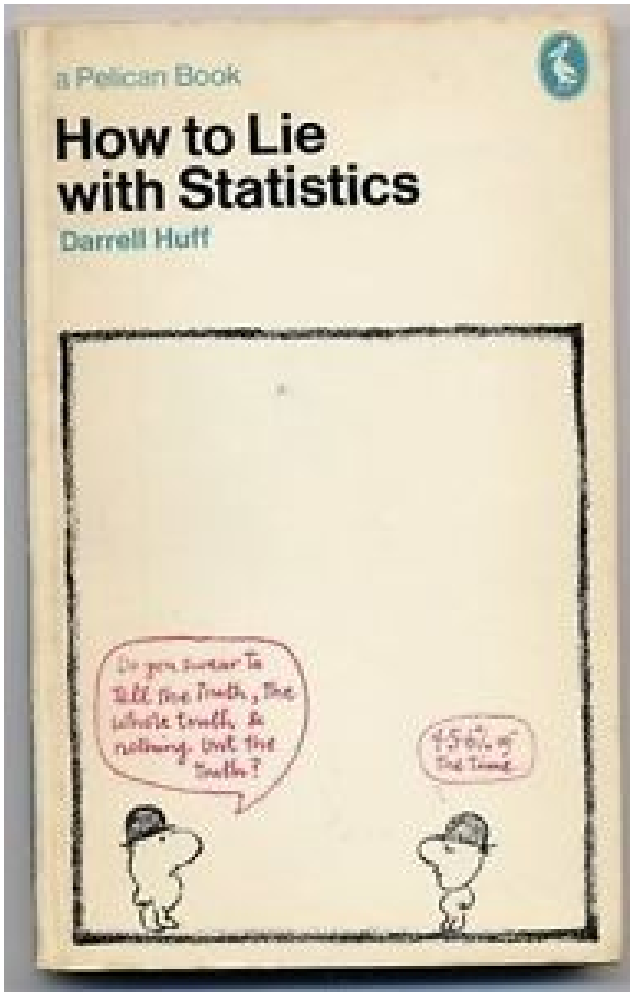
How to Lie with Statistics for physicists and everyone

