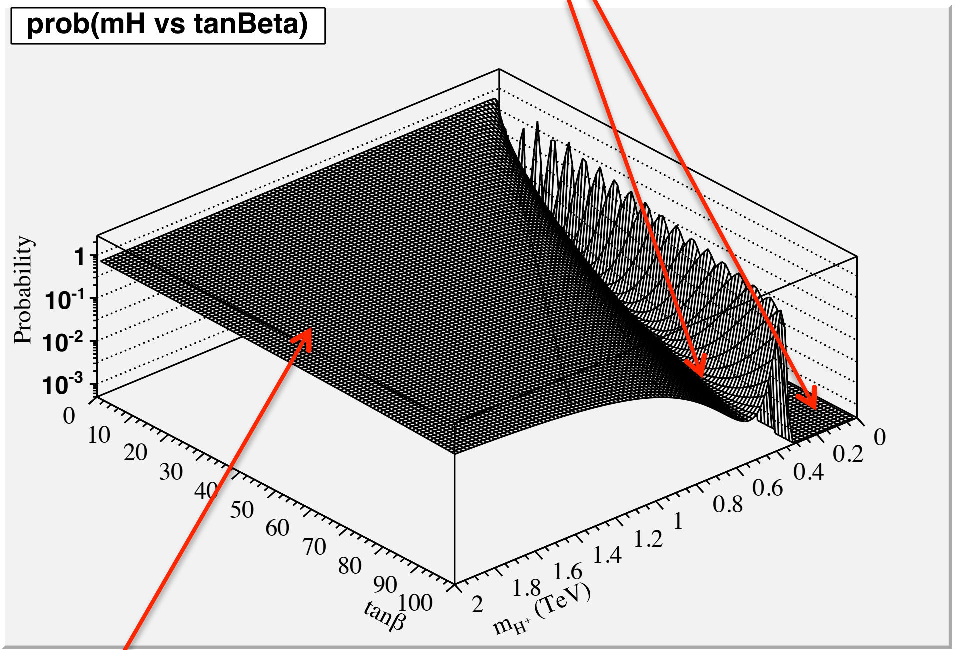
In between we have to decide on a confidence level that we use as a cut-off.

We really want to covert this distribution to a probability: so use the χ^2 probability distribution.

There are 2 parameters and one constraint (the data), so there are 2−1 degrees of freedom, i.e. $\nu = 1$

Method 1: χ^2 approach

Forbidden
(P ~ 0)



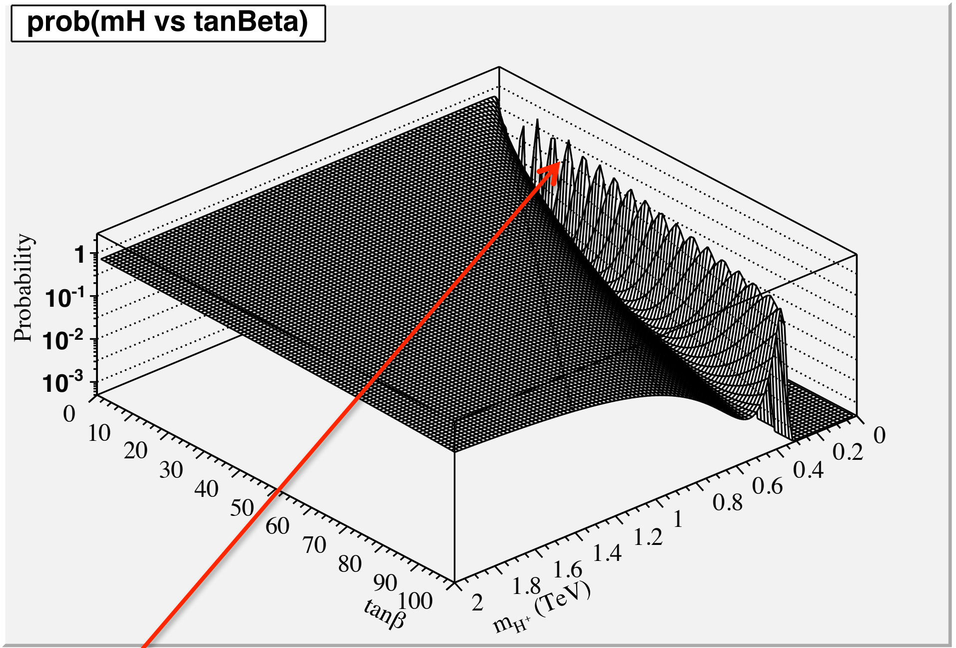
Allowed region
(P ~ 1)

A P ~ 1 means that we have no constraint on the value of the parameters (i.e. they are allowed).

A small value of P, ~0 means that there is a very low (or zero) probability of the parameters being able to take those values (i.e. the parameters are forbidden in that region).

Typically one sets a 1−CL corresponding to 1 or 3 σ to talk about the uncertainty of a measurement, or indicate an exclusion region at that CL.

Method 1: χ^2 approach



Artefact: a remnant of binning the data. For these plots there are 100 x 100 bins. As a result visual oddities can occur in regions where the probability (or χ^2) changes rapidly.

