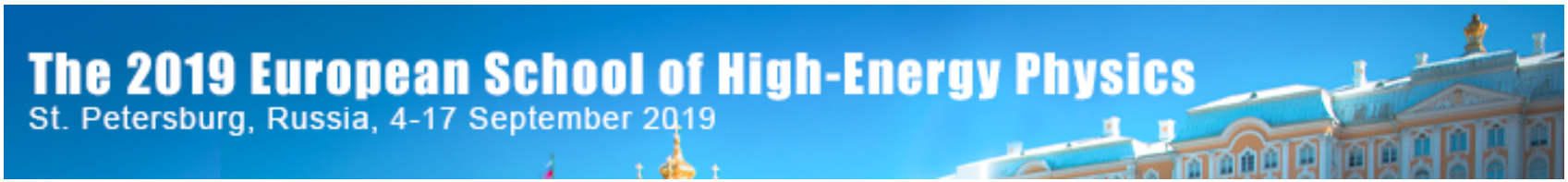
Practical Statistics for Particle Physics
Lecture 3: Setting limits, claiming discoveries

Roger Barlow
The University of Huddersfield

CERN European School on High Energy Physics
St Petersburg

16^{th} September 2019



Lecture 3: Setting Limits and making discoveries

1. Goodness of Fit
   - $p-$values
   - The χ^2 distribution
   - Wilks' Theorerm
   - Toy Monte Carlos and Likelihood for Goodness of Fit
2. Upper Limits
   - Frequentist Confidence
   - Confidence Belts
   - Coverage
   - Bayesian Intervals
   - Feldman-Cousins
   - CL_S
3. Making a Discovery
   - Sigma language
   - The Look Elsewhere Effect
   - Blind Analysis
4. Conclusions

Goodness of fit

An Example of Hypothesis Testing

You have the 'best' fit model - but is it any good?

Fit model of data

Construct some measure of agreement t between them.

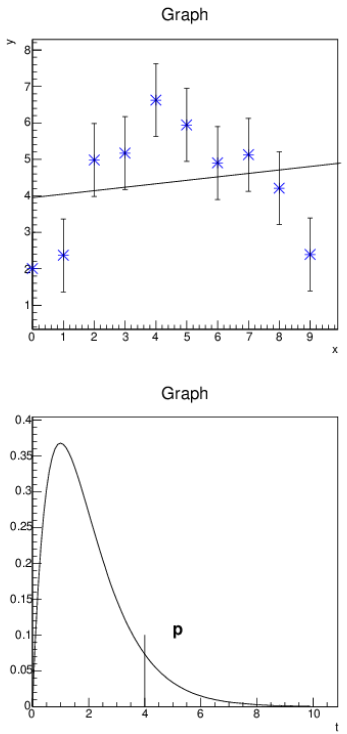
Convention: $t \geq 0$, $t = 0$ is perfect agreement.

Worse agreement $\rightarrow$ larger t

Null hypothesis H_0: The model produced this data.

Construct $p-$value: probability under H_0 of getting a t this bad, or worse.

Usually known algebra - can use simulation ('Toy Monte Carlo')



Is $p-$value the same as α? Sort of. Both are $\int_{bad\ \ region} P(t)\,dt$. But α is a property of a test, p of a particular dataset.

χ^2 Distribution

far and away the most popular measure of (dis)agreement

Total squared scaled differences

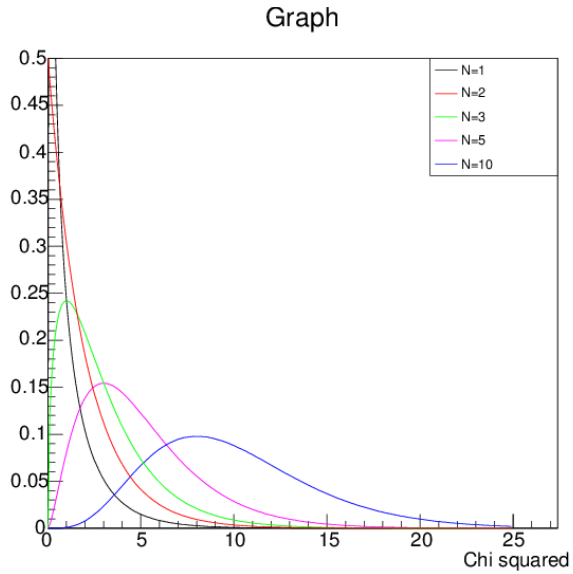
$\chi^2 = \sum_1^N \left(\frac{y_i - f(x_i)}{\sigma_i}\right)^2$

Obviously $\langle \chi^2 \rangle \approx N$.

Turns out to be exact.

$P(\chi^2; N) = \frac{1}{2^{N/2}\Gamma(N/2)} \chi^{N-2} e^{-\chi^2/2}$

To find $p-$value: in ROOT `TMath::Prob(chisquared,ndf)`, in R `1-pchisq(chisquared,ndf)`



Examples

If $N = 10, \chi^2 = 15$ then $p = 0.13$. Probably OK

If $N = 10, \chi^2 = 20$ then $p = 0.03$. Probably not OK

Useful fact

Least-Squares-Fitting the data clearly reduces χ^2. This also follows a χ^2 distribution for $N = N_{data} - N_{parameters}$ 'Degrees of freedom'

χ^2 fitting - comparison

If χ^2 is suspiciously big there are 4 possible reasons

1. Your model is wrong
2. Your data are wrong
3. Your errors are too small
4. You are unlucky

If χ^2 is suspiciously small there are 2 possible reasons

1. Your errors are too big
2. You are lucky

Likelihood and Wilks' Theorem

The Likelihood on its own tells you nothing
(even if you include the constant factors normally omitted in maximisation)
Wilks' Theorem says: Given two nested models, for large N the improvement in $\ln L$ is distributed like χ^2 in $-2\Delta \ln L$, with NDF the number of extra parameters

Example: Model 1 is straight line, Model 2 is quadratic, $NDF = 1$
Run Model 1. Run Model 2. Likelihood increases as more parameters available. If $2\times$ this increase is significantly more than 1 that justifies using Model 2 rather than Model 1.
So works for comparisons, but not absolutely

Important exception

Does not apply if Model 2 contains a parameter which is meaningless under model 1. Model 1 is background, Model 2 is background + unknown Breit-Wigner. (Mass, width and normalisation)

Using Toy Monte-Carlo for Likelihood and goodness of fit

Obvious suggestion: Take the fitted model, run many simulations, plot the spread of fitted likelihoods and use to get $p-$value
This is wrong - J G Heinrich, CDF/MEMO/BOTTOM.CDFR/5630[1]
Test case: model simple exponential $P(t) = \frac{1}{\tau} e^{-t/\tau}$
Then whatever the original sample looks like you get
Log Likelihood $= \sum(-t_i/\tau - \ln\tau) = -N(\overline{t}/\tau + \ln\tau)$
ML gives $\hat{\tau} = \overline{t} = \frac{1}{N}\sum_i t_i$
and this max log likelihood is $\ln L(\hat{\tau}; x) = -N(1 + \ln\overline{t})$
Any distribution with the same $\overline{t}$ has the same likelihood, after fitting.

