Statistics for Particle Physics
Lecture 3:Setting limits and making discoveries

Roger Barlow
Roger.Barlow@cern.ch
Huddersfield University

12th LHC Physics School, NCP, Islamabad

25^{th} August 2023

Contents

1. Setting Limits
   Confidence
   Confidence regions
   Confidence belts
   Limits: Feldman-Cousins
   Limits: CLs
2. Making Discoveries
   Goodness of Fit
   Significance
   The Look Elsewhere effect
   Blind Analysis

Frequentist Confidence

Not allowed

"There is an 80% chance of rain tomorrow"

OK

"The Statement 'It will rain tomorrow' has an 80% chance of being true"

Equivalently

"It will rain tomorrow, with 80% confidence"

We state X with confidence P if X is a member of an ensemble of statements of which at least P are true.

Note that 'at least'. 3 reasons

1. Higher confidences embrace lower ones. If X at 95% then X at 90%
2. Handles cases with integer data where an exact match may not be possible
3. Caters for cases not completely defined

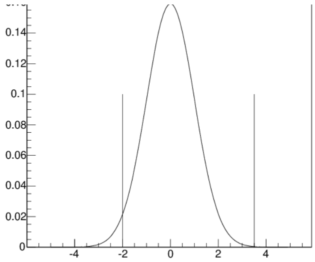
Confidence Regions

also known as Confidence Intervals

Interval $[x_-, x_+]$ such that
$\int_{x_-}^{x_+} P(x)\, dx = CL$
Choice over probability content CL
(68%, 90%, 95%, 99%...)
Choice over strategy

1. Symmetric: $\hat{x} - x_- = x_+ - \hat{x}$
2. Shortest: Interval that minimises $x_+ - x_-$
3. Central: $\int_{-\infty}^{x_-} P(x)\, dx = \int_{x_+}^{\infty} P(x)\, dx = \frac{1}{2}(1 - CL)$
4. Upper Limit: $x_- = -\infty$, $\int_{x_+}^{\infty} P(x)\, dx = 1 - CL$
5. Lower Limit: $x_+ = \infty$, $\int_{-\infty}^{x_-} P(x)\, dx = 1 - CL$

Lots of flexibility!



For the Gaussian (or any symmetric pdf) 1-3 are the same

Measurements are confidence regions

Q: What does it mean to say

$$M_H = 125.10 \pm 0.14 GeV?$$

A: M_H has been measured to be 125.10 with a technique that will give a value within 0.14 GeV of the true value 68% of the time

If we say the true value lies within $\pm\sigma$ we will be correct 68% of the time

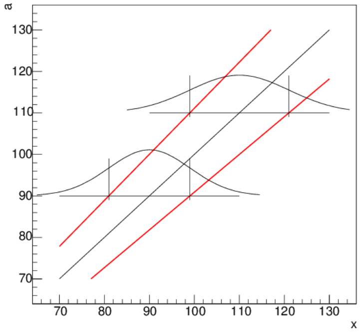
We say $124.96 < M_H < 125.24 GeV$ with 68% confidence.

The statement is either true or false (time will tell) but belongs to a collection of statements of which (at least) 68% are true.

Confidence Belts (1): Gaussian

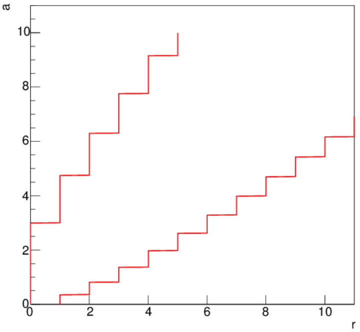
Get $x = 100$ from Gaussian measurement $\sigma = 0.1x$ (10% measurement)
Call (unknown) true value a.
$a = 90$ gives 90 ± 9 but $a = 110$ gives 110 ± 11. Not equivalent...