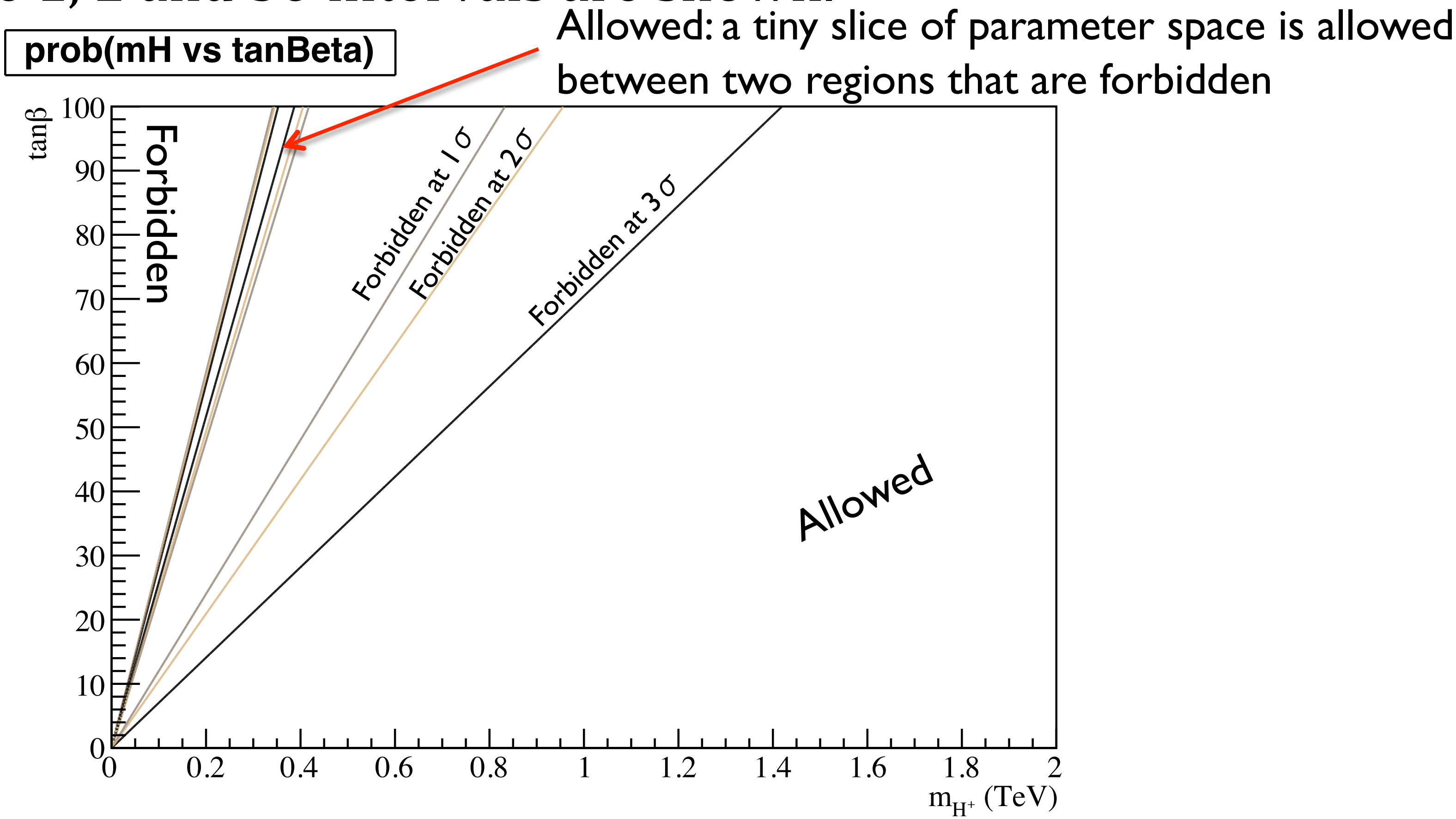
A P ~ 1 means that we have no constraint on the value of the parameters (i.e. they are allowed).

A small value of P, ~0 means that there is a very low (or zero) probability of the parameters being able to take those values (i.e. the parameters are forbidden in that region).

Typically one sets a 1−CL corresponding to 1 or 3 σ to talk about the uncertainty of a measurement, or indicate an exclusion region at that CL.

Method 1: χ^2 approach

- A finer binning can be used to compute a 1-CL distribution. Here 1, 2 and 3σ intervals are shown.



Summary: χ^2 approach

- This is nothing new – it is just a two-dimensional scan solution to a problem.

- It is however more computationally challenging to undertake (excel probably won't be a good fit to solving the problem):
  - 1D problem: N scan points
  - 2D problem: N^2 scan points
  - As N becomes large (e.g. 100 or 1000) the number of sample points becomes very large.
    - i.e. the curse of dimensionality strikes.

  - An MD problem has N^M sample points.
    - e.g. Minimal Super-Symmetric Model (MSSM) has ~160 parameters, so one has a problem with N^{160} sample points. This is currently not a viable computational method to explore the complexities of MSSM.