True measurements... false discoveries

The history of HEP has seen extraordinary claims ...an attempt to analyse why some were not true and what to do to avoid it
Extraordinary claims

• In the last years we have been confronted with extraordinary widely advertised discoveries
  – Higgs Boson
  – Superluminal neutrinos
  – Gravitational waves

• We will attempt to go through the criterias we are using to make ‘discovery’ claims, a little history on how they have developed and see the possible limitations of our approaches
Probability to produce a Higgs boson in the 2 proton collision is $10^{10}$ times smaller than to produce any other final state: choice of decay channel determines $S/B$ ratio and mass resolution.
Success story (II)

\[ pp \rightarrow H \rightarrow \gamma\gamma \]
..and everybody was talking of ‘5σ’
From ‘P-values’ to ‘5-Sigma’

• Statistical significance is a way to report the probability that an experiment obtains data at least as discrepant as those actually observed, under a given "null hypothesis“ $H_0$

• In physics $H_0$ usually describes the currently accepted and established theory

• Given some data $X$ and a suitable test statistic $T$ (a function of $X$), one starts with the $p$-value, i.e. the probability of obtaining a value of $T$ at least as extreme as the one observed, if $H_0$ is true.

**How to go from p-value to "sigma"?** This is done by finding $z$ such that the integral from $z$ to infinity of a unit Gaussian $N(0,1)$ equals $p$:

$$
\frac{1}{\sqrt{2\pi}} \int_z^\infty e^{-\frac{t^2}{2}} dt = p
$$

For example,, a **15.9%** probability is a one-standard-deviation effect; a **0.135%** probability is a three-standard-deviation effect; and a **0.0000285%** probability corresponds to five standard deviations, that is what has become known as the "**five sigma**" 'discovery threshold
Test statistic for H discovery

Likelihoods

\[ \mathcal{L}(\text{data} \mid \mu, \theta) = \text{Poisson} \left( \text{data} \mid \mu \cdot s(\theta) + b(\theta) \right) \cdot p(\tilde{\theta} \mid \theta). \]

And likelihood for background only is for \( \mu = 0 \) and signal + background is for \( \mu = 1 \)

- \( 2 \ln(\mathcal{L}_0 / \mathcal{L}_1) = -2 \left( \ln \mathcal{L}_1 - \ln \mathcal{L}_0 \right) \) then using Wick’s theorem (that for large stats the Log of the Likelihood ratio is distributed like a \( \chi^2 \) and then extract a tail probability \( P(\chi^2, N_{\text{dof}}) \) which can be converted in number of Sigmas
Some caveats

– The whole construction rests on a proper definition of the p-value. Any shortcoming of the properties of p (e.g. a tiny non-flatness of its PDF under the null hypothesis) totally invalidates the meaning of the derived $N\sigma$

  • In particular, using “sigma” units does in no way mean we are implying some kind of Gaussian approximation for our test statistic or in other parts of our problem. Care required here, as many could be led to confusion

– the conversion of p into # of Sigmas is fixed and independent of experimental detail. As such, using $N\sigma$ rather than p is just a shortcut to avoid handling numbers with many digits:

  we prefer to say “5σ” than “0.00000029”
Some history about major discoveries

• A rigorous approach with respect to ‘discovery’ was not always enforced
  – **The J/ψ discovery** (1974): no discussion of significance – the peaks were too big for even bothering discussing significance
  – The τ discovery: discussion about the excess of e-μ events were more about hadron backgrounds
  – The Upsilon discovery involved lots of statistical tests (mainly because of the ‘false’ evidence at 6 GeV – so called ‘Oops Leon’) even if the evidence exceed by far 5σ

Now that the signal (> 8σ) is no longer questionable from statistical objections, systematics must be considered.