Roger Barlow

ISPAD-2019

13th March 2019

Essential reading

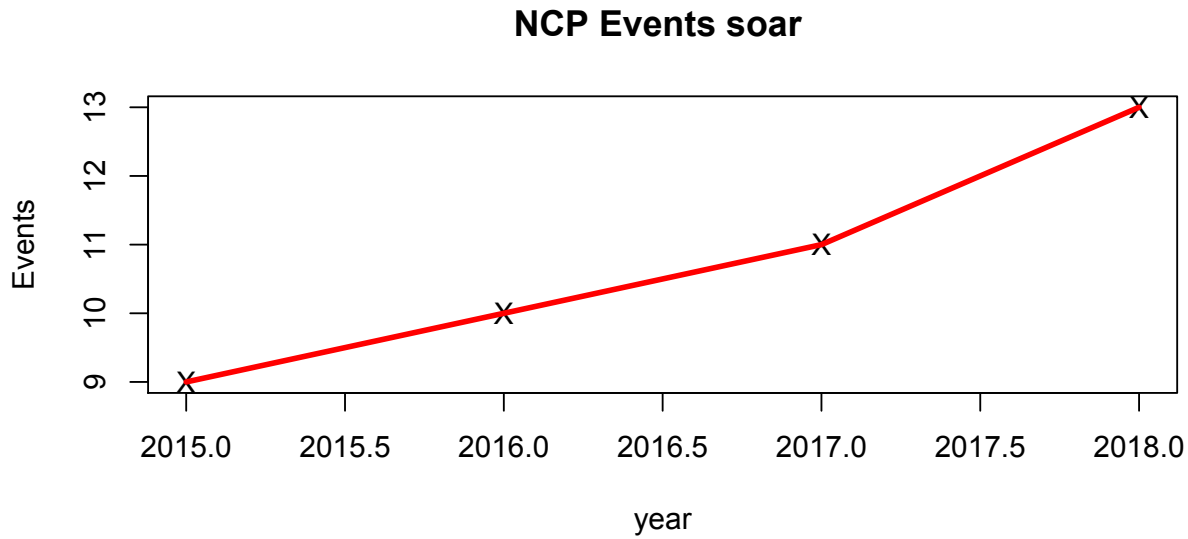


Explains the techniques used by advertisers and politicians

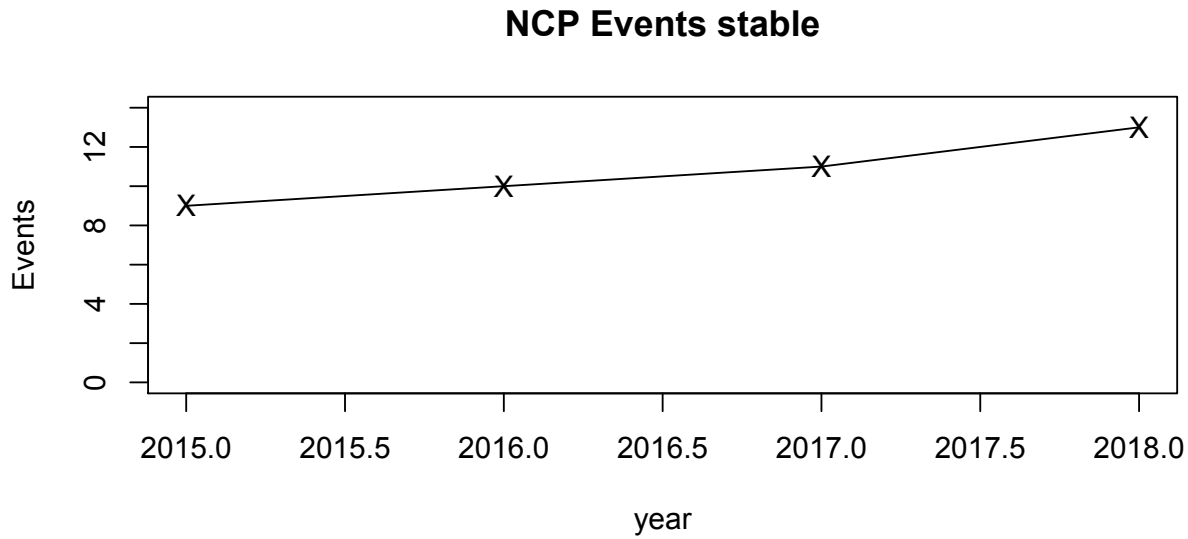
- The suppressed zero graph
- The misleading average
- Selective use of data
- The selective survey
- Respondent bias

Graphs - the power of a suppressed zero

NCP numbers of events show massive increase



NCP numbers of events show modest increase



The irrelevant questionnaire

The university president asks for evidence that there is no discrimination against women in staff appointments in the physics department.

Here are the staff of the physics department

What do you do??

Send round a questionnaire asking "Do you believe that we are treating women fairly in appointments and promotions? "

Tell the university president that your staff overwhelmingly believe that women are treated fairly (by 28 votes to 1)

But surely physicists don't do this sort of thing?

Statistics for physics undergraduates

Statistics 101: possibly the worst class at university

Students want to learn about	Students don't want to learn about
Quarks	Errors
The Higgs Boson	Limits
Graphene	Hypothesis testing
other cool new materials	Mean versus Mode versus Median
Quantum Computing	Student's t distribution
The Big Bang	Analysis of Variance

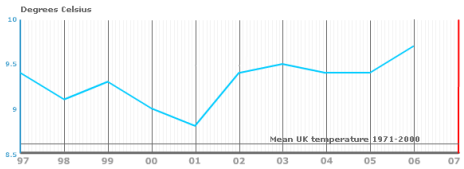
Einstein said "$E = mc^2$" not "$E = mc^{2.01\pm0.02}$"

But after graduation

As PhD students or working in Industry

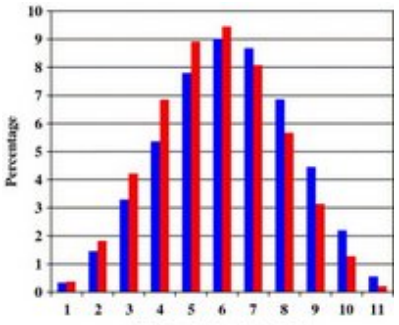
Is there a trend?

This shows the average UK temperature during the years 1997-2006. Is there a rising trend?



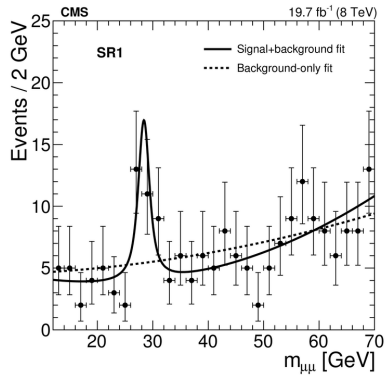
Are these two sets of data compatible?

These are the distributions of SAT scores from male and female candidates. (Details suppressed).



Have we found a new particle?

This shows the effective mass of muon pairs observed by the CMS experiment. Is there a peak at 28 GeV?



Statistics and proof

People say "Statistics can prove anything"
Not true!
Actually statistics can't prove anything

Coin suspected of being double-headed

You toss it once. It comes up heads.
Does that prove it's phony? Of course not.

You toss it 4 times more. Five heads.
Phony? Maybe. But could be chance ($\frac{1}{32}$)

10 times. 10 heads. Almost certainly - but not proven
Only by turning it over and looking at both sides can you *prove* it's got two heads

To be fair to Homer: people can come up with statistics to *support* anything

Hypothesis testing

An exercise in 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

That phony coin 10 tosses - all heads

If the coin is not phony then the chance of this happening is $\frac{1}{1024}$.
Which is small - so small we can (?) rule it out
So the coin is not honest, 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
'The acceptance rates from the two manufacturers were incompatible with being identical' $\rightarrow$ One was better than the other

$p-$values

$\frac{1}{1024}$ was our first example

The $p-$value is the probability of the data being this extreme (or more) under the null hypothesis.

- Patients have a 50% probability of recovering within 1 week. With a new treatment, 10/10 recover. Is the treatment effective?
  Almost certainly. $p-$value for H_0 is $\frac{1}{1024} < 0.1\%$
- Suppose 9/10 recover?
  $p-$value now $P_9 + P_{10} = \frac{11}{1024} \approx 1\%$. Still pretty convincing.
- Suppose we ask: Has the treatment any effect? It could make things worse. $p-$value now includes cases 0 and 1, so 2%. This is a 2-sided rather than a 1-sided result.

The Prosecutor's Fallacy

The $p-$value is the probability of the data being this extreme (or more) under the null hypothesis. This is the conditional probability $P(data|H_0)$
This is not the same as $P(H_0|data)$
... but 'the probability that such data could arise under the standard. model is only 1 in a million ' is inevitably paraphrased by journalists as 'the probability that the standard model is true is only 1 in a million'

"Bloodstains at the scene match the defendant - and only 1 in 1000 have this blood type. Ladies and gentlemen of the jury, statistics proves that the probability of his innocence is only 1 in 1000!"

To get it right use Bayes'Theorem: $P(H_0|data) = \frac{P(data|H_0)}{P(data)} P(H_0)$

For example...

Downloaded from `http://physics.org` on 9th Match 2019

What does the 5 sigma mean?

5 sigma is a measure of how confident scientists feel their results are. If experiments show results to a 5 sigma confidence level, that means if the results were due to chance and the experiment was repeated 3.5 million times then it would be expected to see the strength of conclusion in the result no more than once.

One analogy is trying to find an odd dice in a set 60, just by looking at a summary of all the rolls. Imagine you have one dice that had a 5 on every face in your set of 60. After each roll of the 60 dice, you are told how many 5's were rolled. On average you would expect ten 5's. But as the rolls of the dice are random you might see perhaps nine 5's the first time or twelve 5's the second time. As you continue to roll the dice, you can measure the spread of results around your expected result of ten 5's. One way to measure the spread is the standard deviation, or sigma. The more you roll, the smaller the standard deviation gets. After enough rolls, you might get to a point where the average number of 5's is 11 and your standard deviation is 0.2. You expected a result of 10 but your findings are higher by 5 sigmas. So you can be 99.9999% sure you have a dodgy dice in your set.

Read it and weep.....

The $p-$value and the significance α

In Hypothesis Testing one devises a test variable t, and calculates the critical values for accepting/rejecting H_0 with a certain *significance* α.

Gaussian test

For a one-sided test with $\alpha = .05$, $t_{crit} = \frac{x_{crit} - \mu}{\sigma} = 1.645$

χ^2 test

For 10 degrees of freedom, with $\alpha = 0.10$, $\chi^2_{crit} = 15.99$
If you fit a straight line through 12 points, and χ^2 comes out larger than 15.99, you should try fitting something more complicated

α and p are *similar*, given by same formula, but *different*: α is a property of the test but p is a property of the data

Toolbox

For calculations like these you need a computer and a friendly language - `python`, `Matlab`, `Mathematica`, `R`, `ROOT` and others. But not Excel!

$p-$values are dangerous

John Ioannidis (2005) 'Why Most Published Research Findings Are False'

Someone does a study (psychology, sociology, biochemistry...) looking for an effect
If the $p-$value is below 5% it is 'statistically significant'. They publish.
If the $p-$value is above 5% they ignore it and move on to the next.

If they do 100 studies then even if there is nothing there they'll find 5 (on average) $p-$values that are significant.
A busy research team can generate a stream of misleading publications.
Unlikely to be checked – second observations don't bring much prestige.

Repeating 100 different results in experimental psychology confirmed the original conclusions in only 38 per cent of cases

Some journals (not in physics, afaik) now ban papers with $p-$values.
Actually particle physics has, if anything, the opposite problem

$p-$value trouble and ML

Robots can have human failings

" People have applied machine learning to genomic data from clinical cohorts to find groups, or clusters, of patients with similar genomic profiles.

"But there are cases where discoveries aren't reproducible; the clusters discovered in one study are completely different than the clusters found in another, Why? Because most machine-learning techniques today always say, 'I found a group.' "

Prof. Genevera Allen, AAAS meeting, Feb. 2019

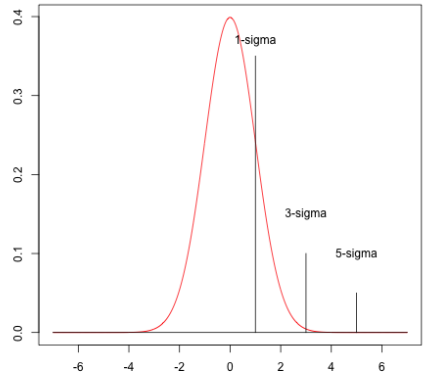
Remember!

Testing and Verifying are as important as Training. Split your dataset! Use the Bootstrap!!

Sigma language

$p-$values don't feel intuitive. 0.0001 and 0.00002 are both 'pretty small'

Codify using properties of Gaussian distribution.
84% lie below one standard deviation (1-sided) .
$p-$value 0.16. Call $p = 0.16$ '1 sigma'
0.023 is '2 sigma', 0.0027 is '3 sigma'
5 sigma 3.5×10^{-6}



