Probability, Likelihood and Confidence

Roger Barlow
Huddersfield University

LHCb-UK Student Talks

20^{th} February 2020

Contents

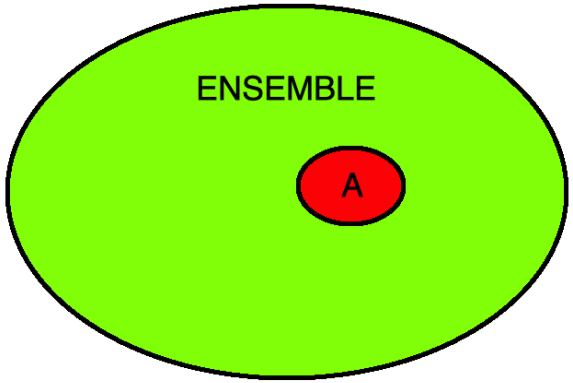
Probability, Likelihood, Confidence - What do they mean?

In normal use - pretty much the same thing

Science has adopted everyday words (like 'work', 'energy', 'charge') and given them exact definitions.

The same thing happens here.

Probability: two definitions



Frequentist

If N_A is the number of times A occurs out of an ensemble (or collective) of N trials, then P, the probability of A, is the limit of $\frac{N_A}{N}$ as $N \to \infty$

Bayesian

Degree of belief. If P is the probability of event A, you will accept a bet on A occurring in a trial if you are offered odds of $1 - P$ to P or better.

Both satisfy the Kolmogorov Axioms: Mathematical Probability.
Both agree with experience with coins, dice, cards, roulette wheels etc.

Frequentist Probability

Important feature

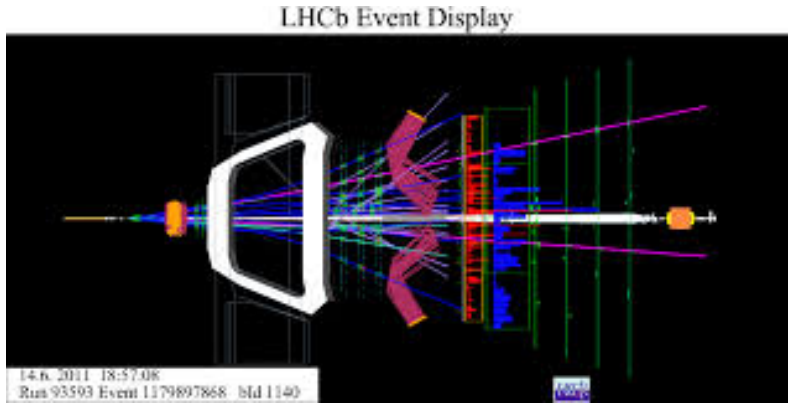
The *Probability of A* depends on the ensemble being considered.

Example: German life insurance companies find they pay out on 1.1% of their 40 year old male clients. (hard numbers: 940 in 85,020).
Your friend Hans is celebrating his 40th birthday.
The probability he will celebrate his 41st is 98.9%
But he does not smoke or drink. For that sample the numbers are (maybe) 198 in 30,120 giving a probability 99.2%
To celebrate his 40th birthday he has bought a motorbike! For that sample the numbers are (maybe) 73 in 5674 giving 98.7%
All these answers are correct.

Probability is a joint property of the event and the ensemble, and there may be more than one possible ensemble

Another example

A muon is produced in a collision. What is the probability it will be correctly identified by the PID?



Calculate using Monte Carlo or tag-and-probe. Take many events, and count how many muons are identified. You will get different results for:

- All muons generated
- High P_T muons
- Muons from B hadron decays
- Muons for which the track is reconstructed
- Muons in your specific η region

All are equally valid. The muon does not have a 'probability of being identified' in its own right.

This is a bit of a conceptual glitch, as coins, dice, cards etc do have this property

What if there is no ensemble?

Obviously, if an event is unique there is no ensemble and there can be no probability.

What is the probability that the gluino exists and has a mass below 1 TeV?

What if there is no ensemble?

Obviously, if an event is unique there is no ensemble and there can be no probability.

What is the probability that the gluino exists and has a mass below 1 TeV?

This is a non-question. Either it does or it doesn't. The probability is either 1 or 0.

What if there is no ensemble?

Obviously, if an event is unique there is no ensemble and there can be no probability.

What is the probability that the gluino exists and has a mass below 1 TeV?

This is a non-question. Either it does or it doesn't. The probability is either 1 or 0.

What is the probability that it will rain here tomorrow?

What if there is no ensemble?

Obviously, if an event is unique there is no ensemble and there can be no probability.

What is the probability that the gluino exists and has a mass below 1 TeV?

This is a non-question. Either it does or it doesn't. The probability is either 1 or 0.

What is the probability that it will rain here tomorrow?