1. Programming error, double counting, etc. will be studied by
• W boson discovery (January 1983): 6 events, no statistical analysis, but discussion about absence of background
• Z boson (May 1983) 4 events, also here discussion to show that backgrounds are negligible
Top: the first modern application of the 5σ criteria

- 1994 the CDF experiment publishes ‘Evidence’ based on a counting excess (2.7σ) in b-tagged single-lepton and di-lepton datasets accumulating in a mass peak which was over 3σ by itself

\[ M = 174 \pm 10^{+13}_{-12} \text{ GeV} \] (now it is 173±0.5)

- One year later CDF and D0 (with 3 times more data) presented counting excesses at the level of 5σ and claimed ‘Discovery’!

The birth of the 5 sigma criteria

“Are There Any Far-out Mesons or Baryons?”, A.H. Rosenfeld in Charles Baltay & Arthur H. Rosenfeld Meson Spectroscopy W.A. Benjamin Inc. 1968

<table>
<thead>
<tr>
<th>Outgoing prongs</th>
<th>Outgoing particles</th>
<th>Number of mass combinations</th>
<th>Events measured (thousands)</th>
<th>Mass combinations (millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>500</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>or 3</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>10</td>
<td>1200</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>or 5</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>56</td>
<td>70</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>or 7</td>
<td>119</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total U.S.</strong></td>
<td></td>
<td></td>
<td>~1,700</td>
<td>~28</td>
</tr>
<tr>
<td><strong>Assume 20% were remeasurements</strong></td>
<td></td>
<td></td>
<td>~1,400</td>
<td>~23</td>
</tr>
<tr>
<td><strong>×1.5 (?) to include other countries</strong></td>
<td></td>
<td></td>
<td>~2</td>
<td>~35</td>
</tr>
<tr>
<td>Divide by 2500 events/histogram; yields 15,000 histograms.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Arthur H. Rosenfeld (Univ. Berkeley)
A comparison with the literature in fact showed a correspondence of his estimate with the number of unconfirmed new particle claims.
First reason for 5σ: stat fluctuation

- Besides rendering pure statistical fluctuation unlikely, the 5σ criteria aims to protect from the fact that if we try hard enough we SHALL find a fluctuation.
- The number of trials required to reach $10^{-7}$ probabilities is of course very large... on the other hand modern experiments are performing a large number of searches... so we tend to correct our significances by estimating the Look Elsewhere Effect which accounts for the reduction of significance due to the trials we made to find an excess.
- The brute force way to estimate the LEE is to simulate a set of experiments under the null (background only) hypothesis and varying all the parameters within their precision and check for the likelihood to have a significant fluctuation: in order to match $10^{-7}$ one has to ‘simulate’ order of $10^7$ experiments for each parameter set ...
- When dealing with some of the searches at LHC this can be practically impossible (the Higgs search implied combination of dozens of different channels with hundreds of nuisance parameters).
- Stimulated by this consideration ‘asymptotic’ methods have been defined to evaluate the LEE. E. Gross and O. Vitells, “Trials factors for the Look-Elsewhere Effect in High-Energy Physics”, arxiv:1005.1891v3, Oct 7th 2010
Notes About the LEE Estimation
courtesy of Tommaso Dorigo

Even if we can usually compute the trials factor by brute force or estimate with asymptotic approximations, there is a degree of uncertainty in how to define it.

If I look at a mass histogram and I do not know where I try to fit a bump, I may consider:

1. the location parameter and its freedom to be anywhere in the spectrum
2. the width of the peak: is that really fixed a priori?
3. the fact that I may have tried different selections before settling on the one I actually end up presenting
4. the fact that I may be looking at several possible final states and mass distributions
5. Different people in the experiment can be doing similar things with different datasets; should I count that in?
6. There is ambiguity on the LEE depending who you are (grad student, experiment spokesperson, lab director...)

Also note that Rosenfeld considered the whole world’s database of bubble chamber images in deriving a trials factor.

