Carnegie-Mellon University, August 14 2020

Frequentist Statistics, the Particle Physicists' Way

Tommaso Dorigo

INFN Padova

Why this talk

- Particle physicists generally 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 these 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 how techniques are used, such that improvements can be made

- The problems of particle physics are quite special. This calls for specialized solutions… You are the right audience to advertise them

Contents

- **Jargon check**
- **What it is that we do**
  - Particle collisions, accelerators, detectors, and all that
- **Searching for a new particle: upper limits and hypothesis testing**
  - Neyman's construction, coverage, and the paradigmatic problem of limits on a bounded parameter
- **A brief history of the five-sigma criterion** in HEP
  - Rosenfeld on exotic baryons
  - Discovery of standard model particles
- **The trouble with five sigma**
  - Ill-quantifiable trials factors: the look-elsewhere effect
  - Systematics
  - The Jeffrey-Lindley paradox
  - Ignoring it: the case of the diphoton bump

Jargon check

When Physicists say…	**Statisticians call it…**
Determine	Estimate
Estimate	Guess
Observable space	Population
Observe	Draw a sample
Data / event	Sample
Uncertainty	Error
Error	Mistake
Systematic	Nuisance parameter

Particle physics in 8 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

- I claim I can say all you need to know about this before you manage to fall asleep

"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

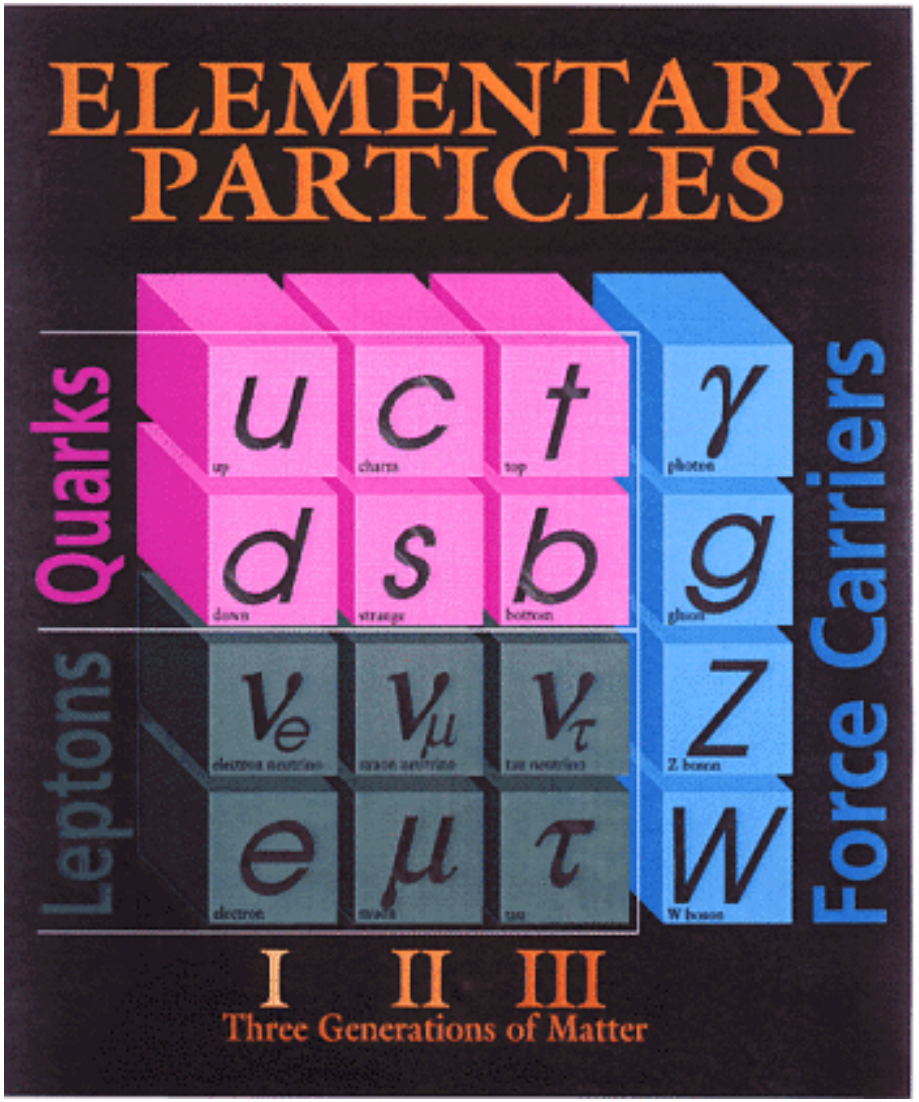
Three families of **quarks**, and three families of **leptons**, are the matter constituents

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** Bosons

The **Higgs boson** is an additional peculiar particle that gives mathematical consistency to the whole construction

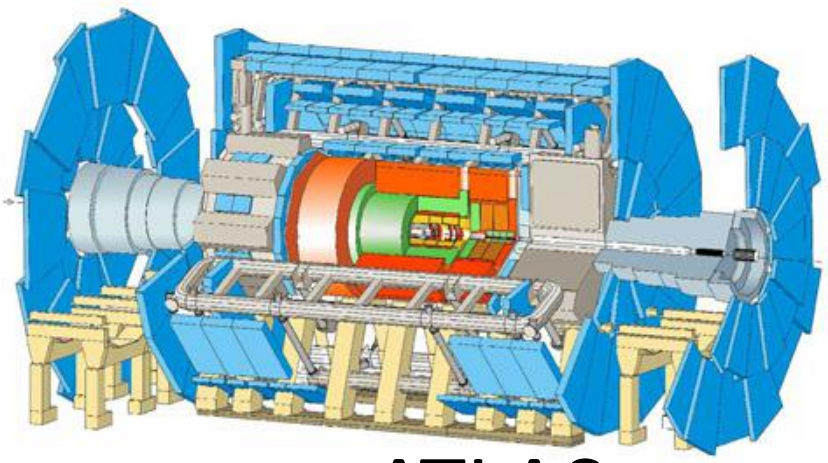


The LHC

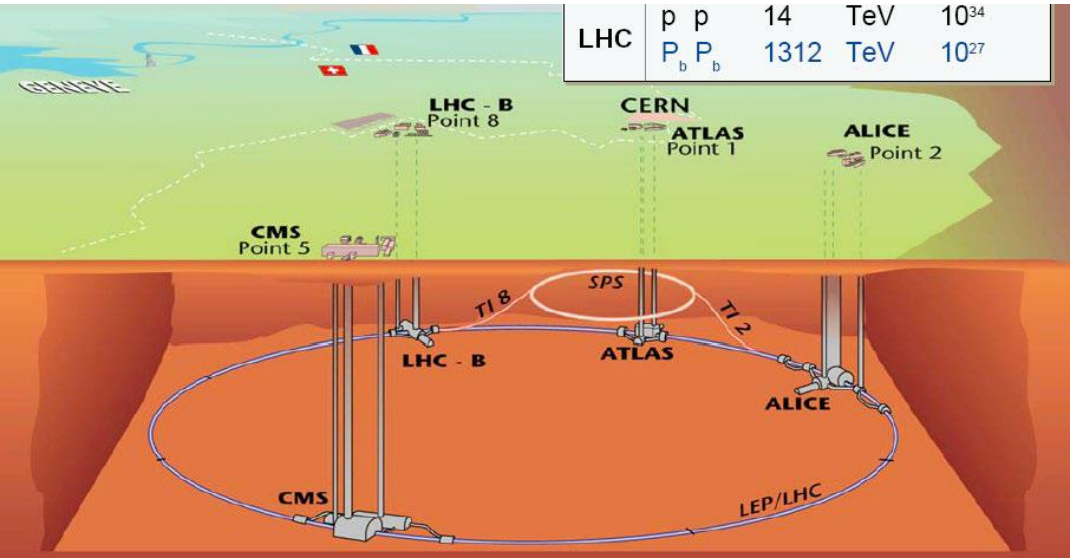
LHC is the **world's largest and most powerful** particle accelerator, built to investigate matter at the shortest length scales

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

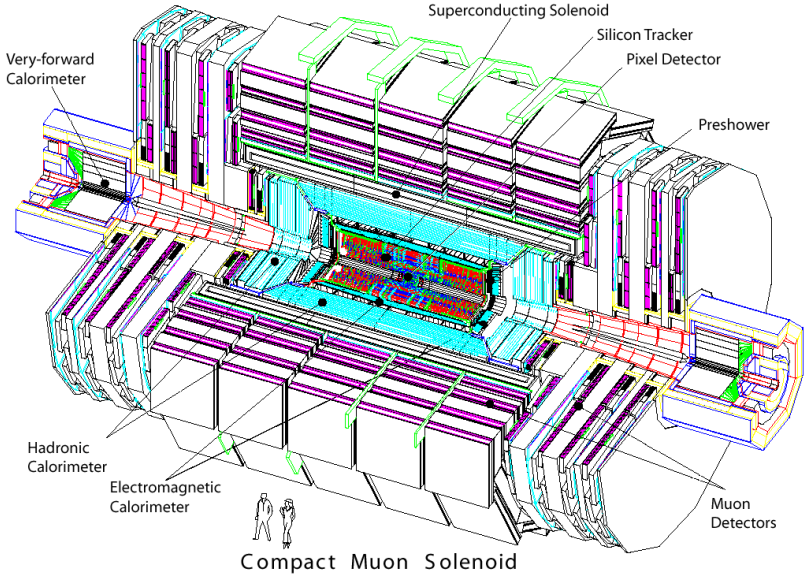
Collisions between protons are created where the beams intersect: 4 caverns are equipped with huge detectors. Two of these (ATLAS and CMS) are multi-purpose «electronic eyes» that try to detect everything that comes out of the collision



ATLAS



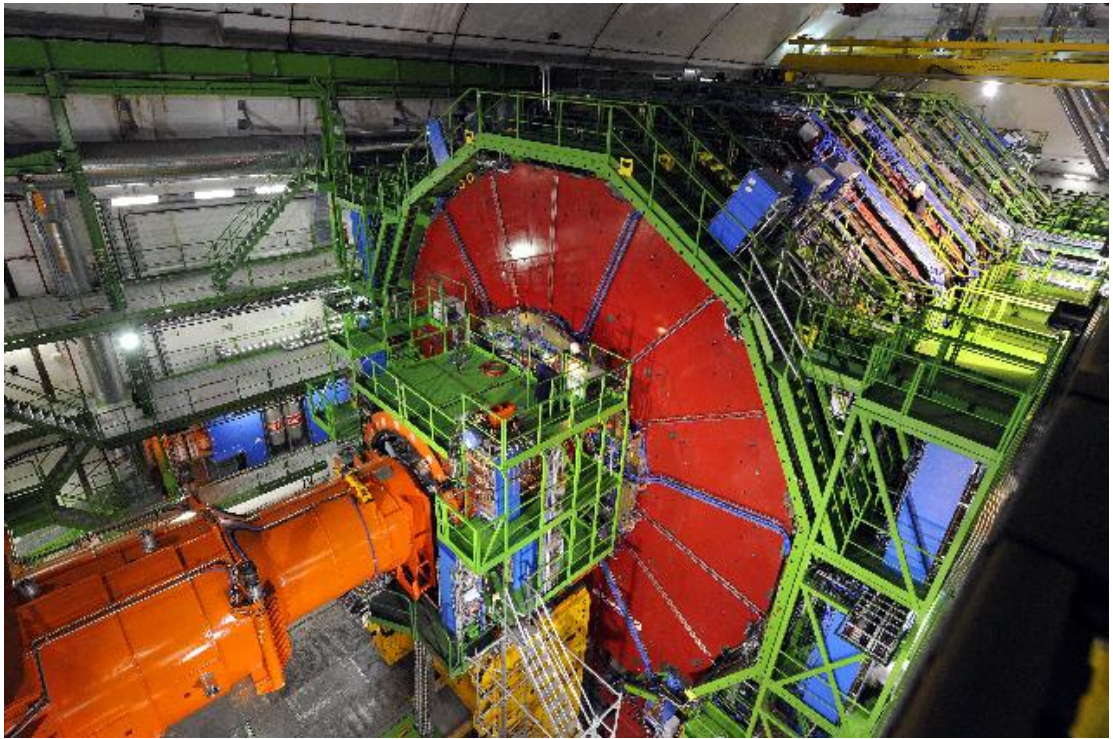
CMS

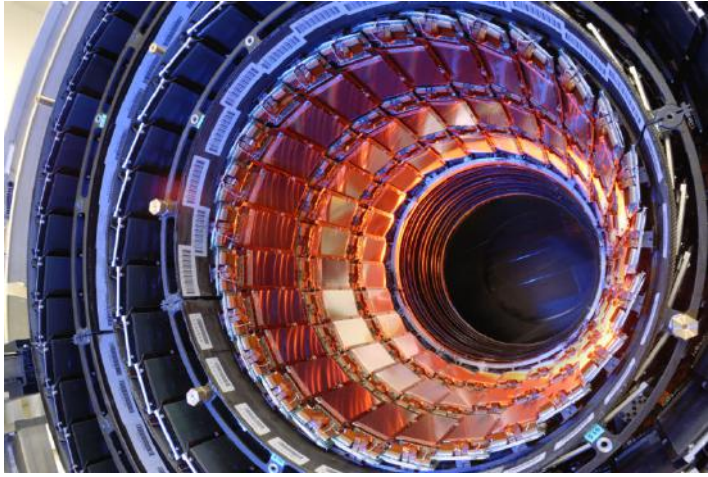


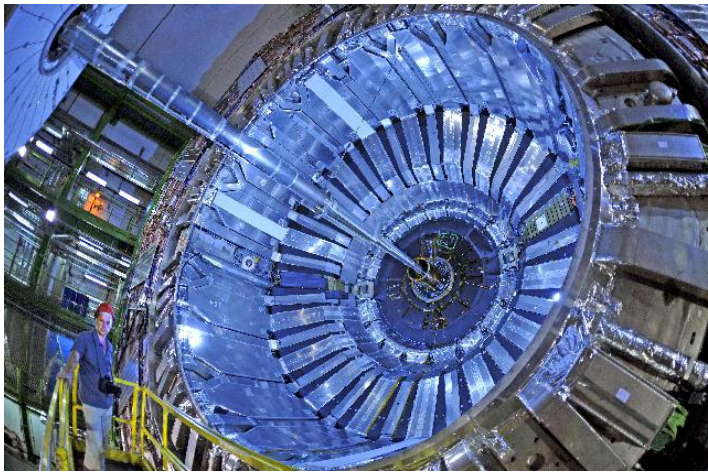
CMS

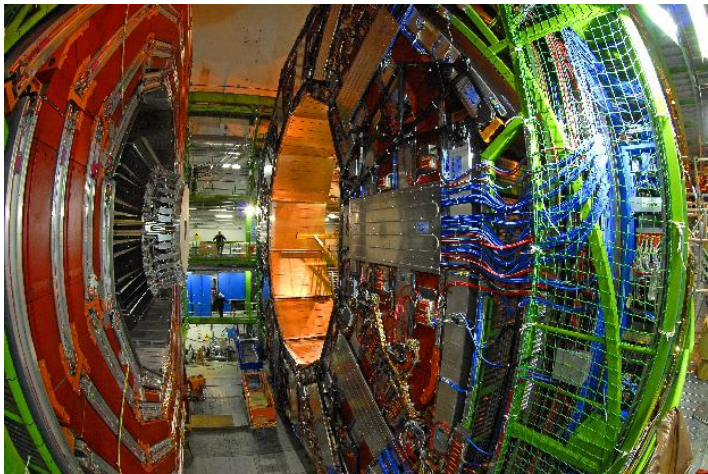
CMS (Compact Muon Solenoid) was built with the specific goal of finding the Higgs boson