$\boxed{\text{Construct a Confidence Belt horizontally and then read it vertically}}$



1. For each a, construct desired confidence interval (here 68% central)
2. The result (x, a) lies inside the belt, with 68% confidence.
3. Measure x (here 100.0)
4. The result (x, a) lies inside the belt, with 68% confidence.
5. Read off a_+ and a_-: 111.1, 90.9

Confidence Belts (2): Poisson



Horizontal axis is discrete

For central 90% confidence require for each a the largest r_{lo} and smallest r_{hi} for which
$\sum_{r=0}^{r_{lo}-1} e^{-a} \frac{a^r}{r!} \leq 0.05$
$\sum_{r=r_{hi}+1}^{\infty} e^{-a} \frac{a^r}{r!} \leq 0.05$
For the second, easier to calculate
$\sum_{r=0}^{r_{hi}} e^{-a} \frac{a^r}{r!} \geq 0.95$

Whatever the value of a, the probability of the result falling in the belt is 90% or more. Proceed as for Gaussian...

Upper Limits

Why all this matters

Many analyses are 'searches for...'
most of these are unsuccessful

But you have to say something! Not just 'We looked but didn't see anything.'

Use upper limit confidence region as way of reporting: 'We see (almost) nothing, so $a \leq a_{hi}$ at some confidence level.'

Example

Simple use case : $P(0; 2.996) = 0.05$ *and* $2.996 \sim 3$. *So if you see 0 events, you can say with 95% confidence that the true value is less than 3.0*
Use this to calculate limit on branching fraction, cross section, or whatever you're measuring

Bayesian ‘credible intervals’

Bayesian has no problems saying ‘It will probably rain tomorrow’ or ‘The probability that $124.85 < M_H < 125.33 GeV$ is 68%’

Downside is that another Bayesian can say ‘It will probably not rain tomorrow’ and ‘The probability that $124.85 < M_H < 125.33 GeV$ is 86%’ with equal validity.

Bayesian has prior (or posterior) belief pdf $P(a)$ and defines region R such that $\int_R P(a)\,da = 90\%$ (or whatever)

Same ambiguity as to choice of content (68%, 90%, 95%...) and strategy (central, symmetric, upper limit...). So Bayesian credible intervals look a lot like frequentist confidence intervals. But...

Two happy coincidences

Gaussian Limits

Bayesian credible intervals on Gaussians, with a flat prior, are the same as Frequentist confidence intervals

F quotes 68% or 95% or ... confidence intervals.

B quotes 68% or 95% or ... credible intervals.

They are numerically the same

Poisson upper limits

The Frequentist Poisson upper limit is given by $\sum_{r=0}^{r=r_{data}} e^{-a_{hi}} a_{hi}^r / r!$

The Bayesian Poisson flat prior upper limit is given by

$\int_0^{a_{hi}} e^{-a} a^{r_{data}} / r_{data}! \, da$

Integration by parts gives a series - same as the Frequentist limit

Bayesian will say : 'I see zero events - the probability is 95% that the true value is 3.0 or less.' Numbers same as for Frequentist even if meaning different...

This is a coincidence - does not apply for lower limits

Limits in the presence of background

When it gets tricky

Typically background N_B and efficiency η, and want $N_S = \frac{N_D - N_B}{\eta}$
(Any uncertainties in η and N_B handled by profiling or marginalising)
Actual number of background events Poisson in N_B.

Straightfoward case

See 12 events, expected background 3.4, $\eta = 1$: $N_S = 8.6$
though error is $\sqrt{12}$ not $\sqrt{8.6}$

Hard case

But suppose you see 4 events. or 3 events. Or zero events...
Can you say $N_S = 0.6$? or -0.4? Or -3.4???

We will look at 4 methods of getting out of this fix

Example

See 3 events with expected background 3.40. What is the 95% limit on N_S?

Method 1: Pure frequentist

$N_D - N_B$ is an unbiassed estimator of N_S and its properties are known
Quote the result. Even if it is non-physical

Argument for doing so

This is needed for balance: if there is really no signal, approx. half of the experiments will give positive values and half negative. If the negative results don't publish, but the positive ones do, people will be fooled.

If $N_D < N_B$, we know that the background has fluctuated downwards. But this cannot be incorporated into the formalism

Upper limit from 3 is 7.75, as $\sum_0^3 e^{-7.75} 7.75^r / r! = 0.05$

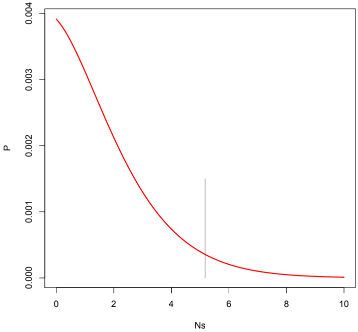
95% upper limit on $N_S = 7.75 - 3.40 = 4.35$

What if N_B were 8.0? Then publish -0.25! For a 95% confidence limit one accepts that 5% of the results can be wrong. This (unlikely) case is clearly one of them. So what?