The bottomline is that while we can always compute a local significance, it may not always be clear what the true global significance is.
The name of the game: systematic errors

• The other reason for the 5σ criteria is to protect against problems with the modelling of the systematics behind a given measurement

• The evaluation of systematic errors is a challenging field:
  – The models used to evaluate the null hypothesis could be flawed/incomplete
  – The subjective prejudices on the way to evaluate them can play a significant role
  – The underlying assumption that the ‘systematic’ error is gaussian is often a rough approximation
Model inadequacy

• When looking for new phenomena the discovery assumes a correct estimation of the null-Hypothesis, i.e. showing that one sees an excess with respect to what is predicted by the model without the new phenomena.

• The limitation of the theoretical modelling (sometime its implementations and/or understanding by the experimental teams) have been the source of ‘false’ claims in the recent history
Examples: quark substructure

- Quark substructure: the imperfect knowledge of the Parton Distribution Functions inside the proton have led to some unjustified excitement.
Example: new particle

- ALEPH observed in 1996 a $4\sigma$ excess of Higgs-like events at 105 GeV in the 4-jet final state of electron-positron collisions at 130-136 GeV. They published the search: 9 events in a narrow mass region with an expected background of 0.7.

- None of the other LEP expts saw anything, but still the run at CM energy of 136 GeV was repeated..and the peak was not confirmed.

In DELPHI we could see some events of this kind if we dropped the cut on possible radiative returns where the photon would ‘hadronize to a $\rho$. 

[Graph of events per 2 GeV/c^2 with data and standard processes compared]
Example: Sbottom ‘discovery’

- CDF (1999) observed a significant excess of events with two or more leptons in dijet events
- ...with characteristics different from B decays
- Evidence disappeared when increasing by orders of magnitude the sample
...but news spread

- Aleph informed about the CDF excess ...found a $3\sigma$ effect in their data (LEPC, July 2000)
- DELPHI showed some problem with the MC simulation of ALEPH and that no excess was present in their data
- Later the signal was understood as an artifact of a wrong MC simulation and miscalibrated electron fake rates
• In 2011 the CDF collaboration showed a large, $4\sigma$ signal at 145 GeV in the dijet mass distribution of proton-antiproton collision events producing an associated leptonic W boson decay.

• The effect grew with data size!

• It was eventually understood to be due to the combination of two nasty background contaminations.

Subjective prejudice

- A certain level of ‘personal’ appreciation in the estimation of Systematic errors is almost unavoidable.
- It can go both ways: adopting criteria which inflate the error (hence with the danger of preventing optimal extraction of information from the data) or having an attitude too optimistic about the understanding of possible systematics ..and so provoking false claims.
- VERY dangerous is the ”N” effect.
In 2011 the OPERA collaboration produced a measurement of neutrino travel times from CERN to Gran Sasso which appeared to go faster than light in vacuum. The effect was at the level of $6\sigma$. It was finally understood to be due to a single large source of systematic uncertainty – a loose cable.

There have been conjectures that the haste with which the result was ‘put out’ was also due to the rumors about an imminent result from a ‘competing’ US collaboration.


...the importance of ‘preliminary’

• We have grown accustomed to have a ‘quick’ presentation of results at conferences or lab seminar which are labelled *preliminary* because not all the ultimate treatment of the data/sophistication of the analysis has been implemented. That often hides the fact that the systematic errors one quotes for these results might be rough estimations stemming sometime from ‘subjective’ judgement of what is still missing to achieve the ultimate result
December 2015
CERN jamboree:
Striking coincidence of an excess at the same mass!
Refrained to make a combination which would have pushed things close to discovery level ..because of the preliminary nature
..and we know the story

- ...more than 400 theoretical papers in 3 months
- ... Excitement in the media
- ... But with 4 times more data ... looks like a fluctuation

CMS, ICHEP 2016

Atlas: D. Charlton, ICHEP 2016

2016 data: no clustering around 730-750 GeV, and 3.8x more data
- 2016 data consistent with 2015 at the 2.7σ level
- Appears that the 2015 excess was a statistical fluctuation
...but the real story is

- that there was never a match between the ATLAS and CMS excesses!

