Padova, Astro@Stats2017 – September 8th 2017

Extraordinary Claims: the 0.000029% Solution

Tommaso Dorigo
INFN Padova

Why this talk

- Physicists and astro-physicists believe they know Statistics well enough to carry out their measurements without external help, and have over time built an arsenal of «standard» methods of inference, not all of which have solid foundations
- It looks fruitful to have a discussion, in order to "bridge the gap" between Statisticians and Physicists on the jargon and on those methods
- In this talk I will focus on a couple of techniques of special interest. In particular I will discuss the much publicized concept that a scientific discovery in physics research requires that an effect be found with a statistical significance exceeding five standard deviations
- Conventions may be a good thing provided one remembers their rationale and their roots. Hence I will offer an historical overview of where the five-sigma criterion comes from and what it was designed to address, before I discuss its limitations

Contents

- **Jargon check**
- **Searching for new phenomena: upper limits and hypothesis testing**
  - Bump hunting and upper limits
  - Neyman's construction and beyond: why we are not Bayesians
  - Being on the same page: significance and related concepts
    (p-value, Wilks' theorem, type-I and type-II error rates)
- **A brief history of the five-sigma criterion**
  - Rosenfeld on exotic baryons
  - Successful and failed applications in recent times
- **The trouble with it**
  - Ill-quantifiable LEE (a.k.a. trials factors – a.k.a. Bonferroni correction)
  - Subconscious Bayes factors
  - Systematics
  - The Jeffrey-Lindley paradox
- **A possible way out: a search-dependent consensus on α**

Jargon Check

Physicists say	Statisticians say
Determine	Estimate
Estimate	Guess
Observable space	Population
Observe	Draw a sample
Data	Sample
Uncertainty	Error
Systematic	Nuisance parameter

What it Is That We Do

- In HEP we have a theory that works wonders – the Standard Model, AKA Electroweak Theory plus QCD - but we believe it is incomplete and to some extent unsatisfactory.
- In astro-HEP we also have distinct expectations for observable phenomena

- So we look for new physics processes: things that standard physics does not predict
  – New matter particles
  – signals of new phenomena
- We also measure known processes with the utmost precision, in the attempt of finding a significant difference with theory calculations
  – In particular we are keen of "measuring zero" and "measuring unity"

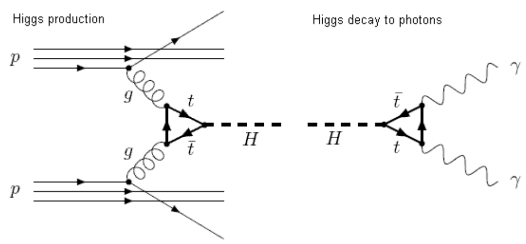
- We thus make extensive use of
  – Hypothesis testing
  – Point and interval estimation

I will discuss these

- In our analyses of the data we also frequently employ
  – Unfolding techniques
  – Machine-learning classification and regression
  – Goodness-of-fit tests

...Each of which would be worth a separate talk or two

Bump Hunting Example: Search for a New Particle



- The search for a new particle **usually** involves a theoretical model which predicts it

From the model we may infer the expected signature of the signal

- Monte Carlo methods allow us to produce simulated datasets that teach us how the signal looks like

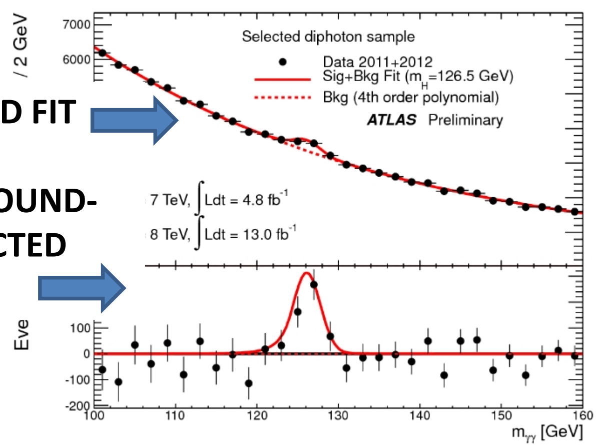
- A data selection isolates a sample where we try to evidence the particle footprint – typically a narrow bump on a smooth background

- A test of hypotheses allows to derive p(data|H0)

  – Let us remind ourselves how that is done

DATA AND FIT

BACKGROUND-SUBTRACTED DATA



Type-I and Type-II Error Rates

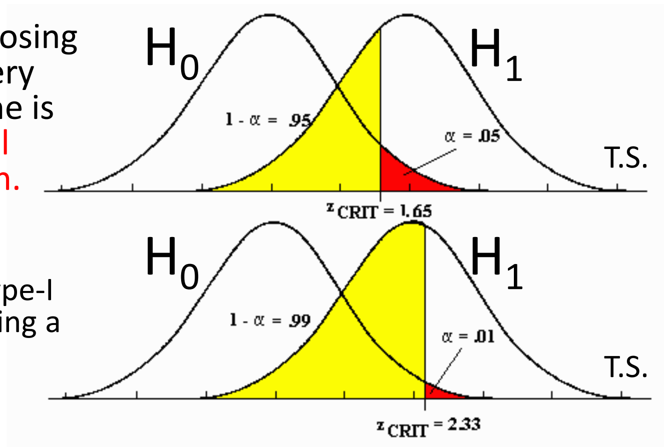
In the context of hypothesis testing the type-I error rate α is the probability of rejecting the null hypothesis when it is true.

Testing a simple null hypothesis versus a composite alternative (*e.g.* μ=0 versus μ>0) at significance level α is **dual** to asking whether 0 is in the confidence interval for μ at confidence level 1-α.

Strictly connected to α is the concept of "power" (1-β), where β is the type-2 error rate, defined as the probability of accepting the null, even if the alternative is instead true.

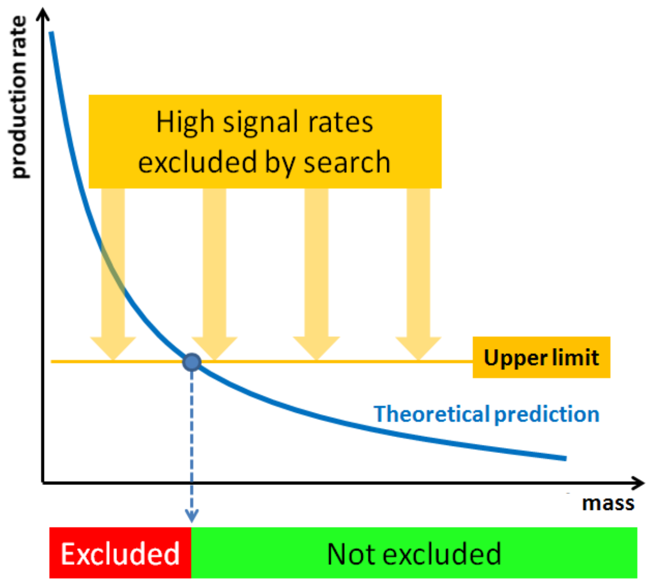
Once the test statistic is defined, by choosing α (*e.g*. to decide a criterion for a discovery claim, or to set a confidence interval) one is automatically also choosing β. In general there is no formal recipe for the decision.

A stricter requirement for α (*i.e*. a smaller type-I error rate) implies a higher chance of accepting a false null (yellow region), *i.e*. smaller power.



And What If There Is No Signal ?

- If we do not see a signal we can exclude the new physics model
- More often the model depends on an unknown parameter, and we exclude ranges of its value
  - Typically this is the mass of the particle
- We can e.g. derive lower limits on the particle mass from upper limits on the signal strength, by comparing those to a theoretical model



Luckily, the lower mass limit is useful information, worth a publication !

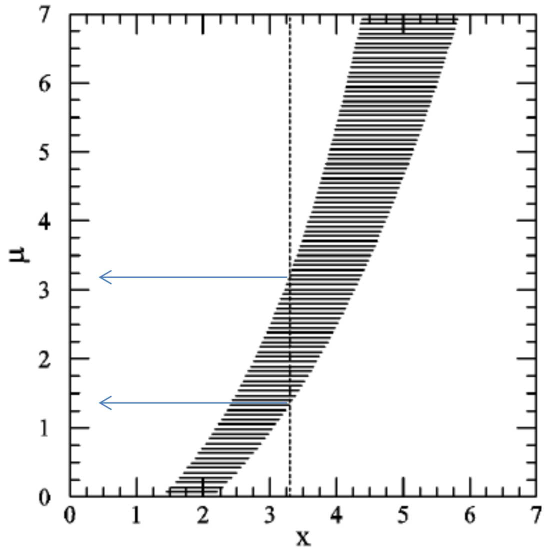
Neyman's Confidence Interval Recipe

- Specify a model which provides the probability density function of a particular observable x being found, for each value of the unknown parameter of interest: *$p(x|\mu)$*
- *Choose a Type-I error rate α (e.g. 32%, or 5%)*
- For each μ, draw a horizontal *acceptance interval $[x_1,x_2]$ such that*

  $$p\,(x \in [x_1,x_2] \mid \mu) = 1 - \alpha.$$

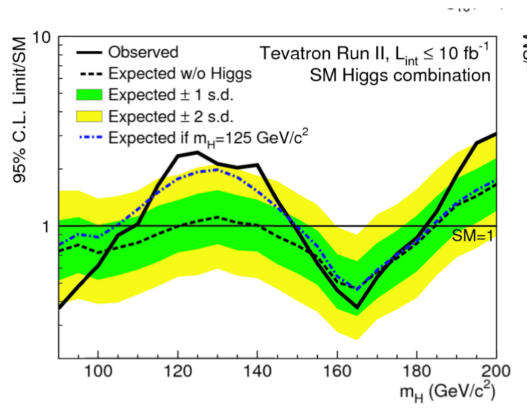
  ***There are infinitely many ways of doing this**: an ordering principle is required to well-define*
  - *for upper limits, integrate the pdf from –inf to x*
  - *for lower limits do the opposite*
  - *or choose central intervals, or shortest intervals…*
- Upon performing an experiment, you measure *$x=x^*$. You can then draw a vertical line through it.*

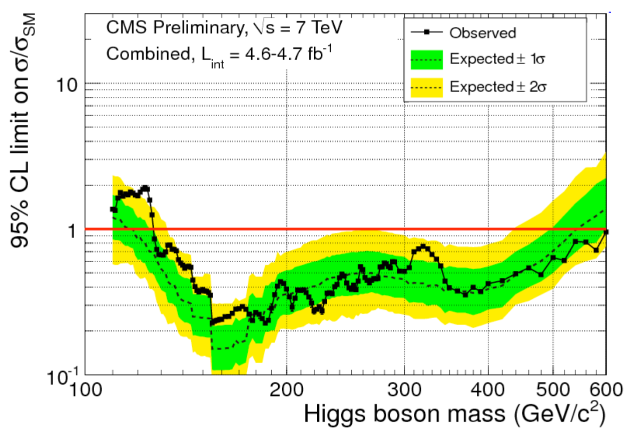
→ The vertical *confidence interval $[\mu_1,\mu_2]$ (with Confidence Level C.L. = 1 -α) is the union of all values of μ for which the corresponding acceptance interval is intercepted by the vertical line.*



The Problem Is Relevant in HEP…

- To give you the flavour of the relevance of the problem of setting correct upper limits, suffices to tell the story of the Higgs search
- For a long time (from the seventies through 2011) all we could say was where the particle could *not* be
- The competition (also for funding) centred for a while on who best refined that information rather than on who came up with the actual observation of the particle





On Coverage

- For physicists, coverage is a **very important property** of classical intervals
  - We especially like the fact that coverage is preserved even if we collect results produced by different experiments adopting different methods
  - We usually frown at the introduction of **a subjective input** in our results
  - Also note that we work with parameters which describe physical reality → we dislike speaking of the probability of a physical constant having this or that value
  - This has led to preference of classical over Bayesian techniques

However…

- Often physical quantities must fulfil constraints that **restrict the space** of possible true values
  - This has brought back Bayesian methods from the window
  - Let me discuss shortly the simplified "template" case and mention the studies that Physicists have made of them to overcome the difficulties

Typical Study: Measuring the Rate of a Process

- The typical search for a new physics process determines an interval for μ, a signal strength, where $\mu=0$ is the null hypothesis and μ is confined to be non-negative

- In this case, the Neyman construction may return as a confidence interval of size α the **empty set** if e.g. backgrounds under-fluctuate
  - In contrast with Neyman's own prescriptions !