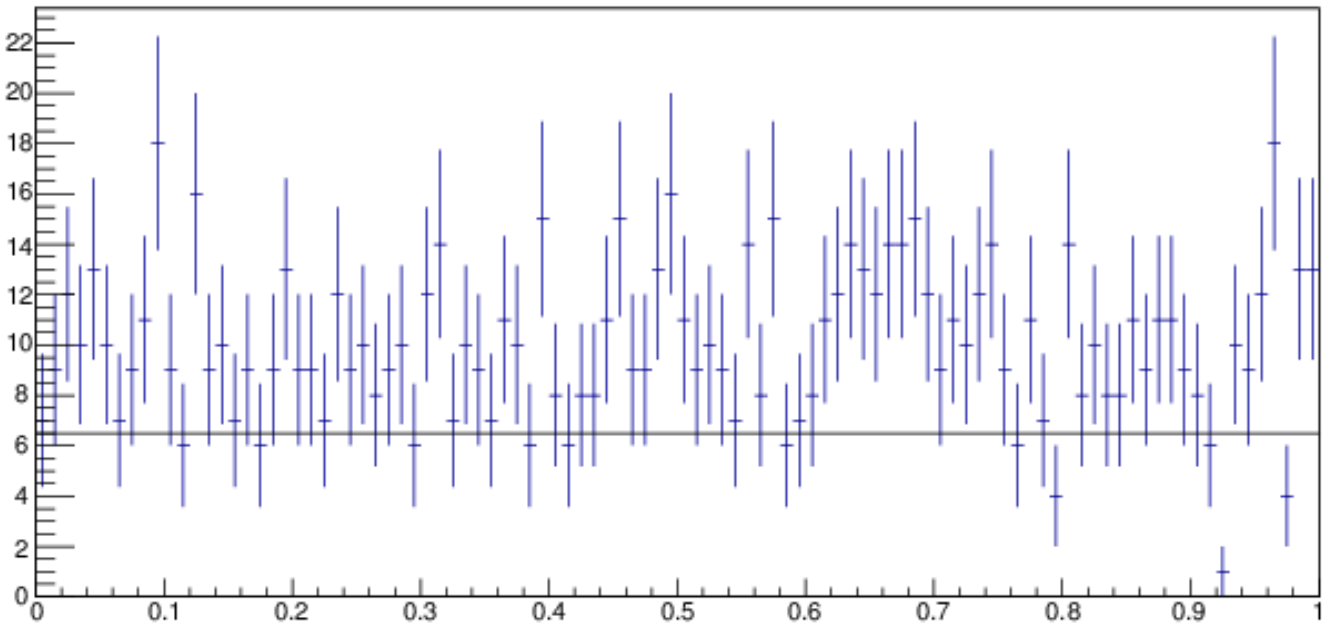
Obvious but maybe necessary point

There is (almost certainly) no actual Gaussian distribution involved. This is nothing but language games

In particle physics 5 sigma needed to 'claim discovery'
3 sigma is just 'evidence of'

That's well below 5%
Seems excessively strict...
But demanded by (1) logic and (2) history

The Look Elsewhere Effect



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

With a collaboration of 1000 people...

This can be compensated for to some extent. What can't be calculated is the number of plots that are drawn in the hope of finding something.

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'.

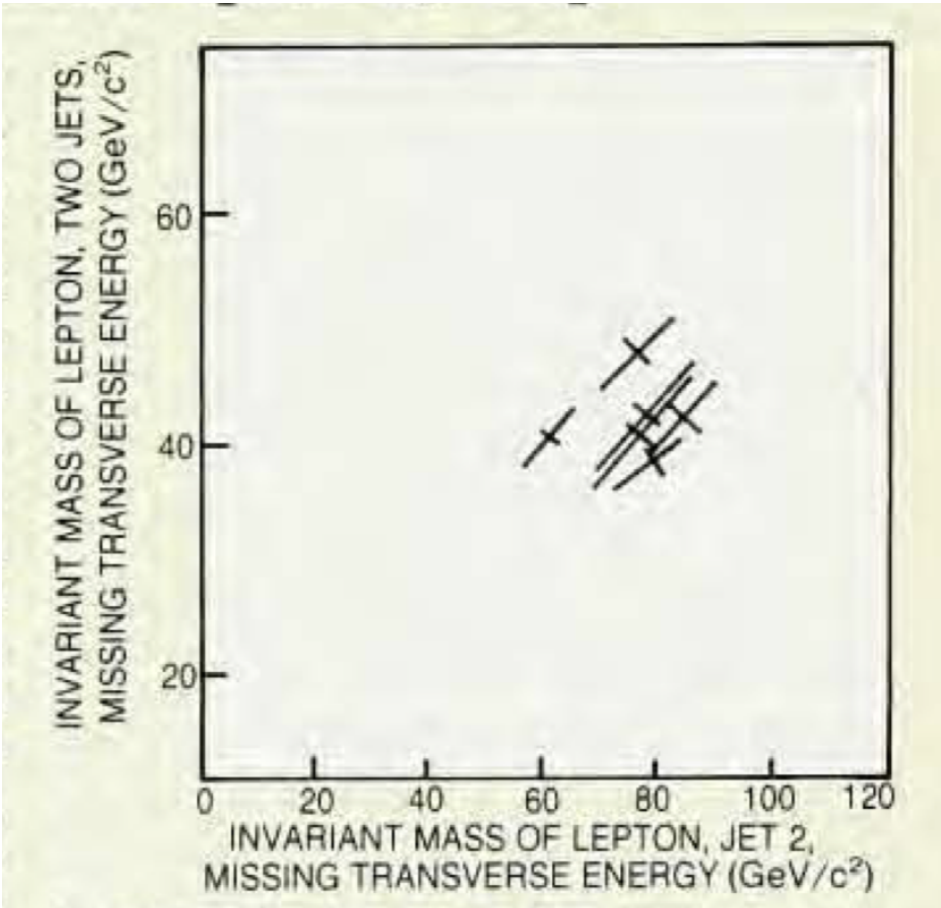
Some experiments have formal apparatus for doing this.