What you can do: Histogram the $p(x_i; \hat{a})$ values. This should be flat (almost- the fitting will distort it).
If not enough data - cumulative plot should be straight line. Use max deviation as test statistic. Apply K-S test or use toy Monte Carlo.

[1] Many thanks to Jonas Rademacker for pointing this out

Frequentist Confidence

What is the probability that it will rain tomorrow?
There is only one tomorrow.
It will either rain or not rain.
The probability $N_{rain}/N_{tomorrows}$ is either 0 or 1.
P_{rain} is "unscientific" [von Mises]

Frequentist Confidence

What is the probability that it will rain tomorrow?
There is only one tomorrow.
It will either rain or not rain.
The probability $N_{rain}/N_{tomorrows}$ is either 0 or 1.
P_{rain} is "unscientific" [von Mises]
This is unhelpful

Frequentist Confidence

What is the probability that it will rain tomorrow?
There is only one tomorrow.
It will either rain or not rain.
The probability $N_{rain}/N_{tomorrows}$ is either 0 or 1.
P_{rain} is "unscientific" [von Mises]
This is unhelpful

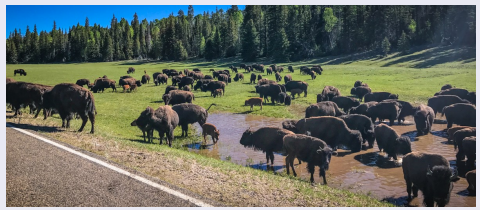
Suppose the forecast says it will rain.
Studies show this forecast is correct 90% of the time
The statement 'It will rain tomorrow' has a 90% probability of being true.
We can say 'It will rain tomorrow' with 90% confidence.
(Note how this depends on the ensemble used.)
We state X with confidence P if X is a member of an ensemble of statements of which at least P are true.
Note that 'at least'. 2 reasons

1. Higher confidences embrace lower ones. If X at 95% then X at 90%
2. Caters for composite hypotheses, with unknown parameters

Measurements

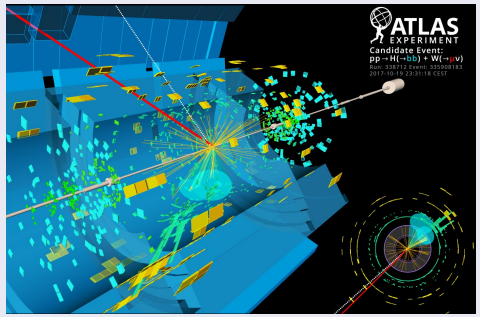
Example

Q:A herd of bison have a mean weight of 1250 kg and a standard deviation of 240 kg. What is the probability that particular bison has a mass $1010 < M_B < 1490 kg$*?*
A: 68%



Example

M_H *has been measured as* 125.18 ± 0.16 *GeV*
Q: What can we was about the probability that $125.02 < M_H < 125.34 GeV$
A: Nothing. There is only one M_H *- Future experiments will determine it to very high precision - and it either is in the range or not.*



What frequentists can say about the Higgs mass

or any other measurement

M_H has been measured with a technique that will give a value within 0.16 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

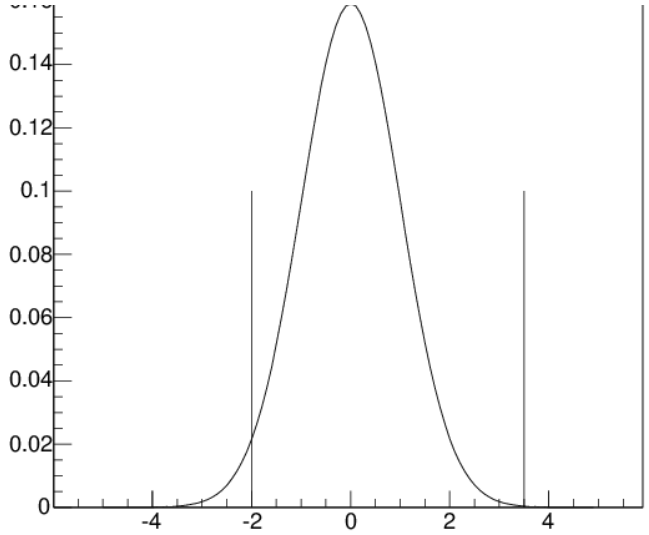
We say: $125.02 < M_H < 125.34 GeV$ with 68% confidence.
The statement is either true or false (time will tell) but belongs to a collection of statements of which (at least) 68% are true.

Confidence 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: $x_+ = \infty$, $\int_{-\infty}^{x_-} P(x)\,dx = 1 - CL$



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

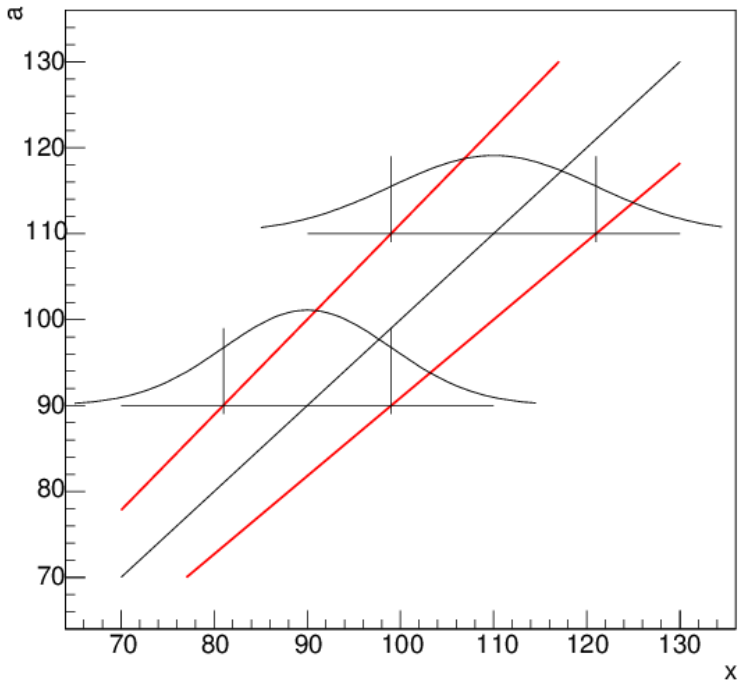
Confidence Belts

Measured $x = 100$ from Gaussian measurement $\sigma = 10$, say [90,110] is 68% central confidence region

Bit more complicated: $x = 100$ from Gaussian measurement $\sigma = 0.1x$ (10% measurement)

90 gives 90 ± 9 but 110 gives 110 ± 11. 90 and 110 not equidistant.

Confidence Belts are constructed horizontally and read 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
4. The result (x, a) lies inside the belt, with 68% confidence.
5. Read off a_+ and a_-: 111.1, 90.9

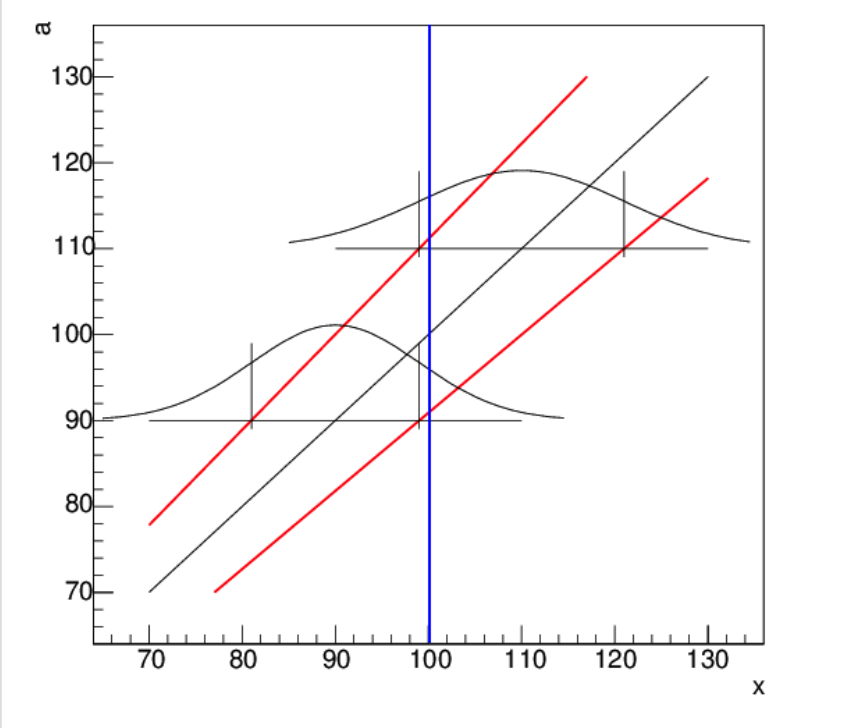
Confidence Belts

Measured $x = 100$ from Gaussian measurement $\sigma = 10$, say [90,110] is 68% central confidence region

Bit more complicated: $x = 100$ from Gaussian measurement $\sigma = 0.1x$ (10% measurement)

90 gives 90 ± 9 but 110 gives 110 ± 11. 90 and 110 not equidistant.

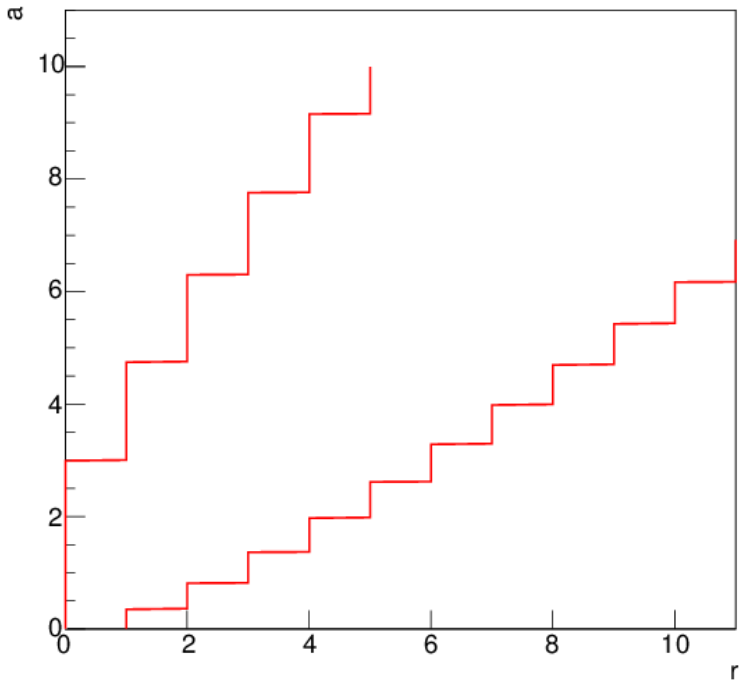
Confidence Belts are constructed horizontally and read vertically



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
4. The result (x, a) lies inside the belt, with 68% confidence.
5. Read off a_+ and a_-: 111.1, 90.9

Confidence Belts for the Poisson Distribution

Almost the same idea



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

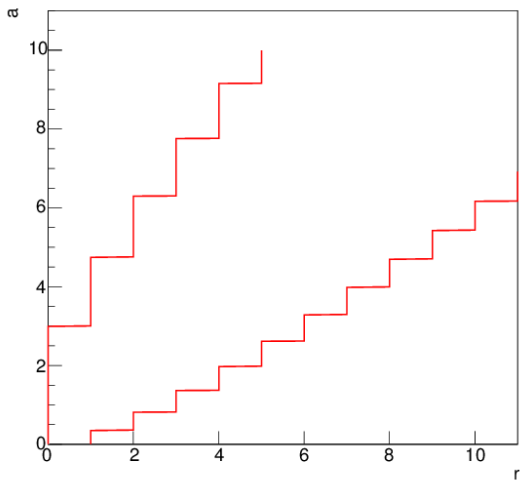
$\sum_{r=0}^{r_{hi}} e^{-a}\frac{a^r}{r!} \geq 0.95$

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

Coverage

The probability, given a, that the statement '$a_{lo} \leq a \leq a_{hi}$' will be true
May exceed the quoted confidence level ('overcover') but should never be less ('undercover')

Example: suppose $a = 3.5$ and we want a 90% central limit
There is a probability $e^{-3.5}$ =3% of getting 0 events, leading to $a_{hi} = 3.0$, which is wrong
There is a probability $3.5e^{-3.5}$ =11% of getting 1 event, leading to $a_{hi} = 4.7$, which is right
...
There is a probability $3.5^7 e^{-3.5}/7!$ =4% of getting 7 events, leading to $a_{lo} = 3.3$, which is right
There is a probability $3.5^8 e^{-3.5}/8!$ =2% of getting 8 events, leading to $a_{lo} = 4.0$, which is wrong
Total 'right' probability 94%. - 4% overcoverage



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 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 $125.02 < M_H < 125.34 GeV$ is 68%'

Downside is that another Bayesian can say 'It will probably not rain tomorrow' and 'The probability that $125.02 < M_H < 125.34 GeV$ is 86%' with equal validity.

Bayesian has posterior (or prior) 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 they mean something subtley different.

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 also say : 'I see zero events - the probability is 95% the true value 3.0 or less.'

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 $N_S = \frac{N_D - N_B}{\eta}$
(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 $\sum_0^3 e^{-7.75}7.75^r/r! = 0.05$

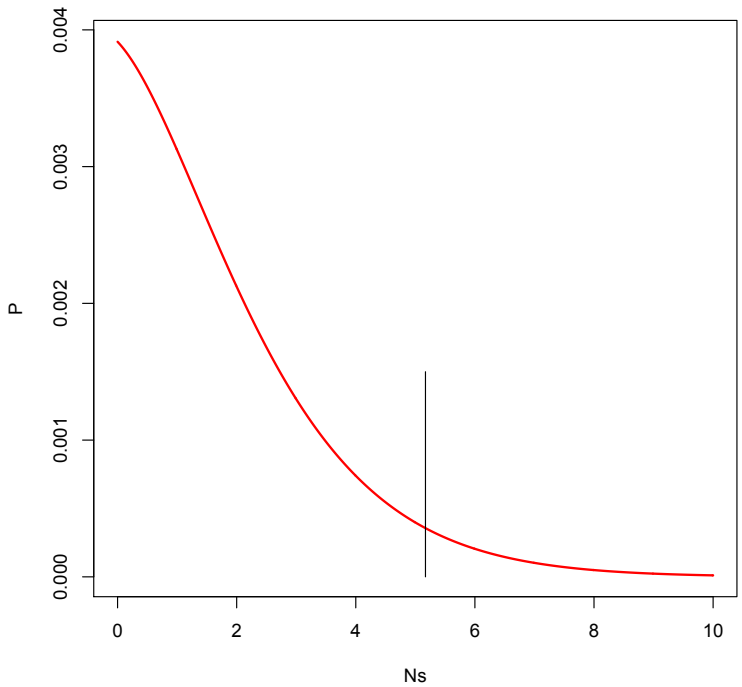
95% upper limit on $N_S = 7.75 - 3.40 = 4.35$

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

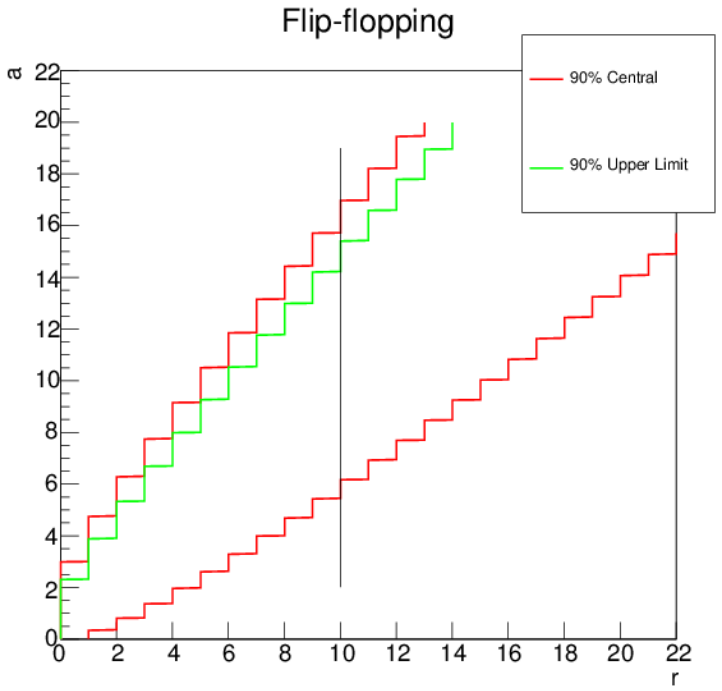
Assign a uniform prior to N_S, for $N_S > 0$, zero for $N_S < 0$.
The posterior is then just the likelihood,
$P(N_S|N_D, N_B) = e^{-(N_S+N_B)} \frac{(N_S+N_B)^{N_D}}{N_D!}$
Required Limit from integrating $\int_0^{N_{hi}} P(N_S)\, dN_S = 0.95$

$P(N_S) \propto e^{-(N_s+3.40)} \frac{(N_s+3.4)^3}{3!}$
Limit is 5.17



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

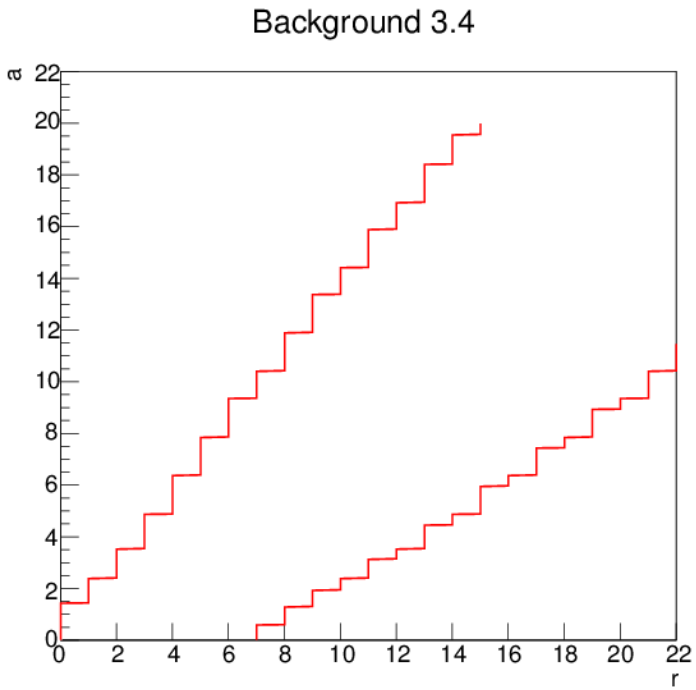
Plot $r \equiv N_D$ horizontally as before, but N_S vertically. So different $N_B \rightarrow$ different plot. Probability values $P(r; N_s) = e^{-(N_s+N_B)} \frac{(N_S+N_B)^r}{r!}$

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

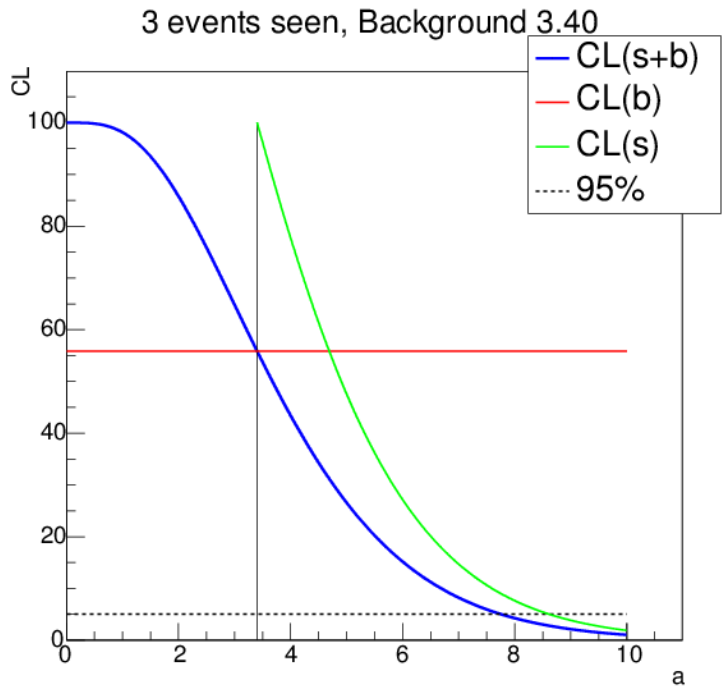
Have to re-compute confidence belt specifically for each background number. Not a problem.

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.

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

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

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

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)	5.17
Feldman-Cousins	4.86
CL_s	5.21

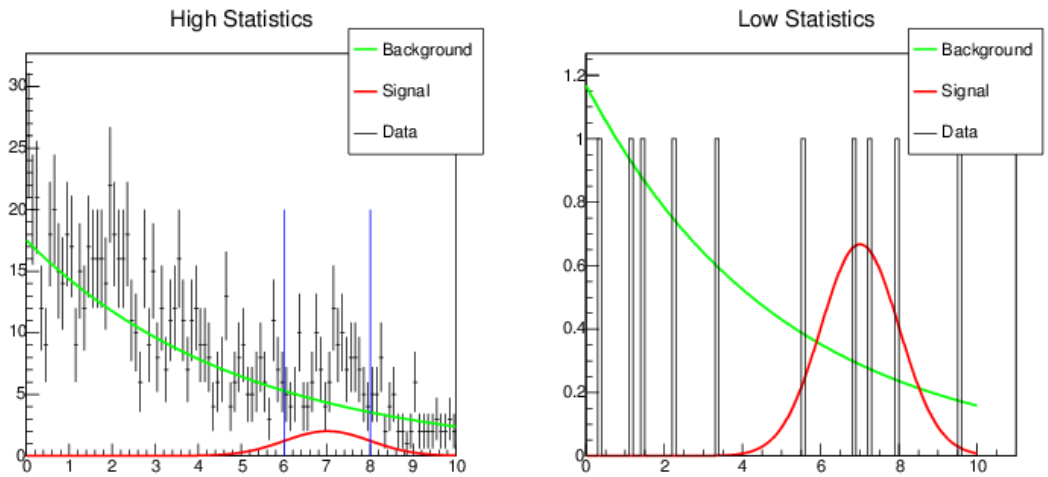
Take your pick!

All are correct. (Well, not wrong.)

Golden Rule

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

Extension: not just numbers



Simple counting not (usually) exploiting full information

Better: Likelihood

$lnL_{s+b} = \sum_i \ln N_s S(x_i) + N_b B(x_i)$ $\quad$ $lnL_b = \sum_i \ln N_b B(x_i)$

Look at L_{s+b}/L_b, or $-2\ln(L_{s+b}/L_b)$

Get confidence quantities from simulations/data

Extension: From numbers to masses

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

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

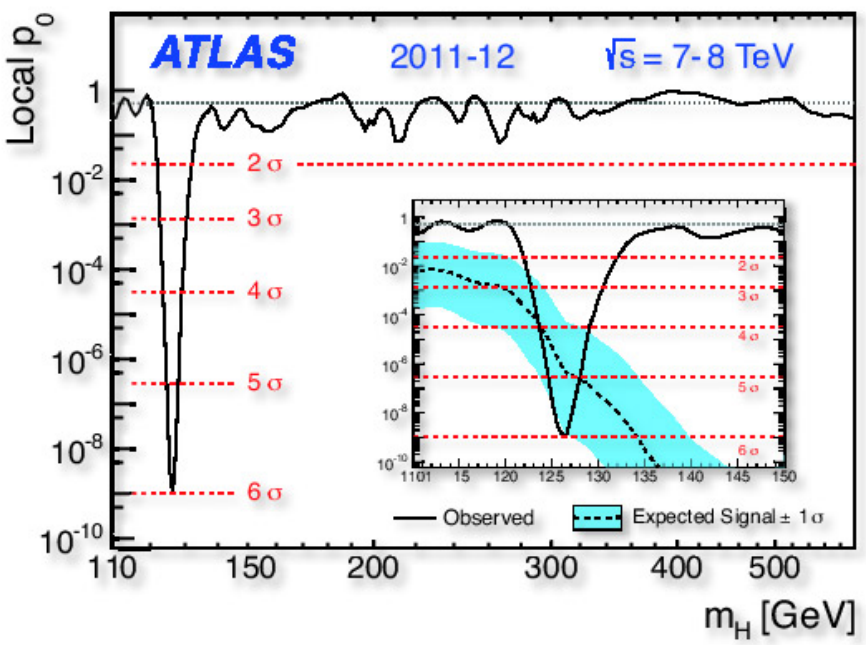
or limits on cross sections

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

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

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'

Brazilian Band plots

Green-and-yellow 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

f

Brazilian Band plots

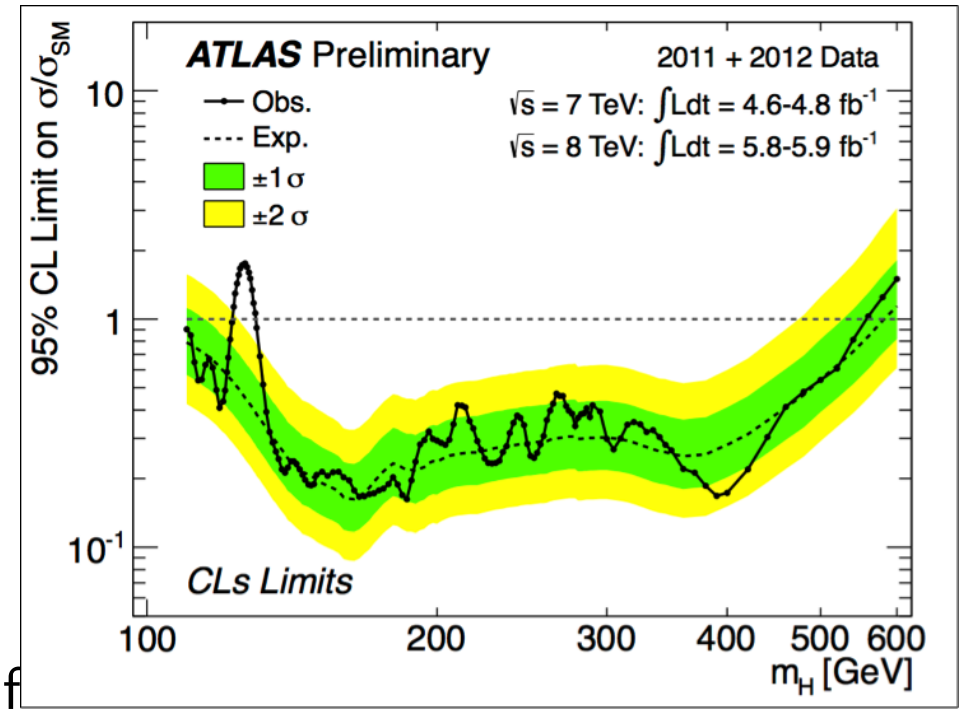
Green-and-yellow plots

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



Claiming a discovery

Remember Hypothesis testing?

To claim a discovery, show that your data can't be explained without it
Quantified by $p-$value. Probability of getting a result this extreme (or worse) under the null hypothesis/Standard Model.
Not 'The probability that the Standard Model is correct'

Sigma language

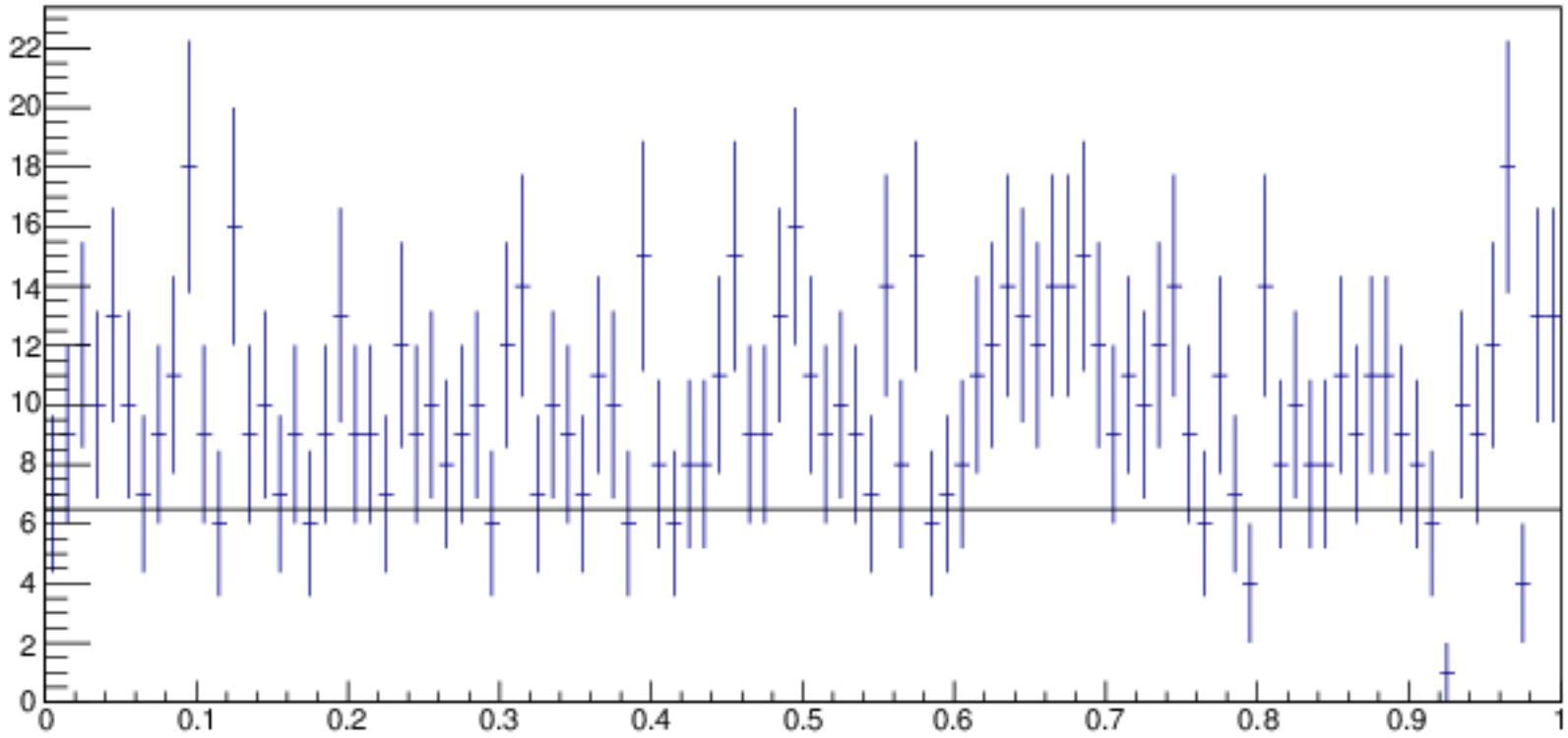
Often translated into Gaussian-like language: the probability of a result more than 3σ from the mean is 0.27%... a $p-$value of 0.0027 is a '3 σ effect' (or 0.0013 depending on 1-tailed or 2-tailed. Both are used.)

3 sigma 0.0013 'Evidence for' 5 sigma 0.0000003 'discovery of'

Some journals (Psychology) refuse to publish papers giving $p-$values
Why? Do lots of studies. Some will have low $p-$values (5% below 0.05 etc). Publish those and bin the rest...

The "Look Elsewhere" effect

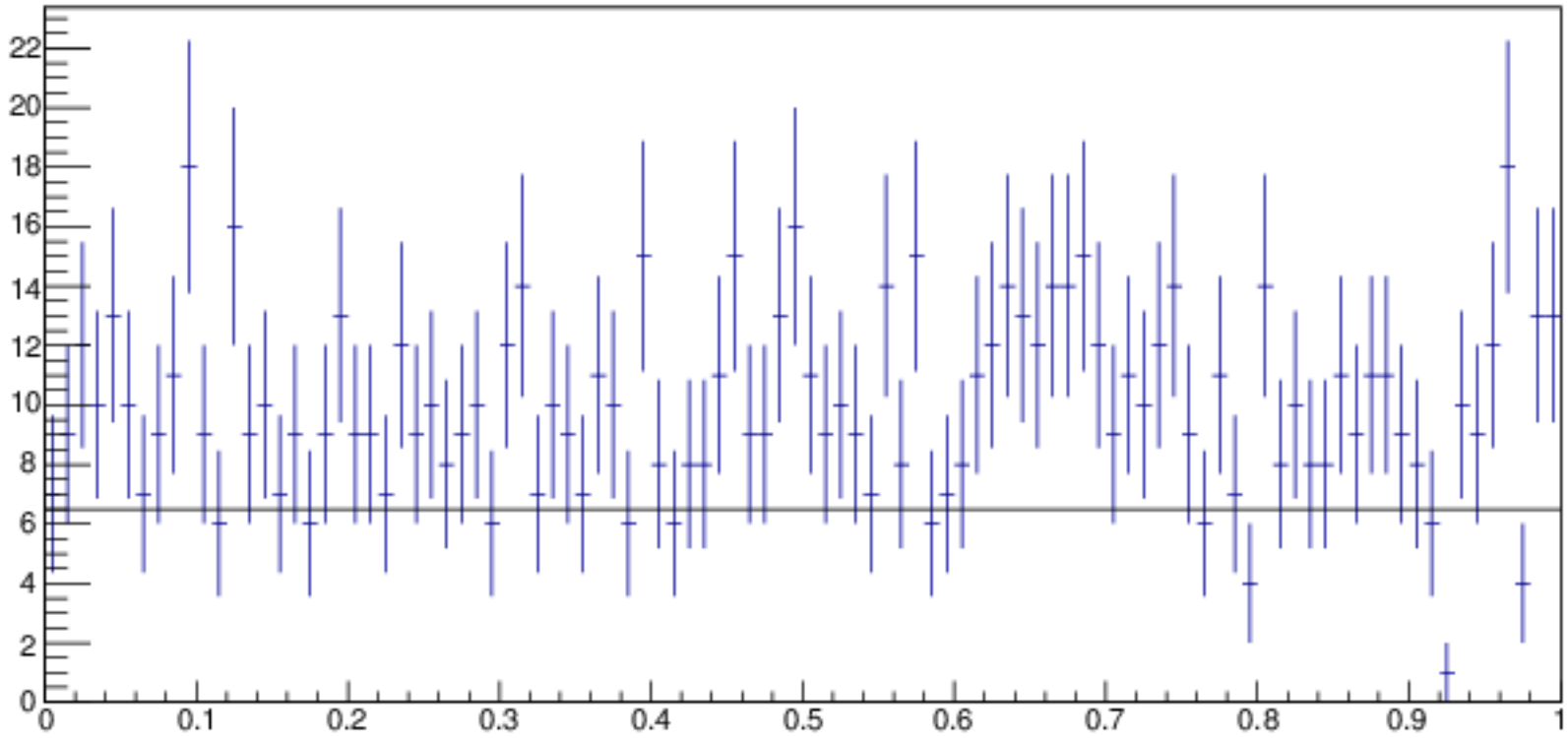
Why 5 σ. Isn't that excessive?



How many peaks are there in this plot?

The "Look Elsewhere" effect

Why 5 σ. Isn't that excessive?



How many peaks are there in this plot?

None

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'.

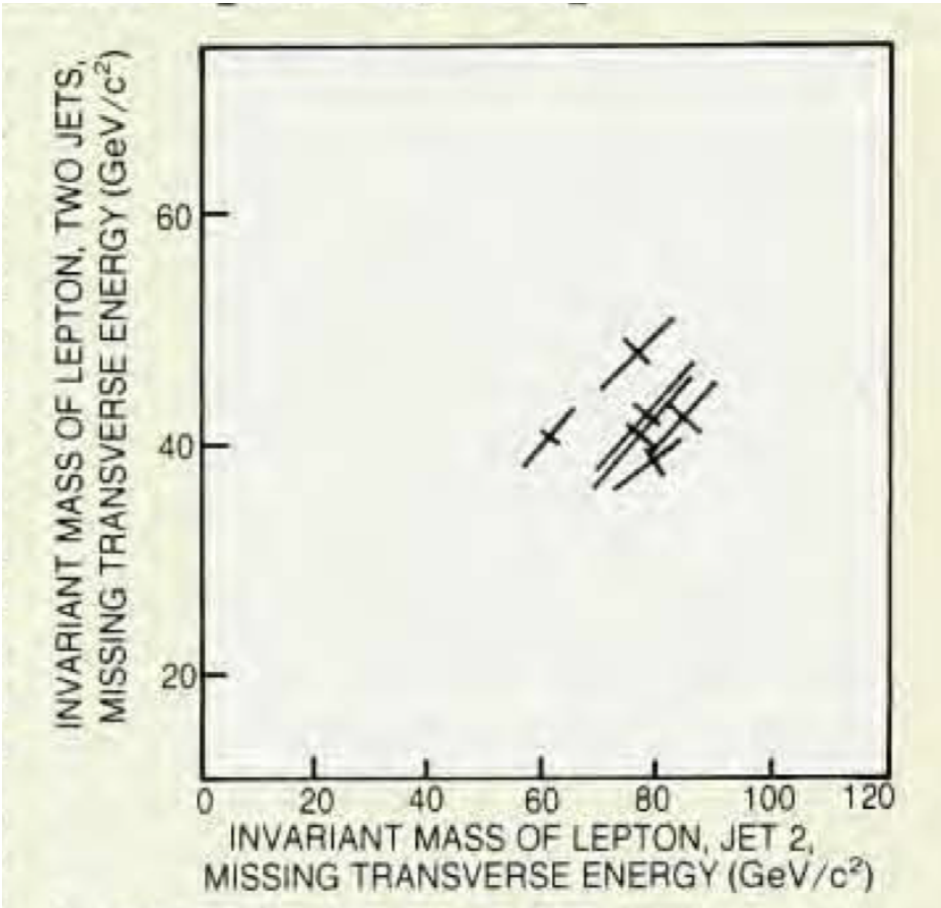
Some experiments have formal apparatus for doing this.

The top quark 'discovery' at UA1

$W \to t\bar{b}$ 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 $\zeta(8.3)$

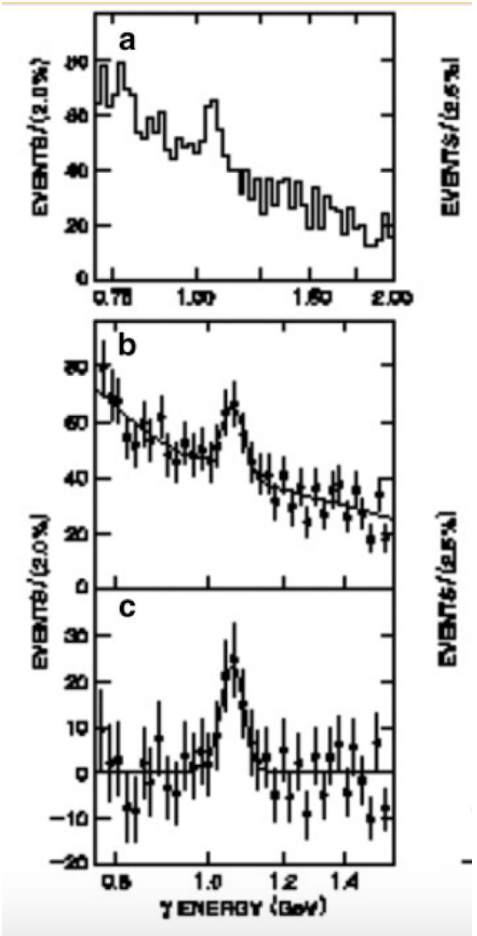
"Discovered" in 1984 by the Crystal Ball experiment at DESY.

e^+e^- storage ring (DORIS) with energy 9.46 GeV, the mass of the Υ meson (which is a $b\overline{b}$ 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 $\zeta(8.3)$

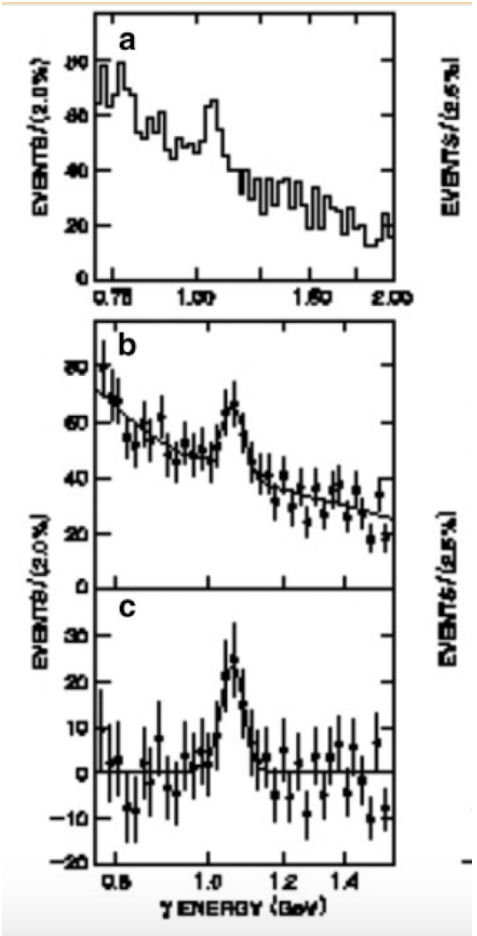
"Discovered" in 1984 by the Crystal Ball experiment at DESY.

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



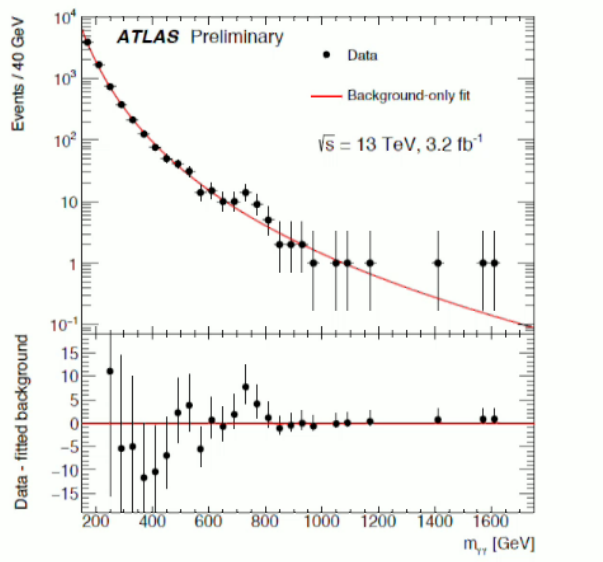
When more data was taken (in 1985) the peak went away.

The $F(750)$

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

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

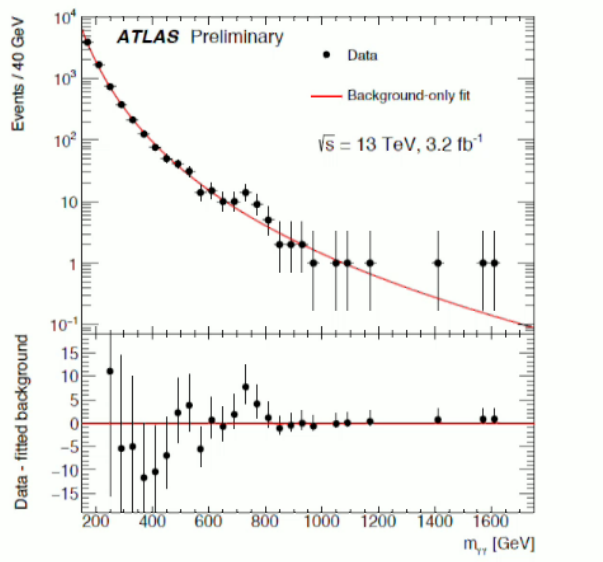
3.6 sigma in ATLAS, 2.6 sigma in CMS



The $F(750)$

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

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

Conclusions

Statistics is a tool for doing physics.

A good physicist understands their tools.

Read books and conference proceedings, go to seminars, talk to people, experiment with the data, and understand what you are doing.

And you will succeed.

Have a great time!