The problem has been called "What to do when you **know** you're in the wrong 10%"

- A Bayesian solution exists: use a flat prior for μ, null for $\mu<0$
  - This is often used in HEP, although its improper nature creates problems in certain cases
  - Also, non-invariance of prior over reparametrizations is a unwelcome feature, along with usual criticism of prior dependence and error control
  - Lots of literature exists on the topic; Bayesians offer Jeffreys' priors, Bernardo's objective priors, etcetera… But these have not taken roots in HEP, a bit more in astro-HEP

- Many modifications of Neyman's recipe have been offered to avoid null intervals and produce a coherent treatment – let us give a quick look

What to Do When You Know You Are in the Wrong 10%

The classical problem is that of a Gaussian-resolution measurement x (with σ=1) of a quantity μ constrained to be non-negative:

$$P(x|\mu) = \frac{1}{\sqrt{2\pi}} \exp(-(x-\mu)^2/2)$$

Neyman's recipe for 90% upper limits is then to take μ_{UL}**=x+1.28** (labeled 1 below)

There are a number of recipes that can be compared on this paradigmatic problem. While the «diagonal line» method of Neyman is easy to criticize (x<-3 means that one is excluding physical reality at 99.9%CL), others have less clear shortcomings.

Method 4) is a Bayesian result with a step-function prior. It provides meaningful results that converge to the classical result for x>0: $X_{UL} = x\Phi^{-1}(1.0 - 0.1\Phi(\mathrm{x}))$

6) is McFarlane's «loss of confidence» method: the more x is negative, the less we can trust the measurement.

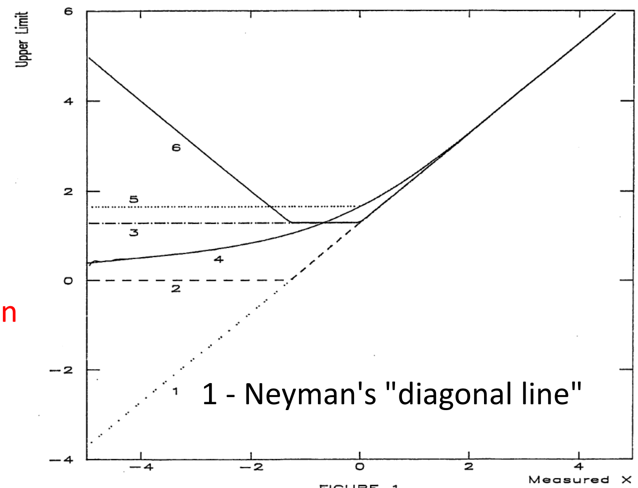


FIGURE 1

Statistical Significance: What It Is

- Statistical significance reports 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 may obtain a **p**-value as the probability of obtaining a value of T at least as extreme as the one observed, if H_0 is true. A way to do that is e.g. Wilks' theorem (discussed later).

  p can then be converted into the corresponding number of "sigma," *i.e.* standard deviation units from a Gaussian mean. This is done by finding **x** such that the integral from **x** to infinity of a unit Gaussian N(0,1) equals **p**:

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

- According to the above recipe, 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 - "**five sigma**" in jargon.

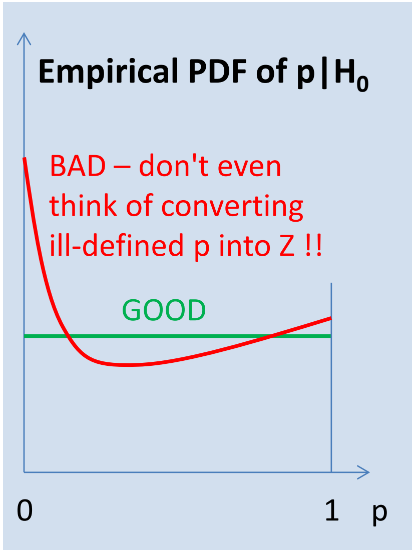
Notes

The convention is to use a "one-tailed" Gaussian: we do not care about departures of x from the mean in the *un-interesting direction*

The conversion of p into σ is independent of experimental detail. Using Nσ rather than p is **a shortcut**: we prefer to say "5σ" than "0.00000029" just as we prefer to say "a nanometer" instead than "0.000000001 meters" or "a Petabyte" instead than "1000000000000 bytes"

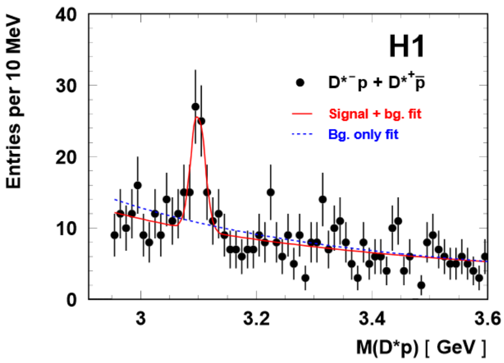
In particular, using "sigma" units does in no way mean we are operating some kind of Gaussian approximation anywhere in the problem

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 Nσ



Again, note: the "*probability of the data*" is not used. What is used is the probability of a subset of the possible outcomes of the experiment, defined by the outcome actually observed (**as much** or **more** extreme)

An Important Ingredient: Wilks' Theorem

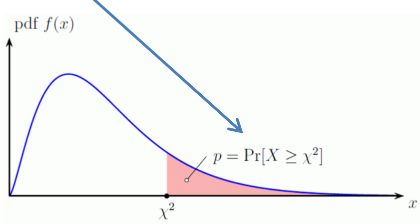


A common method to derive a significance from a likelihood fit is the one of invoking **Wilks' theorem**

One has a likelihood under the null hypothesis, L_0 (e.g., a background-only fit), and a likelihood for an alternative, L_1 (a signal+background fit)

- One takes $\mathbf{-2\,(\ln L_1 - \ln L_0) = -2\,\Delta(\ln L)}$ and interprets it as a χ^2 value - i.e. one sampled from a chisquare distribution of the relevant N_{dof}
- $P(\chi^2, N_{dof})$ can then be obtained as a "**tail probability**", and from it one gets a Z-value – the number of sigma.

This is only applicable when the two hypotheses are connected by H_0 being a particular case of H_1 (i.e., $H_0 == H_1$ when some of the H_1 parameters are fixed to special values): they must be **nested models**.



The Birth of the Five-Sigma Criterion

Arthur H. Rosenfeld (Univ. Berkeley)

Far-Out Hadrons

- In 1968 Arthur Rosenfeld wrote a paper titled "*Are There Any Far-out Mesons or Baryons?*“ **[1]**. In it, he demonstrated that the number of claims of discovery of exotic particles published in scientific magazines **agreed with the number of statistical fluctuations** that one would expect in the analyzed datasets.

  (“Far-out hadrons” are hypothetical particles, defined as ones that do not fit in SU(3) multiplets. In 1968 quarks were not yet fully accepted as real entities, and the question of the existence of exotic hadrons was important.)

- Rosenfeld pointed his finger at large trial factors coming into play due to the massive use of combinations of observed particles in deriving mass spectra containing potential resonances:

  “[...] This reasoning on multiplicities, extended to all combinations of all outgoing particles and to all countries, leads to an estimate of 35 million mass combinations calculated per year. How many histograms are plotted from these 35 million combinations? A glance through the journals shows that a typical mass histogram has about 2,500 entries, so the number we were looking for, h is then 15,000 histograms per year [...]"

More Rosenfeld

"[...] Our typical 2,500 entry histogram seems to average 40 bins. This means that therein a physicist could observe 40 different fluctuations one bin wide, 39 two bins wide, 38 three bins wide... This arithmetic is made worse by the fact that when a physicist sees 'something', he then tries to enhance it by making cuts..."

(We shall get back to the last issue later)

"In summary of all the discussion above, I conclude that each of our 150,000 annual histograms is capable of generating somewhere between 10 and 100 deceptive upward fluctuations [...]".

That was indeed a problem! Rosenfeld concluded:

*"To the theorist or phenomenologist the moral is simple: **wait for nearly 5σ effects**. For the experimental group who has spent a year of their time and perhaps a million dollars, the problem is harder... go ahead and publish... but they should realize that any bump less than about 5σ **calls for a repeat of the experiment**."*

What 5σ May Do For Us

- Setting the bar at 5σ for a discovery claim undoubtedly removes the large majority of spurious signals due to statistical fluctuations
  - The trials factor required to reach 10^{-7} probabilities is of course very large, but the large number of searches being performed in today's experiments makes up for that
  - Nowadays we call this "**LEE**", for "look-elsewhere effect".
  - 50 years after Rosenfeld, we do not need to compute the trials factor by hand: we can estimate a "global" as well as a "local" p-value using brute force computing, or advanced tricks (**more later**).
- The other reason at the roots of the establishment of a high threshold for significance is the ubiquitous presence in our measurements of unknown, or ill-modeled, systematic uncertainties
  - To some extent, a 5σ threshold protects systematics-dominated results from being published as discoveries

Protection from trials factor and unknown or ill-modeled systematics are the rationale of the 5σ criterion

It is to be noted that the criterion has no basis in professional statistics literature, and is **totally arbitrary**, no less than the 5% threshold often used for the type-I error rate of research in medicine, biology, cognitive sciences, *et cetera*.

How 5σ Became a Standard
1: the Seventies

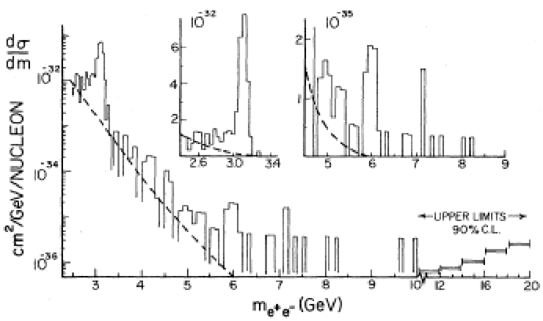
A lot has happened in HEP since 1968. In the seventies the gradual consolidation of the SM shifted the focus of particle hunts from random bump hunting to more targeted searches

Let us have a look at a few important searches to understand how the 5σ criterion gradually became a standard

- **The J/ψ discovery** (1974): no question of significance – the bumps were too big for anybody to bother fiddling with statistical tests
- **The τ discovery** (1975-1977): no mention of significances for the excesses of (eμ) events; rather a very long debate on hadron backgrounds.

- **The Oops-Leon(1976)**: "Clusters of events as observed occurring anywhere from 5.5 to 10.0 GeV appeared less than 2% of the time[8]. Thus the statistical case for a narrow (<100 MeV) resonance is strong although we are aware of the need for a confirmation."[2]

In footnote 8 they add: "An equivalent but cruder check is made by noting that the "continuum" background near 6 GeV and within the cluster width is 4 events. The probability of observing 12 events is again <=2%"
Note that P(μ=4;N>=12) = 0.00091, so this does include a x20 trials factor.



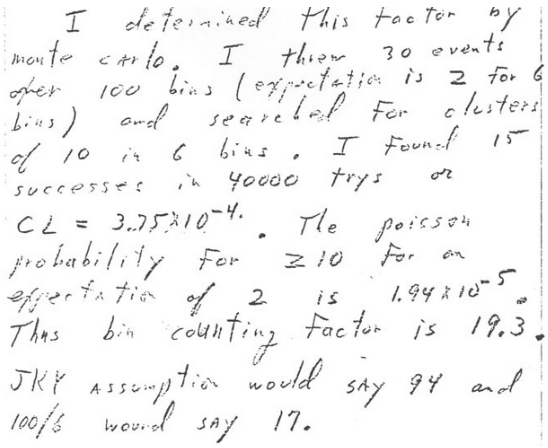
The *Real* Upsilon

The Upsilon discovery (1977): burned by the Oops-Leon, the E288 scientists waited more patiently for more data after seeing a promising 3σ peak

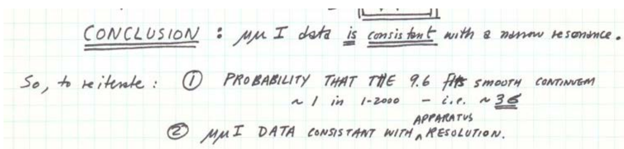
They did statistical tests to account for the trials factor (comparing MC probability to Poisson probability)

- Even after obtaining a peak with very large significance (>>5σ) they continued to investigate systematical effects
- Final announcement claims discovery but does not quote significance, noting however that the signal is "statistically significant"[3]

Nov 19th 1976

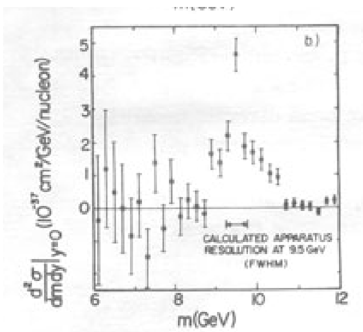

I determined this factor by monte carlo. I threw 30 events over 100 bins (expectation is 2 for 6 bins) and searched for clusters of 10 in 6 bins. I found 15 successes in 40000 trys or CL = 3.75×10⁻⁴. The poisson probability for ≥10 for an expectation of 2 is 1.94×10⁻⁵. Thus bin counting factor is 19.3. JKY assumption would say 94 and 100/6 would say 17.

Nov 21st 1976

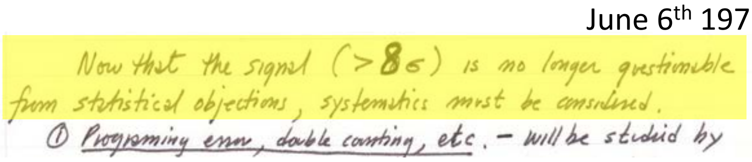

CONCLUSION: μμ I data is consistent with a narrow resonance.

So, to reiterate: ① PROBABILITY THAT THE 9.6 IS SMOOTH CONTINUUM ~1 in 1-2000 – i.e. ~3σ

② μμ I DATA CONSISTANT WITH APPARATUS RESOLUTION.

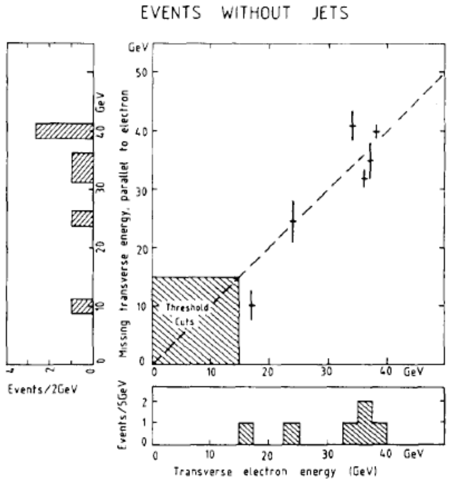


June 6th 1977

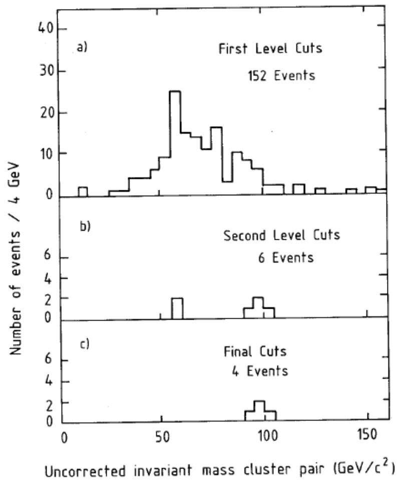

Now that the signal (>8σ) is no longer questionable from statistical objections, systematics must be considered.

① Programming error, double counting, etc. – will be studied by

The W and Z Bosons

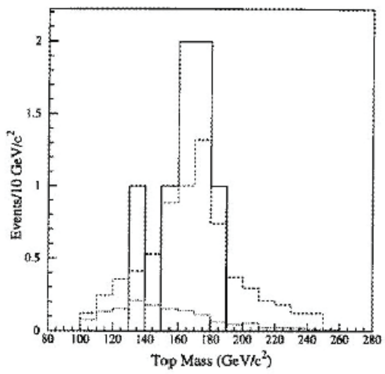


- The W discovery was announced on January 25th 1983 based on 6 striking events.
- No statistical analysis is discussed in the discovery paper[4], which however tidily rules out backgrounds as a source of the signal
  - There was no trials factor to account for: the signature was unique and predetermined; further, theory prediction for the mass (82+-2 GeV) was matched well by the measurement (81+-5 GeV).

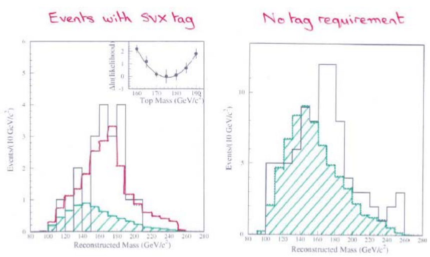


- The Z was "discovered" shortly thereafter, with an official CERN announcement made in May 1983 based on 4 events.
  - Also for the Z no trials factor was applicable
  - No mention of statistical checks in the paper[5], except notes that the various background sources were negligible.

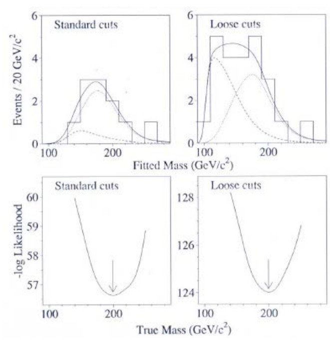
The Top Quark Discovery



- In 1994 the CDF experiment had a serious excess (2.7σ) in a counting experiment, plus a towering mass peak at a value not far from the theory-preferred value
  - the mass peak was over 3σ by itself;
    Nonetheless the paper describing the analysis spoke of "evidence" for top quark production[6]

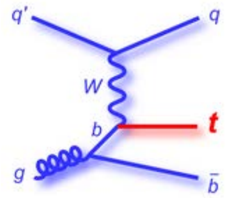


- One year later CDF and DZERO[7] both presented 5σ significances based on their counting experiments, obtained by analyzing 3x more data



The top quark was thus the first particle discovered by a willful application of the "5σ" criterion

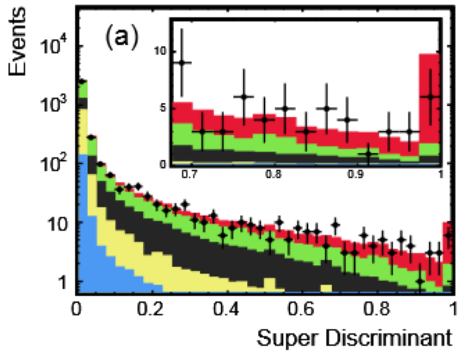
Following the Top Quark...

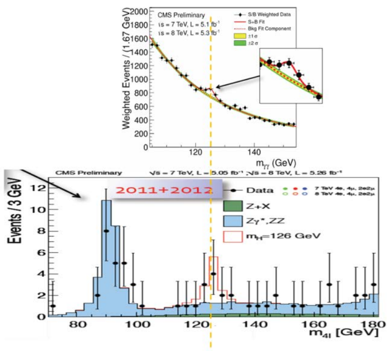


- **Since 1995, the requirement of a p-value below $3*10^{-7}$ slowly but steadily became a standard**. Two striking examples of searches that diligently waited for a 5-sigma effect before claiming discovery are:
  - **Single top quark production**: electroweak processes yielding top quarks are harder to detect than strong pair-production processes, and took 14 more years to be seen.

    CDF and DZERO competed for almost a decade in their search, obtaining 2-sigma, then 3- and 4-sigma effects, and only resolving to claim observation in 2009 [8], when clear 5-sigma effects had been observed.
  - In 2012 the **Higgs boson** was claimed by ATLAS and CMS [9]. Note that the two experiments had mass-coincident >3σ evidence in their data 6 months earlier, but the 5σ recipe was followed diligently.

    It is precisely the Higgs search what brought the five-sigma criterion to the attention of media





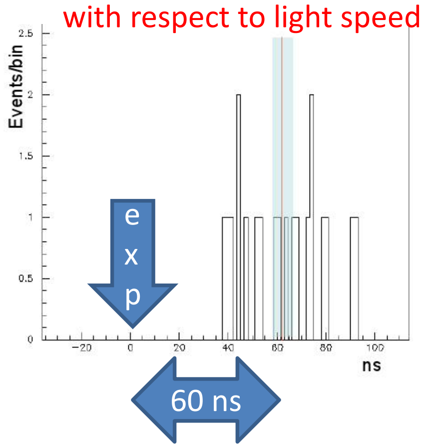
Discoveries that Weren't

- Since 1995 (but also before then) discoveries of new physics were claimed from >5σ effects
  - The protective nature of the magic number 0.00000029 is illusory when large unknown nuisances play in
- I could make a long list (see backup) but one example suffices

In 2011 the OPERA collaboration produced a measurement of neutrino travel times from CERN to Gran Sasso which appeared smaller by 6σ than the travel time of light in vacuum[15].

The effect spurred lively debates, media coverage, checks by the nearby ICARUS experiment and dedicated beam runs.
It was finally understood to be due to a single large source of systematic uncertainty – a loose cable[16]

Measured time advance with respect to light speed



A look Into the Look-Elsewhere Effect

- The discussion above clarifies that a compelling reason for enforcing a small test size as a prerequisite for discovery claims is the presence of large trials factors, aka LEE

- The LEE was a concern 50 years ago; nowadays we have enormously more CPU power. But the complexity of our analyses has also grown considerably
  - Take the Higgs discovery: CMS combined in a global likelihood dozens of final states with hundreds of nuisance parameters, partly correlated, partly constrained by external datasets, often non-Normal.
    → we still occasionally cannot compute the trials factor satisfactorily by brute force!

A study by E. Gross and O. Vitells[19] demonstrated in 2010 how it is possible to estimate the trials factor in most experimental situations, without resorting to throwing toys

Trials Factors

The situation is the one of a hypothesis test when a nuisance parameter is present only under the alternative hypothesis. The regularity conditions under which Wilks' theorem applies are then **not satisfied.**

Let us consider a particle search when the mass is unknown. The null hypothesis is that the data follow the background-only model **b(m)**, and the alternative hypothesis is that they follow the model **b(m)+ μs(m|M)**, with **μ** a signal strength parameter and **M** the particle's true mass, which here acts as a nuisance parameter only present in the alternative.

μ=0 corresponds to the null, μ>0 to the alternative.

One then defines a test statistic encompassing all possible particle mass values,

$$q_0(\hat{m}_H) = \max_{m_H} q_0(m_H)$$

This is the maximum of the test statistic for the bgr-only hypothesis H_0, across the many tests performed at the various possible masses being sought. The problem consists in assigning a p-value to the maximum of q_0(m) in the entire search range.

One can use an asymptotic "regularity" of the distribution of the above q to get a global p-value by using the technique of Gross and Vitells.

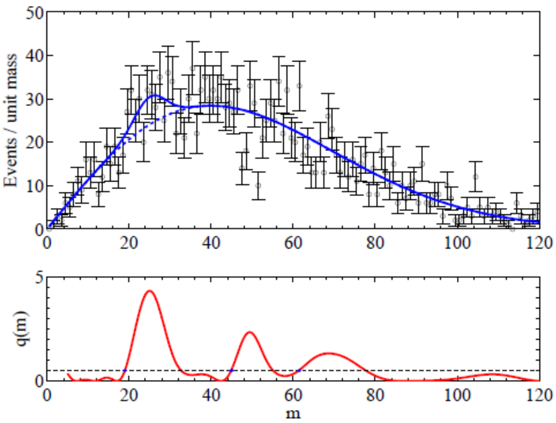
Local Minima and Upcrossings

One counts the number of "upcrossings" of the distribution of the test statistic, as a function of the nuisance parameter (mass). Its wiggling tells how many independent places one has been searching in.

The number of local minima in the fit to a distribution is closely connected to the freedom of the fit to pick signal-like fluctuations in the investigated range

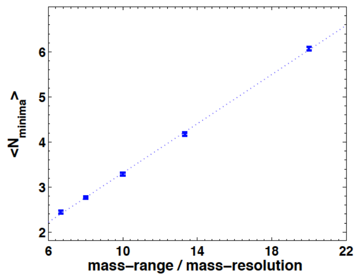
The number of times that the test statistic (below, the likelihood ratio between H_1 and H_0) crosses some reference line can be used to estimate the trials factor. One estimates the global p-value with the average number N_0 of upcrossings from a minimal value of the q_0 test statistic (for which $p=p_0$) by the formula

$$p_b^{global} = P(q_0(\hat{m}_H) > u) \le \langle N_u \rangle + \frac{1}{2} P_{\chi_1^2}(u)$$



The number of upcrossings can be best estimated using the data themselves at a low value of significance, as it has been shown that the dependence on Z is a simple negative exponential:

$$\langle N_u \rangle = \langle N_{u_o} \rangle \, e^{-(u-u_o)/2}$$



Notes About the LEE Estimation

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. My colleagues 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...)

The bottomline is that while we can always compute a local significance, it may not always be clear what the true global significance is.

Systematic Uncertainties

- Systematic uncertainties affect any physical measurement and it is sometimes quite hard to correctly assess their impact.