This is a non-question. Either it does or it doesn't. The probability is either 1 or 0.

Technically this is correct. But not helpful.

From probability to confidence

The probability of rain tomorrow is either 0 or 1

From probability to confidence

The probability of rain tomorrow is either 0 or 1
Suppose some weather forecast predicts rain

From probability to confidence

The probability of rain tomorrow is either 0 or 1
Suppose some weather forecast predicts rain
You check the past accuracy of these forecasts and ascertain that they are right (say) 8 times out of 10.

You can't say "There is an 80% chance of rain tomorrow"

You can say "The statement 'It will rain tomorrow' has an 80% chance of being true."
This is not a statement about the rain, but about the forecast as a member of an ensemble of forecasts. You say "It will rain tomorrow" with 80% confidence

Note 1: There may be different ensembles. Perhaps the track record is 90% for forecasts of rain.
Note 2: Or the forecaster may say 'Looking at conditions lie today, rain fell in 80% of cases.' Or they ran many simulations and 80% gave rain.....

Confidence

Definition

Confidence: A statement is true at confidence CL if it is a member of an ensemble of statements of which at least CL are true

Reasons for that 'at least'

- Higher CL statements embrace lower. If a statement is true at 99% CL it is true at 95% CL
- If you want to make a statement at some CL but no ensemble has that value, you go for the next one up
- When dealing with composite models it tells you to deal with extra parameters by considering the worst case

Confidence and measurement

What does the measurement '$x = 100 \pm 10$' mean?

Not 'there is a 68% probability that x lies between 90 and 110'

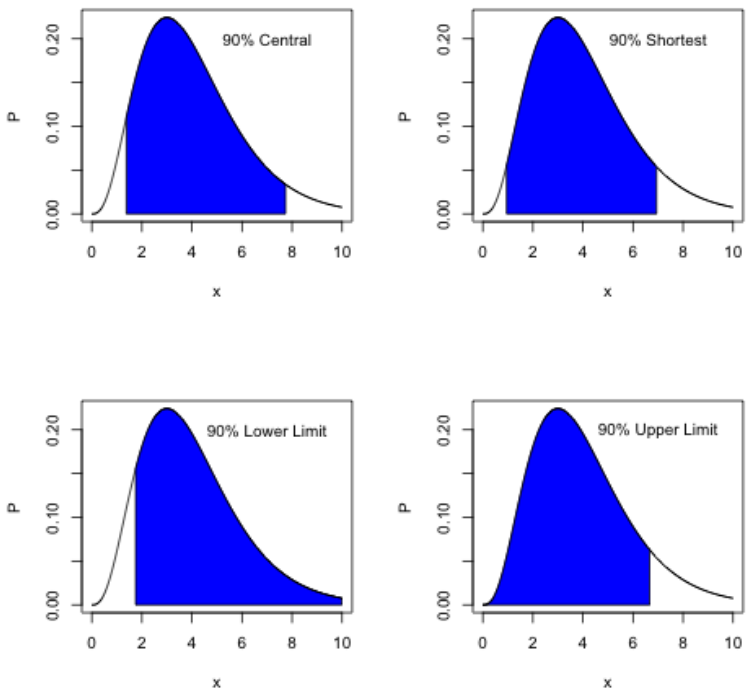
x has been measured with some device which gives a result distributed about the true value according to a Gaussian distribution with standard deviation 10

' x lies between 90 and 110, with 68% confidence'

If you go through life and whenever you see a measurement you assume the true values lies within the error, you will be right (at least) 68% of the time – and wrong (up to) 32% of the time..

Confidence regions

You have a choice about what CL to use. 68% or 90% or 95% or....
You also have a choice about how to choose the limits.

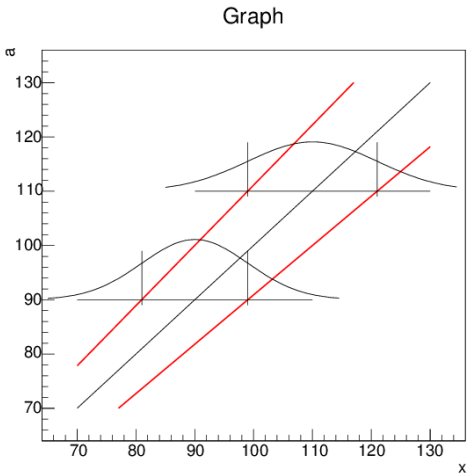


- Equidistant
- Central (Equal probability)
- Shortest length
- Lower Limit
- Upper Limit

Confidence Bands (or belts)

More subtle: a proportional Gaussian

You measure $x = 100$ with some device which has an error of 10%
Could be from a true value of 90, a 1.1 σ upward fluctuation
Could be from a true value of 110, a 0.9 σ downward fluctuation