- Along with ATLAS, it is arguably one of the most complex machines ever built by mankind
- Hundreds of millions collisions take place every second in its core, and each produces signals in tens of millions of electronic channels. These data are read out in real time and stored for offline analysis



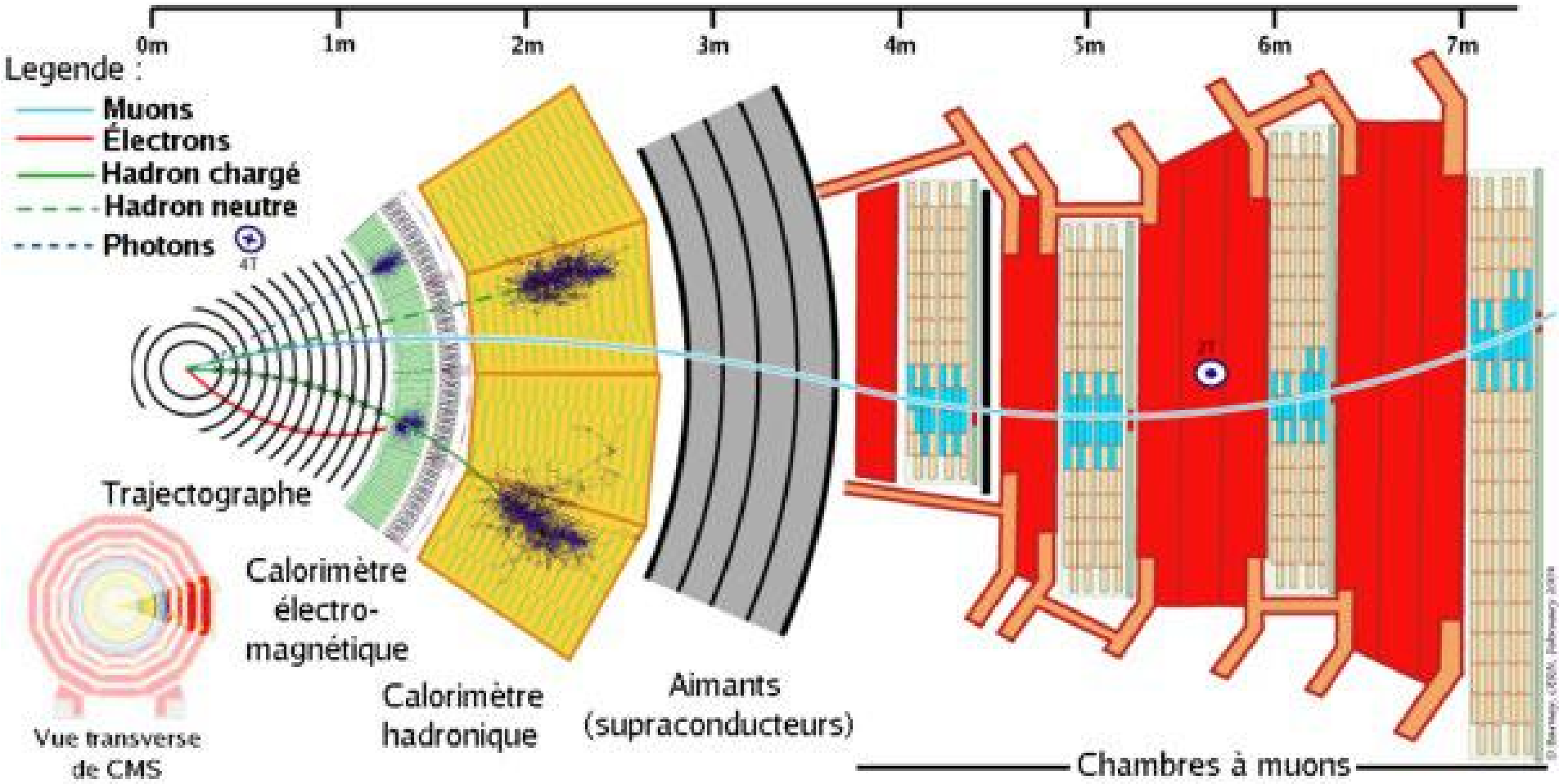






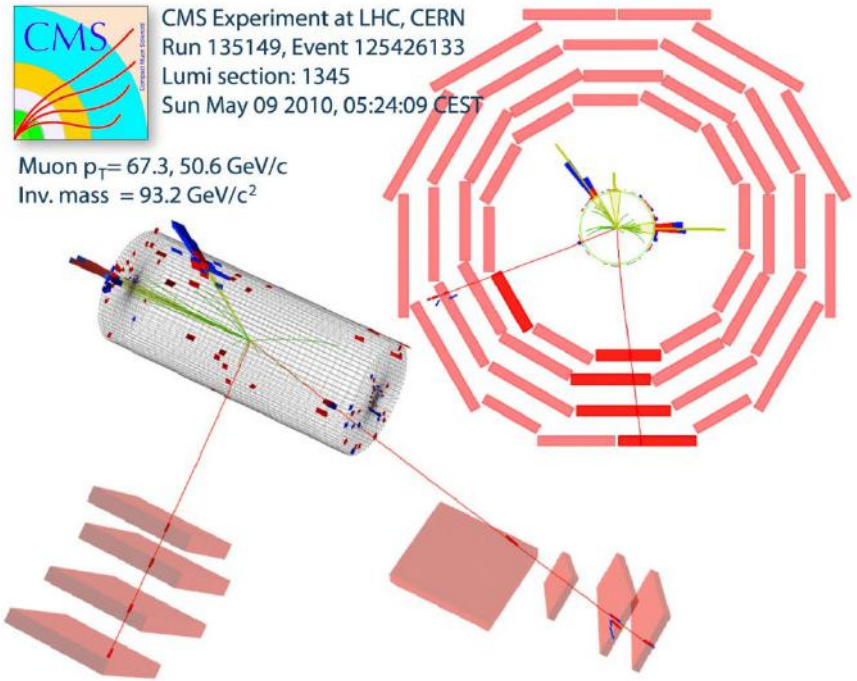
How we detect particles

Charged particles are detected in the **tracker**, through the ionization they leave in 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

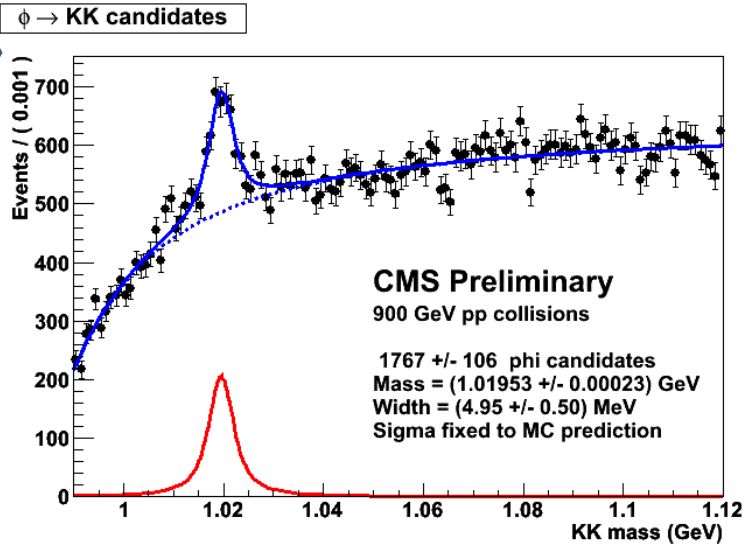


How we see a collision

A reconstruction of the O(10M) electronic signals provides us a «view» of the created objects: using their characteristics we build O(100) high-level variables which we compare to theoretical models after a further compression (usually into a 1-dim test statistic) → then we do measurements and inference



This is a huge dimensionality reduction…



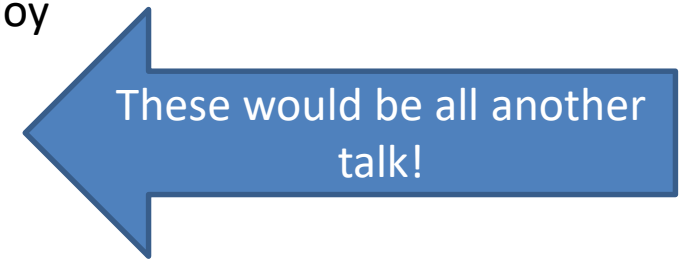
What we do with it

- We have a theory that allows us to calculate predicted probabilities for the possible physics processes, to extreme accuracy– but **we believe it is incomplete and to some extent unsatisfactory**.

- So we look for new physics processes: things that the standard model does not predict
  - New matter particles
  - New force carriers
- We also measure with precision known processes, in the attempt of finding a significant difference with model calculations

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

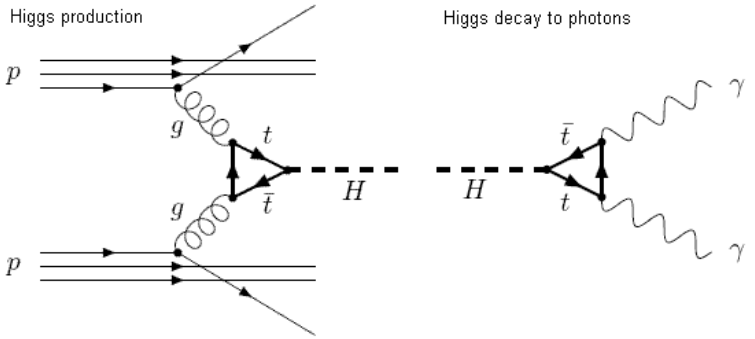
- In our analyses of the data we also frequently employ
  - Unfolding techniques
  - Machine-learning-driven dimensionality reduction
  - Goodness-of-fit tests

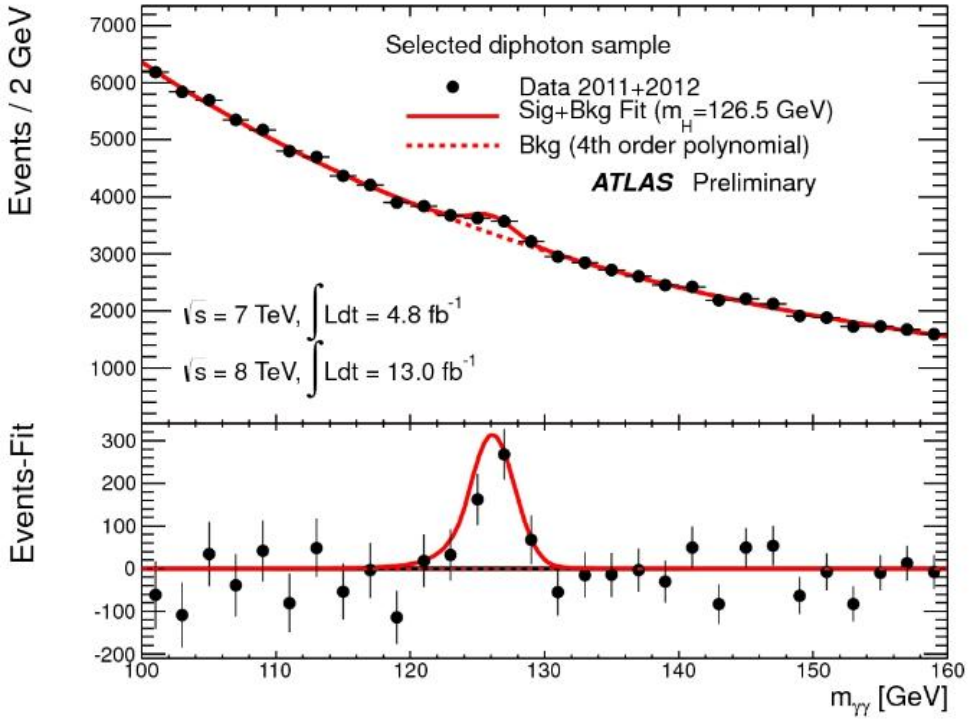


Example: new particle searches

The typical search for a new particle involves a model which predicts it

- Monte Carlo generators use the model 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
- Typically we attempt to reconstruct the **particle mass** from the measured features. As mass is a unique attribute of the particle, a histogram may then display a narrow bump on a smooth background
- A test of hypotheses allows to derive **$p(data|H_0)$**
  - More on that later

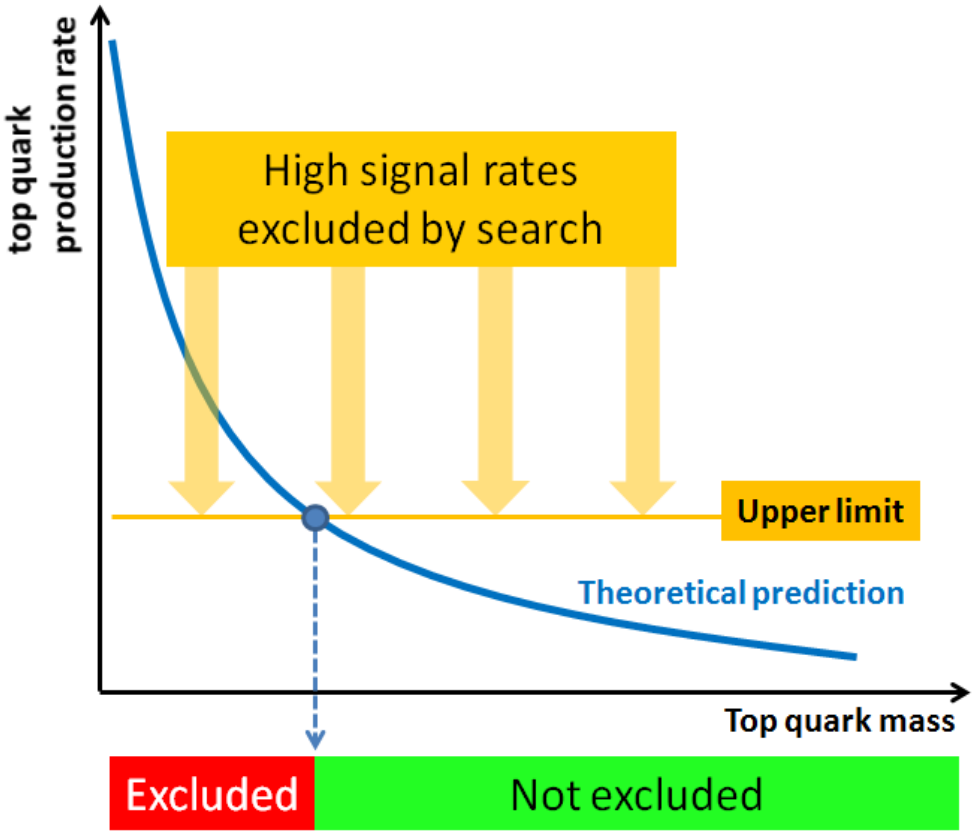




And what if there is no signal ?