  Often one sizes up the typical range of variation of an observable due to the imprecise knowledge of a nuisance parameter at the 1-sigma level; then one stops there and assumes that the probability density function of the nuisance be Gaussian.

  → if however the PDF has larger tails, it makes the odd large bias much more frequent than estimated

- Indeed, the potential harm of large non-Gaussian tails of systematic effects is one arguable reason for sticking to a 5σ significance level even when we can somehow cope with the LEE.
- However, the safeguard that the criterion provides to mistaken systematics is not always sufficient.

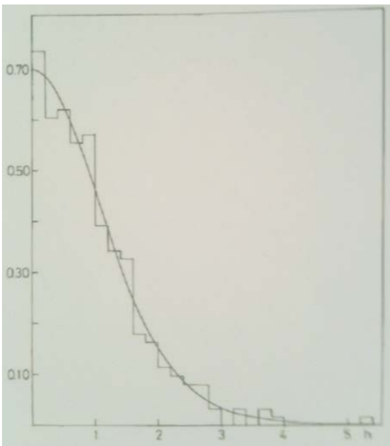
A HEP Study of Residuals

A study of the measurement of particle properties in 1975 revealed that residuals were **not Gaussian in fact**. Matts Roos *et al.* [20] considered the difference between true and measured values of kaon and hyperon mean life and mass measurements, and concluded that these seemed to all have a similar shape, well described by a Student distribution $S_{10}(h/1.11)$:

$$S_{10}\left(\frac{x}{1.11}\right) = \frac{315}{256\sqrt{10}}\left(1+\frac{x^2}{12.1}\right)^{-5.5}$$

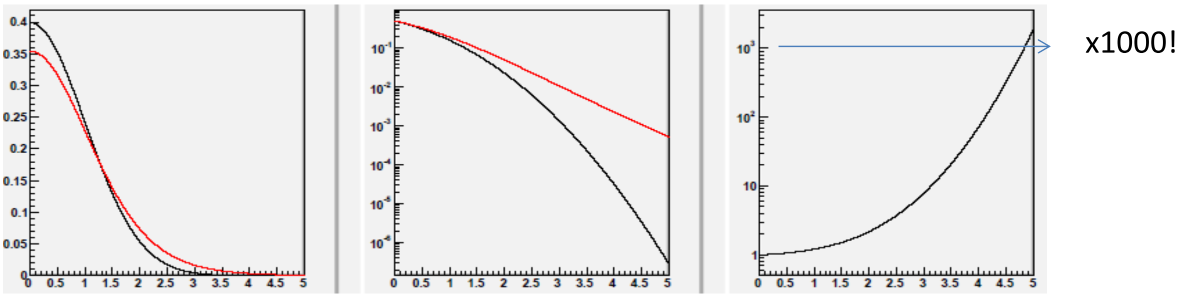
Of course, one cannot extrapolate to 5-sigma the behaviour observed by Roos and collaborators in the bulk of the distribution; however, one may consider this as evidence that the uncertainties evaluated in experimental HEP may have a significant non-Gaussian component

The distribution of residuals of 306 measurements in [20]



Black: a unit Gaussian; red: the $S_{10}(x/1.11)$ function

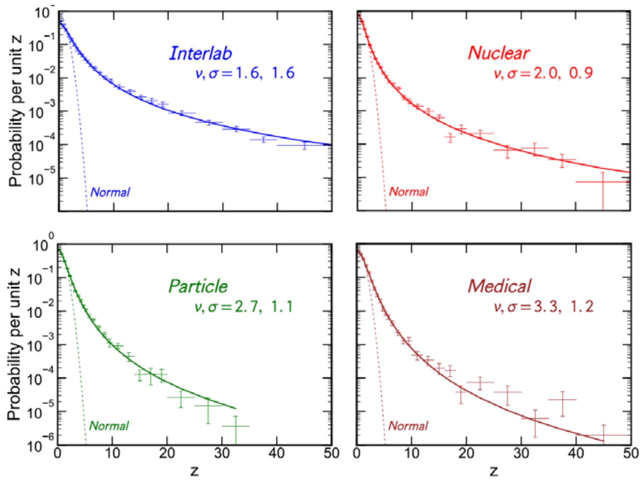
Left: 1-integral distributions of the two functions. Right: ratio of the 1-integral values as a function of z



A Bigger, Newer Study of Residuals

- David Bailey (U. Toronto) recently published an **article [26]** where use of large datasets is made (all of RPP, Cochrane medical and health database, Table of Radionuclides)
  - 41,000 measurements of 3200 quantities studied

- The methodology is similar to that of Roos et al., but some shortcuts are made, and data input automation prevents more vetting (e.g. correlations not properly accounted for)



Results are quite striking - we seem to have ubiquitous Student-t distributions in our Z values, with large tails – almost Cauchy-like.

The "Subconscious Bayes Factor"

Louis Lyons calls this way the ratio of prior probabilities we subconsciously assign to two hypotheses

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

– The p<0.000029% criterion is then applied to the distribution of λ under H_0 to claim a discovery

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

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

where p(data|H) are the likelihoods, and π are the priors of the hypotheses

In that case, if our prior belief in the alternative, π_1, were low, we would still favor the null even with a large evidence λ against it.

- The above is a Bayesian application of Bayes' theorem, while HEP physicists prefer to remain in Frequentist territory. Lyons however notes that "*this type of reasoning does and should play a role in requiring a high standard of evidence before we reject well-established theories: there is sense to the oft-quoted maxim 'extraordinary claims require extraordinary evidence'* " **[21]**.

The Heart of the Matter: the "Point Null" and the Jeffreys-Lindley Paradox

All what we have discussed so far makes sense strictly in the context of classical statistics. One might well ask what is the Bayesian view of the problem

The issue revolves around the existence of a null hypothesis, H_0, on which we base a strong belief. It is quite special to physics and astrophysics that **we usually do believe in our "point null"** – a theory which works for a specific value of a parameter, known with arbitrary accuracy; in other sciences a true "point null" hardly exists

The fact that we must often compare a null hypothesis for which a parameter has a very specific value to an alternative which has a continuous support for the parameter under test bears on the definition of a prior belief for the parameter. Bayesians speak of a "probability mass" at $\theta=\theta_0$.

The use of probability masses in priors in a simple-vs-composite test throws a monkey wrench in the Bayesian calculation, as it can be proven that no matter how large and precise is the data, Bayesian inference **strongly depends** on the scale over which the prior is non-null – that is, on the **prior belief** of the experimenter.

The Jeffreys-Lindley paradox [22] may bring Frequentists and Bayesians to draw **opposite conclusions** on some data when comparing a point null to a composite alternative. This fact bears relevance to the kind of tests we are discussing, so let us give it a look.

The Paradox

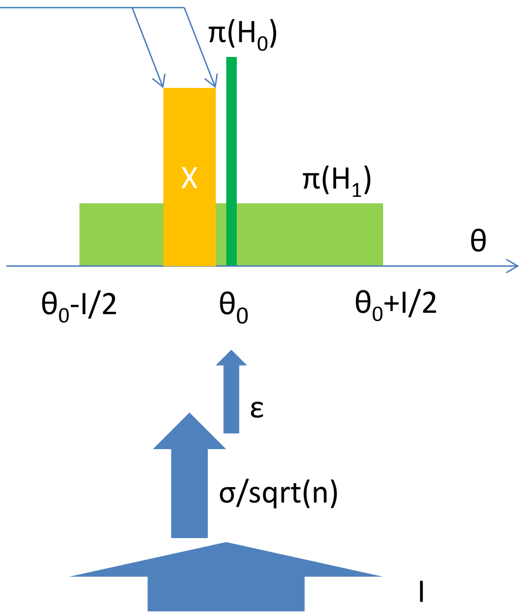
Take $X_1 \ldots X_n$ i.i.d. as $X_i|\theta \sim N(\theta,\sigma^2)$, and a prior belief on θ constituted by a mixture of a point mass p at θ_0 and (1-p) uniformly distributed in $[\theta_0\text{-}l/2, \theta_0\text{+}l/2]$.

In classical hypothesis testing the "critical values" of the sample mean delimiting the rejection region of H_0: $\theta=\theta_0$ in favor of H_1: $\theta<>\theta_0$ at significance level α are

$$\bar{X} = \theta_0 \pm (\sigma/\sqrt{n}) z_{\alpha/2}$$

where $z_{\alpha/2}$ is the significance corresponding to test size α for a two-tailed normal distribution

Given the above, it can be proven that the **posterior probability that H_0 is true** conditional on the data in the critical region (i.e. excluded by a classical α-sized test) **approaches 1 as the sample size becomes arbitrarily large**.



As evidenced by Bob Cousins[23], the paradox arises if there are three different scales in the problem, **ε << σ/sqrt(n) << l**, i.e. the width of the point mass, the measurement uncertainty, and the scale **l** of the prior for the alternative hypothesis

The three scales are usually independent in HEP!!

Notes on the JL Paradox

- The paradox is often used by Bayesians to criticize the way inference is drawn by frequentists:
  - Jeffreys: "***What the use of [the p-value] implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred"*** **[24]**
  - Alternatively, the criticism concerns the fact that no mathematical link between p and P(H|x) exists in classical HT.
- There is no clear Bayesian substitute to the Frequentist p-value for reporting exp. results
  - Bayesians prefer to cast the HT problem as a Decision Theory one, where by specifying the loss function allow a quantitative and well-specified (although subjective) recipe to choose between alternatives
  - Bayes factors, which describe by how much prior odds are modified by the data, cannot factorize out the subjectivity of the prior belief when the JLP applies: even asymptotically, **they retain a dependence on the scale of the prior of H_1.**
- In their debates on the JL paradox, Bayesian statisticians have blamed the concept of a "point mass", as well as suggested n-dependent priors. There is a large body of literature on the subject
  - As the source if the problem is assigning to the null hypothesis a non-zero prior, **statisticians tend to argue that "the precise null" is never true**. However, physicists do believe in their point nulls !!

- **From a Frequentist perspective, the JL paradox draws attention to the fact that a fixed level of significance does not cope with a situation where the amount of data increases, which is common in HEP.**

  In summary, the issue is an active research topic and is not resolved. I have brought it up here to show how the trouble of defining a test size α in classical hypothesis testing is not automatically solved by moving to Bayesian territory.

So What To Do With 5σ ?

To summarize the points made above:

- the LEE can be estimated analytically as well as computationally; experiments in fact now routinely produce "global" and "local" p-values and Z-values
  - What is then the point of protecting from large LEE ?
- In any case sometimes the trials factor is 1 and sometimes it is enormous; a one-size-fits-all is then hardly justified – **it is illogical to penalize an experiment for the LEE of others**

- the impact of systematic uncertainties also varies widely from case to case

- The cost of a wrong claim, as image damage or backfiring of media hype, can vary dramatically

- Some claims are intrinsically less likely to be true –*i.e.* we have a subconscious Bayes factor at work. It depends if you are discovering an unimportant new meson or a violation of physical laws

So why a fixed discovery threshold ?

- One may take the attitude that any claim is anyway subject to criticism and independent verification anyway; and it is good to just have a "reference value" for the level of significance of the data – a «tradition», a useful standard

Lyons' Table

My longtime CDF and CMS colleague Louis Lyons considered several known searches in HEP and astro-HEP, and produced a table where for each effect he listed several "inputs":

1. the degree of surprise of the potential discovery
2. the impact for the progress of science
3. the size of the trials factor at work in the search
4. the potential impact of unknown or ill-quantifiable systematics

He could then derive a "reasonable" significance level that would account for the different factors at work, for each considered physics effect **[21]**

- The entries in Lyons' table are entirely debatable. The message is however clear: **physicists should beware of a "one-size-fits-all" standard**.

I have slightly modified Lyons' original table to reflect my personal bias

Table of Searches for New Phenomena and "Reasonable" Significance Levels

	Search	Surprise level	Impact	LEE	Systematics	Z-level
Observed	Neutrino osc.	Medium	High	Medium	Low	4
Observed	Bs oscillations	Low	Medium	Medium	Low	4
Observed	Single top	Absent	Low	Absent	Low	3
Observed	$B_s \rightarrow \mu\mu$	Absent	Medium	Absent	Medium	3
Observed	Higgs search	Medium	Very high	Medium	Medium	5
Observed	Grav. waves	Low	High	Huge	High	7
Still being sought	SUSY searches	High	Very high	Very high	Medium	7
Still being sought	Pentaquark	High	High	High	Medium	7
Still being sought	G-2 anomaly	High	High	Absent	High	5
Still being sought	H spin >0	High	High	Absent	Low	4
Still being sought	4th gen fermions	High	High	High	Low	6
Still being sought	V>c neutrinos	Huge	Huge	Absent	Very high	THTQ
Still being sought	Direct DM search	Medium	High	Medium	High	5
Still being sought	Dark energy	High	Very high	Medium	High	6
Still being sought	Tensor modes	Medium	High	Medium	High	5

Conclusions

- Physicists use profusely the technique of hypothesis testing and derive upper limits and intervals from their data
  - The specificities of the problems call for specialized solutions. Largely these have not been offered yet

- In this talk I could only scratch the surface of some of the issues… The relative debates have lasted >=30-years, with no sign of resolution yet

- I argue that it is important that we continue to remind ourselves and educate our researchers on the roots of the conventions we use
  - hence this talk, e.g.

- I also argue that sub-fields of research targeting specific questions (e.g.: "*Can we directly detect dark matter?", "Is there Supersymmetry in LHC data?", "Is there an optical counterpart of that gravitational wave signal?*") **could agree on numbers** that **differ from 0.00000029** to validate a discovery

- **No p-value saves you from unknown unknowns. That is Science.**

Thank you for your attention!

References

[1] A. H. Rosenfeld, "*Are there any far-out mesons and baryons?,*" In: C.Baltay, AH Rosenfeld (eds) Meson Spectroscopy: A collection of articles, W.A. Benjamin, New York, p.455-483.

[2] D. C. Hom et al., "*Observation of High-Mass Dilepton Pairs in Hadron Collisions at 400 GeV*", Phys. Rev. Lett. 36, 21 (1976) 1236

[3] S. W. Herb et al., "*Observation of a Dimuon Resonance at 9.5-GeV in 400-GeV Proton-Nucleus Collisions*", Phys. Rev. Lett 39 (1977) 252.

[4] G. Arnison et al., "Experimental Observation of Isolated Large Transverse Energy Electrons with Associated Missing Energy at sqrt(s)=540 GeV, Phys. Lett. 122B, 1 (1983) 103.

[5] G. Arnison et al., "Experimental Observation of Lepton Pairs of Invariant Mass Around 95 GeV/c2 at the CERN SpS Collider", Phys. Lett. 126B, 5 (1983) 398.

[6] F. Abe et al., "Evidence for Top Quark Production in p anti-p Collisions at s**(1/2) = 1.8 TeV", Phys. Rev. D50 (1994) 2966.

[7] F. Abe et al., "*Observation of Top Quark Production in p anti-p Collisions with the Collider Detector at Fermilab*", Phys. Rev. Lett. 74 (1995) 2626; S. Abachi et al., "*Observation of the Top Quark*", Phys. Rev. Lett. 74 (1995) 2632.

[8] V.M. Abazov et al., "Observation of Single Top-Quark Production", Phys. Rev. Lett. 103 (2009) 092001; T. Aaltonen et al., "Observation of Electroweak Single Top Quark Production", Phys. Rev. Lett. 103 (2009) 092002.

[9] J. Incandela and F. Gianotti, "Latest update in the search for the Higgs boson", public seminar at CERN. Video: http://cds.cern.ch/record/1459565; slides: http://indico.cern.ch/conferenceDisplay.py?confId=197461.

[10] S. Park, "Searches for New Phenomena in CDF: Z', W' and leptoquarks", Fermilab-Conf-95/155-E, July 1995.

[11] J. Berryhill et al., "Search for new physics in events with a photon, b-tag, and missing Et", CDF/ANAL/EXOTIC/CDFR/3572, May 17th 1996.

[12] D. Acosta et al., "*Study of the Heavy Flavor Content of Jets Produced in Association with W Bosons in p anti-p Collisions at s**(1/2) = 1.8 TeV*", Phys. Rev. D65, (2002) 052007.

[13] D. Buskulic et al., "*Four-jet final state production in e^e collisions at centre-of-mass energies of 130 and 136 GeV*", Z. Phys. C 71 (1996) 179.

[14] A. Aktas et al., "*Evidence for a narrow anti-charm baryon state*", Phys. Lett. B588 (2004) 17.

[15] T. Adam et al., "*Measurement of the neutrino velocity with the OPERA detector in the CNGS beam*", JHEP 10 (2012) 093.

[16] T. Adam et al., "*Measurement of the neutrino velocity with the OPERA detector in the CNGS beam using the 2012 dedicated data*", JHEP 01 (2013) 153.

[17] T. Aaltonen et al., "*Invariant Mass Distribution of Jet Pairs Produced in Association with a W Boson in p anti-p Collisions at s**(1/2) =1.96 TeV*", Phys. Rev. Lett. 106 (2011) 71801.

[18] T. Aaltonen et al., "*Invariant-mass distribution of jet pairs produced in association with a W boson in p pbar collisions at sqrt(s) = 1.96 TeV using the full CDF Run II data set*", Phys. Rev. D 89 (2014) 092001.

[19] E. Gross and O. Vitells, "*Trials factors for the Look-Elsewhere Effect in High-Energy Physics*", arxiv:1005.1891v3, Oct 7th 2010

[20] M. Roos, M. Hietanen, and M.Luoma, "*A new procedure for averaging particle properties*", Phys.Fenn. 10:21, 1975

[21] L. Lyons, "*Discovering the significance of 5σ*", arxiv:1310.1284v1, Oct 4th 2013

[22] D.V. Lindley, "*A statistical paradox*", *Biometrika, 44 (1957) 187-192.*

[23] R. D. Cousins, "*The Jeffreys-Lindley Paradox and Discovery Criteria in High-Energy Physics*", arxiv:1310.3791v4, June 28th 2014, to appear in a special issue of Synthese on the Higgs boson

[24] H. Jeffreys, "*Theory of Probability*", 3rd edition Oxford University Press, Oxford, p.385.

[25] G. K. Karagiannidis and A. S. Lioumpas, A. S., "*An improved approximation for the Gaussian Q-function*." Communications Letters, IEEE, 11(8), (2007), 644

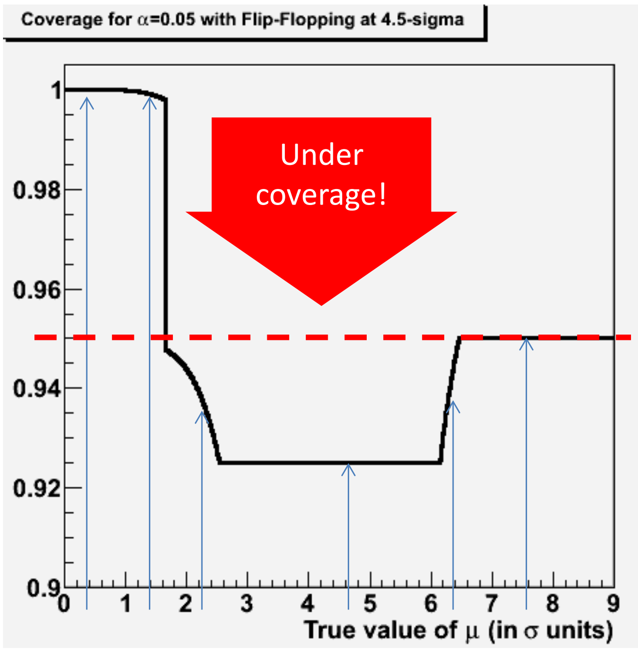
Flip-Flopping

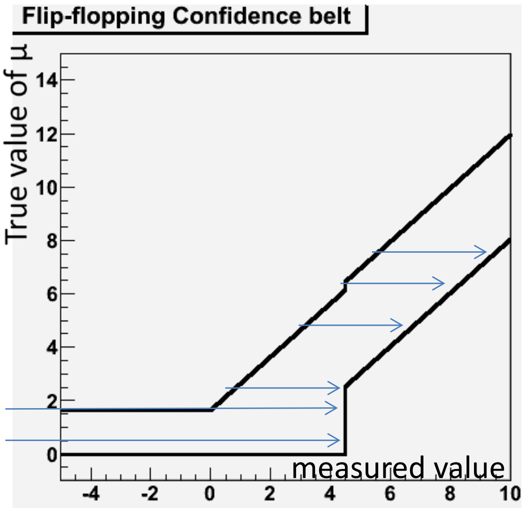
- One additional issue is the fact that physicists usually do not say beforehand whether they will set an upper limit on a quantity or claim a discovery of its non-null value
  - All they pre-define is the size of their UL test and the size of their discovery-level test
  - Typical sentence in papers: "*since we observe no significant signal, we proceed to derive upper limits…*"
- This is called «flip-flopping», and can be shown to yield under-coverage in the Neyman construction
- Suppose e.g. that we take μ_{UL} = max(x,0) +1.28 at 90%CL for the Gaussian-resolution measurement of a non-negative μ
- Upon finding x>5 (say) we have an «observation-level» significance and rather than quoting the upperl imit, we proceed to claim discovery, quoting a two-sided interval for μ: [x-1.64,x+1.64]
- This undercovers! (see next slide)

Flip-Flopping, Illustrated

- E.g. α=0.05, Disc. Threshold =4.5

The issue of Flip-Flopping and the empty set problem can be cured in the frequentist setting by the recipe advocated by G. Feldman and R. Cousins in 1998, based on a likelihood-ratio ordering of the acceptance intervals.
The FC technique is widely used in HEP





JLP Example: Charge Bias of a Tracker

- Imagine you want to investigate whether your detector has a bias in reconstructing positive versus negative particles. You count how many positive and negative particles you have reconstructed in a set of n=1,000,000 events.
- You get n^+=498,800, n^-=501,200. You want to test the hypothesis that R=0.5 with a size α=0.05.
- Bayesians will **need a prior** to make a statistical inference: their typical choice would be to **assign equal probability** to the chance that R=0.5 and to it being different (R<>0.5): say a "*point mass*" of p=1/2 at R=0.5, and a uniform distribution of the remaining p=1/2 in [0,1]
- We are in high-statistics regime and away from 0 or 1, so Gaussian approximation holds for the Binomial. The probability to observe a number of positive tracks n^+ can then be written, with $x=n^+/n$, as $N(x,\sigma)$ with $\sigma^2=x(1-x)/n$.

  The posterior probability that R=0.5 is then

$$P\left(R=\frac{1}{2}\,|\,x,n\right) \approx \frac{1}{2}\frac{e^{-\frac{(x-\frac{1}{2})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}} \Bigg/ \left[\frac{1}{2}\frac{e^{-\frac{(x-\frac{1}{2})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}} + \frac{1}{2}\int_0^1 \frac{e^{-\frac{(x-R)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}}\,dR\right] = 0.97816$$

from which a Bayesian concludes that there is **no evidence against R=0.5,** and actually the data strongly supports the null hypothesis (P>>α)

JLP Charge Bias: Frequentist Solution

Frequentists will not need a prior, and just ask themselves how often a result "at least as extreme" as the one observed arises by chance, if the underlying distribution is N(R,σ) with R=1/2 and $\sigma^2=x(1-x)/n$ as before.

One then has

$$P(x \le 0.4988 \mid R = \frac{1}{2}) = \int_0^{0.4988} \frac{e^{-\frac{(t-\frac{1}{2})^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dt = 0.008197$$

$$\Rightarrow P'(x \mid R = \frac{1}{2}) = 2 * P = 0.01639$$

(we multiply by two since we would be just as surprised to observe an excess of positives as a deficit).

From this, frequentists conclude that the detector is biased, since there is a less-than 5% probability, P'<α, that a result as the one observed could arise by chance.

A frequentist thus draws the **opposite conclusion** of a Bayesian from the same data

Particle Physics in Six Slides

- My goal today is to explain how statistical problems are handled in particle physics

→ but I need first to explain the general framework of these problems

- Let's see if I manage to say all you need to know about this in six slides

"Particles, particles, particles."