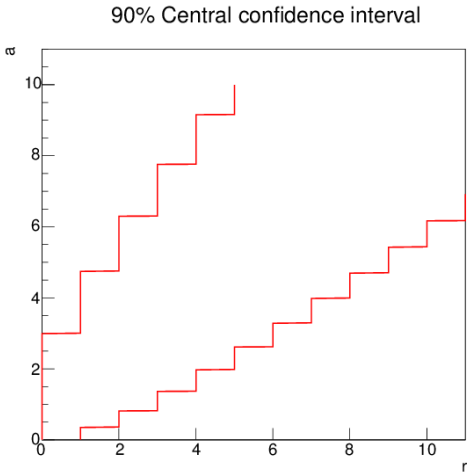
Construct Horizontally: for each true value *a* construct the desired confidence region - here 68% central
The probability that a measurement will lie in the band is 68%
Take the measurement - get a value of x
The probability that a measurement will lie in the band is 68%
Read Vertically: upper and low limits for *a* here are 90.9 and 111.1 .

'*a* lies between 90.9 and 111.1, with 68% confidence'

Note that the lower limit for *a* comes from the upper limit line, and vice versa. Makes sense if you think about it.

More Complicated: Poisson

$P(r;a) = e^{-a}\frac{a^r}{r!}$



Similar to previous example as standard deviation increases with true value

Different in that observable is discrete (0,1,2,3...) rather than continuous

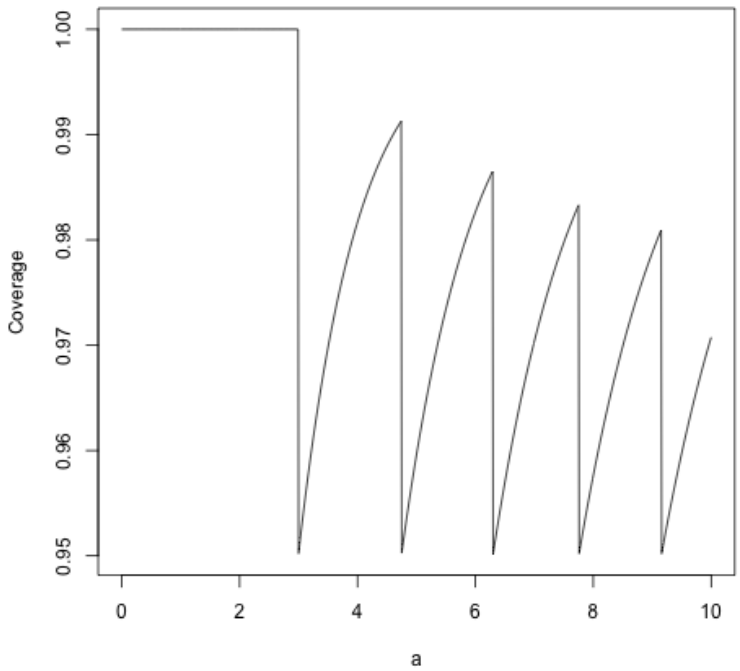
Need to invoke the 'at least' in the definition of confidence. Want at most 5% above and below the band

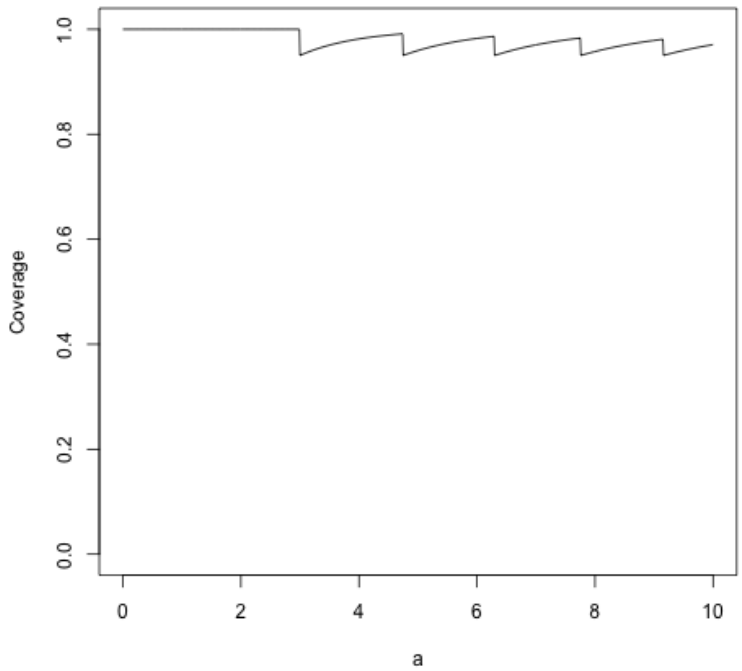
Example: consider $a = 3.9$. Want to construct probability 0.05
$P(0;3.9) = 0.020$: OK
But adding $P(1;3.9) = 0.079$ would take it over the limit.
Only at $a = 4.744$ do we get $P(0) + P(1) = 0.0087 + 0.0413 = 0.05$

This line is interesting/important as it is also the 95% upper limit
Breaks at a=2.996,4.744,6.296,7.754,9.153...