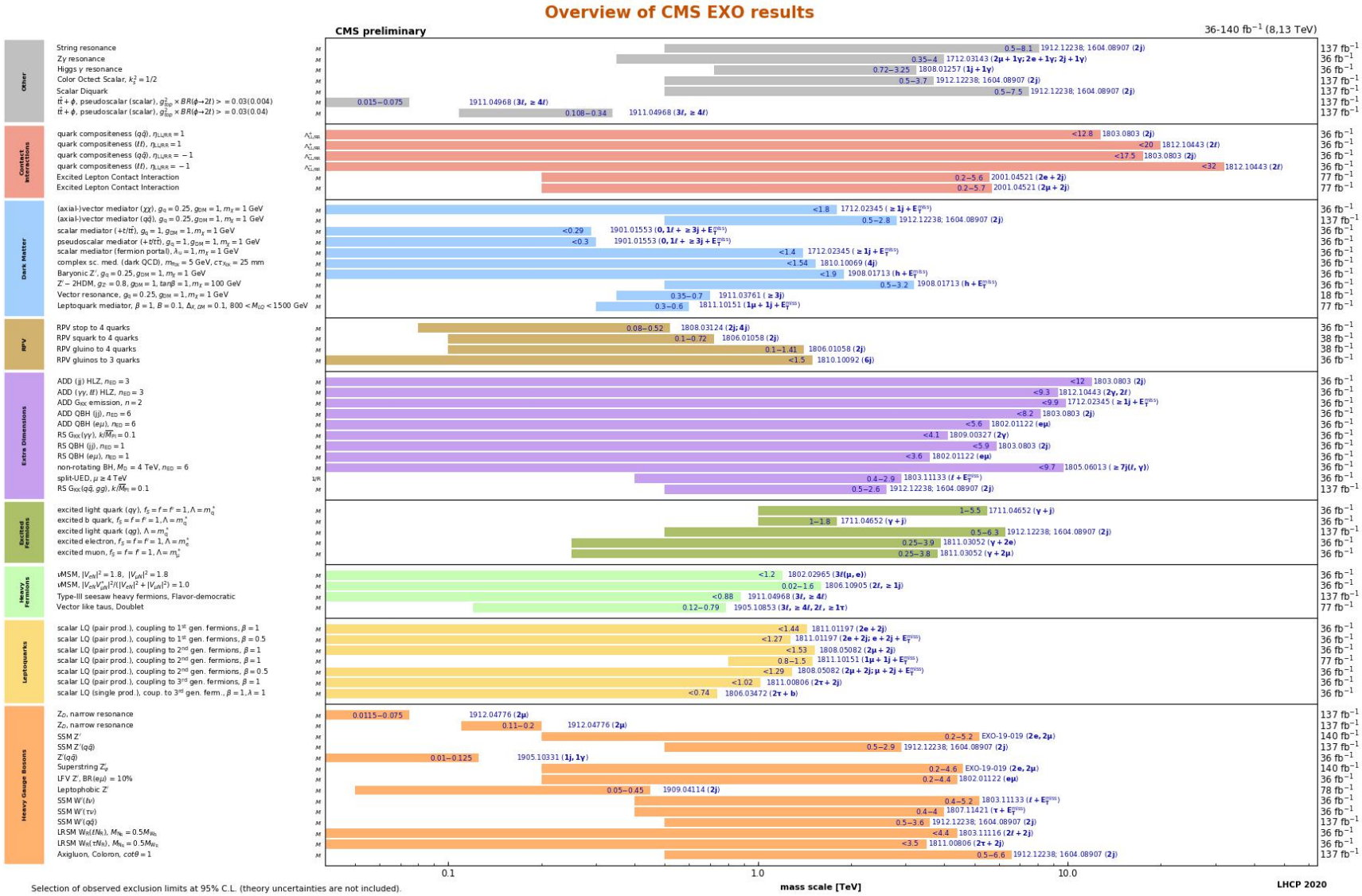
If we do not see a signal we can **exclude the new physics model**

- More often the model is composite, so we exclude a range of values of the relevant nuisance parameter
  - Often this is, again, 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, even a lower mass limit is useful information, **worth a publication!**

CMS Limits on exotic particles circa 2020



Neyman's Confidence interval recipe

Let us review the original recipe for frequentist CIs…

We 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|μ)***

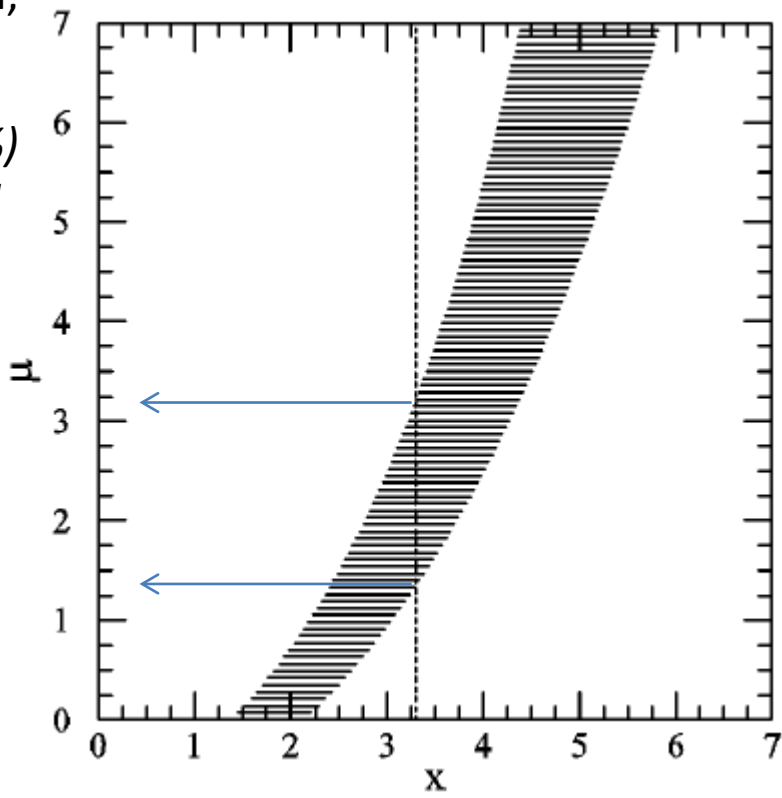
- *We also choose a Type-I error rate α (e.g. 32%, or 5%)*
- For each μ, we draw a horizontal *acceptance interval* $[x_1,x_2]$ *such that*

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

  There are infinitely many ways of doing this*: an ordering principle is required to well-define*
  - *for upper limits, integrate the pdf from x to inf*
  - *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.*



This procedure guarantees coverage

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
- We instead try to avoid the introduction of **a subjective input** in our results
- Also note that we work with parameters that describe physical reality – we hate to speak of the probability of a physical constant having this or that value

(although we fancy a flutter now and then!)

physicsworld

mathematics and computation

- (…) **pace** of possible true values
  - This has brought back Ba…

MATHEMATICS AND COMPUTATION | ANALYSIS

Physicists who fancy a flutter

01 Dec 2006

The age-old practice of betting on science is alive and well among modern physicists. **Martin Griffiths** spoke to a few of the gambling fraternity

Indeed, particle physics seems to provide rich pickings for physics gamblers, and the Large Hadron Collider (LHC), which is due to switch on next year, is inspiring plenty of speculation. Tommaso Dorigo, a particle physicist at Padova University in Italy, has bet $1000 on his weblog (dorigo.wordpress.com) that no physics "beyond the Standard Model" will be discovered at the LHC by the end of 2010. The bet was taken up by fellow particle physicist Gordon Watts and by string theorist Jacques Distler.

Long bets

But why do physicists like to gamble? In Dorigo's case, it acts as a sort of insurance policy: although he thinks he will win the bet, he says he would be much happier if he had to pay out.

Coverage, or the Lack Thereof

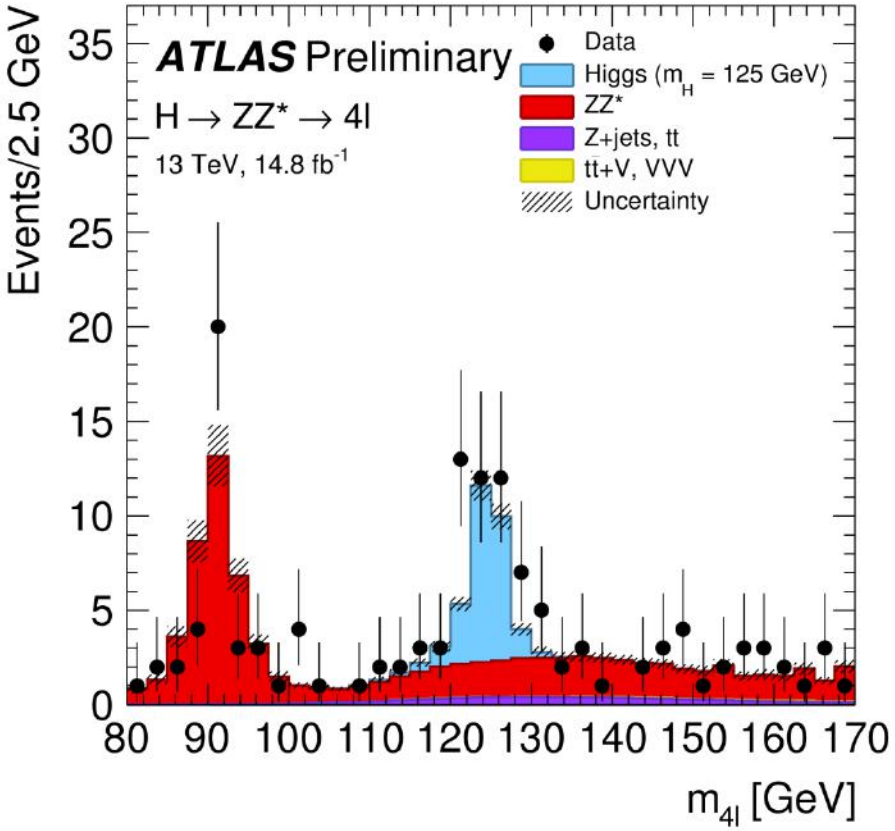
To see how we're affectionate with coverage, but we also are likely to neglect it, let us consider a typical HEP graph: event counts in a mass histogram, with sqrt(N) bars

As you see, data (**black points**) get compared to models (full histograms). Funnily, physicists attach uncertainty bars to the points: these only refer to the MLE of the rate in the bin, which of course is equal to the observed count.

What are those uncertainty bars supposed to mean, anyway? They report central intervals that "cover" at 68.3%.

But do they ?

Alas, usually they don't, as the Gaussian approximation for the Poisson distribution breaks down quite miserably for small N



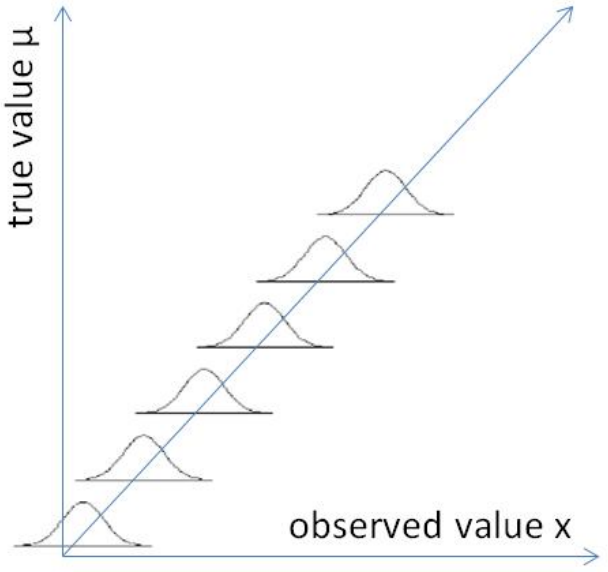
Of course, a solution exists: it was obtained in the fifites by Garwood, who used Neyman's construction for the Poisson distribution

Where It Gets Murky

If the parameter you are measuring is **bounded** (e.g. a mass or a process rate, which are >0) Neyman's recipe needs a fix.

Take e.g. μ>0 measured by P(x|μ) = N(μ,1):

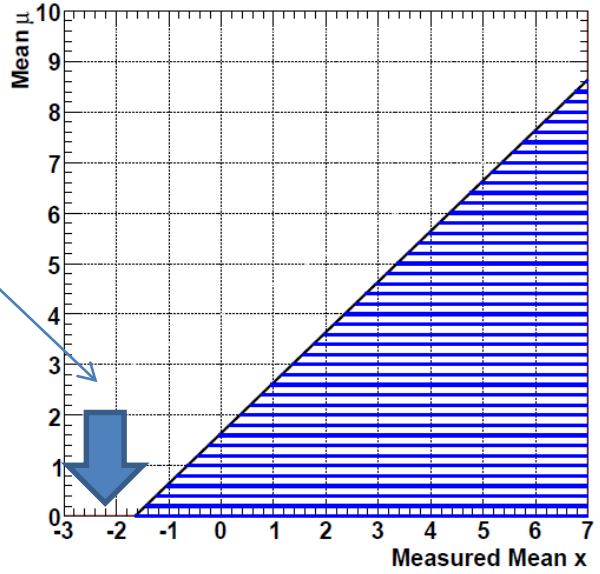
$$P(x|\mu) = \frac{1}{\sqrt{2\pi}} \exp(-(x-\mu)^2/2)$$



The classical method for UL at α=0.05 produces upper limit **μ<x+1.64σ**

- for x<-1.64 this results in the **empty set**!, in violation of one of Neyman's own demands (confidence set does not contains empty sets)

→ «*what do you do when you know you're in the wrong 5%*» problem



Can it be fixed? Yes!

Is there general agreement on how to deal with it? No!

Bounded μ Problem: Proposed Solutions

The graph illustrates various choices for confidence belts one can construct for the bounded parameter problem

The most principled among classical constructions is the one provided by Feldman and Cousins[1] in 1998

Bayesians have their own solution too

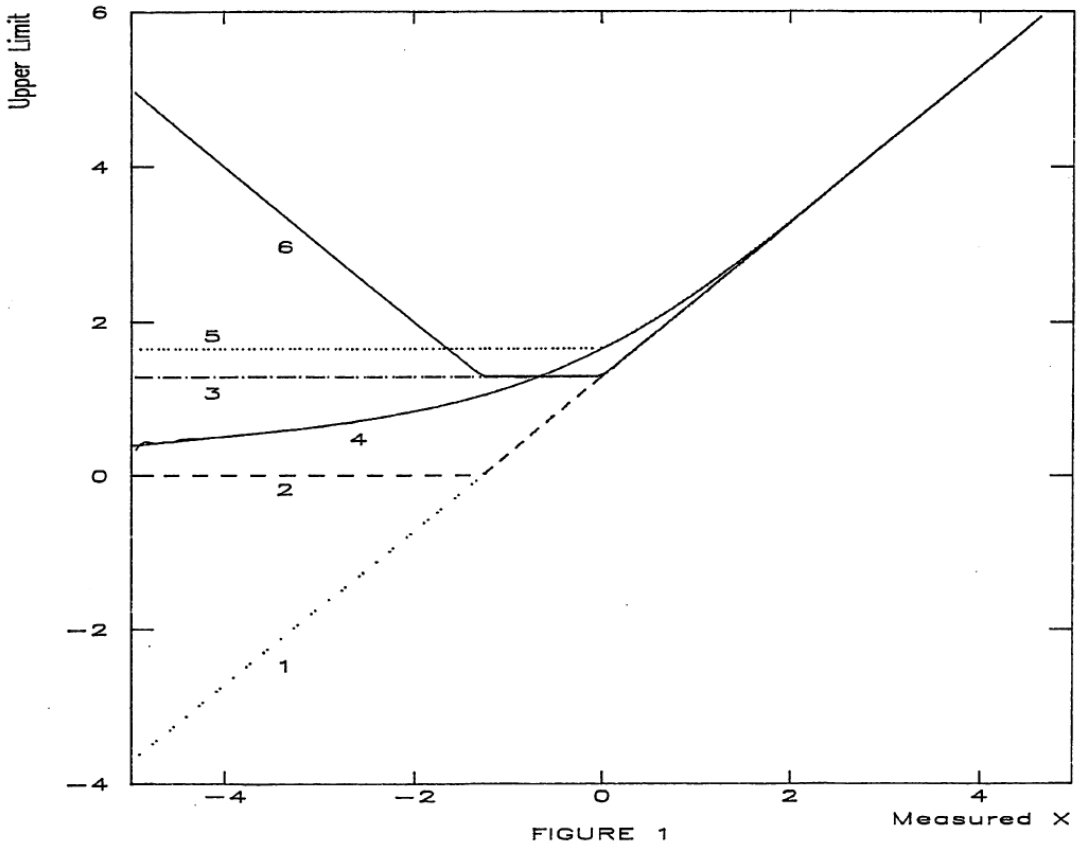


FIGURE 1

(1) Neyman's recipe for 90% upper limits: μ_{UL}=x+1.28.
(2) Hacked Neyman (cap at zero)
(4) Bayesian solution: step-function prior
(6) Mc Farlane's "loss of confidence"

Food for Thought: Relevant Subsets

Neyman's method applied to Gaussian measurement with known **σ** of a parameter with unknown positive mean **μ** yields upper limits at 95% CL in the form $\mu_{UL}=x+1.64\sigma$. **The procedure guarantees coverage, and yet...**

- Yet one can devise a betting strategy against it, at nominal 19:1 odds, using no more information than observed x, and **be guaranteed to win** in the long run!
  - How ? *Just choose a real constant k: bet that the interval does not cover when x<k, pass otherwise*.
  - For k<-1.64 this wins EVERY bet! For larger k, advantage is smaller but still >0

Surely then, the procedure is not making the best inference on the data?