The Standard Model

A misnomer – it is not a model but a full-blown theory which allows us to compute the result of subatomic processes with high precision

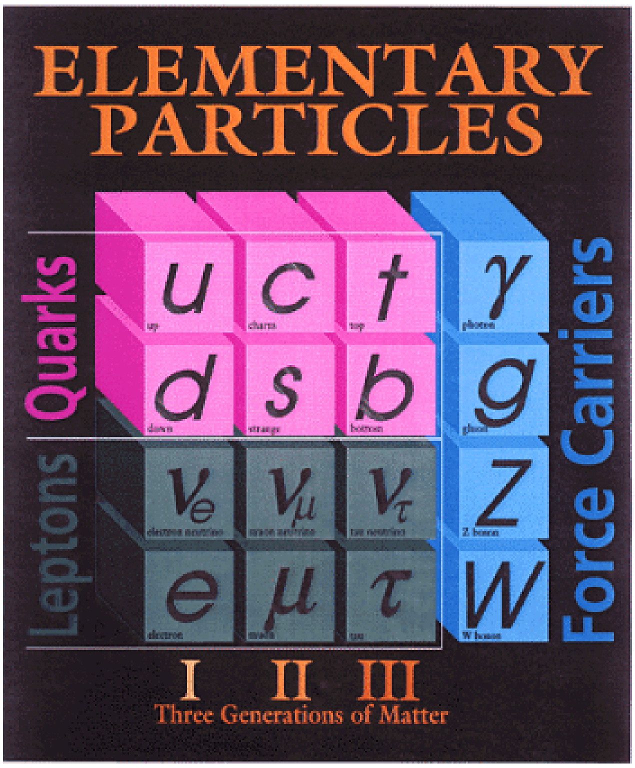
Three families of quarks, and three families of leptons, are the matter constituents

Strong interactions between quarks are mediated by 8 gluons, g

Electromagnetic interactions between charged particles are mediated by the photon, γ

The weak force is mediated by W and Z

Gravity is not included in the model

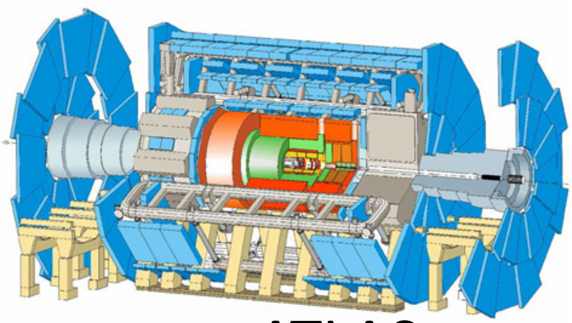


The LHC

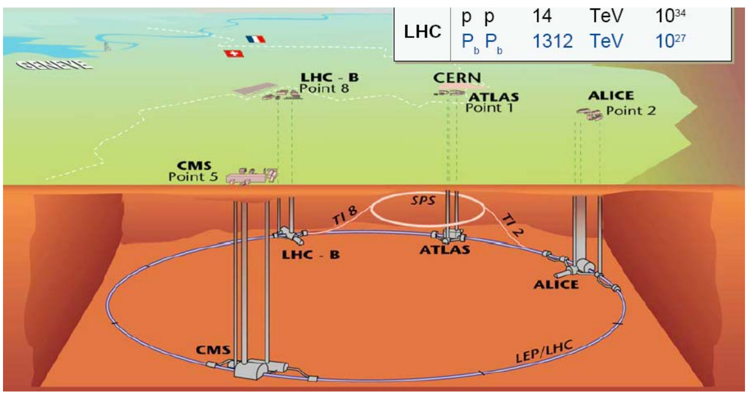
LHC is the largest and most powerful particle accelerator, built to investigate matter at the shortest distances

It resides in a 27km long tunnel 100 meters underground near Geneva

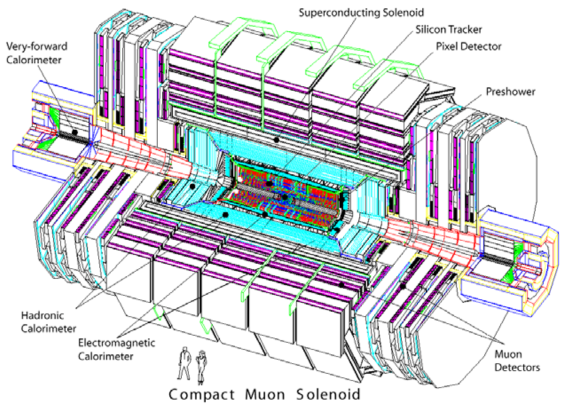
Collisions between protons are created where the beams intersect: the caverns are equipped with huge detectors. Two of these are multi-purpose «electronic eyes» that try to detect everything that comes out of the collision



ATLAS

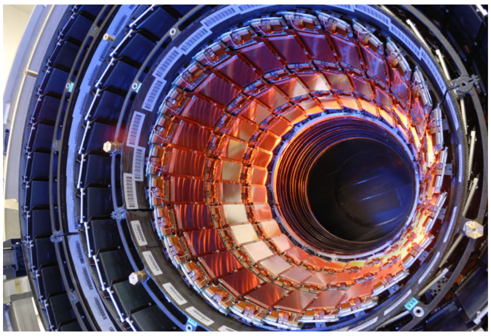


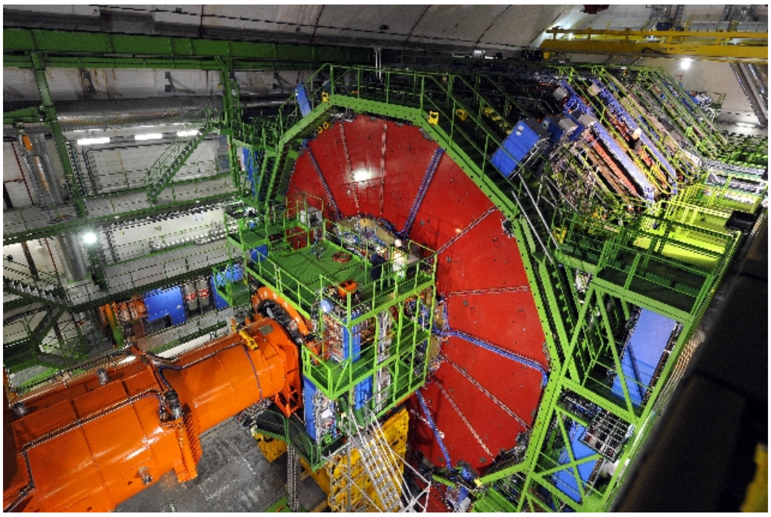
CMS

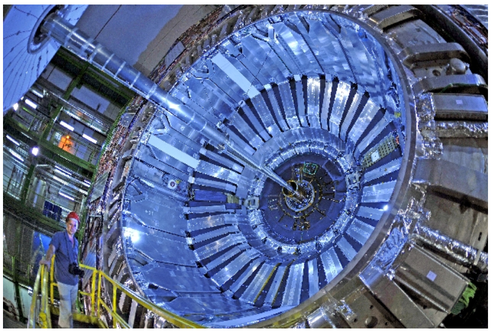


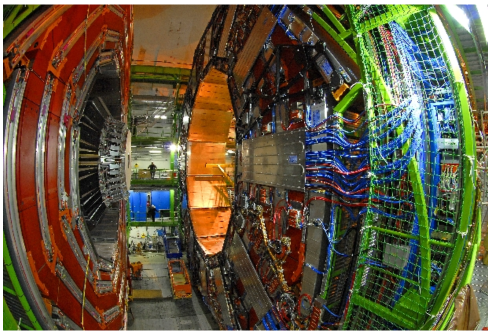
CMS

- CMS (Compact Muon Solenoid) was built with the specific goal of finding the Higgs boson
- Along with ATLAS, it is arguably the most complex machine ever built by mankind
- Hundreds of millions collisions take place every second in its core, and each produces signals in hundreds of millions of electronic channels. These data are read out in real time and stored for offline analysis



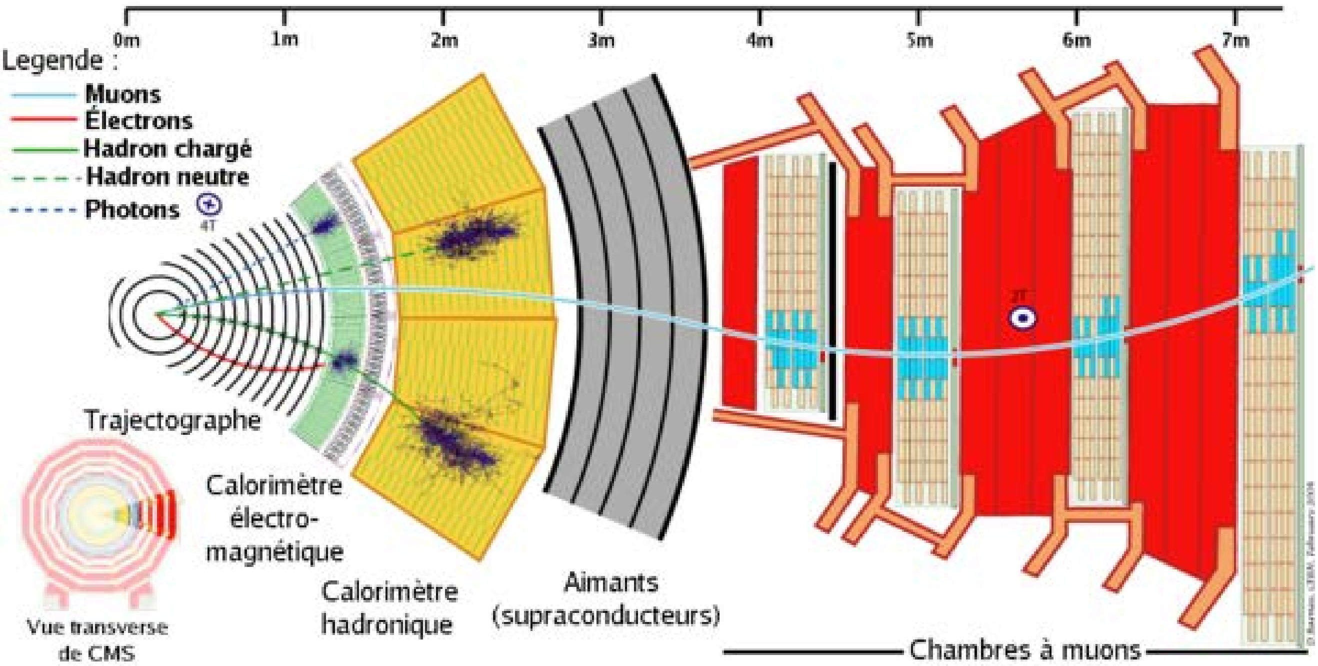






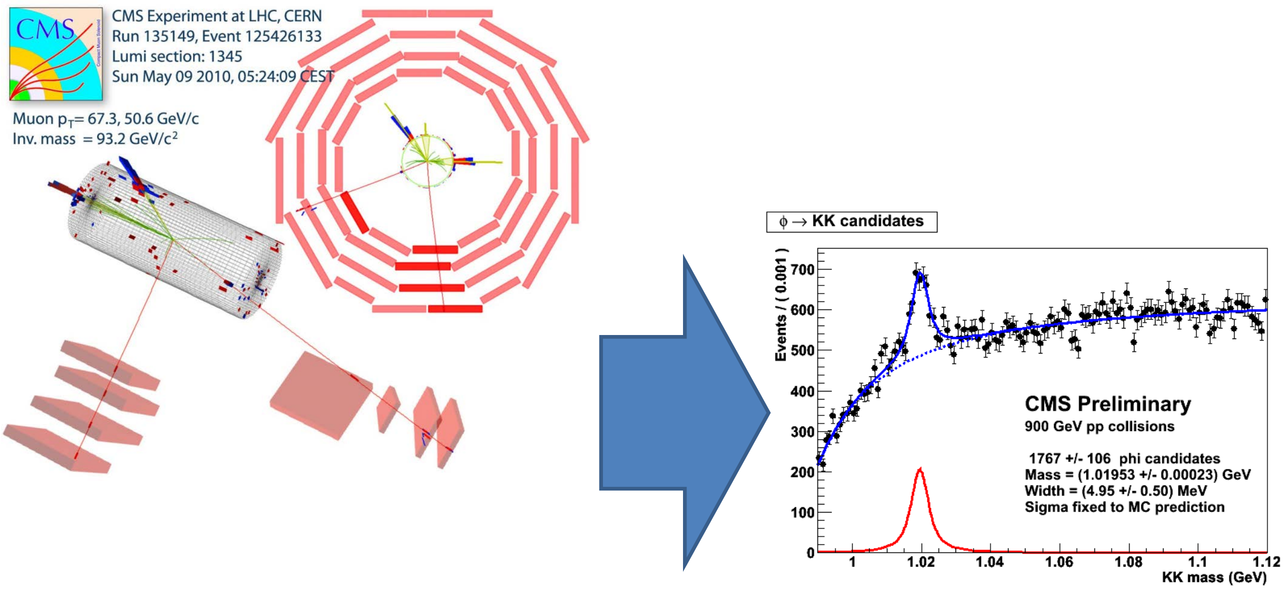
How We Detect Particles

Charged particles are tracked in the inner section, through the ionization they leave on silicon; a powerful magnet bends their trajectories, allowing a measurement of their momentum
Then calorimeters destroy both charged and neutral ones, measuring their energy
Muons are the only particles that can traverse the dense material and get tracked outside



How We See a Collision

A reconstruction of the electronic signals provides us a «view» of the created objects. Using their characteristics we build high-level variables which we compare to theoretical models, for measurements and searches



Gerry Lynch and GAME

- Rosenfeld’s article also cites the half-joking, half-didactical effort of his colleague Gerry Lynch at Berkeley:

> “My colleague Gerry Lynch has instead tried to study this problem ‘experimentally’ using a ‘Las Vegas’ computer program called Game. Game is played as follows. You wait until a unsuspecting friend comes to show you his latest 4-sigma peak. You draw a smooth curve through his data (based on the hypothesis that the peak is just a fluctuation), and punch this smooth curve as one of the inputs for Game. The other input is his actual data. If you then call for 100 Las Vegas histograms, Game will generate them, with the actual data reproduced for comparison at some random page. You and your friend then go around the halls, asking physicists to pick out the most surprising histogram in the printout. Often it is one of the 100 phoneys, rather than the real ‘4-sigma’ peak.”