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 $\zeta(8.3)$

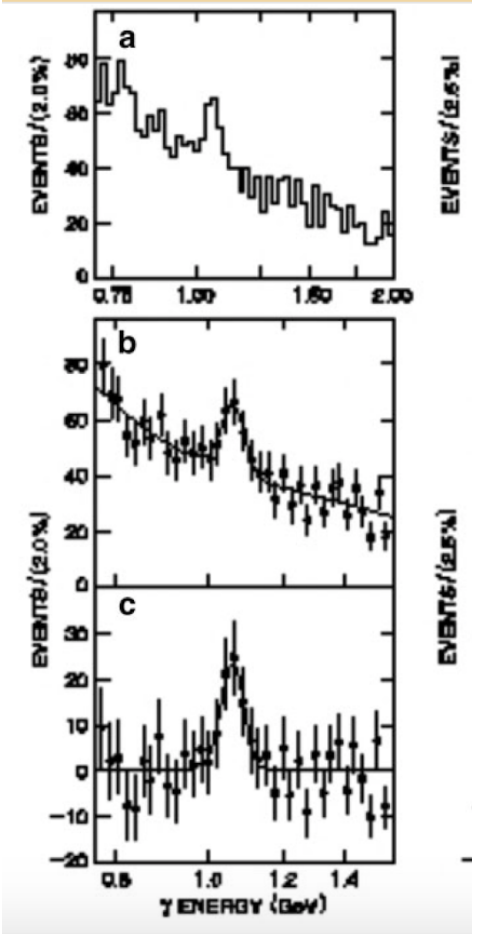
"Discovered" in 1984 by the Crystal Ball experiment at DESY.

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

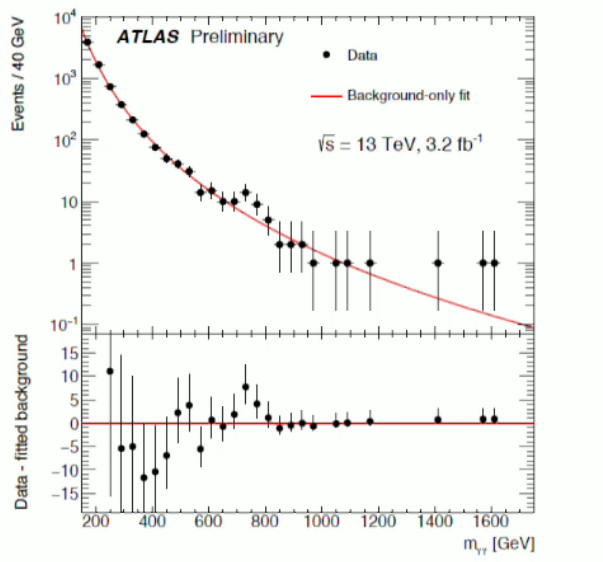


When more data was taken (in 1985) the peak went away.

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

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

Conclusions

- Statistics is a tool for surviving in the 21st century
- And a tool for doing physics
- Tools should be well looked after,
- They must be used carefully and skilfully
- $p-$values are essential part of the tool-kit
- Be fully aware of what they are
- Use them for illumination, not support
- Be honest, be careful – but not too careful