Issue is discussed in paper by R. Cousins[2]

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 (now deprecated): "*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

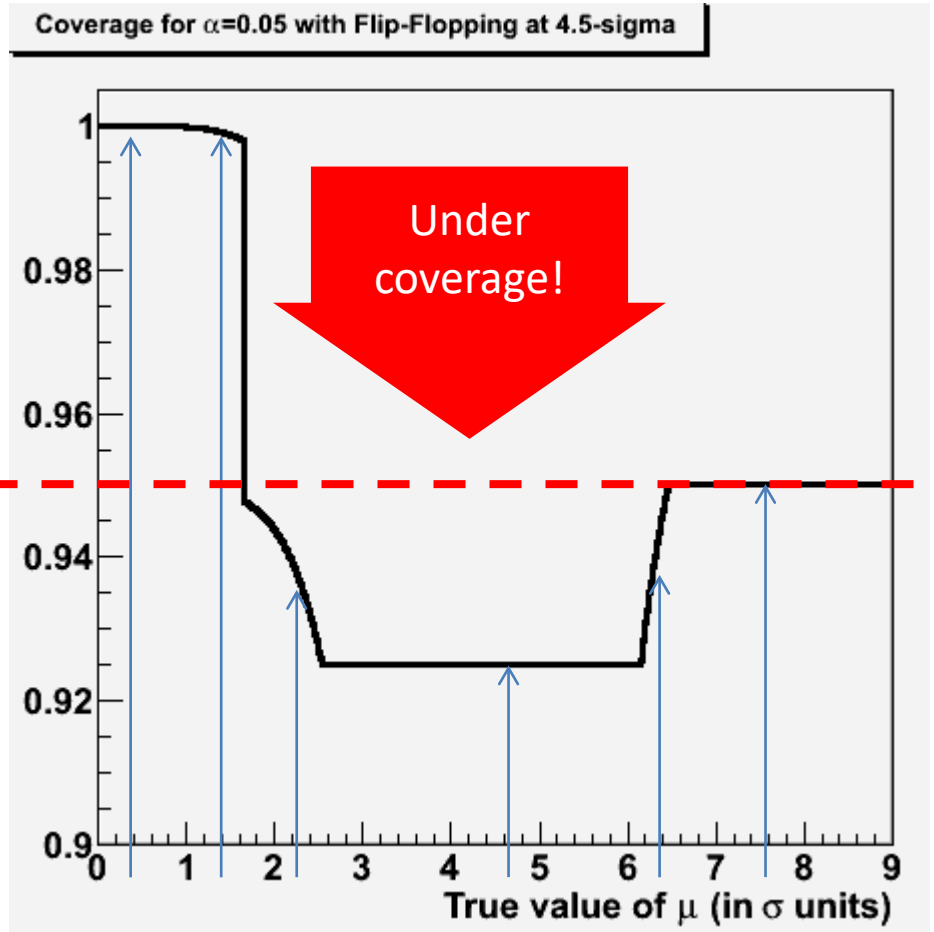
Suppose e.g. that we take μ_{UL} = max(x,0) +1.28 at 90%CL for the Gaussian-resolution measurement of a non-negative μ

- 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]

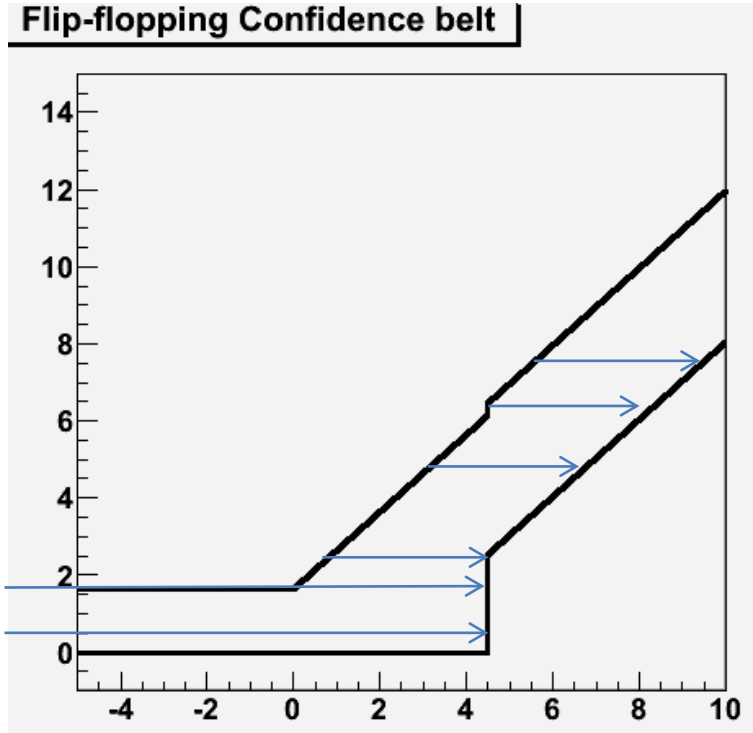
- **This undercovers!** (see next slide)

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 [1], based on a likelihood-ratio ordering of the acceptance intervals.
The FC technique is widely used in HEP



Statistical significance: What we mean

By Statistical significance we mean a way to report the probability that an experiment obtains data at least as discrepant as those actually observed, under a given null H_0

- In physics H_0 usually describes the currently accepted and established theory (but there are exceptions).
- Given some data X and a suitable test statistic T 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.

  p can always 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**" for insiders.

Notes

The alert observer will no doubt notice a few facts:

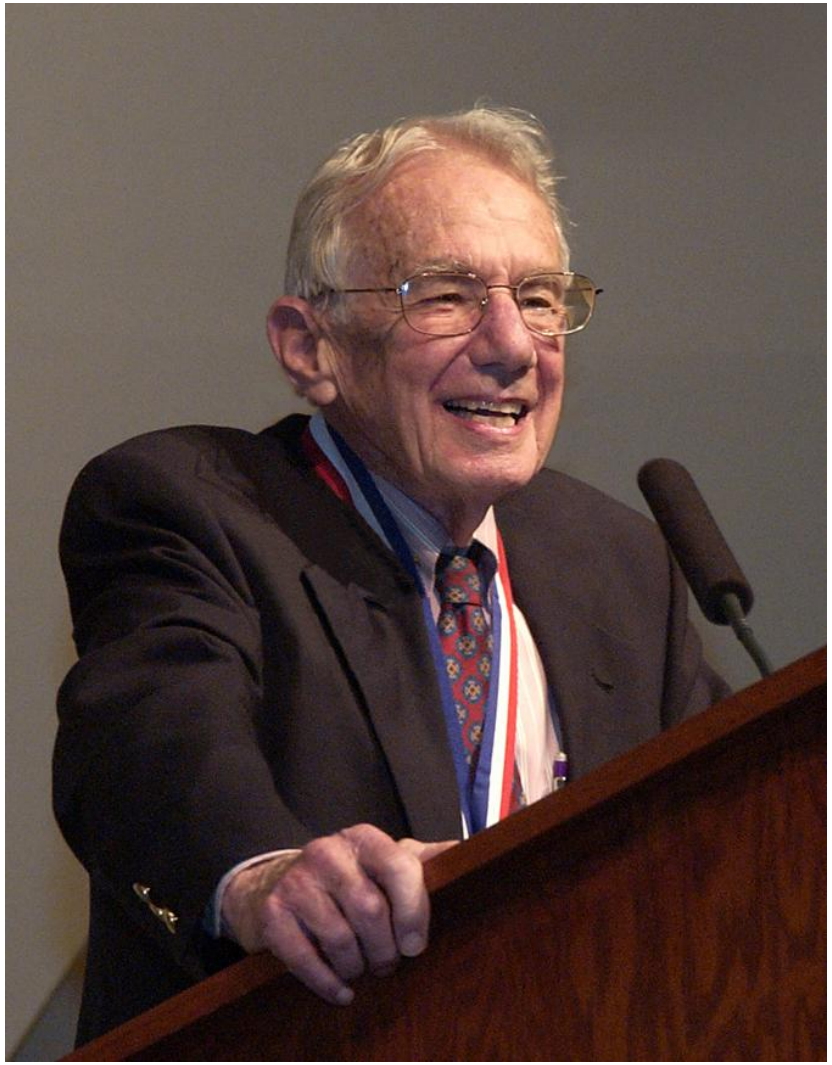
- the convention is to use a "one-tailed" Gaussian: we do not consider departures of x from the mean in the *un-interesting direction*
  - Hence "negative significances" are mathematically well defined, but we do not care about those

- the conversion of p into σ is fixed and independent of experimental detail. As such, using Nσ rather than p is **just a shortcut** to avoid handling numbers with many digits: 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"

- 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

- The "*probability of the data*" has no bearing on the concept, and 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)

The Birth of the Five-Sigma Criterion



Arthur H. Rosenfeld (Univ. Berkeley)

Careless particle hunters

- In 1968 A. Rosenfeld wrote a paper titled "*Are There Any Far-out Mesons or Baryons?*"**[3]**. In it, he demonstrated that the number of claims of discovery of those exotic particles published in scientific magazines agreed reasonably well with the **number of statistical fluctuations** that one would expect in the analyzed datasets.

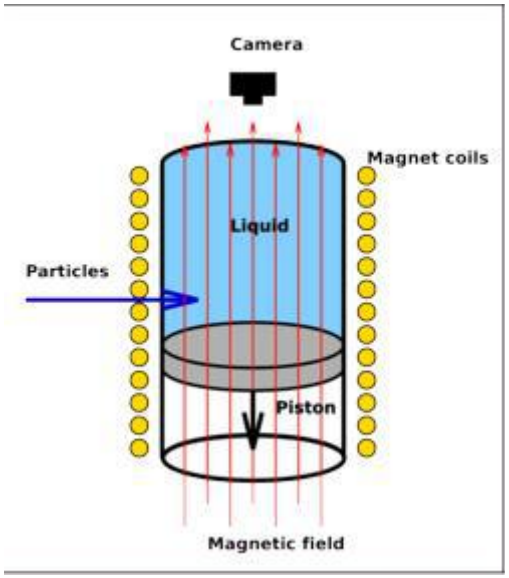
  ("Far-out hadrons" are hypothetical particles which can be defined as ones that do not fit in SU(3) multiplets. In 1968 quarks were not yet fully accepted as real entities, and the question of the existence of exotic hadrons was important.)

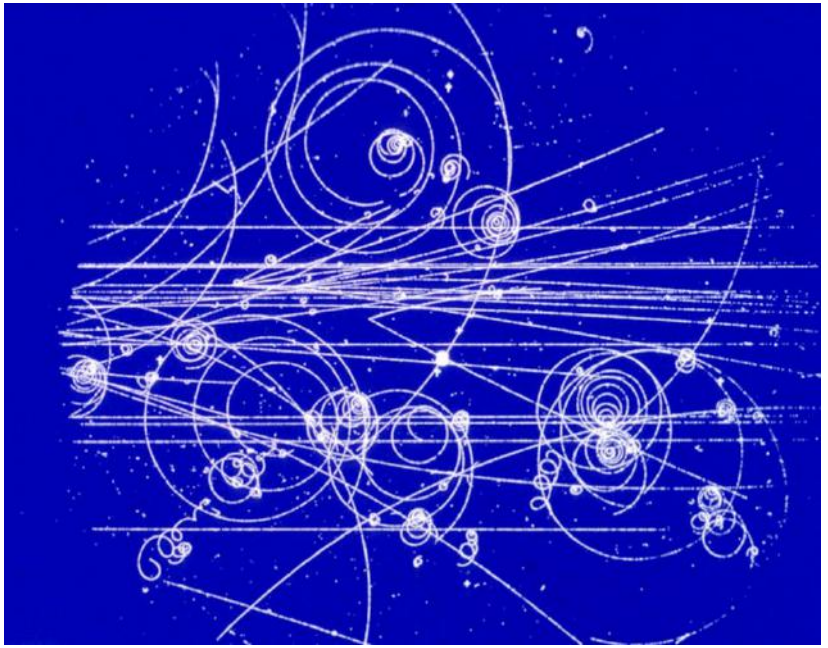
- Rosenfeld examined the literature and pointed his finger at large trial factors coming into play due to the massive use of combinations of observed particles to derive mass spectra containing potential resonances:

  > "[...] 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 (Our annual surveys also tells you that the U.S. measurement rate tends to double every two years, so things will get worse)."

Footnote:
Bubble chamber physics

A bubble chamber is a vessel filled with a gas in a phase of superheating. The passage of charged particles ionizes the gas and bubbles are formed along the path





By measuring the tracks in a magnetic field, one determines their momentum. The mass of a particle decaying into others can be determined from the daughters' momenta

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..."

(I will 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! A comparison with the literature in fact showed a correspondence of his eyeballed estimate with the number of unconfirmed new particle claims.

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.**"

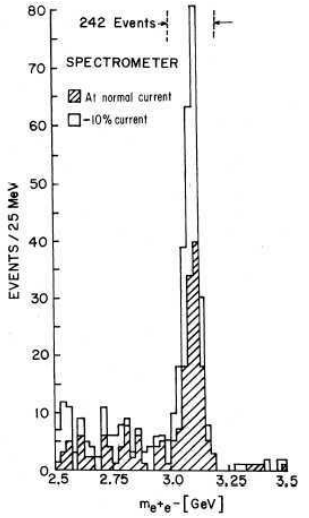
What 5σ May Do For You

- Setting the bar at 5σ for a discovery claim undoubtedly removes the large majority of spurious signals due to statistical fluctuations
- Nowadays we call this "**LEE**", for "look-elsewhere effect".
- The other reason at the roots of the establishment of a high threshold for significance has been 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 is the rationale behind the 5σ criterion

Still, the criterion has no basis in professional statistics literature, and is considered **totally arbitrary** by statisticians, no less than the 5% threshold often used for the type-I error rate of research in medicine, biology, social sciences, *et cetera*.

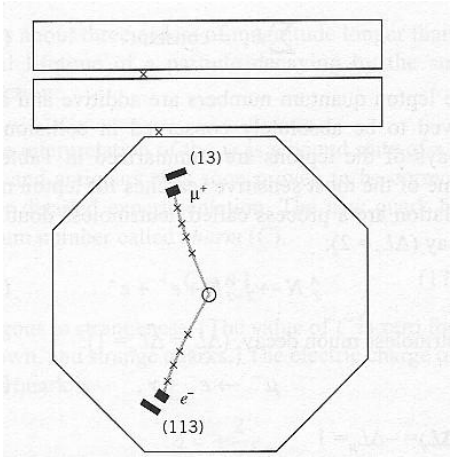
How 5σ Became a Standard in HEP: 1 - the Seventies



In the seventies the gradual consolidation of the SM shifted the focus from random bump hunting to more targeted searches

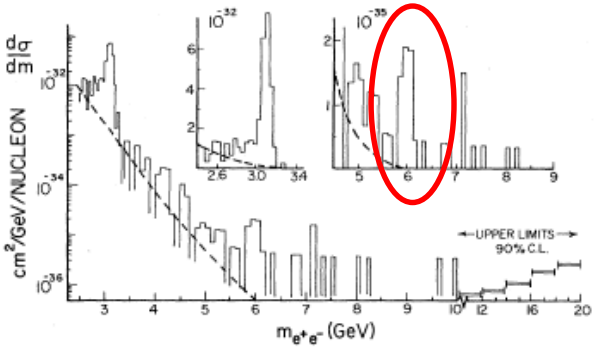
Let us check a few important searches to understand how the 5σ criterion gradually became a standard

- **The J/ψ discovery** (1974): no question of significance – the bumps were too big to call for fiddling with hypothesis tests



- **The τ discovery** (1975-1977): no mention of significances for the observed excess of (eμ) events; rather a very long debate on possible 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.*"



In a footnote 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 20x trials factor.

The Real Upsilon

The Upsilon discovery (1977): burned by their Oops-Leon, the authors waited more patiently for more data after seeing a promising 3σ peak at 9.5 GeV

- They did many statistical tests to account for the trials factor
- 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"

Nov 19th 1976

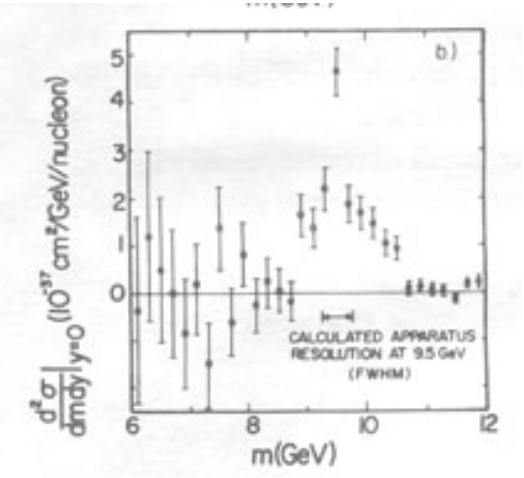
I determined this factor by monte carlo. I threw 30 events over 100 bins (expectation is 2 for 6 bins) and searched for clusters of 10 in 6 bins. I found 15 successes in 40000 trys or CL = 3.75×10^{-4}. The poisson probability for ≥10 for an expectation of 2 is 1.94×10^{-5}. This bin counting factor is 19.3. JKP 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 fits 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

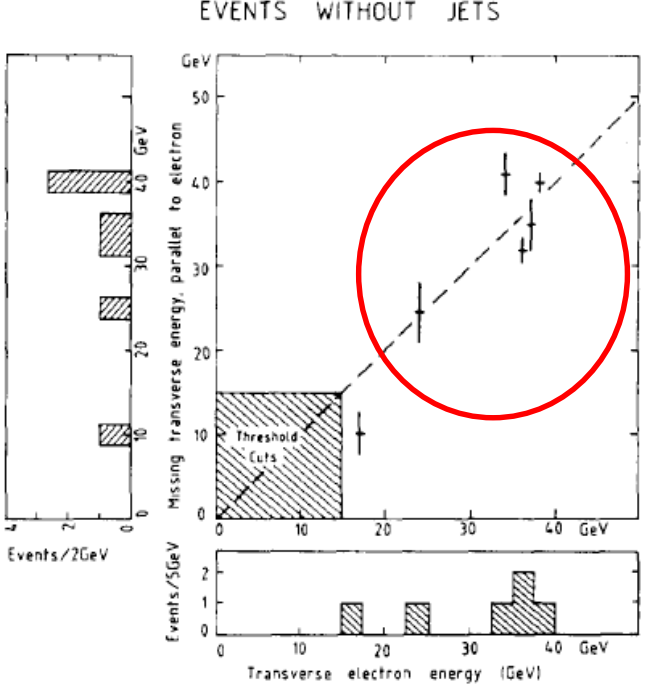
The W and Z Bosons

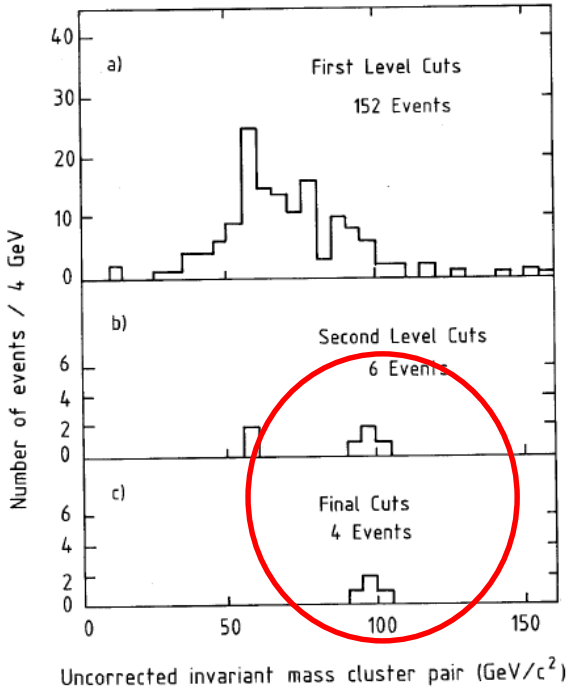
The 1983 W discovery was announced based on 6 events with all the required features

- No statistical analysis is discussed in the discovery paper, 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; theory prediction for W mass (82+-2 GeV) was matched well by the measurement (81+-5 GeV).

The Z was discovered shortly thereafter, with an official CERN announcement based on 4 events

- Also for the Z no trials factor was applicable
- No mention of statistical checks in the paper, except notes that backgrounds were negligible





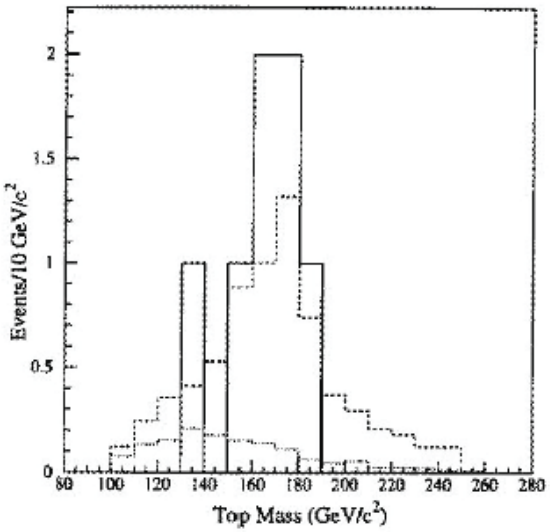
The Top Quark Discovery

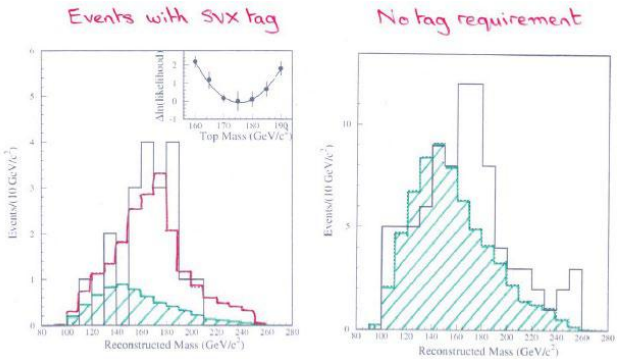
- In 1994 the CDF experiment had a serious counting excess (2.7σ) in b-tagged single-lepton and dilepton datasets, plus a mass peak at a value compatible with theory predictions
  - the mass peak, or corresponding kinematic evidence, was over 3σ by itself

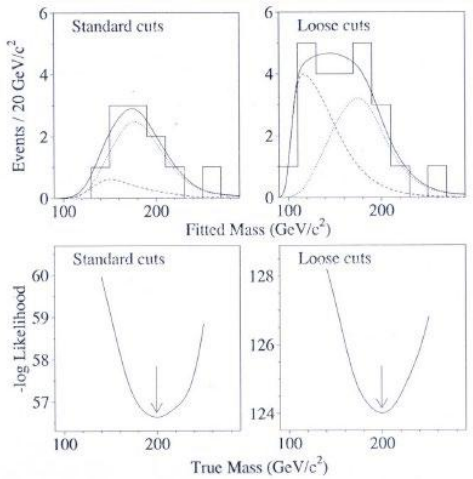
  The paper describing the analysis (120-pages long) spoke of "**evidence**" for top quark production

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

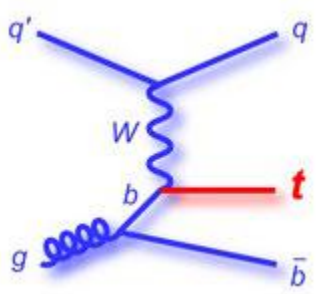
  The top quark was thus the first particle discovered by a willful application of the "5σ" criterion





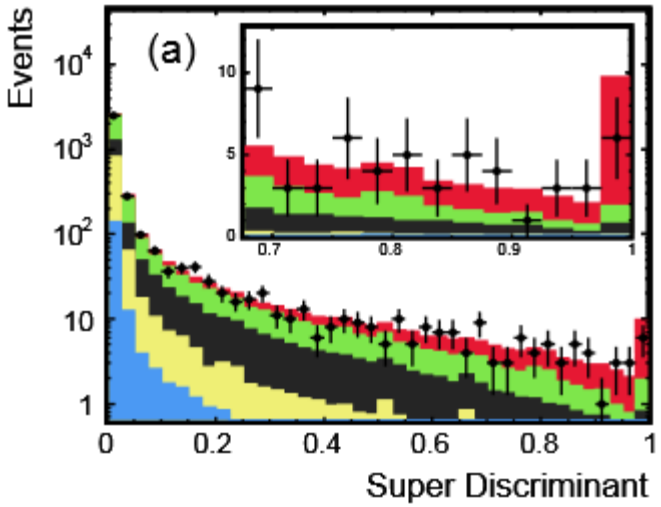


Following the Top Quark…

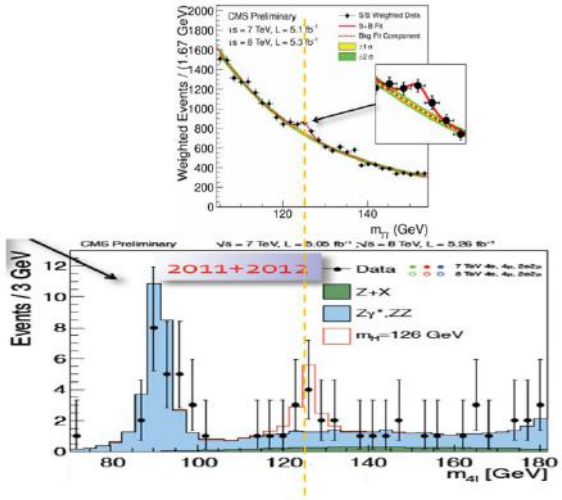


- Since 1995, the requirement of a p-value below $3*10^{-7}$ slowly but steadily became a standard.
- Striking examples of searches that diligently waited for a 5-sigma effect before claiming discovery:

1) Single top quark production: harder to detect than strong pair-production processes; it took 14 more years to be seen. CDF and DZERO claimed observation in 2009, over clear 5-sigma effects, using MVA methods



2) In 2012 the **Higgs boson** was claimed by **ATLAS and CMS**. Note that the two experiments had coherent >3σ evidences in their data 6 months earlier, but the 5σ recipe was followed diligently.



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

A look into the Look-Elsewhere Effect

The above discussion clarifies that a 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, so we can correct p-values for it. 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[4] 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)+ μ s(m|M)**, with **μ** a signal strength parameter, **S(m)** the signal model, 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 hypothesis (only background), 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, 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(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.

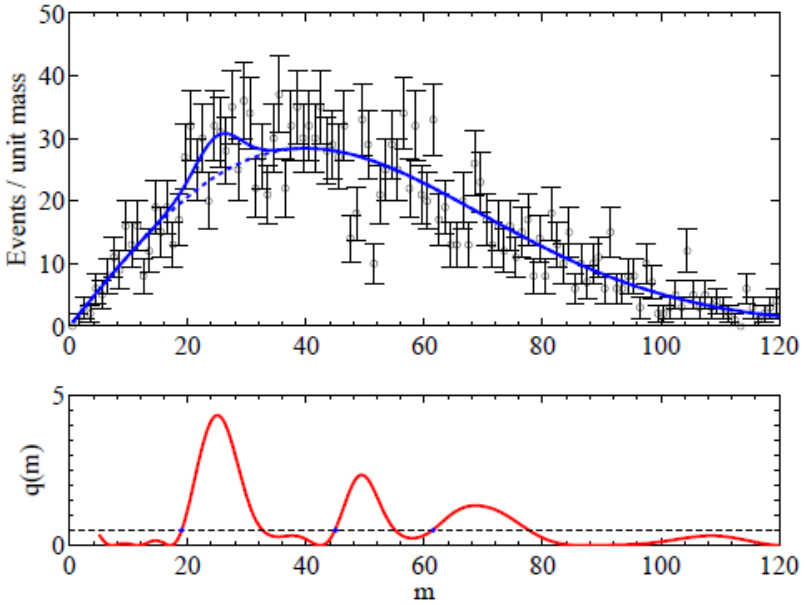
Local minima and upcrossings

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

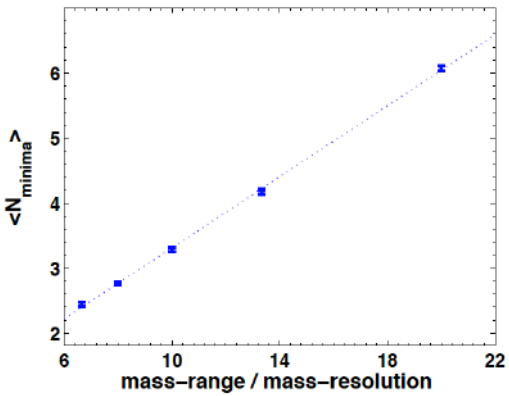
The number of times that the test statistic (below, the likelihood ratio between H_1 and H_0) crosses some reference line can be used to estimate the trials factor. One estimates the global p-value with the average number N_0 of upcrossings from a minimal value of the q_0 test statistic (for which $p=p_0$)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

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



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 might 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 at the 1-sigma level the typical range of variation of an observable due to the imprecise knowledge of a nuisance parameter; 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.

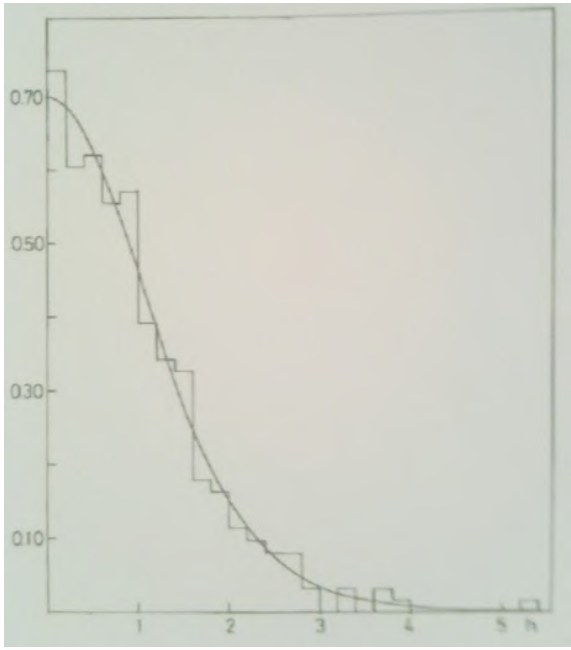
A study of residuals

A study of the measurement of particle properties in 1975 revealed that residuals were **not Gaussian in fact**. Matts Roos *et al.* **[5]** 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 $S_{10}(h/1.11)$:

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

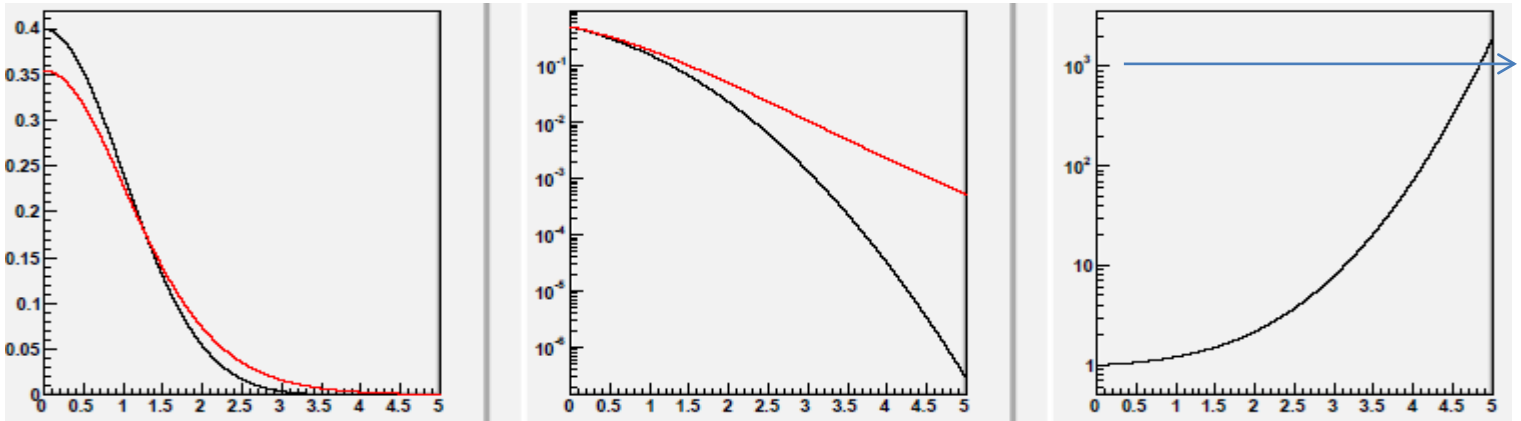
Of course, one cannot extrapolate to 5-sigma the behaviour observed by Roos and collaborators in the bulk of the distribution; however, one may consider this as evidence that the uncertainties evaluated in experimental HEP may have a significant non-Gaussian component

*The distribution of residuals of 306 measurements in **[5]***



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

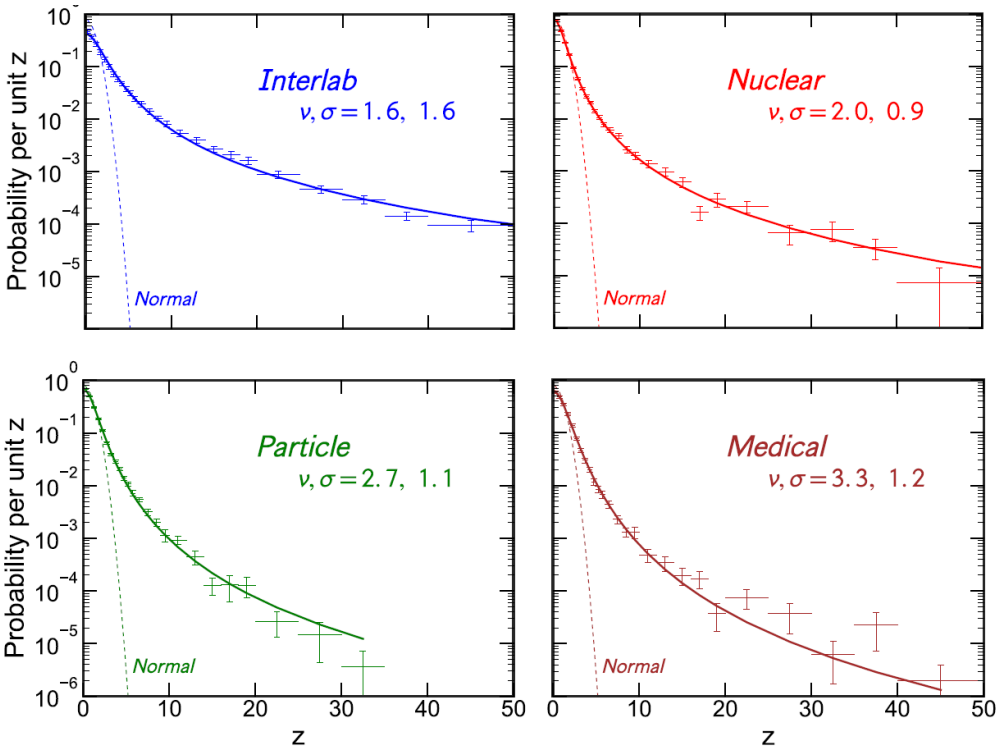
Left: 1-integral distributions of the two functions. Right: ratio of the 1-integral values as a function of z



x1000!

A Bigger, Meaner Study of Residuals

- David Bailey (U. Toronto) recently published an article[6] 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.

Going ~~Postal~~ Bayesian: The Jeffreys-Lindley Paradox

So what happens if one tries to move to Bayesian territory ?

Consider a null hypothesis, H_0, on which we base a strong belief. In physics **we do believe in our "point null"** – a theory valid for a specific value θ_0 of a parameter θ (say the photon mass being 0); in other sciences a true "point null" hardly exists

Comparing a point null $\theta=\theta_0$ to an alternative which has a continuous support for θ, we need to suitably encode this in a prior belief. Bayesians use a "probability mass" at $\theta=\theta_0$ for H_0.

The use of probability masses to encode priors for a **simple-vs-composite test** throws a monkey wrench in the Bayesian paradigm, 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[7] arises as frequentists and Bayesians 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

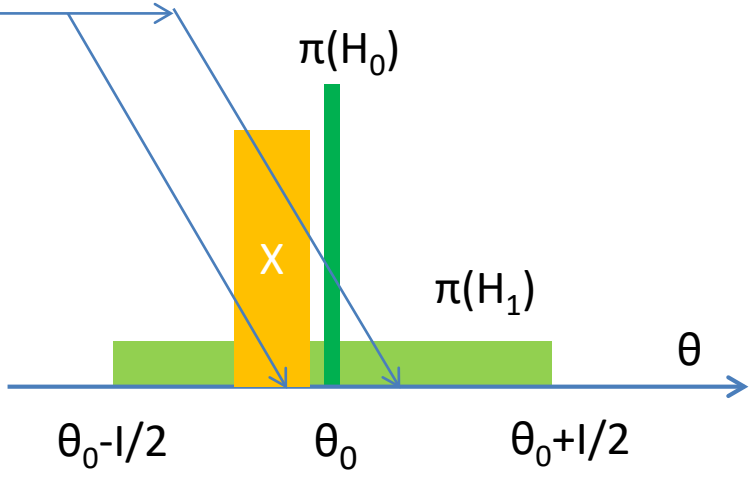
Take $X_1 \ldots X_n$ i.i.d. as $X_i|\theta \sim N(\theta,\sigma^2)$, and a prior belief on θ constituted by a mixture of a point mass **p** at θ_0 and **(1-p)** uniformly distributed in $[\theta_0\text{-}I/2, \theta_0\text{+}I/2]$.

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

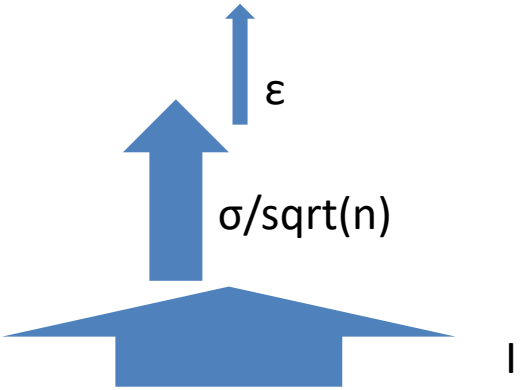
$$\bar{X} = \theta_0 \pm (\sigma/\sqrt{n}) z_{\alpha/2}$$

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

The **paradox** is that the **posterior probability that H_0 is true,** conditional on seeing data in the *critical region* (i.e. ones which exclude H_0 in a classical α-sized test) **approaches 1 (not α, NB!) as the sample size becomes arbitrarily large**.



As evidenced by R. Cousins[8], the paradox arises if there are three independent scales in the problem, **ε << σ/sqrt(n) << I**, i.e. the width of the point mass, the measurement uncertainty, and the scale **I** of the prior for the alternative hypothesis



Common situation in HEP!!

Notes on the JL Paradox

- The paradox has been 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***" **[9]**

- Still, the Bayesian approach offers no effective substitute to the p-value
  - **Bayes factors**, which describe by how much prior odds are modified by the data, do not factor out the subjectivity of the prior when the JLP applies: even asymptotically, they **retain a dependence on the scale of the prior of H_1.**

- In JLP debates, Bayesians have blamed the concept of a "point mass", or suggested n-dependent priors. Their final line of defence is to **argue that "the precise null" is never true.**
  - However, we do believe our point nulls in HEP and astro-HEP!!

There is a large body of literature on the subject. The issue is an active research topic and is **not resolved**.

→ The trouble of picking α in classical hypothesis testing is not automatically solved by moving to Bayesian territory.

So What to Do With 5σ ?

To summarize:

- the LEE can be estimated; experiments now routinely produce "global" and "local" p-values and Z-values
  - What is then the point of protecting from large LEE ?
  - Trial factor can be anything from 1 to enormous; a one-size-fits-all is hardly justified – **it is illogical to penalize an experiment for the LEE of others**
- Impact of systematic uncertainties varies widely; sometimes one has control samples to check their tails, but not always.
- The cost of a wrong claim, as backfiring of media hype, can vary dramatically
- Some claims are intrinsically less likely to be true («extraordinary claims require extraordinary evidence»)

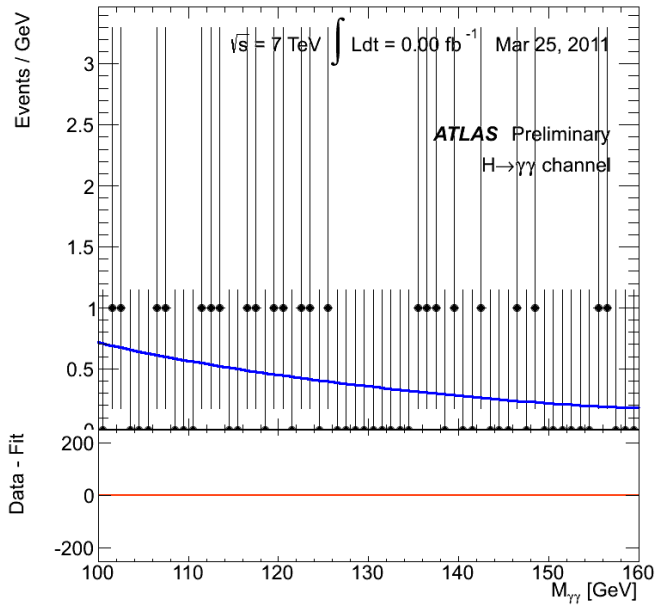
So why a fixed discovery threshold ?

- Any claim is anyway subject to criticism and independent verification, and the latter is always more rigorous when the claim is steeper and/or more important
- It is good to just have a "**reference value**" for the level of significance of the data – a «tradition», a useful standard

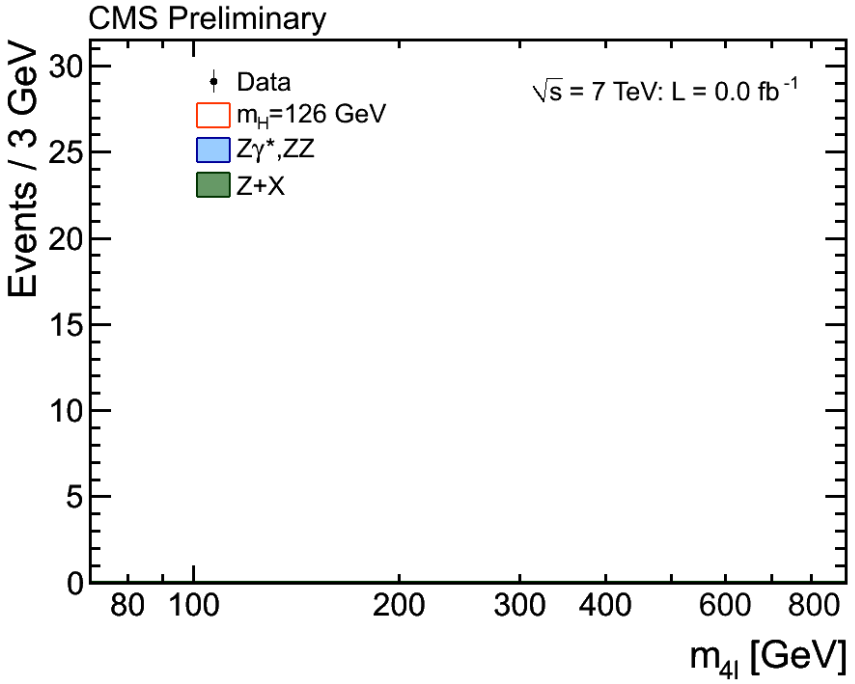
What it yielded

In July 2012, ATLAS and CMS both reported 5-sigma combined significances for observed departures from a model nobody believed (the SM without a Higgs), thereby establishing observation of the Higgs boson

ATLAS animation, H→γγ candidates



CMS animation, H→4 lepton candidates

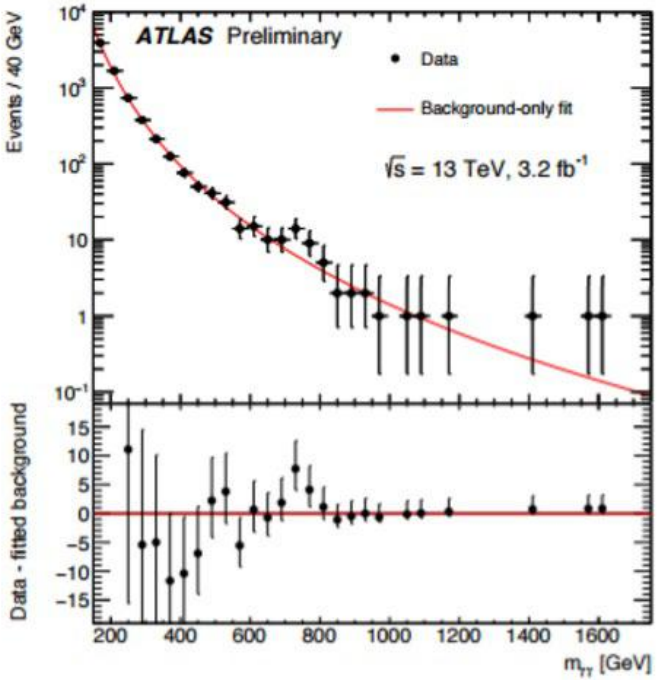


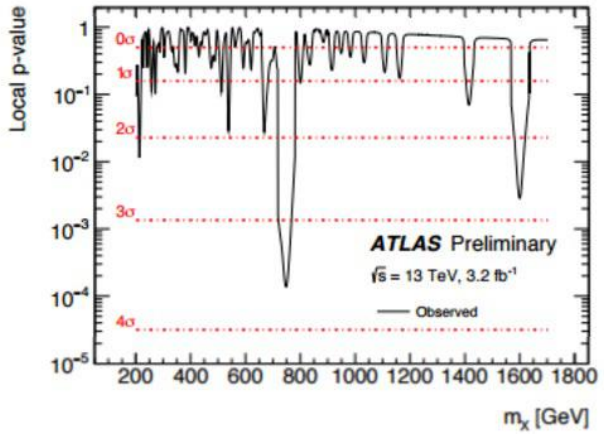
Rather than going through that now oldish, well-known result, let us see a more recent, **failed application** of the discussed hypo testing machinery: the «750 GeV diphoton affair»

The Case Of The Photon Pairs

In December 2015 ATLAS and CMS announced evidence for a 750 GeV particle decaying to photon pairs

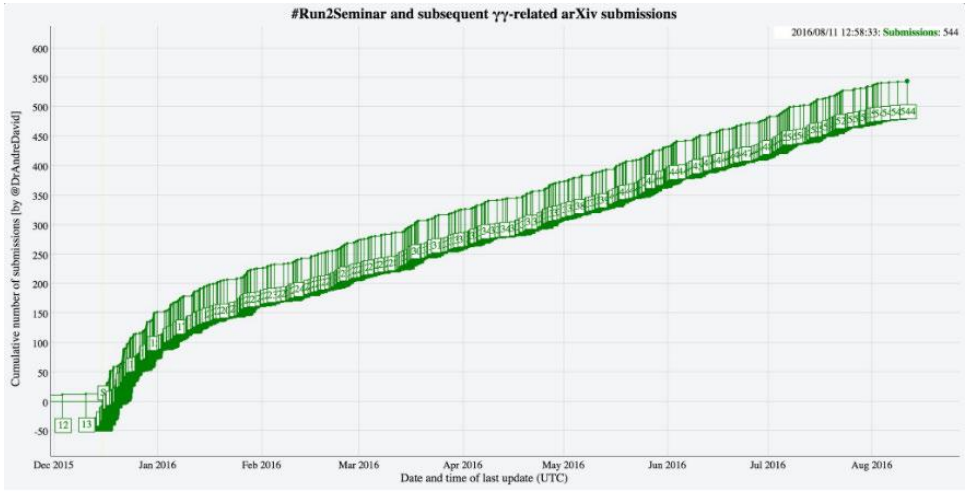
- Significance in the 4-sigma ballpark
  - ATLAS 3.6σ alone, CMS 2-sigmaish evidence
  - Conflicting evidence on width
- Theorists jumped at it, proposing interesting and less interesting scenarios to fit it in
- Experiments set out to search for it in other ways and with additional data





Phenomenologists' feeding frenzy

In the matter of 8 months the Cornell arxiv got flooded with over **550 new papers** that tried to explain the diphoton excesses of ATLAS and CMS



Bets were offered and accepted on the nature of the new particle, with various odds

In the process, we learned that finding new physics will not teach us much per se – one needs to then characterize it quite well to sort out what underlying theory can be responsible for it!

Some of the proposed explanations:

Two higgs doublets
Seesaw vectorlike fermions
Closed strings
Neutrino-catalyzed
Indirect signature of DM
Colorful resonances
Resonant sneutrino
SU(5) GUT
Inert scalar multiplet
Trinification
Dark left-right model
Vector leptoquarks
D3-brane
Deflected-anomaly SUSY breaking
Radion candidate
Squarkonium-Diquarkonium
R-parity violating SUSY
Gravitons in multi-warped scenario

750-GeV Bump Interpretation Summary

1 - It seems quicker to say what a 750 GeV bump cannot be:

Not the Lochness monster, which has an evident 3-bump structure

Not Mickey Mouse, who clearly has a non-Gaussian tail

2 – The signal clearly inspired the creativity of theorists, but it also forced them to work around the clock.
Best title in arXiv Preprint server for a while:

"**How the gamma-gamma Resonance Stole Christmas**"

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. It is remarkable (and probably also suboptimal) that some of the seminal studies addressing this issue come from physicists!
- In this talk I could only scratch the surface of some of the issues… The debates are 30-years long (but I know you statisticians have your own!)
- Statisticians have in some cases offered help to HEP experimentalists, yet more of it is welcome.
- And there is a large potential bonus – we publish hundreds of high-citation papers, and we cite the statistical techniques.
  - One example: the Feldman-Cousins paper, which is basically a rediscovery of 1.5 pages in Kendall & Stuart, has >3200 citations.
  - Another one: a paper on asymptotic formulae for likelihood ratio tests (G.Cowan et al., *Eur. Phys. J. C* 71 (2011) 1554) has >2800.

So…

So… Come and work with us!

&&

Thank you for your attention!

References

[1] G. Feldman and R. D. Cousins, "*A Unified Approach to the Classical Statistical Analysis of Small Signals*", Phys. Rev. D 57 (1998) 3873.

[2]] R. D. Cousins, "*Negatively Biased Relevant Subsets Induced by the Most-Powerful One-Sided Upper Confidence Limits for a Bounded Physical Parameter*", arXiv:1109.2023 (2011).

[3] 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.

[4] E. Gross and O. Vitells, "*Trials factors for the Look-Elsewhere Effect in High-Energy Physics*", arxiv:1005.1891v3 [physics.data-an](2010), Eur. Phys. J. C 70 (2010) 525.

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

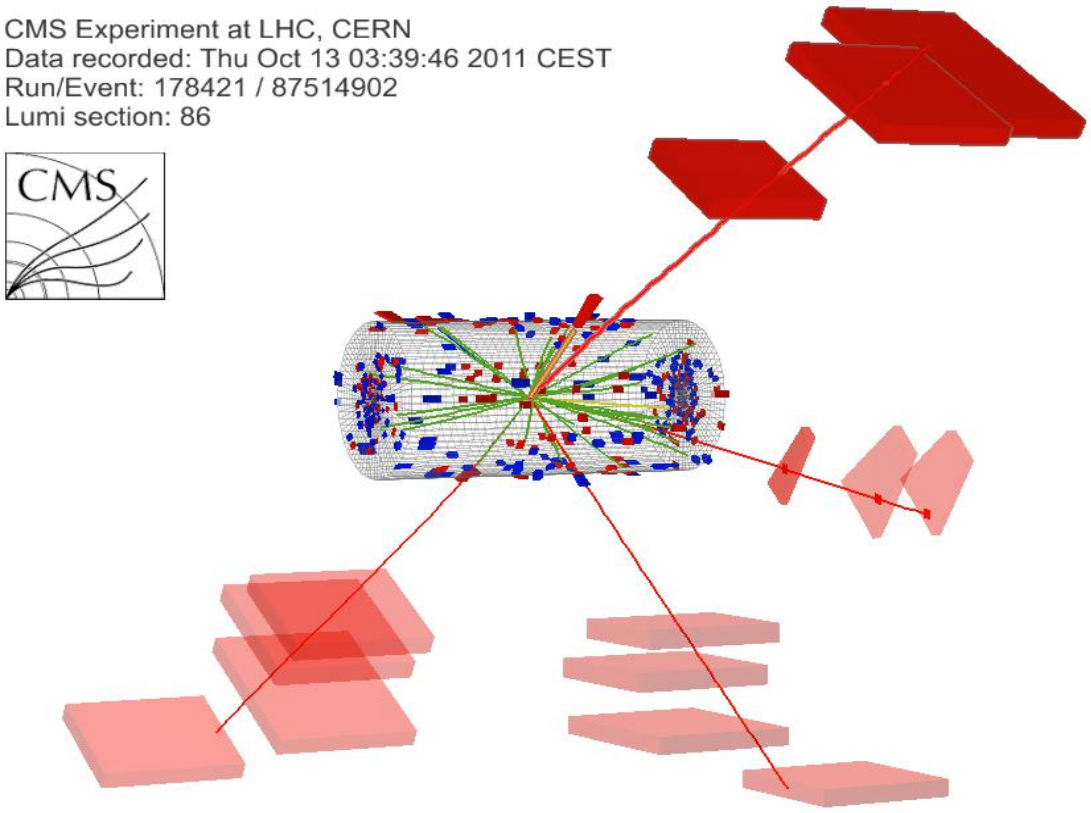
[6] D. Bailey, "*Not Normal: the uncertainties of scientific measurements*", ArXiv:1612.00778 [stat.AP] (2016), Royal Society Open Science 4 (2017) 160600.

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

[8] R. D. Cousins, "*The Jeffreys-Lindley Paradox and Discovery Criteria in High-Energy Physics*", arxiv:1310.3791v4; Synthese 194 (2017) 2, 395.

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

Higgs Discovery: a case study



Nuts and Bolts of Higgs Combination

1) One writes a global **likelihood function**, whose parameter of interest **μ** is called «signal strength modifier». If **s** and **b** denote signal and background, and **θ** is a vector of systematic uncertainties, one can generically write for a single channel:

$$\mathcal{L}(\text{data}\,|\,\mu,\theta) = \text{Poisson}\,(\,\text{data}\,|\,\mu\cdot s(\theta)+b(\theta)\,)\cdot p(\tilde{\theta}|\theta)$$

Note that **θ** has a "prior" coming from a hypothetical auxiliary measurement.
In the LHC combination of Higgs searches, nuisances are treated in a frequentist way by taking for them the likelihood which would have produced as posterior, given a flat prior, the PDF one believes the nuisance is distributed from.

In **L** one may combine many different search channels where a counting experiment is performed as the product of their Poisson factors:

$$\prod_i \frac{(\mu s_i + b_i)^{n_i}}{n_i!}\, e^{-\mu s_i - b_i}$$

or from a unbinned likelihood over **k** events, factors such as:

$$k^{-1}\prod_i (\mu S f_s(x_i) + B f_b(x_i)) \cdot e^{-(\mu S+B)}$$

2) One then constructs a profile likelihood test statistic $\mathbf{q_\mu}$ as

$$\tilde{q}_\mu = -2\ln\frac{\mathcal{L}(\text{data}|\mu,\hat{\theta}_\mu)}{\mathcal{L}(\text{data}|\hat{\mu},\hat{\theta})}$$

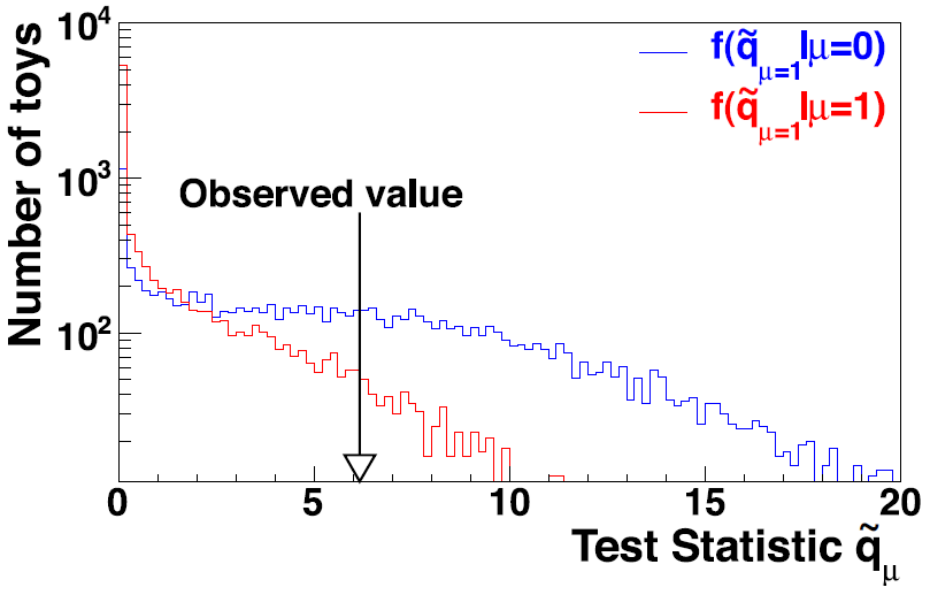
Note that the denominator has **L** computed with the values of $\boldsymbol{\mu}$^ and $\boldsymbol{\theta}$^ that globally maximize it, while the numerator has $\boldsymbol{\theta=\theta}$^$_{\boldsymbol{\mu}}$ computed as the conditional maximum likelihood estimate, given $\boldsymbol{\mu}$.

A constraint is posed on the MLE $\boldsymbol{\mu}$^ to be confined in $\mathbf{0 \le \mu}$^$\mathbf{\le \mu}$: this avoids negative solutions for the cross section, and ensures that best-fit values *above* the signal hypothesis $\boldsymbol{\mu}$ are not counted as evidence against it.

3) ML values $\boldsymbol{\theta}_{\boldsymbol{\mu}}$^ for H_1 and $\boldsymbol{\theta}_{\mathbf{0}}$^ for H_0 are then computed, given the data and $\boldsymbol{\mu=0}$ (bgr-only) or $\boldsymbol{\mu>0}$ hypotheses

4) Pseudo-data is generated for the two hypotheses, using the above ML estimates of the nuisance parameters. With the data, one constructs the pdf of the test statistic given a signal of strength μ (H_1) and μ=0 (H_0).

This recipe has good coverage properties.



5) With pseudo-data one can then compute the integrals defining p-values for the two hypotheses. For the signal plus background hypothesis H_1 one has

$$p_\mu = P(\tilde{q}_\mu \geq \tilde{q}_\mu^{obs} \,|\, \text{signal+background}) = \int_{\tilde{q}_\mu^{\text{obs}}}^{\infty} f(\tilde{q}_\mu|\mu, \hat{\theta}_\mu^{\text{obs}})\, d\tilde{q}_\mu$$

and for the null, background-only H_0 one has

$$1 - p_b = P(\tilde{q}_\mu \geq \tilde{q}_\mu^{obs} \,|\, \text{background-only}) = \int_{q_0^{\text{obs}}}^{\infty} f(\tilde{q}_\mu|0, \hat{\theta}_0^{\text{obs}})\, d\tilde{q}_\mu$$

6) Finally one **can compute the value called CL_s as**

$$CL_s = p_\mu/(1-p_b)$$

CL_s is thus a "modified" p-value, in the sense that it describes how likely it is that the value of test statistic is observed under the alternative hypothesis by also accounting for how likely the null is: the drawing incorrect inference based on extreme values of p_μ is "damped", and cases when one has no real discriminating power, approaching the limit $f(q|\mu)=f(q|0)$, are prevented from excluding the alternate hypothesis.

7) We can then **exclude H_1 when $CL_s < \alpha$**. In the case of Higgs searches, all mass hypotheses $H_1(M)$ for which $CL_s<0.05$ are said to be excluded (one would rather call them "disfavoured"…)

Significance in the Higgs search

To test for the significance of an excess of events, given a M_h hypothesis, one uses the bgr-only hypothesis and constructs a modified version of the q test statistic:

$$q_0 = -2 \ln \frac{\mathcal{L}(\text{data}|0, \hat{\theta}_0)}{\mathcal{L}(\text{data}|\hat{\mu}, \hat{\theta})} \quad \text{and } \hat{\mu} \geq 0.$$

- This time we are testing any $\mu>0$ versus the H_0 hypothesis. One builds the distribution $f(q_0|0,\theta_0\text{^}^{obs})$ by generating pseudo-data, and derives a p-value corresponding to a given observation as

$$p_0 = P(q_0 \geq q_0^{obs}) = \int_{q_0^{obs}}^{\infty} f(q_0|0, \hat{\theta}_0^{\text{obs}})\, dq_0.$$

One then converts p into Z using the relation

$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)\, dx = \frac{1}{2} P_{\chi_1^2}(Z^2)$$

where p_{χ}^2 is the survival function for the 1-dof χ^2.

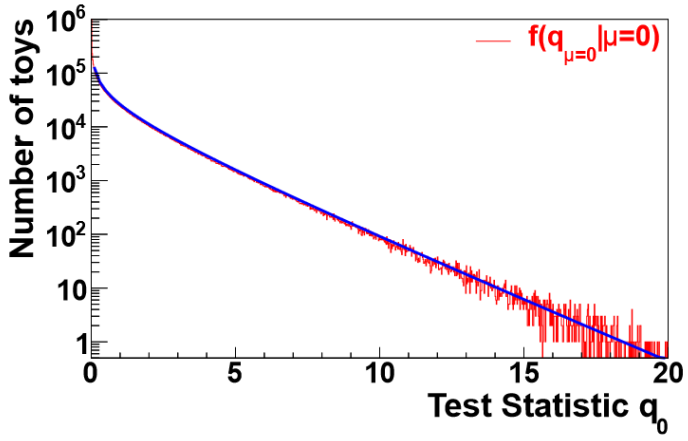
Asymptotic formula

- Often it is impractical to generate large datasets given the complexity of the search (dozens of search channels and sub-channels, correlated among each other). One then relies on a very good asymptotic approximation:

$$p^{estimate} = \frac{1}{2}\left[1 - \mathrm{erf}\left(\sqrt{q_0^{\mathrm{obs}}/2}\right)\right]$$

- The derived p-value and the corresponding Z value are "local": they correspond to the specific hypothesis that has been tested (a specific M_h) as q_0 also depends on M_h (the search changes as M_h varies)

- When dealing with many searches, one needs to get a global p-value and significance, i.e. **evaluate a trials factor.**
- This can be done using the techniques discussed earlier.



Type-I and type-II error rates

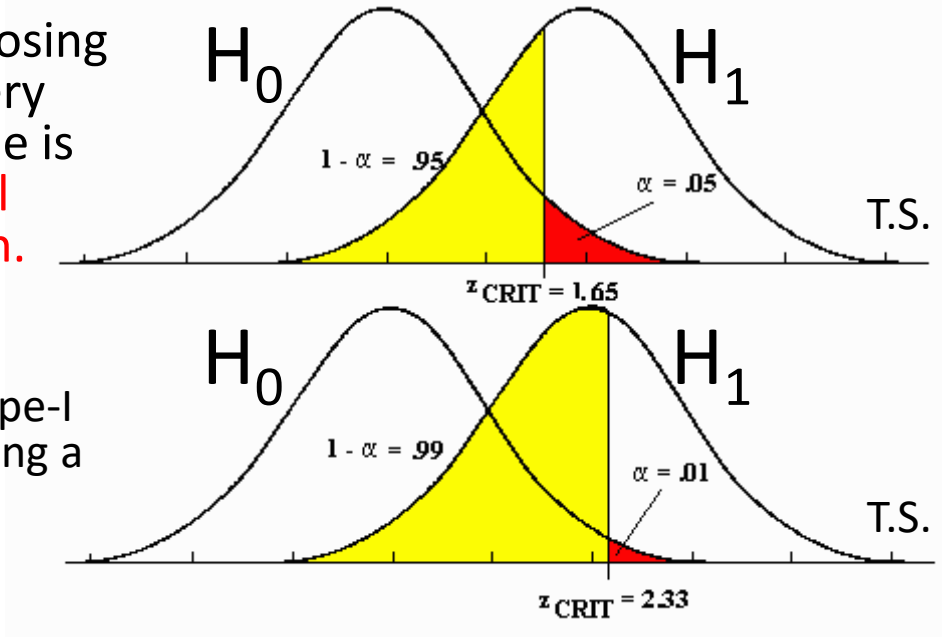
In the context of hypothesis testing the type-I error rate α is the probability of rejecting the null hypothesis when it is true.

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.

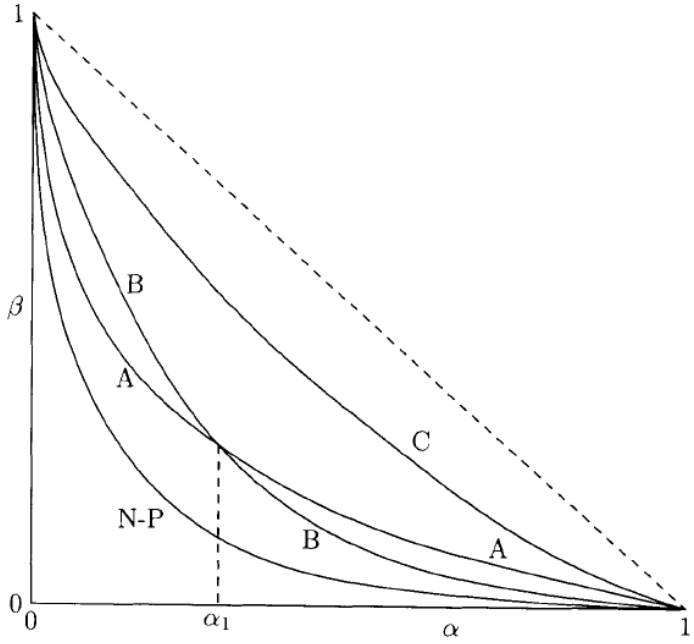
Strictly connected to α is the concept of "power" (1-β), where β is the type-2 error rate, defined as the probability of accepting the null, even if the alternative is instead true.

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.



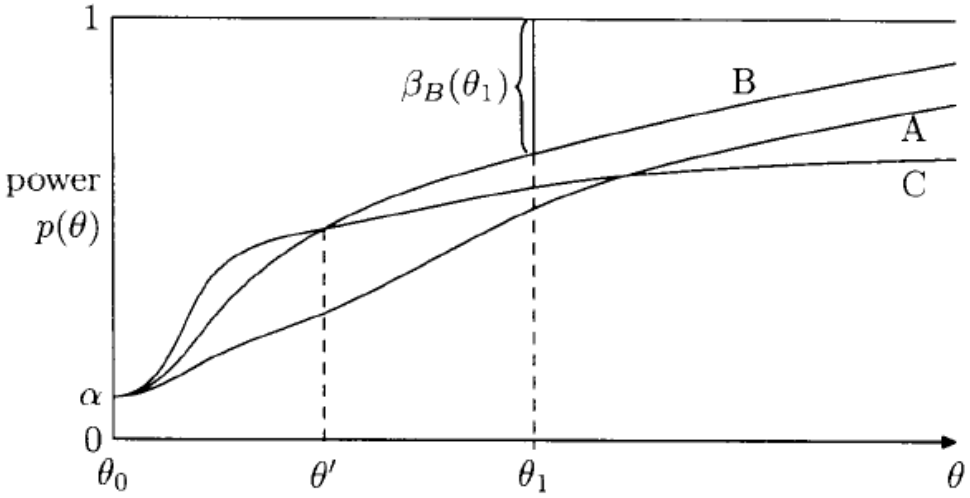
Alpha vs Beta and power graphs



- Where to stay in the curve provided by your analysis method highly depends on **habits in your field**
- What makes a difference is the test statistic.

  The N-P likelihood-ratio test outperforms others for simple-vs-simple HT, as dictated by the Neyman-Pearsons lemma: higher power 1-β for any α.

As data size increases, the power curve (shown below) becomes closer to a step function



The power 1-β of a test usually depends on the parameter of interest: different methods may have best performance in different parameter space points

JLP Example: Charge Bias of a Tracker

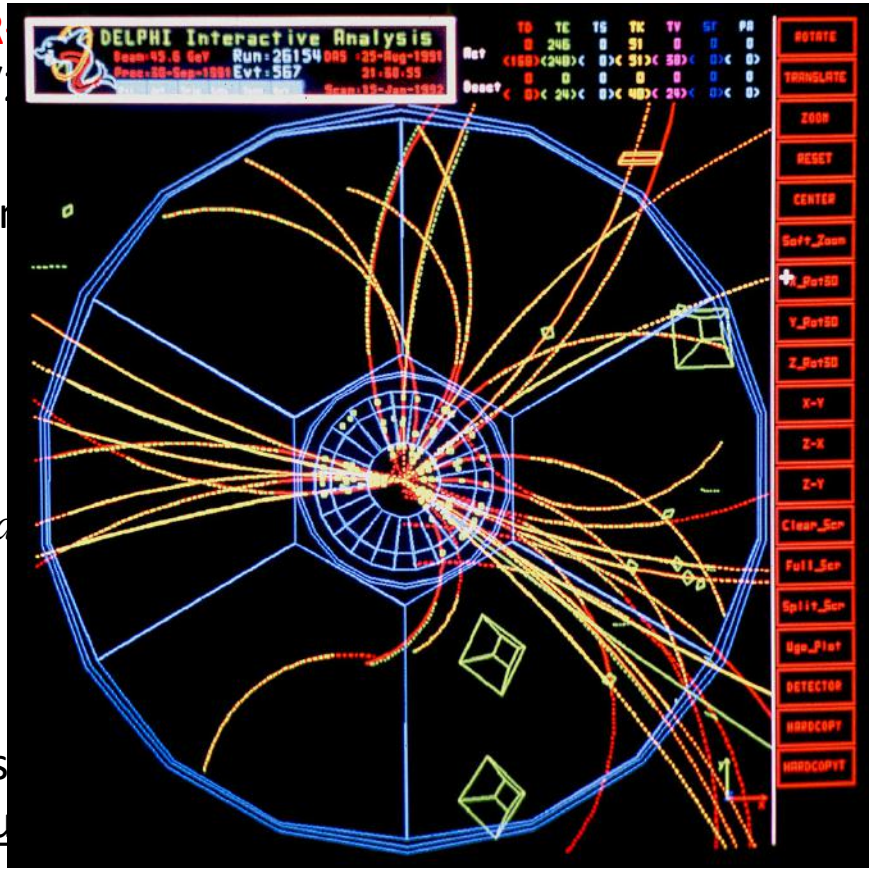
Imagine you want to investigate whether your detector has a bias in reconstructing positive versus negative curvature, say at a lepton collider (e^+e^-). You take a unbiased set of collisions, and count positives and negatives in a set of n=1,000,000.

- You get n^+=498,800, n^-=501,200. You want to test the hypothesis that the fraction of positive tracks, say, is R=0.5 with a size **α=0.05**.
- Bayesians will **need a prior π(R)**: a typical choice would be to **assign equal probability** to the chance that R=0.5 and to it being different (R
  and a uniform distribution of the remaining p=1/
- We are in high-statistics regime and away from 0
  for the Binomial. The probability to observe a nu
  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

$$P(R=\frac{1}{2}\,|\,x,n) \approx \frac{\frac{1}{2}\frac{e^{-\frac{(x-\frac{1}{2})^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}}{\frac{1}{2}\frac{e^{-\frac{(x-\frac{1}{2})^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}+\frac{1}{2}\int_0^1 \frac{e^{-\frac{(x-R)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}dR}$$

from which a Bayesian concludes that there is
and actually the data strongly supports the nu



JLP Charge Bias: Frequentist Solution

Frequentists calculate 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 σ^2=x(1-x)/n

One then has

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

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

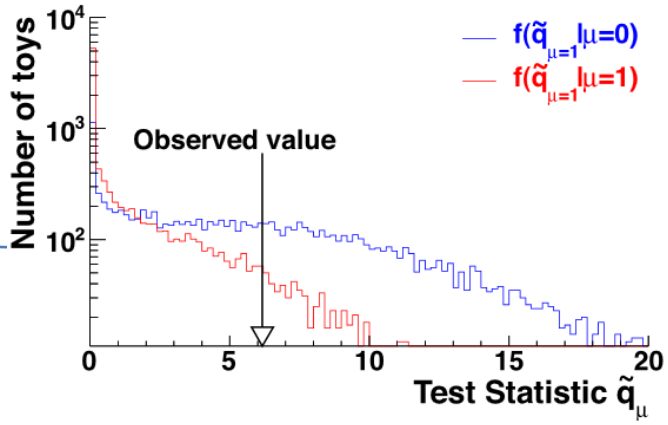
(we multiplied by two since we would be just as surprised to observe an excess of positives as a deficit).

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

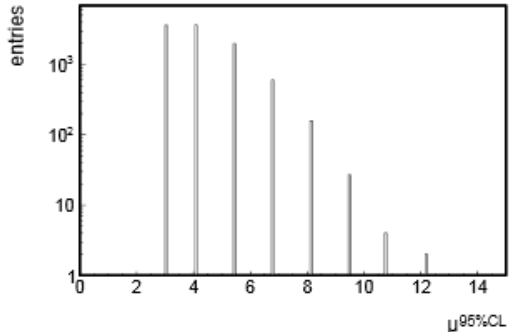
A frequentist thus draws the **opposite conclusion** of a Bayesian from the same (large body of) data !

Derivation of expected limits

One starts with the **background-only hypothesis μ=0**, and determines a distribution of possible outcomes of the experiment with toys, obtaining the CLs test statistic distribution for each investigated Higgs mass point

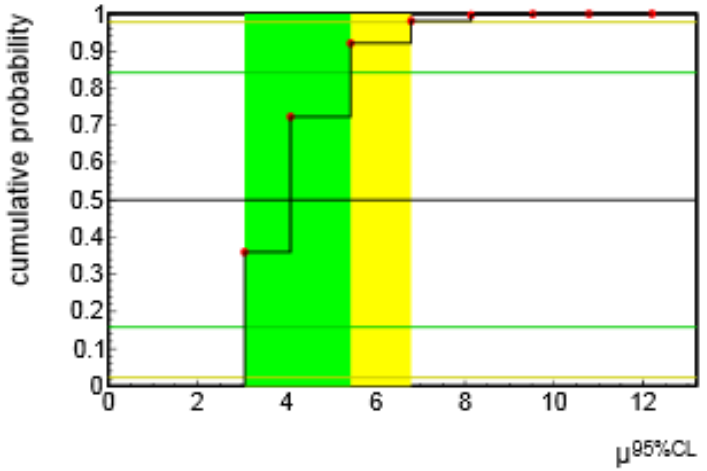


From CLs one obtains the PDF of upper limits μ^{UL} on μor each M_h. [*E.g. on the right we assumed b=1 and s=0 for μ=0, whereas μ=1 would produce <s>=1*]

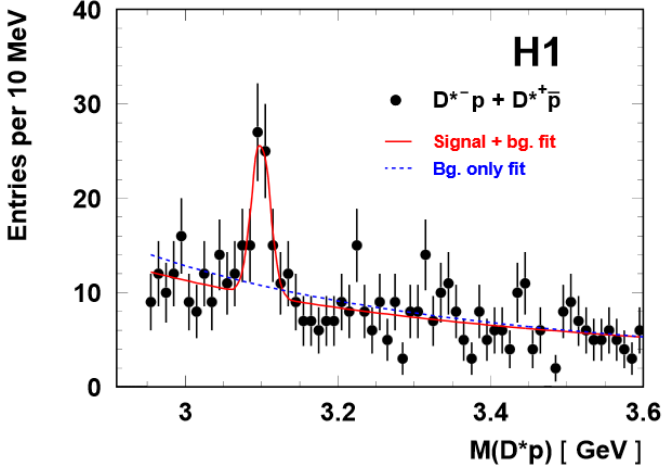


Then one computes the cumulative PDF of μ^{UL}

Finally, one can derive the median and the intervals for μ which correspond to 2.3%, 15.9%, 50%, 84.1%, 97.7% quantiles. These define the “expected-limit bands” and their center.



An important ingredient: Wilks' Theorem



- An almost ubiquitous method to derive a significance from a likelihood fit is the one of invoking **Wilks' theorem**
  - that is, many physicists invoke it although they're not aware!

- One has a likelihood under the null hypothesis, L_0 (say, 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 chisquare

- $P(\chi^2)$ can then be obtained, and from it a Z-value
  - But people regularly forget that this is only applicable when the two hypotheses are connected by H_0 being a particular case of H_1 (fixing of one parameter): they must be **nested models**.
  - In most cases this is not so: we routinely test a H_1 where one of the parameters is not present in H_0 (mass m for σ=0).
  - Fortunately, often even when the regularity conditions demanded by the theorem are not met, the asymptotic properties of $\Delta\ln L$ are good enough