- Obviously particle physicists in the ‘60s were more “bump-happy” than we are today. The proposal to raise to 5-sigma of the threshold above which a signal could be claimed was an attempt at reducing the flow of claimed discoveries, which distracted theorists and caused confusion.

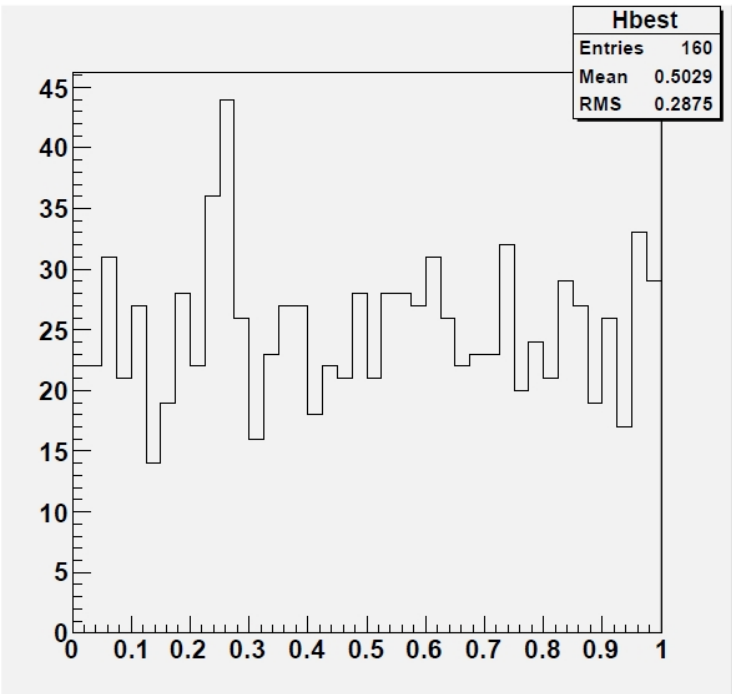
Let's Play GAME

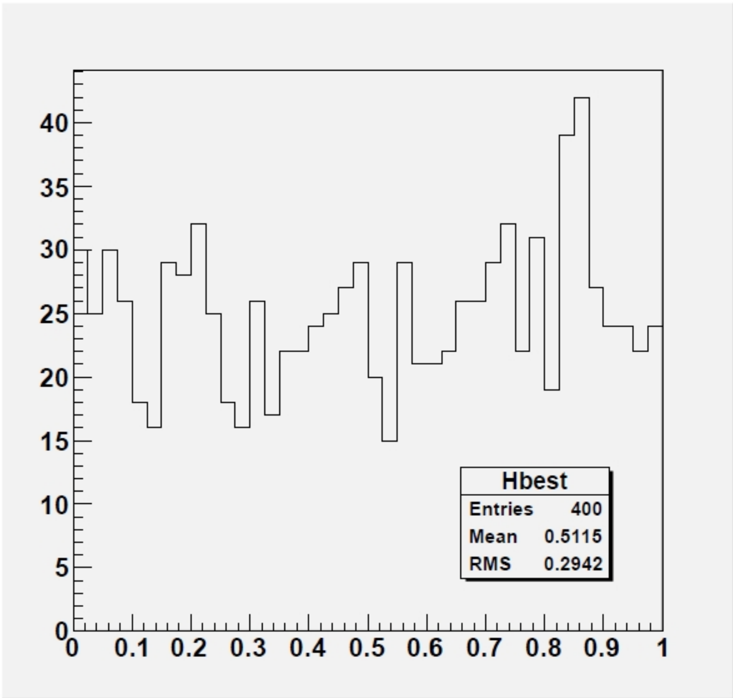
It is instructive even for a hard-boiled statistician to play GAME.

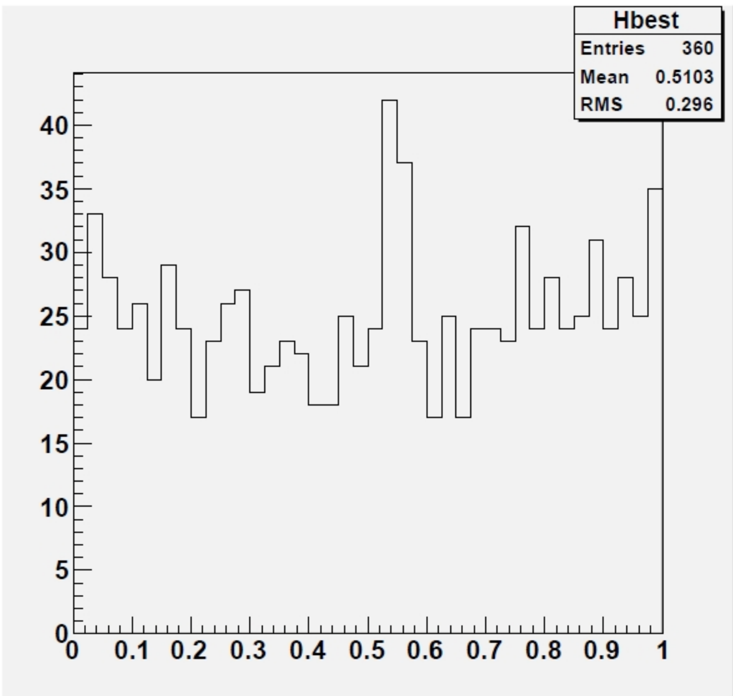
In the following slides are shown a few histograms. Some of them are created by an automated procedure **as the one** containing "the most striking" peak **among a set of 100** drawn from a smooth distribution, but one of them might be a true signal...

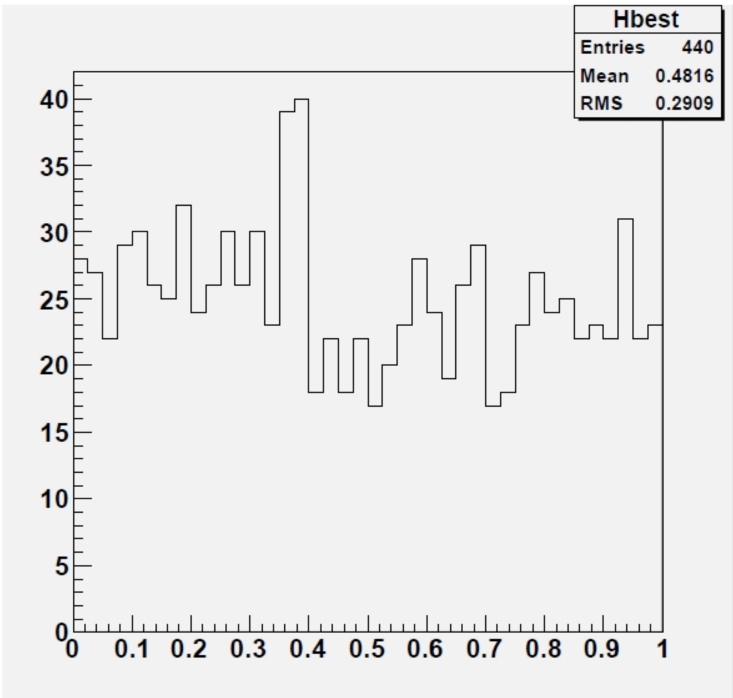
> *Details: 1000 entries; 40 bins; the "best" histogram in each set of 100 is the one with most populated adjacent pair of bins (in the first five slides) or triplets of bins (in the second set of five slides)*

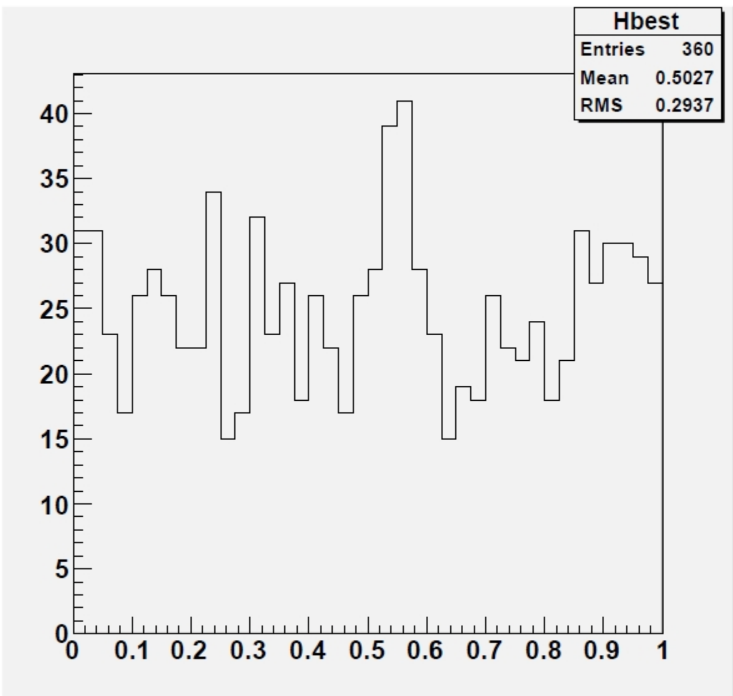
You are asked to consider **what you would tell your student if she came to your office with such a histogram**, claiming it is the result of an optimized selection for some doubly charmed baryon, say, that she has been looking for in her research project.

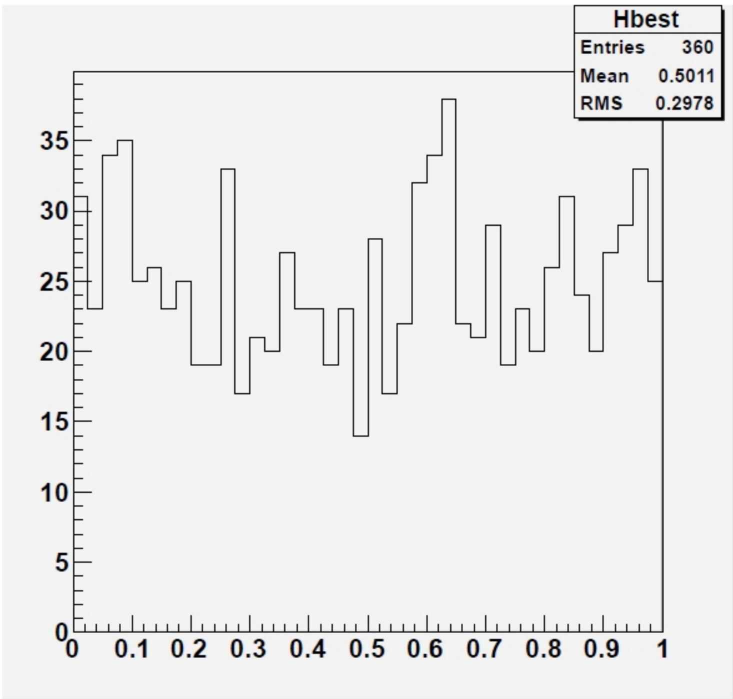


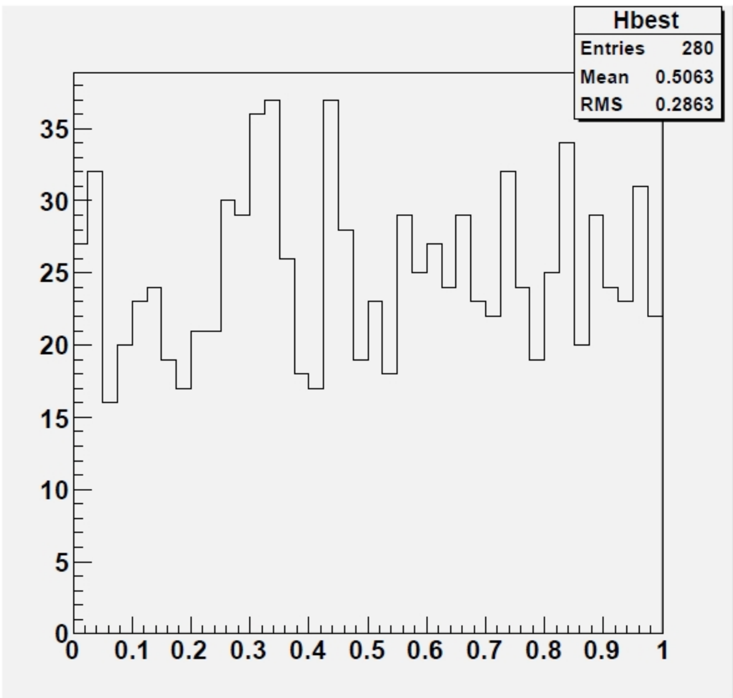

Hbest
Entries 400
Mean 0.5115
RMS 0.2942
0 5 10 15 20 25 30 35 40
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1


Hbest
Entries 360
Mean 0.5103
RMS 0.296
0 5 10 15 20 25 30 35 40
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1


Hbest
Entries 440
Mean 0.4816
RMS 0.2909
40
35
30
25
20
15
10
5
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1


Hbest
Entries 360
Mean 0.5027
RMS 0.2937
40
35
30
25
20
15
10
5
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1


Hbest
Entries 360
Mean 0.5011
RMS 0.2978
35
30
25
20
15
10
5
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1


Hbest
Entries 280
Mean 0.5063
RMS 0.2863
35
30
25
20
15
10
5
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Notes on GAME

All of the histograms are fake!

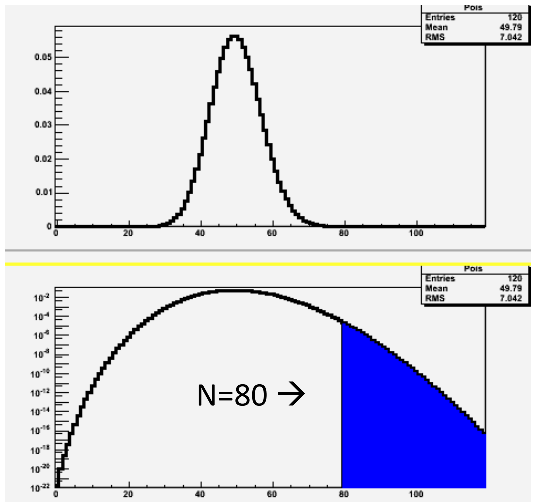
Each of them is the best one in a set of a hundred; yet some of the isolated signals have p-values corresponding to 3.5σ - 4σ effects

[*As the 2-bin bumps contain N=80 evts with an expectation of* $\mu=2*1000/40=50$, *and* $p_{Poisson}(\mu=50;N>=80)=5.66*10^{-5}$ → $Z=3.86\sigma$]

Why so large significance?

Because the bump can appear anywhere (x39) in the spectrum – we did not specify beforehand where we would look because we admit 2- as well as 3-bin bumps as "interesting" (also, we could extend the search to wider structures without penalty)

One should also mention the overlooked fact that **researchers finding a promising "bump" will usually modify the selection *a posteriori*, voluntarily or involuntarily enhancing it**. This makes the trials factor quite hard to estimate *a priori*



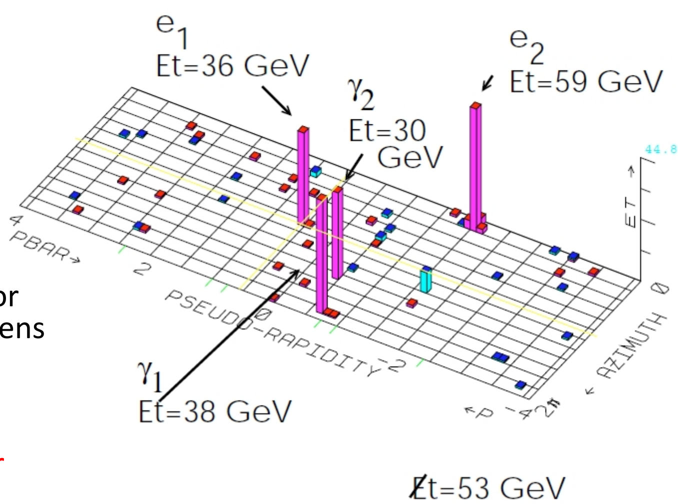
P(N|μ=50) in linear (top) and semi-log scale (bottom)

Discoveries that Were Not - 1

In April 1995 CDF collected an event which featured two clean electrons, two clean photons, large missing transverse energy, and nothing else

It could be nothing! No SM process appeared to come close to explain its presence. Possible backgrounds were estimated below 10^{-7}, a 6-sigma find

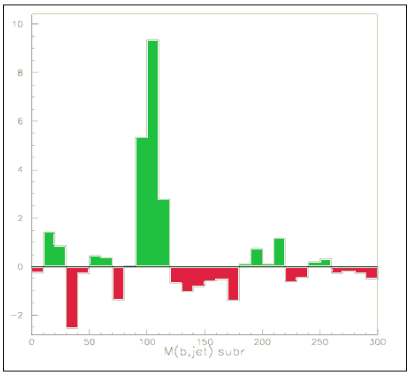
- The observation[10] caused a whole institution to dive in a 10-year-long campaign to find "cousins" and search for an exotic explanation; it also caused dozens of theoretical papers and revamping or development of SUSY models
- In Run 2 no similar events were found; DZERO never saw anything similar either



Discoveries that Were Not - 2

In 1996 CDF found a clear resonance structure at 110 GeV

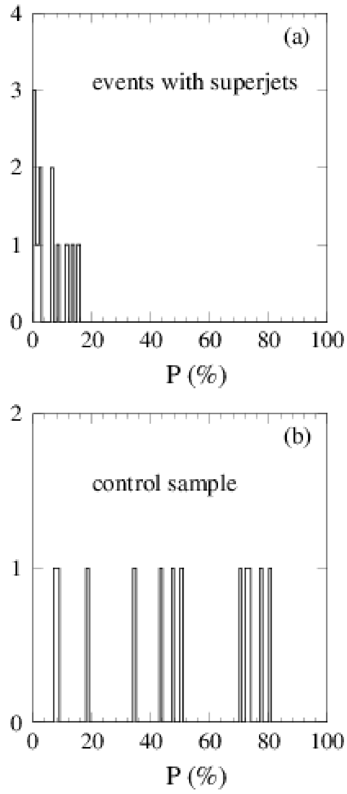
- The signal [11] had almost 4σ significance and looked quite good – but there was no compelling theoretical support for the state, no additional evidence in orthogonal samples, and the significance did not pass the threshold for discovery. It was soon archived.



In 1998 CDF observed 13 "superjet" events; a 3σ excess from background expectations (4+-1 events) but weird kinematics

Checking a "complete set" of kinematical variables yielded a significance in the 6σ ballpark

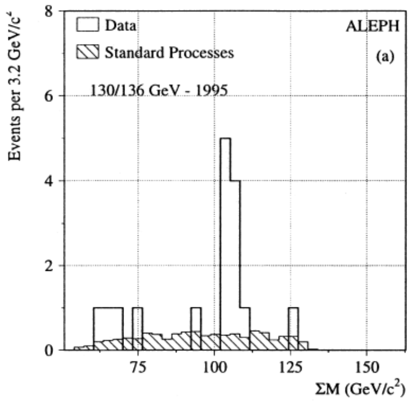
The analysis was published [12]only after a fierce, three-year-long fight within the collaboration; no similar events appeared in the x100 statistics of Run II.



Discoveries that Were Not - 3

1996 was a prolific year for particle ghosts in the 100-110 GeV region.

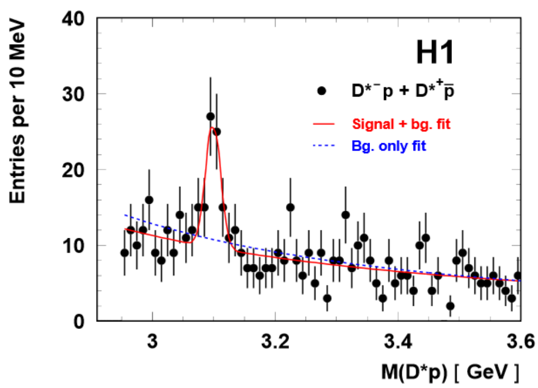
ALEPH also observed a 4σ-ish excess of Higgs-like events at 105 GeV in electron-positron collisions. They published the search[13], which found 9 events in a narrow mass region with a background of 0.7, estimating the effect at the 0.01% level



In 2004 H1 published a pentaquark signal of 6 sigma significance[14]. The prominent peak was indeed suggestive, however it was not confirmed by later searches.

In the paper they write that "*From the change in maximum log-likelihood when the full distribution is fitted under the null and signal hypotheses, corresponding to the two curves shown in figure 7, the statistical significance is estimated to* ***be p=6.2σ***"

Note: H1 worded it "Evidence" in the title ! This was a wise departure from blind application of the 5-sigma rule...

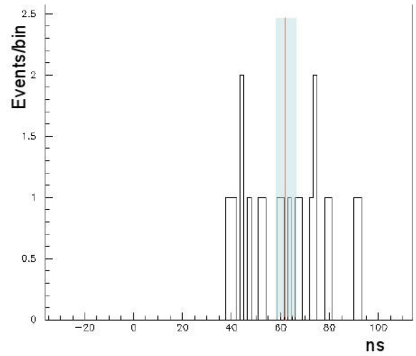


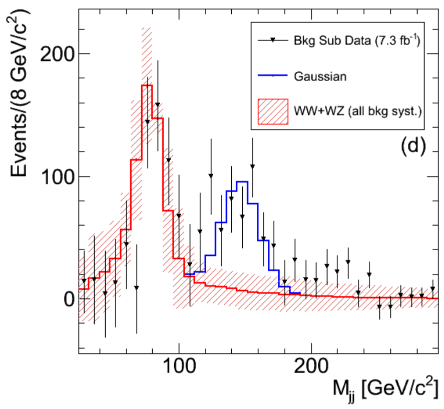
Discoveries that Were Not - 4

A mention has also to be made of two more recent, striking examples:

- In 2011 the OPERA collaboration produced a measurement of neutrino travel times from CERN to Gran Sasso which appeared smaller by 6σ than the travel time of light in vacuum[15]. The effect spurred lively debates, media coverage, checks by the nearby ICARUS experiment and dedicated beam runs. It was finally understood to be due to a single large source of systematic uncertainty – a loose cable[16]

- Also in 2011 the CDF collaboration showed a large, 4σ signal in the dijet mass distribution of proton-antiproton collision events [17]. The effect grew with data size and was systematical in nature; indeed it was later understood to be due to the combination of two nasty background contaminations[18].





An Almost Serious Table

Given the above information, an intriguing pattern emerges...

Claim	Claimed Significance				Verified or Spurious
Top quark evidence					
Top quark observation					
CDF bbγ signal					
CDF eeggMEt event					
CDF superjets					
Bs oscillations					
Single top observation					
HERA pentaquark					
ALEPH 4-jets					
LHC Higgs evidence					
LHC Higgs observation					
OPERA v>c neutrinos					
CDF Wjj bump					

An Almost Serious Table

Given the above information, an intriguing pattern emerges…

Claim	Claimed Significance				Verified or Spurious
Top quark evidence	3				true
Top quark observation			5		true
CDF bbγ signal					
CDF eeggMEt event					
CDF superjets					
Bs oscillations			5		True
Single top observation			5		True
HERA pentaquark					
ALEPH 4-jets					
LHC Higgs evidence	3				True
LHC Higgs observation			5		True
OPERA v>c neutrinos					
CDF Wjj bump					

An Almost Serious Table

Given the above information, an intriguing pattern emerges…

Claim	Claimed Significance				Verified or Spurious
Top quark evidence	3				True
Top quark observation			5		True
CDF bbγ signal		4			False
CDF eeggMEt event				6	False
CDF superjets				6	False
Bs oscillations			5		True
Single top observation			5		True
HERA pentaquark				6	False
ALEPH 4-jets		4			False
LHC Higgs evidence	3				True
LHC Higgs observation			5		True
OPERA v>c neutrinos				6	False
CDF Wjj bump		4			False