Method 2: Go Bayesian

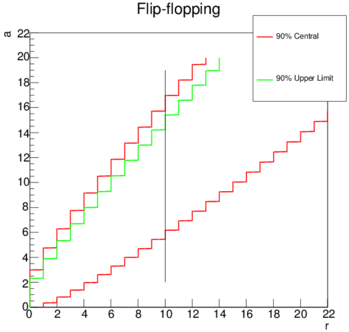
Assign a uniform prior to N_S, for $N_S > 0$, zero for $N_S < 0$.
The posterior is then just the likelihood,
$P(N_S|N_D, N_B) = e^{-(N_S+N_B)} \frac{(N_S+N_B)^{N_D}}{N_D!}$
Required Limit from integrating $\int_0^{N_{hi}} P(N_S)\, dN_S = 0.95$

$P(N_S) \propto e^{-(N_s+3.40)} \frac{(N_s+3.4)^3}{3!}$
Limit is 5.21



Method 3: Feldman-Cousins 1: Motivation

The Unified Approach



In principle, can use 90% central or 90% upper limit, and the probability of the result lying in the band is at least 90%.

In practice, you would quote an upper limit if you get a low result, but if you get a high result you would quote a central limit. Flip-flopping.

Break shown here for $r = 10$

Confidence belt is the green one for $r < 10$ and the red one for $r \geq 10$.

Probability of lying in the band no longer 90%. Undercoverage. Method breaks down if used in this way

Method 3: Feldman-Cousins 2: Method

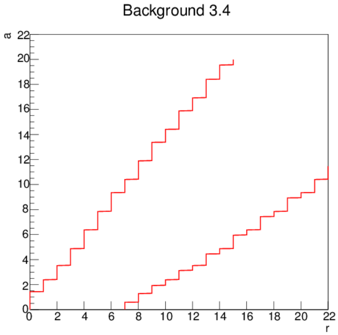
Plot $r \equiv N_D$ horizontally as before, but N_S vertically. So different $N_B \rightarrow$ different plot. Probability values $P(r; N_s) = e^{-(N_s+N_B)} \frac{(N_S+N_B)^r}{r!}$

For any N_S have to define region R such that $\sum_{r \epsilon R} P(r; N_s) \geq 90\%$.

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

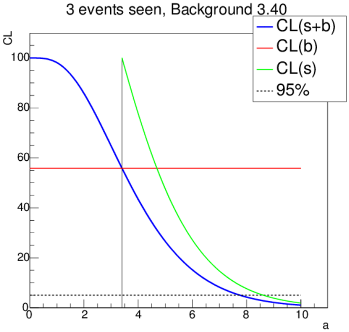
Better suggestion: For each r, compare $P(r; N_s)$ with the largest possible value obtained by varying N_S. This is either at $N_S = r - N_B$ (if $r \geq N_B$) or 0 (if $r \leq N_B$) Rank on the ratio

Method 3: Feldman-Cousins 3: Example



Flip-flopping incorporated! Coverage is correct.
For $r = 3$ get limit 4.86

Method 4: 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

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

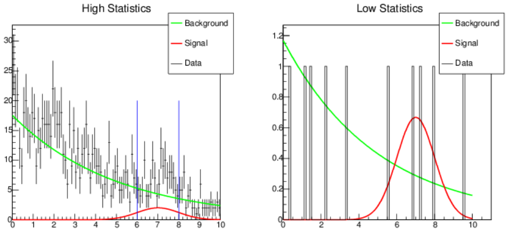
Apply as if confidence level $1 - CL_s$

Result larger than strict frequentist ('conservative') ('over-covers')

In our example 8.61 for $s+b$, 5.21 for s

CL_S Extension: not just numbers

In this simple example (just counting) CL_S is the same as Bayesian
But simple counting does not (usually) exploit the full information



Better: Likelihood
$lnL_{s+b} = \sum_i \ln N_s S(x_i) + N_b B(x_i)$ $\qquad lnL_b = \sum_i \ln N_b B(x_i)$
Look at L_{s+b}/L_b, or $-2\ln\left(L_{s+b}/L_b\right)$
Get confidence quantities from simulations/data

Summary so far

Given 3 observed events, and an expected background of 3.4 events, what is the 95% upper limit on the 'true' number of events?

Answers:

Strict Frequentist	4.35
Bayesian (uniform prior)	5.21
Feldman-Cousins	4.86
CL_s	5.21

Take your pick!

All are correct. (Well, not wrong.)

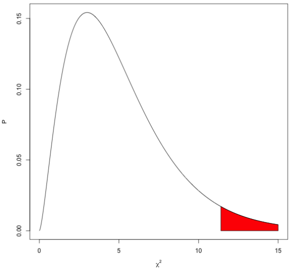
Golden Rule

Say what you are doing, and if possible give the raw numbers

Goodness of Fit

"Goodness of fit'" described by
$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - f_i}{\sigma_i}\right)^2$
(Other measures occur, but χ^2 overwhelmingly most popular)
Expect $\chi^2 \approx N$. In detail want p-value.
For example: Have $N = 5$ but get $\chi^2 = 11.3$
p=0.046. (Z=1.69 σ)

If the f_i have been fitted to the data, use
$N_{DF} = N - N_{params}$



Wilks' Theorem

$-2\ln(L/L_0)$ behaves like χ^2
where L_0 is the log likelihood for a basic model and L has extra term(s)
So L cannot answer the question "does the data fit" but can answer "does adding a signal term really help?"

Making Discoveries.

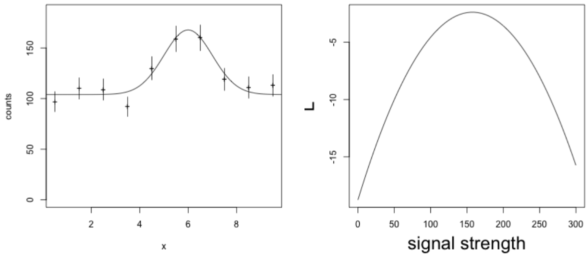
Hypothesis testing and the double-negative

Using statistics to support a statement you have to show that the opposite statement is not supported. Construct the Null Hypothesis H_0 that the effect you're interesting in does not exist

Suppose you bet a coin will come up heads, and lose 10 times running

If the coin is fair (H_0) then the chance of this happening is $\frac{1}{1024}$.
We say with 99.9% confidence that an honest coin will not let you lose 10 times running (p-value 0.001 or 3.1 sigma)
Which is small - so small we can (?) rule it out
So the coin is not fair. Hence it must be phony

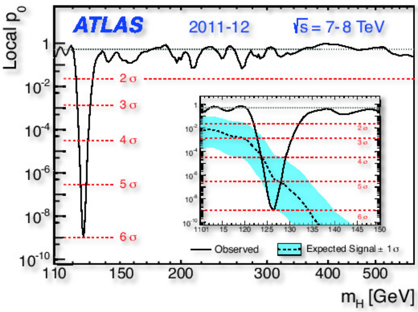
If your experiment succeeds, it does so by ruling out H_0
'The new drug produces more cures than would occur naturally' $\rightarrow$ the new drug works
'The peak in the mass distribution is too large to be a background fluctuation' $\rightarrow$ there is a new particle



You fit data to a model with a flat background and a Gaussian peak
Parameters specified: $\mu = 6, \sigma = 1$. Only the size (IF ANY) is unknown
You get the log likelihood shown in right hand plot. Then either

1. Find best value S from peak, error σ from $\Delta \ln L = -\frac{1}{2}$, and express significance as S/σ standard deviations, or
2. Note change in $\ln L$ from $S = 0$ to $S = \hat{S}$ and apply Wilks' theorem to get equivalent χ'^2, and thus p-value, and thus Z

Significance plots



For each M_H (or whatever): find signal and plot CL_s (or whatever) significance of signal

Small values indicate: unlikely to get a signal this large just from background

Often also plot expected (from MC) significance assuming signal hypothesis is true. Better measure of 'good experiment'

Green-and-yellow plots

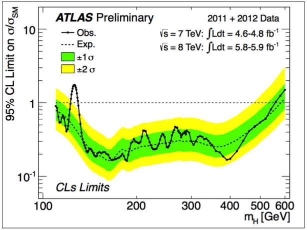
also known as "Brazilian flag plots"

Basically same data, but fix *CL* at chosen value (here 95%)

At this value, find limit on signal strength and interpret as σ/σ_{SM}

Again, plot actual data and expected (from MC) limit, with variations.

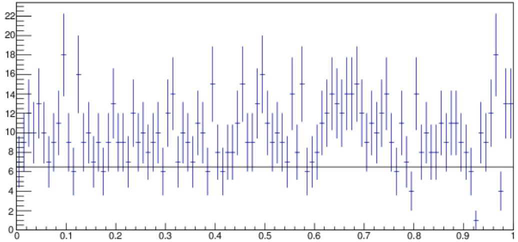
If there is no signal, 68% of experiments should give results in the green band, 95% in the yellow band



Calculations using 'Azimov dataset'
Essential reading: "Asymptotic formulae for likelihood-based tests of new physics" G Cowan, K Cramer, E Gross, O Vitells, arXiv:1007.1727v3, Eur.Phys.J.C71:1554,2011

The Look Elsewhere Effect

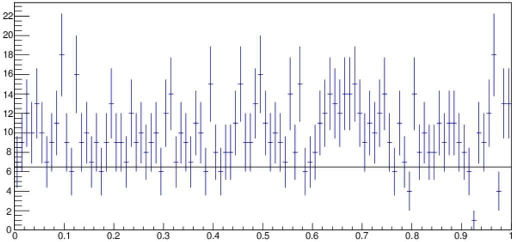
5 sigma needed to 'claim discovery'. 3 sigma is just 'evidence of'
Seems excessive ... p-value 2.9×10^{-7}. Due to (1) logic and (2) history



How many peaks can you see in this plot?

The Look Elsewhere Effect

5 sigma needed to 'claim discovery'. 3 sigma is just 'evidence of'
Seems excessive ... p-value 2.9×10^{-7}. Due to (1) logic and (2) history



How many peaks can you see in this plot? Actually there are NONE
With 100 bins, 1% probabilities are liable to happen

Local and global significance

This can be compensated for to some extent. What can't be calculated is the number of plots drawn by 1000+ collaborators hoping for a discovery.

Blind Analysis

"It was easy - I just got a block of marble and chipped away anything that didn't look like David."

Michaelangelo Buonarotti(attrib.)

Maybe good way of creating sculpture - but very bad way of doing physics

To resist temptation, devise cuts *before* looking at the data. Use Monte Carlo simulations, and/or data in 'sidebands'. Only when cuts are optimised do you 'open the box'.

Some experiments have formal apparatus for doing this.

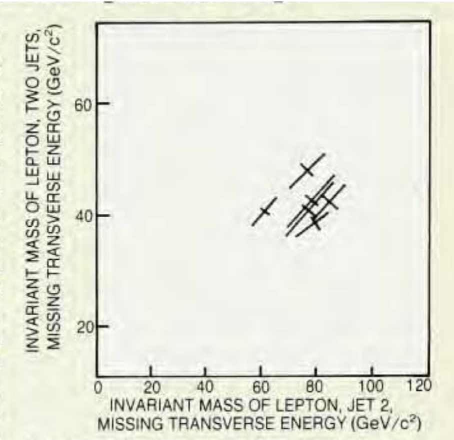
Why are we so cautious? And why do we insist on 5 sigma?

The top quark 'discovery' at UA1

$W \to t\overline{b}$ and $t \to b\ell^{\pm}\nu$

2 b jets, charged lepton, missing energy

Find 6 events. Plot total mass against $b\ell^{\pm}\nu$ mass (ν from missing energy/momentum)
W mass in right place
t mass around 40 GeV

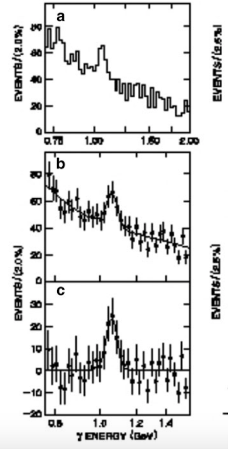


Turned out to be background - and very creative selection cuts

The $\zeta(8.3)$

"Discovered" in 1984 by the Crystal Ball experiment at DESY.

e^+e^- storage ring (DORIS) with energy 9.46 GeV, the mass of the Υ meson (which is a $b\bar{b}$ bound state)



Measure energy of photons

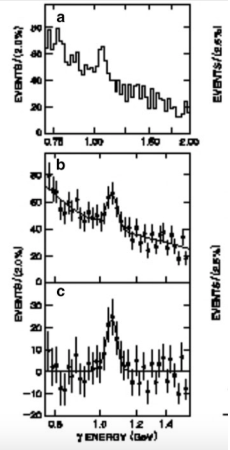
Single energy peak seen!!

Signals $e^+e^- \to \Upsilon \to \zeta\gamma$
4.2 sigma effect
Plots show (a) raw data , (b) fit, and (c) background-subtracted fit

The $\zeta(8.3)$

"Discovered" in 1984 by the Crystal Ball experiment at DESY.

e^+e^- storage ring (DORIS) with energy 9.46 GeV, the mass of the Υ meson (which is a $b\bar{b}$ bound state)



Measure energy of photons

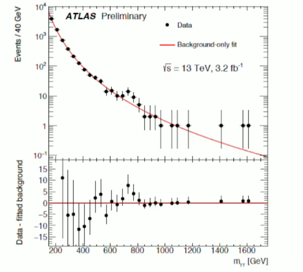
Single energy peak seen!!

Signals $e^+e^- \to \Upsilon \to \zeta\gamma$
4.2 sigma effect
Plots show (a) raw data , (b) fit, and (c) background-subtracted fit
When more data was taken (in 1985) the peak went away.

The $F(750)$

"Discovered" in 2015 by the ATLAS and CMS experiments at the LHC.

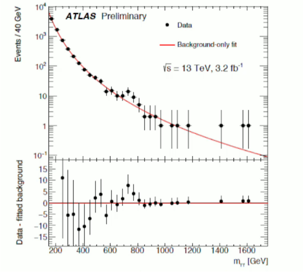
Invariant mass of pairs of high energy photons from proton proton collisions (Hence the name 'digamma')



3.6 sigma in ATLAS, 2.6 sigma in CMS

The $F(750)$

"Discovered" in 2015 by the ATLAS and CMS experiments at the LHC.



Invariant mass of pairs of high energy photons from proton proton collisions (Hence the name 'digamma')

3.6 sigma in ATLAS, 2.6 sigma in CMS

When more data was taken (in 2016) the peak went away

We need 5 sigma to keep ourselves honest.

Further Reading

Books

- R.B., Statistics: A guide to the use of statistical methods in the physical sciences
- Glen Cowan, Statistical Data Analysis
- Louis Lyons, Statistics for Nuclear and Particle Physicists,
- Olaf Behnke et al, Data Analysis in High Energy Physics
- Ilya Narsky and Frank Porter, Statistical Analysis techniques in Particle Physics
- Gerhard Bohm and Günter Zech, Introduction to Statistics and Data Analysis for Physicists
- Fred James, Statistical Methods in Experimental Physics

Papers

- R.B., in CERN yellow report: Proceedings of the 2018 Asia– Europe–Pacific School of High-Energy Physics, Quy Nhon, Vietnam (2020)
- The PDG review of Particle Physics, sections 39 and 40
- PHYSTAT conference proceedings

Conclusions

- Statistics is a tool for doing physics
- Tools should be well looked after
- They must be used carefully and skilfully
- You become familiar with them through using them
- Be honest, be careful – but not too careful
- You have got tremendous opportunities
- Good luck!

Backup

Method 3: Feldman-Cousins 4: Discussion

There are two arguments raised against the method
It deprives the physicist of the choice of whether to publish an upper limit or a range. Could be embarrassing if you look for something weird and are 'forced' to publish a non-zero result. *But isn't this the point?*

If two experiments with different N_B get the same small N_D, the one with the higher N_B will quote a smaller limit on N_S. The worse experiment gets the better result!
But for an event with large background to get a small number of events is much less likely.

Extension: From numbers to masses

Limits on Numbers-of-events/signal strength may translate to limits on Branching Ratios

$$BR = \frac{N_s}{N_{total}}$$

or limits on cross sections

$$\sigma = \frac{N_s}{\int \mathcal{L} dt}$$

These may translate to limits on other parameters, depending on the theory

In some cases (e.g. M_H) these parameters also affect detection efficiency, and may require changing strategy (hence different backgrounds)
Need to repeat analysis for all (of many) M_H values