Coverage

Much discussed





Probability - as function of a - that a measurement will give a result for which the constructed confidence interval includes a

Here for 95% Poisson upper limit

Should never undercover: go below desired CL (here 0.95)

Is allowed to overcover: go above desired CL, though this is inefficient

Bayesian Probability

Advantage: no restriction on topics: rain tomorrow, existence of gluino, whatever... are all valid topics

Disadvantage: no reason why my $P(A)$ and your $P(A)$ should be the same.

Hence also known as 'Subjective Probability'
Makes great use of Bayes' Theorem

Theorem (Bayes' Theorem)

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

Actually applies to Frequentist probability too

Proof.

$$P(A|B)P(B) = P(A\&B) = P(B|A)P(A)$$ □

Bayes' Theorem - general

Example (Trivial)

A card is drawn from a pack. It is black. What is the probability it is the ace of spades?

$$P(A\spadesuit|black) = \frac{P(black|A\spadesuit)}{P(black)} P(A\spadesuit) = \frac{1}{1/2} \times \frac{1}{52} = \frac{1}{26}$$

Example (More useful)

A beam which is 90% π^+ and 10% K^+ passes through a Cherenkov counter which has a 95% chance of producing a signal for a π^+ and 2% for a K^+. If some particle gives no signal, what is the probability that it is a K^+?

$$P(K^+|nosignal) = \frac{P(nosignal|K^+)}{P(nosignal)} P(K^+) = \frac{0.98}{0.1\times0.98+0.9\times0.05} \times 0.1 = 68.5\%$$

Lemma

$$P(B) = P(B|A) \times P(A) + P(B|\overline{A}) \times (1 - P(A))$$

Bayes' Theorem - Bayesian

$$P(Theory|Data) = \frac{P(Data|Theory)}{P(Data)} \times P(Theory)$$

Why is this Bayesian and not Frequentist?

Bayes' Theorem - Bayesian

$$P(Theory|Data) = \frac{P(Data|Theory)}{P(Data)} \times P(Theory)$$

Why is this Bayesian and not Frequentist? Because in Frequentist Probability there is no such thing as $P(Theory)$ let alone $P(Theory|Data)$

In Bayesian Probability they are known as your Prior and Posterior belief.

Behaves sensibly: observation of data that the theory say are likely boosts your belief in the theory, though this is moderated by the probability it could happen anyway. Likewise if $P(Data|Theory)$ is small or zero, so is $P(Theory|Data)$.
We use this all the time in our daily lives, making life decisions based on posterior beliefs and reward/penalty functions.

Priors and Posteriors - from numbers to distributions

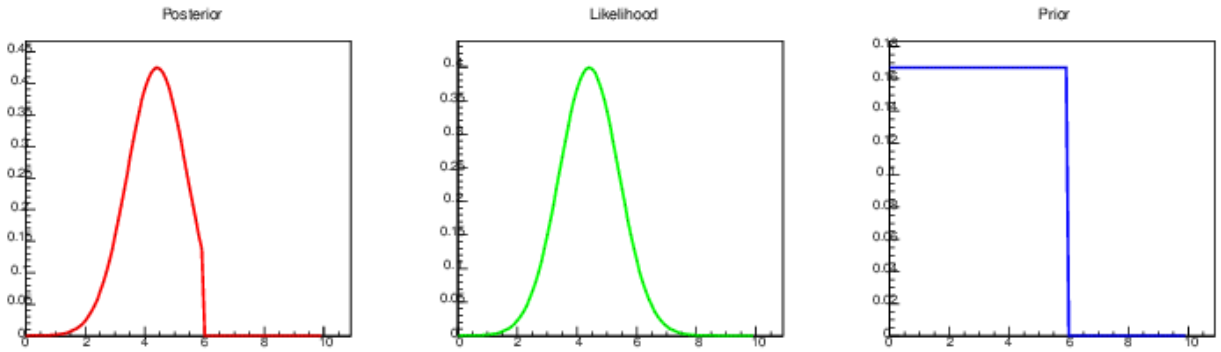
Generalise from 'the probability that the gluino exists and has a mass below 1 TeV' to 'the probability distribution function for the gluino mass' $P(M_{\tilde{g}})$

Then in the light of some data (which presumably is not in agreement with light $M_{\tilde{g}}$ but is compatible with heavier values) you get a posterior $P(M_{\tilde{g}}|Data) = \frac{P(Data|M_{\tilde{g}})}{P(Data)} P(M_{\tilde{g}})$

Notice that the denominator does not contain $M_{\tilde{g}}$ so it is easier to write $P(M_{\tilde{g}}|Data) \propto P(Data|M_{\tilde{g}})P(M_{\tilde{g}})$
and normalise (to 1, or, strictly, to your overall belief that $M_{\tilde{g}}$ is somewhere)

Posteriors and Priors

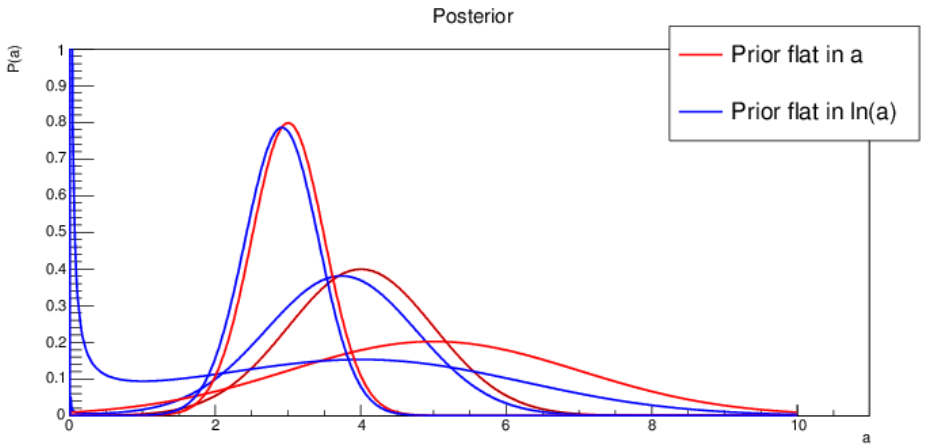
Replace specific $M_{\tilde{g}}$ by general a, and data by x (or, if integers, r)



Sometimes prior cannot be integrated ("improper"). Example: a uniform from 0 to infinity.....

Don't worry, normalising the posterior takes care of that

Posteriors depend on Priors



Shows posteriors from 3 measurements. $3 \pm 0.5, 4 \pm 1, 5 \pm 2$
Red curves are from a prior flat in a (which is often plausible)
Blue curves are from a prior flat in $\ln a$ (which is often plausible)
Bernstein-von Mises theorem says that if you have lots of data the posterior is independent of the prior. But we do not usually have this luxury - certainly in interesting cases.

Making a virtue out of ignorance: 'I know nothing so I take a prior uniform in a ' is dishonest. Uniform in a is not uniform in $\ln a$ or $\cos a$ or $\sqrt{a}$ or ...

Credible Intervals

Bayesians do not need all the confidence apparatus

From the posterior distribution can say 'The probability that a lies in the region is 68%', or whatever.

Call these 'credible intervals': look like confidence regions but are subtly different

Same freedom of choice about the number to quote and the region (central, shortest, upper limit, lower limit...)

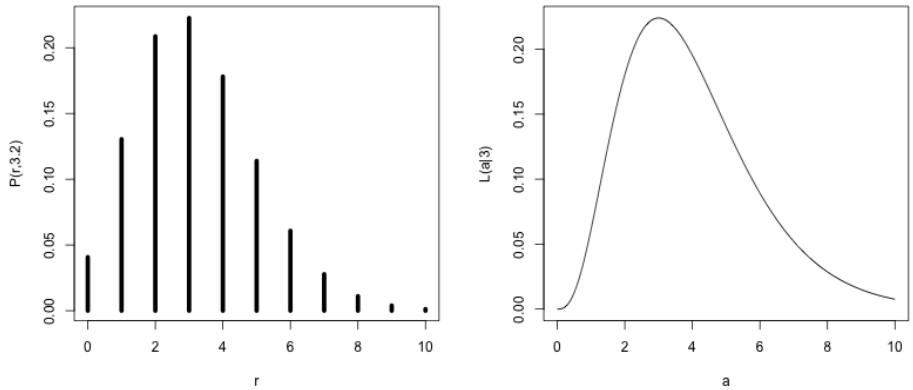
Coverage is in principle irrelevant for Bayesians - undercoverage is not illegal - but it has proved an interesting thing to study

The Likelihood

Has already appeared as $P(Data|Theory)$
Probability of a data sample: $L(x_1, x_2....|a) = \prod_i P(x_i|a)$
Use in Fisher information, Minimum Variance Bound, expectation values for a sample... including Maximum Likelihood Estimation

$L(data|a)$ does not tell you the probability of a particular value for a
Bayesian: Because you need to multiply it by the prior.
Frequentist: because the 'probability of a' is meaningless
Example: single Poisson measurement

$$P(r; a) = e^{-a}\frac{a^r}{r!} = L(a; r)$$



The Likelihood Principle

Says that all the information about an experimental result is contained in the likelihood function $L(a|data)$

Sounds plausible

But standard Frequentist interpretations of results break this as they also include outcomes of experiments that didn't happen

So not so plausible

However, publishing the complete likelihood function (as opposed to just peak value and + - errors) is a good thing

The Upper Limit problem - I

You are searching for a new particle (or decay mode, or ...)
You see r events. What can you say - specifically if $r = 0$?

Frequentist.

If the true value is a_{hi} or more, the probability of getting r events or less is
$\sum_0^r e^{-a_{hi}} \frac{a_{hi}^r}{r!} = 0.05$
If $r = 0$, $a_{hi} = 2.996$: the true value is at most 2.996,with 95% confidence

Bayesian

Use likelihood to construct the 95% credible interval $\int_0^{a_{hi}} e^{-a} \frac{a^r}{r!}\, da = 0.95$
Integration by parts gives
$\left[-e^{-a} \frac{a^r}{r!}\right]_0^{a_{hi}} + \int_0^{a_{hi}} e^{-a} \frac{a^{r-1}}{(r-1)!}\, da = 0.95$
Repetition gives $\sum_0^r e^{-a_{hi}} \frac{a_{hi}^r}{r!} = 0.05$
Same as Frequentist formula!
If $r = 0$, $a_{hi} = 2.996$: the true value is at most 2.996,with 95% probability

Agreement is a fortunate co-incidence

From the statement about r, obtain statements about cross-section, branching ratio, mass or whatever.

The Upper Limit problem - II

Suppose there is also a known background B

Example (1 Straightforward)

$B = 1.2$, observe 4 events
Upper limit 9.2 on $N_T = N_S + N_B$ at 95% confidence/probability
So upper limit 8.0 on S

Example (2 Suspicious)

$B = 1.2$, observe 0 events
Upper limit 3.0 on $N_T = N_S + N_B$ at 95% confidence/probability
So upper limit 1.8 on S

Example (3 Crazy)

$B = 3.2$, observe 0 events
Upper limit 3.0 on $N_T = N_S + N_B$ at 95% confidence/probability
So upper limit -0.2 on S

Straightforward approaches

What's happening? In cases (2) and (3) there is clearly a downward fluctuation in the background. But there is no way to fold that knowledge into the standard Frequentist analysis.

Ultra-strict frequentist: publish non-physical upper limit
Justification: There will be upward and downward fluctuations. This happens to be a downward fluctuation. If we suppress it, that will bias the published results.

Or just go Bayesian. Prior for S which is zero for negative S. Tempting. (Though you should still check for robustness under different priors)

But there are ways through this. Feldman-Cousins and CL_s

The Feldman-Cousins unified method

Works by attacking a somewhat different problem

Suppose you observe r events with expected background B
If $r >> B$ you will want to publish a 2-sided measurement

Example

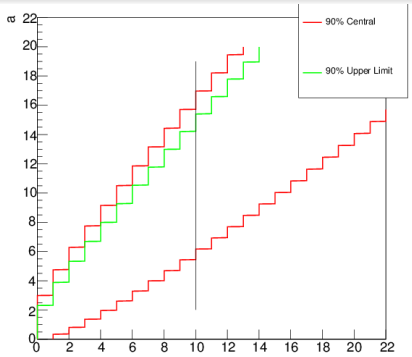
You observe 25 events with a background of 1.2, you report a measurement of 23.8 ± 5.0 events

If $r \sim B$ you will want to publish an upper limit

Example

You observe 4 events with a background of 1.2, you report an upper limit of 8.0 signal events, at 95% confidence. (Example (1) on previous slide)

'Flip-flopping'. Transition point will depend on plausibility. But it's somewhere.
You've broken your confidence band

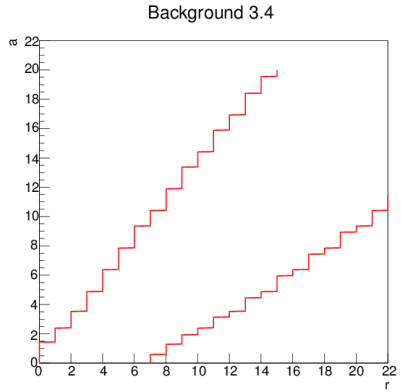


Feldman-Cousins Method

Plot r horizontally as before, but S vertically. So different $B \rightarrow$ different plot. Probability values $P(r;S) = e^{-(S+B)}\frac{(S+B)^r}{r!}$

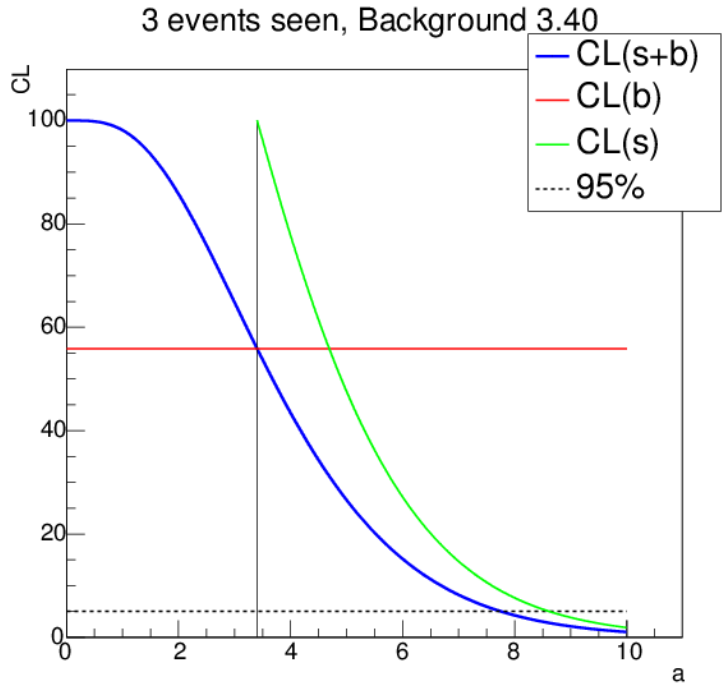
For any S have to define region R such that $\sum_{r\epsilon R} P(r;S) \geq 95\%$.

First suggestion: rank r by probability and take them in order (would give shortest interval). Drawback: outcomes with $r << B$ will have small probabilities and all S will get excluded. But such events happen - want to say something constructive, not just 'This was unlikely'



Better suggestion: For each r, compare $P(r;S)$ with the largest possible value obtained by varying S. This is either at $S = r - B$ (if $r \geq B$) or 0 (if $r \leq B$) Rank on the ratio
Valid confidence belt. Flip-flop smooth. Limit always sensible.

CL_S



CL_{s+b}: Probability of getting a result this small (or less) from $s + b$ events. Same as strict frequentist.

CL_b: CL_{s+b} for $s = 0$ - no signal, just background

$$CL_s = \frac{CL_{s+b}}{CL_s}$$

Apply as if confidence level $1 - CL_s$

Result larger than strict frequentist ('conservative') ('over-covers')
In this example 8.61 for $s + b$, $s = 8.61 - 3.40 = 5.21$
As opposed to strict frequentist $s = 7.75 - 3.40 = 4.35$
CL_s is neither frequentist nor Bayesian,. But it gives sensible numbers and is widely used in Particle Physics. But has no sound statistical foundation.

Extension: beyond simple counting

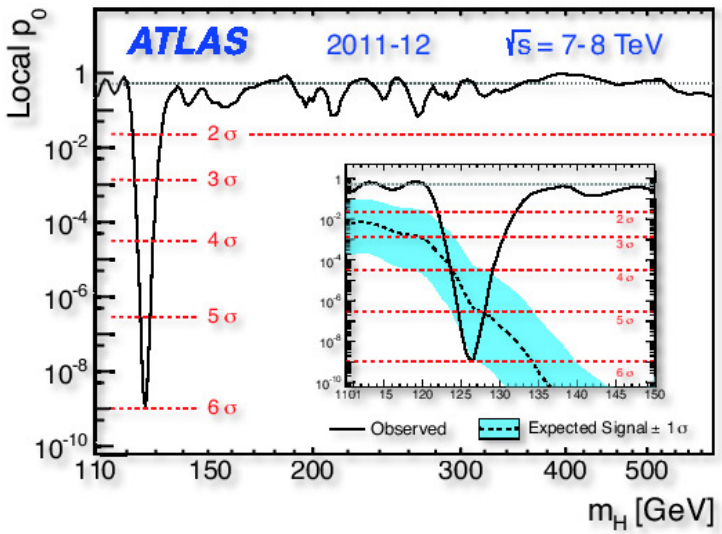
There's typically a lot more information in an event that just its existence: masses, submasses, b-quark tagging probabilities, NN outputs....
You also may want to combine results from several channels. And different experiments.
Tools were developed during (especially) the Higgs search at LEP
A simple accept/reject criterion does not make full use of that information
Define some test statistic Q and replace Poisson formula by some $P(Q|a)$
Usual choice the Likelihood Ratio $Q = L(\{x\}|a)/L(\{x\}|0)$
for N Channels/experiments

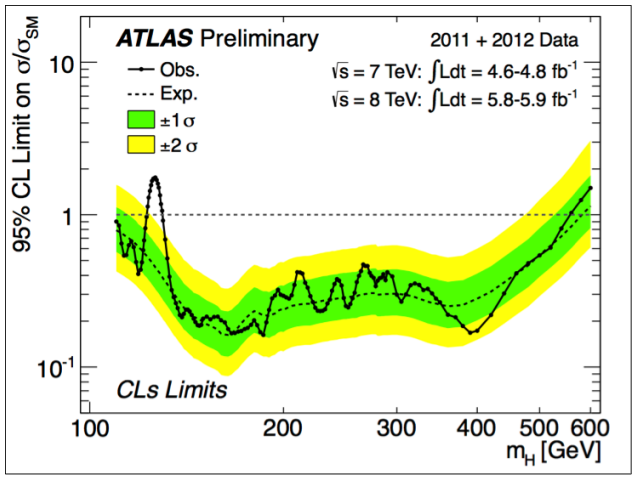
$$Q = \frac{\prod_{i=1}^{N} e^{-(s_i+b_i)} \frac{(s_i+b_i)^n_i}{n_i!} \prod_{j=1}^{n_i} \frac{s_i S_i(x_{ij})+b_i B_i(x_{ij})}{s_i+b_i}}{\prod_{i=1}^{N} \frac{e^{-b_i} b_i^{n_i}}{n_i!} \prod_{j=1}^{n_i} B_i(x_{ij})} \qquad \text{with known } s_i(a)$$

Bonus: $-2 \ln Q$ behaves (roughly) like χ^2 (Wilks' theorem)

Putting it all together

Scan the parameter of interest (e.g. M_H). Calculate the number(s) of expected events $s_i(M_H)$. Will depend on cross section and on efficiency. The S function depends on M_H. The B functions and the b_i background numbers will also, if your analysis strategy adapts.
Calculate Q as a function of M_H,and the equivalent CL_S or whatever
Can establish a limit, at a given confidence level, or can calculate the $p-$value for the SM prediction





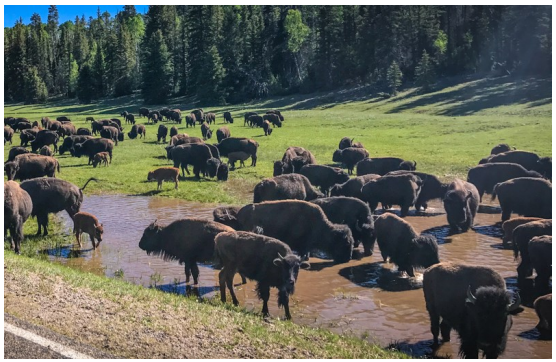
Also simulate what you'd expect to see with no signal (dotted line) and spread (green and yellow bands), using Monte Carlo

Conclusions

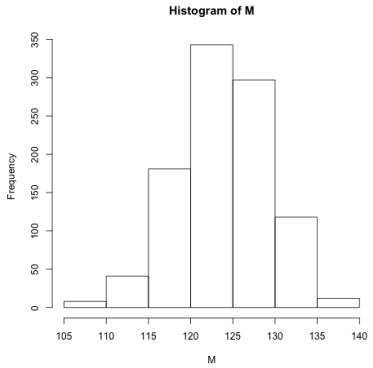
1. Frequentist Probability
   - Confidence
   - Coverage
2. Bayesian Probability
   - Bayes Theorem
   - Priors and Posteriors
   - Credible Intervals
   - Likelihood
3. The Upper Limit Problem
   - Simple counting
   - Upper limits with background
     - Feldman-Cousins
     - CL_S
   - Beyond simple counting

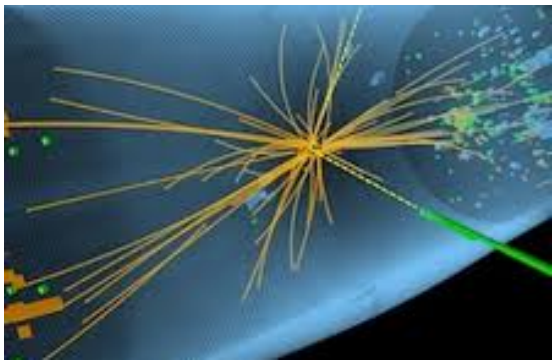
Backup slides

Aside: the difference between a bison and a boson

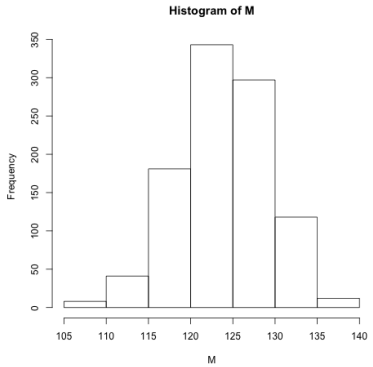


An experimenter measures the masses of a sample of bison. They plot the results in a histogram





An experimenter measures the masses of a sample of bosons. They plot the results in a histogram



Histograms look the same (apart from units etc.) but are different.
The spread in the bison histogram is due to the different bison masses.
The spread in the boson histogram is due to measurement uncertainties.