Diphoton Searches

Localised excess seen in 2015 ATLAS data
- 2.1σ global (3.9σ local) significance at 750 GeV (spin-0 search), width ∼50 GeV
- After reprocessing, new 2016 reconstruction → 3.4σ local, at ∼730 GeV

From D. Charlton presentation at ICHEP 2016, Chicago...and from discussion with the analyzers the major change was the calibration of the ECAL

Importance of making sure that the detector response is fully understood: for example history of Higgs search at the end of LEP when DELPHI ‘significant’ candidates became perfect WW after final calibration/alignment
Moriond: the time for excesses

“The Moriond Workshops play an extremely important role in speculative/controversial issues. They provide a forum for those working in the field to meet, present papers, and have both formal and informal discussions and criticism. For a discussion of the role that the Moriond workshops played in another controversial episode, that of the fifth force, a proposed modification of Newton’s law of gravity, see Franklin (1993a).
• Minds are ‘bayesian’ in nature: we have (most of the time subconscious) priors about the probabilities we assign to different hypothesis

When comparing a “background-only” $H_0$ hypothesis with a “background+signal” one $H_1$ one often uses the likelihood ratio $\lambda = L_1/L_0$ as a test statistic.

However, what would be more relevant to the claim would be the ratio of the probabilities:

$$\frac{P(H_1 \mid data)}{P(H_0 \mid data)} = \frac{p(data \mid H_1)}{p(data \mid H_0)} \times \frac{\pi_1}{\pi_0} = \lambda \frac{\pi_1}{\pi_0}$$

where $p(data \mid H)$ are the likelihoods, and $\pi$ are the priors of the hypotheses.

if our prior belief in the alternative, $\pi_1$, were low, we would still favor the null even with a large evidence $\lambda$ against it.
Example: new physics in Flavour-land

Trust in ‘deviations’ depends on reliability of theoretical expectations

Modified Ligeti Plot from Gilad Perez (SEARCH 2016)
# Table of Searches for New Phenomena and “Reasonable” Significance Levels

L. Lyons, “Discovering the significance of 5σ”, arxiv:1310.1284v1

<table>
<thead>
<tr>
<th>Search</th>
<th>Surprise level</th>
<th>Impact</th>
<th>LEE</th>
<th>Systematics</th>
<th># of σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutrino osc.</td>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
<td>4</td>
</tr>
<tr>
<td>Bs oscillations</td>
<td>Low</td>
<td>Medium</td>
<td>Medium</td>
<td>Low</td>
<td>4</td>
</tr>
<tr>
<td>Single top</td>
<td>Absent</td>
<td>Low</td>
<td>Absent</td>
<td>Low</td>
<td>3</td>
</tr>
<tr>
<td>(B_s \to \mu\mu)</td>
<td>Absent</td>
<td>Medium</td>
<td>Absent</td>
<td>Medium</td>
<td>3</td>
</tr>
<tr>
<td>Higgs search</td>
<td>Medium</td>
<td>Very high</td>
<td>Medium</td>
<td>Medium</td>
<td>5</td>
</tr>
<tr>
<td>SUSY searches</td>
<td>High</td>
<td>Very high</td>
<td>Very high</td>
<td>Medium</td>
<td>7</td>
</tr>
<tr>
<td>Pentaquark</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>Medium</td>
<td>7</td>
</tr>
<tr>
<td>G-2 anomaly</td>
<td>High</td>
<td>High</td>
<td>Absent</td>
<td>High</td>
<td>5</td>
</tr>
<tr>
<td>H spin &gt;0</td>
<td>High</td>
<td>High</td>
<td>Absent</td>
<td>Low</td>
<td>4</td>
</tr>
<tr>
<td>4th gen fermions</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>6</td>
</tr>
<tr>
<td>(V&gt;c) neutrinos</td>
<td>Huge</td>
<td>Huge</td>
<td>Absent</td>
<td>Very high</td>
<td>THTQ</td>
</tr>
<tr>
<td>Direct DM search</td>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
<td>High</td>
<td>5</td>
</tr>
<tr>
<td>Dark energy</td>
<td>High</td>
<td>Very high</td>
<td>Medium</td>
<td>High</td>
<td>6</td>
</tr>
<tr>
<td>750 GeV boson</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>6</td>
</tr>
<tr>
<td>Grav. waves</td>
<td>Low</td>
<td>High</td>
<td>Huge</td>
<td>High</td>
<td>7</td>
</tr>
</tbody>
</table>
An aside: Bayesian vs frequentists

• The approach to discovery is different depending on the approach one has:
  – Bayesian: compares posterior probability of a assumed prior
  – Frequentist: uses P values

• The Jeffreys-Lindley paradox states that an assumed ‘null’ prior will always be favored when getting high statistics

• Frequentists and Bayesians draw opposite conclusions on large data when comparing a null-hypothesis to a composite alternative
We have laid the Keystone of the Std Model Cathedral ...

What we will do is to get a better ‘picture’ ie. measure better the characteristic of the Std Model

Is this all left to do?

Not the first time that the issue is posed:
Lord Kelvin (1900)
There is nothing new to be discovered in physics now. All that remains is more and more precise measurement.
Some of the known unanswered questions:

• The elephant in our ‘research’ room has been Gravity: the difficulty to reconcile Quantum Mechanics and Gravity has been a Theoretical Nightmare since ~ 100 years.

• Dark Matter is another cloud in the Standard Model sky (more on this later)

• Why we have essentially only Baryonic matter and not anti-baryons in the universe is another blemish on the Std Model

• …and we should be ready to deal with surprises: it would not be the first time in the field of High Energy Particle Physics that Nature has shown phenomena which we had not anticipated
• Why three families of Fundamental particles?
• What is the structure of the Neutrino sector: is a signature for physics beyond the STD model hidden in the neutrino transformations?
• The standard model itself seems to indicate that something is missing:
  – What is allowing the Mass of the Higgs boson to be as low as measured? If there is nothing
  – else the standard model has to be valid up to Planck Mass where Quantum Mechanics and Gravity HAVE to come together

\[
\text{Physical Higgs Mass}^2 = \text{Bare Mass}^2 + \frac{\alpha}{\pi} \left[ \right]_\text{tuned to 1 part in } 10^{32} \quad O(M_{\text{Planck}})^2
\]

..there must be more than the Std. Model!
There is more to Nature than the STD model construction.
Relations between theory and experiment (as seen by theorists)

A defendable picture when you have very tight predictions: e.g. Higgs boson, rare decays rate
..as seen by experimentalists

...This is like the situation we are now!
Are $5\sigma$ a safe ‘bet’?

• Not really...example H1 ‘evidence’ for pentaquark
• Despite the thing being quoted at the 6.2 Sigmas level they were smart enough to use the word ‘Evidence’ in the title
• ...so be prepared for some possible fake peak at $5\sigma$ in the future of the LHC running

Summary

Understanding of Nature behaviour has always required ever improved tools and measurement devices.

The complexity of today’s instruments and the sophistication of the measurements we are doing requires a rigorous approach to understand the detectors we are using and the backgrounds we are expecting.

We have developed a deep understanding of the pit-falls to avoid from the errors of the past.

And the fundamental principle of the necessity of having more than one experiment able to perform the same measurement has been proven over and over again.

We are ready to exploit fully our data and make discovery if nature will be kind to us!
Acknowledgement

- Thanks for many discussions with T. Dorigo, L. Lyons, D. Treille and A. Read. If you want to know more and much deeper I recommend Dorigo’s book: