Oxford University, April 24th 2018

INFN
Istituto Nazionale
di Fisica Nucleare

FUNDAMENTAL STATISTICS
FOR
DISCOVERY
IN
FUNDAMENTAL PHYSICS

Tommaso Dorigo
INFN-Padova

Contents

- Confidence Intervals for Physicists
  - Coverage, and the lack thereof
  - Conditioning, relevant subsets, and ancillarity
- The Five-Sigma Criterion: Still a Good Idea ?
  - A bit of history: from Rosenfeld to the LHC
  - The unknown unknowns and their misbehaviour
  - Counting wiggles: how to get 2000 citations
- Going Bayesian and the Related Thorns: The Jeffreys-Lindley Paradox

Confidence Intervals for Physicists

- At the heart of the activities of experimental physicists is what we call *measurement*
  → point estimation + interval estimation

- **Point estimation** can be awfully complicated, but it is almost always non-controversial
- **Interval estimation** is way more complex – and it is what we really care about
  - experimental design: minimize expected uncertainties on parameters of interest
  - BSM searches: "does it agree with the SM?"

- The core question we should always be asking ourselves is "do my uncertainty bars **cover** at the stated confidence level ?"

What's Coverage ?

Suppose we use N data {x}, distributed as f(x,θ), to measure a parameter θ. An estimator based on {x} is used for this.

The value $\theta^* \pm \sigma^*_{\theta^*}$ is finally reported. **What does this mean ?**

- It means that in **repeated** estimates based on the same number N of observations of x, θ^* would distribute according to a pdf $G(\theta^*)$ centered around the true value θ with a true standard deviation σ_{θ^*}, **respectively estimated** by θ^* and $\sigma^*_{\theta^*}$

- *In the large sample limit G() is a (multi-dimensional) Gaussian function*

Unfortunately, in most interesting cases for physics G() is not Gaussian, the large sample limit does not hold, 1-sigma intervals do not cover 68.3% of the time the true parameter, and we have better be a bit more tidy in constructing intervals.

But we need to have a hunch of the pdf $f(x;\theta)$ to start with!

Coverage, or the Lack Thereof

Let us consider a typical HEP graph: event counts in a mass histogram, with sqrt(N) bars

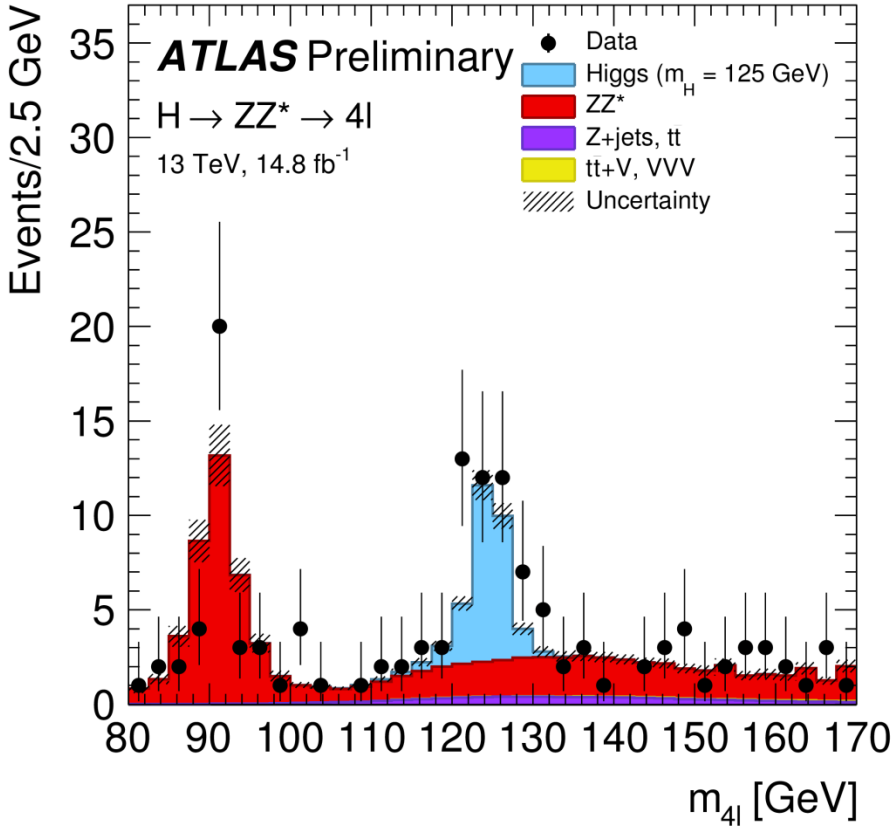
(Note: statisticians never plot their data this way...)

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

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

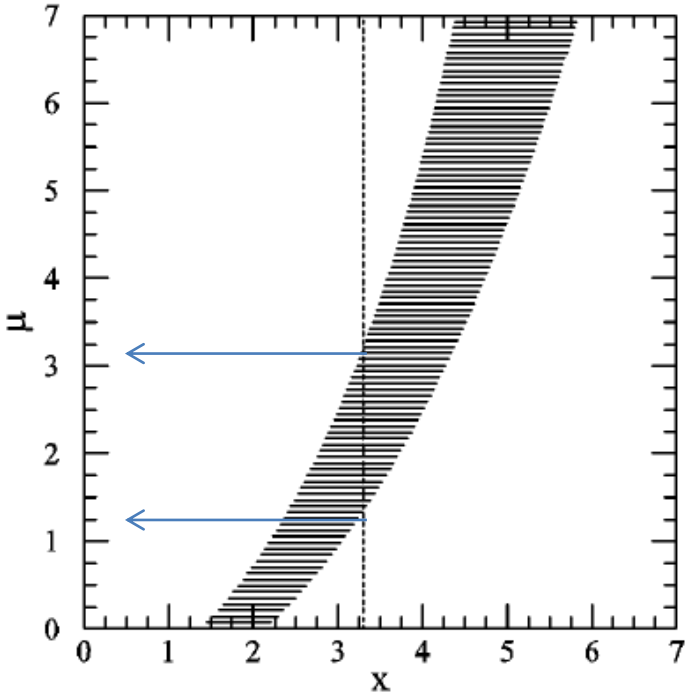
Suppose somebody says x is in [a,b] with 68.3% confidence, but in fact the way a,b were determined makes the confidence level of [a,b] to be, e.g., only 50%.

→ That would be a quite significant misrepresentation of the information content of the measurement !



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

Neyman's Confidence Interval Recipe



1 - Specify a model $p(x|\mu)$

2 - Choose a Type-I error rate α (e.g. 31.7%, or 5%)

3 - 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 *!*

- for UL, integrate the pdf from x to infinity

- for LL, do the opposite

- or choose central intervals, or shortest intervals...

In general: ***an ordering principle*** *is needed to well-define.*

4 - 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.*

Note: the recipe is designed to cover correctly. Thus, one could not, on average, win money by betting that the result of a measurement does not contain the true value, by using payoff odds inverse to the stated type-I error rate (eg. 5% → 19:1)

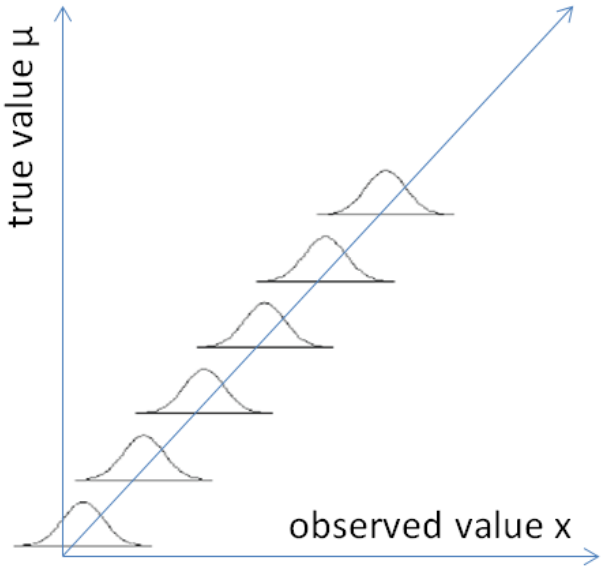
Where It Gets Murky

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.

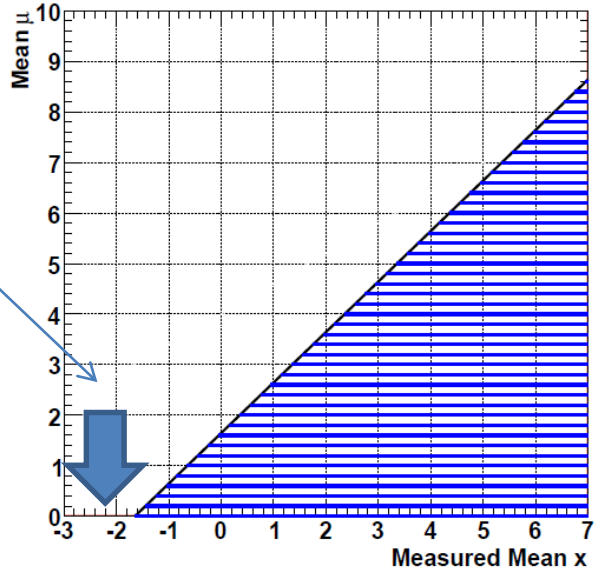
Take e.g. μ>0 measured by P(x|μ) = N(μ,1):

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

The classical method for 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)



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

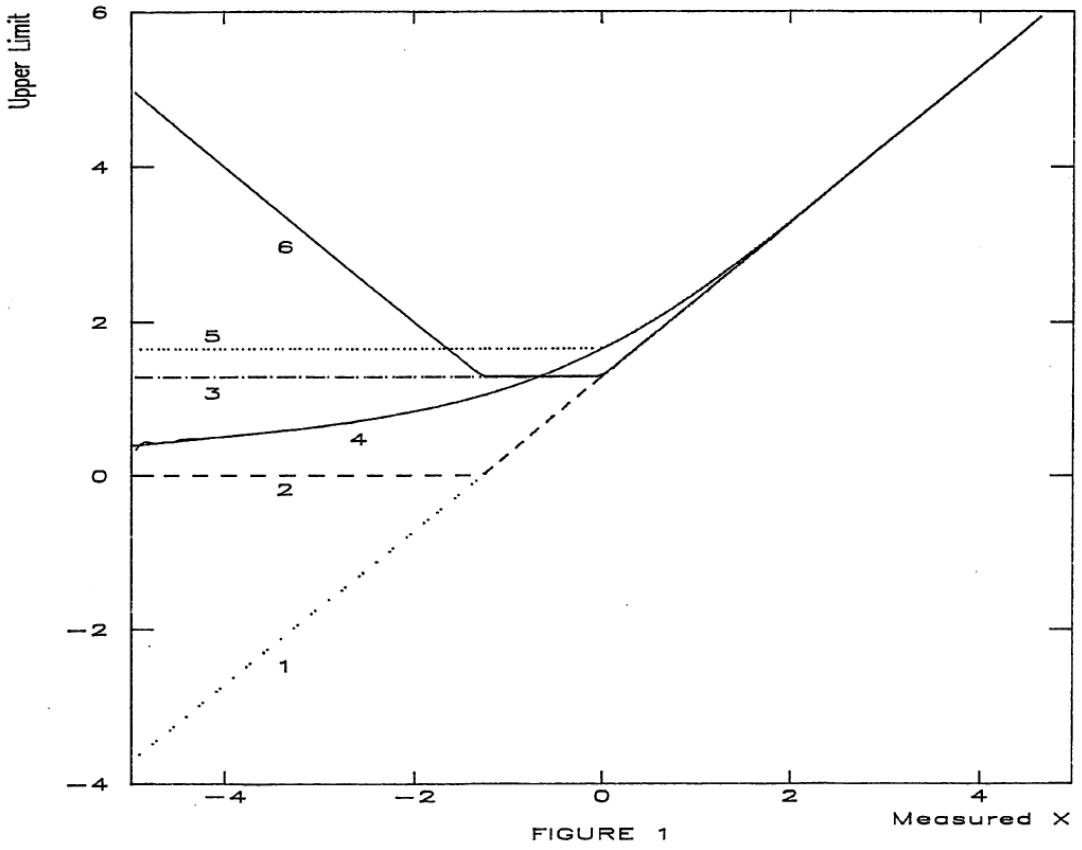


FIGURE 1

(1) Neyman's recipe for 90% upper limits: μ_{UL}=x+1.28.
(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 19:1 odds, using no more information than the 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 is still >0.

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

Conditioning and Ancillary Statistics

In the bounded parameter problem, the flaw of being subject to winning bet strategies can be amended by adding a horizontal line or interval (such that any c.i. will contain that value of μ), but it feels like a hack

In other cases one can identify ancillary statistics and use them to partition the space into **relevant subsets**.

- "Ancillary statistic": f(data) yielding information about the *precision* of the estimate of the parameter of interest, but **no** information about the parameter's value.

- Most typical case in HEP: **branching fraction** measurement. With N_A, N_B event counts in two channels one finds that

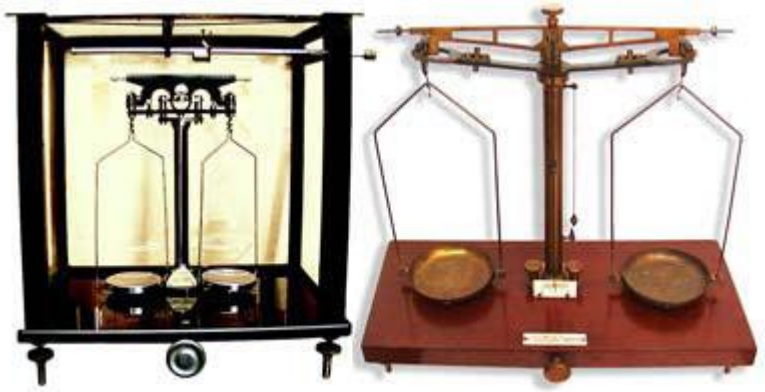
$$P(N_A,N_B) = \text{Poisson}(N_A) \times \text{Poisson}(N_B) =$$
$$= \text{Poisson}(N_A+N_B) \times \text{Binomial}(N_A \mid N_A+N_B)$$

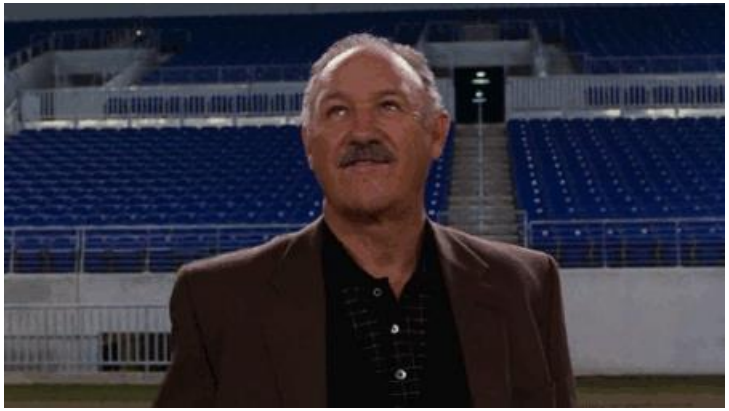
By using the second expression, one may ignore the ancillary statistic N_A+N_B, since all the information on the BR is in the conditional binomial factor

→ by restricting the sample space, the problem is simplified.

Cox Weighting Procedure

Things get even more intriguing in the famous example by **B. Cox[2]**:



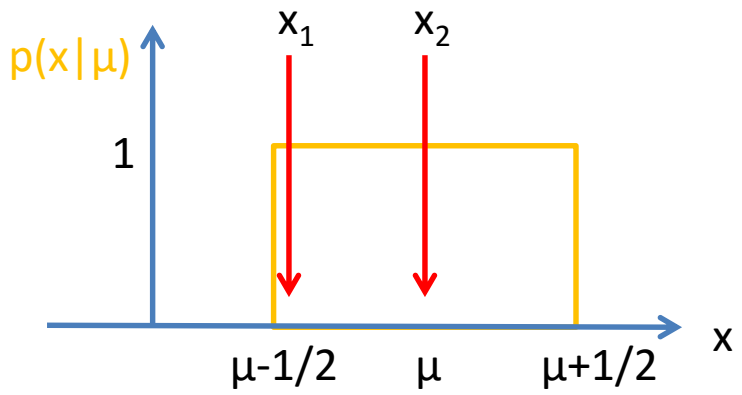


Flip a coin to decide whether to use a 10% scale (if you get tails) or a 1% scale (if you get heads) to measure an object's weight. Which error do you quote for your measurement, upon getting heads ?

Of course the knowledge of your device allows you to estimate that your precision is 1% - but a full NP construction (which is unconditional on the outcomes) would require you to include the coin flipping in the procedure!

Locating the Box

- Another example:
  Find μ using x_1, x_2 sampled from
  p(x|μ) = Uniform [μ-1/2, μ+1/2]



Suppose e.g. that μ=1, and take the two datasets,
A: {0.99,1.01} ; B: {0.6,1.4}.

- NP procedures maximizing power in the unconditional space yield the same confidence interval for both data sets A and B; however, B restricts the set of possible μ to [0.9,1.1] while A only restricts it to [0.51,1.49] !
- **There exists in fact an ancillary statistics $|x_1-x_2|$** which carries no information on μ, yet it can be used to divide the sample space in subsets where inference can be different.
- See **R. Cousins[3]** for more discussion

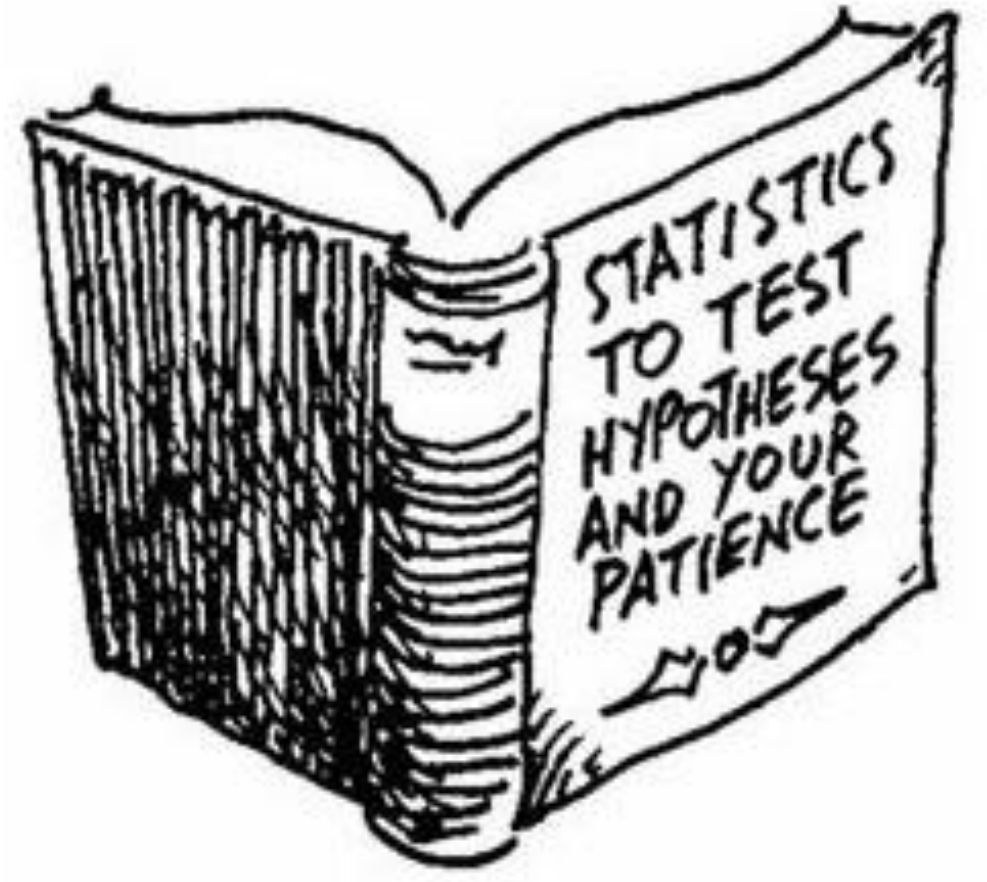
Relevant Subsets: Take-Away Bit

Point made: *The quality of your inference depends on the breadth of the "whole space" you are considering*. The more you can restrict it, the better (i.e. the more relevant) your inference becomes

- Ancillary statistics are not easy to find, but they are quite useful!

→Look for ancillary statistics in your everyday measurements!

Hypothesis Testing in Three Slides



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 **data X** and a **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.

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 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.

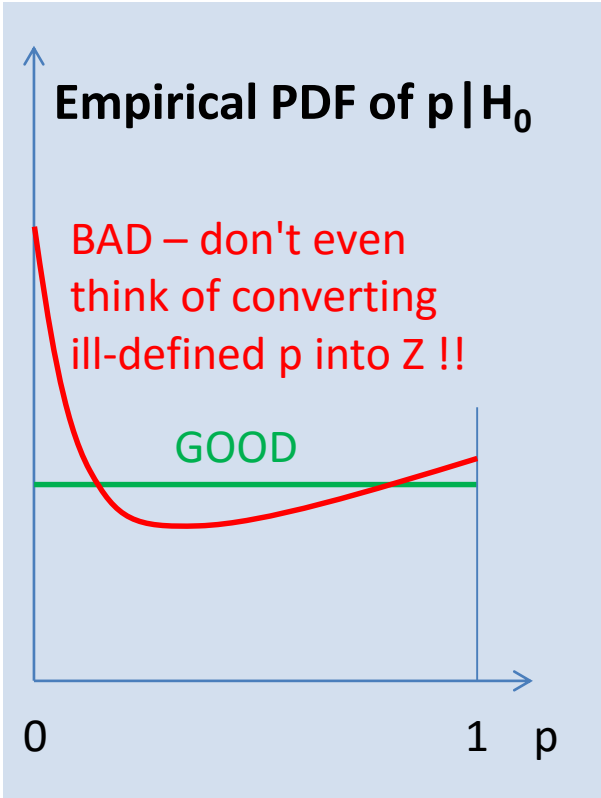
Notes

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 just **a shortcut, nothing more** !

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



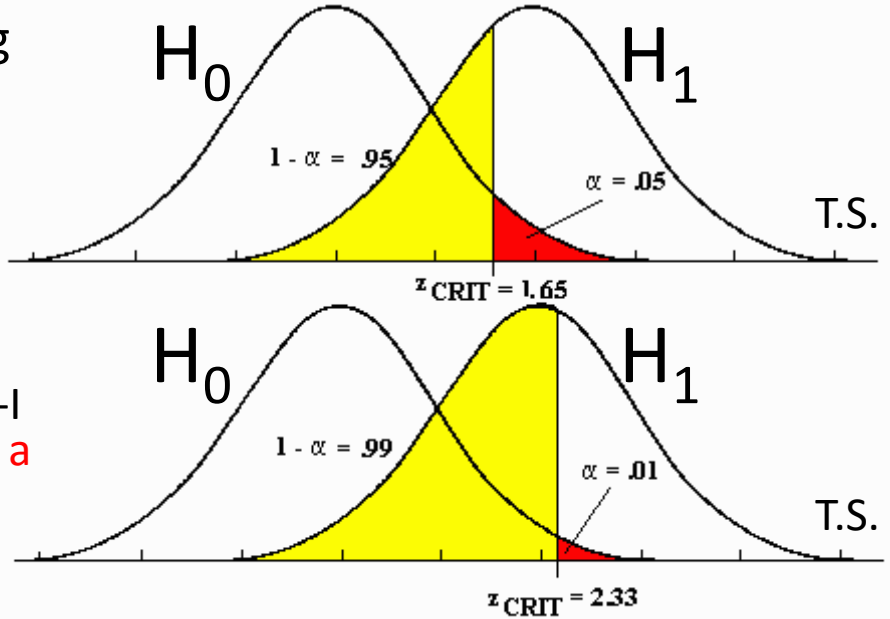
Type-I and Type-II Errors

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

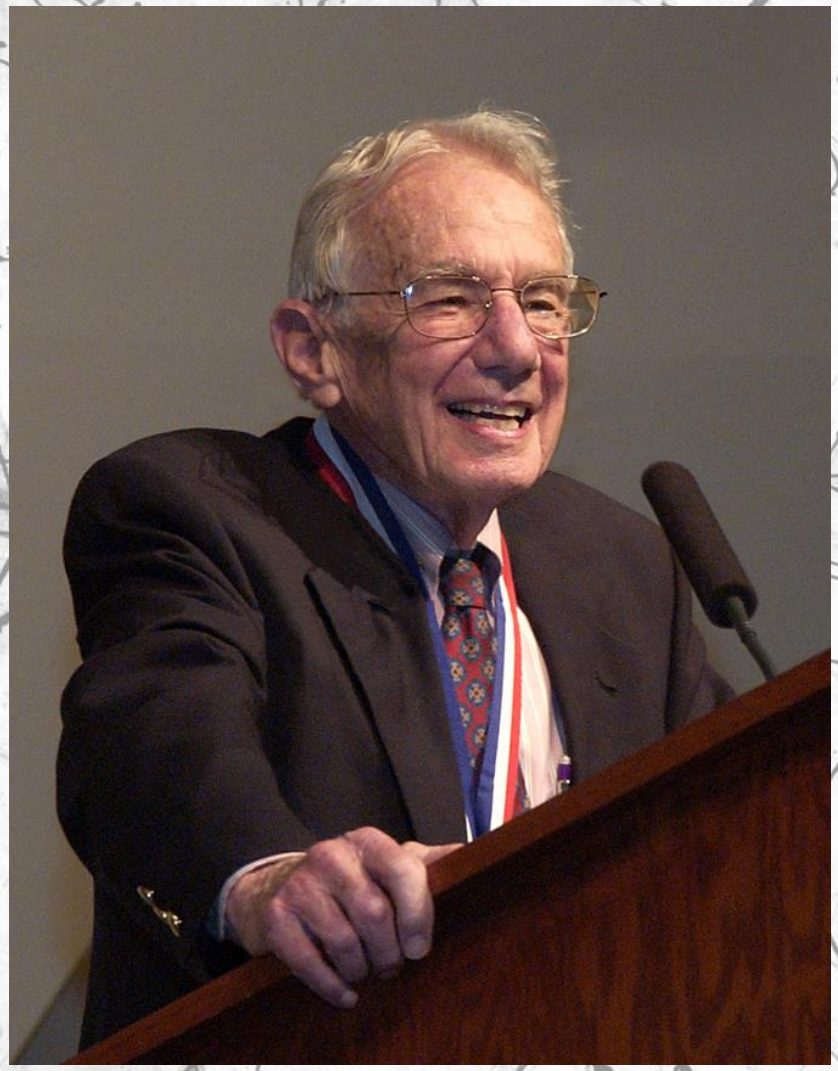
Strictly connected to α is the concept of "power" (1-β), where β is the type-2 error rate, defined as the probability of accepting the null when the alternative is instead true.

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. There is no formal recipe to guide this choice.

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.



The Birth of the Five-Sigma Criterion



Arthur H. Rosenfeld (Univ. Berkeley)

Far-Out Hadrons

- In 1968 Rosenfeld wrote a paper titled *"Are There Any Far-out Mesons or Baryons?"* **[4]**, where he demonstrated that the number of published claims of discovery of exotic particles **agreed with the number of statistical fluctuations** that one would expect in the analyzed datasets.

- The issue: 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..."

"[...] 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****."*

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 […]

  When a friend comes showing his latest 4-sigma peak,

  You draw a smooth curve […] (based on the hypothesis that the peak is just a fluctuation) [and] call for 100 Las Vegas histograms […]

  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."

- The proposal to raise to 5-sigma of the threshold above which a signal could be claimed was an earnest attempt at reducing the flow of claimed discoveries, which distracted theorists and caused confusion.

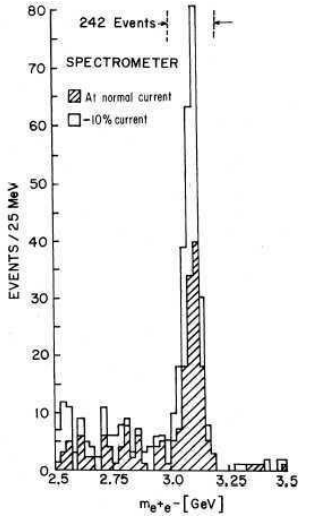
What 5σ May Do For You

- 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*.

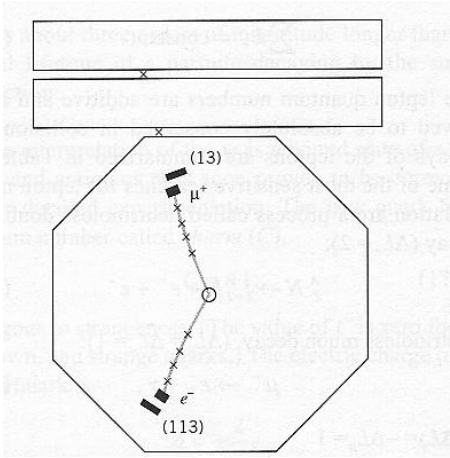
How 5σ Became a Standard in HEP: 1 - the Seventies



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

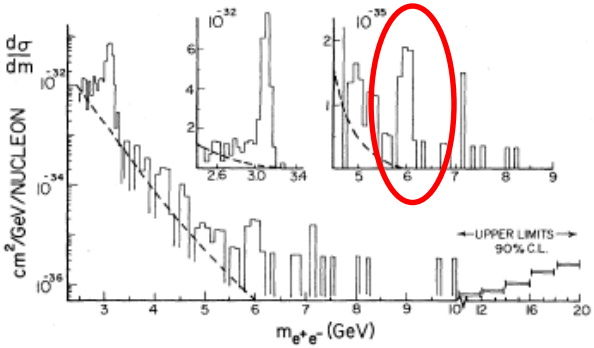
Let us check a few important searches to understand how the 5σ criterion gradually became a standard

- **The J/ψ discovery** (1974): no question of significance – the bumps were too big to fiddle with stat 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.*"**[5]**



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 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"**[6]**

Nov 19th 1976

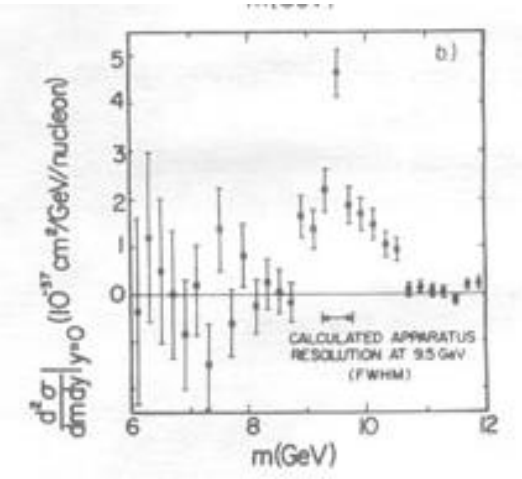
I determined this factor by monte carlo. I threw 30 events over 100 bins (expectation is 2 for 6 bins) and searched for clusters of 10 in 6 bins. I found 15 successes in 40000 trys or CL = 3.75×10^{-4}. The poisson probability for ≥10 for an expectation of 2 is 1.94×10^{-5}. This bin counting factor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

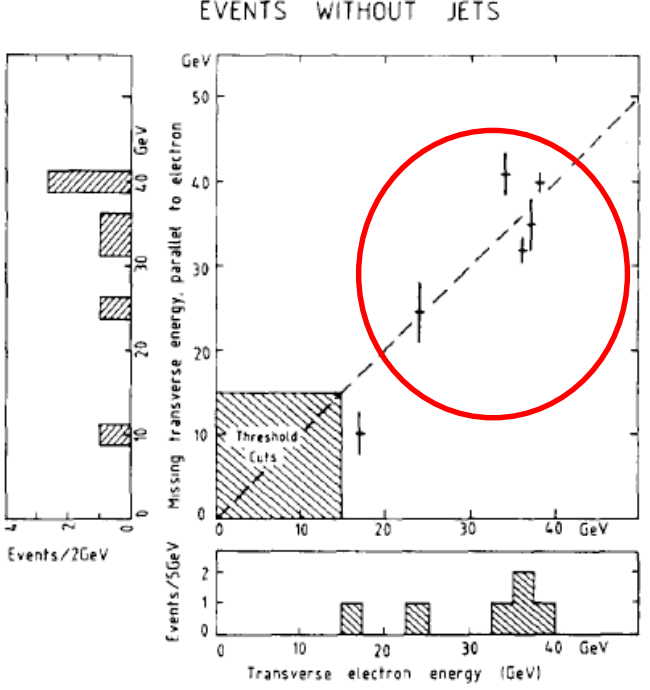
① Programming error, double counting, etc. – will be studied by

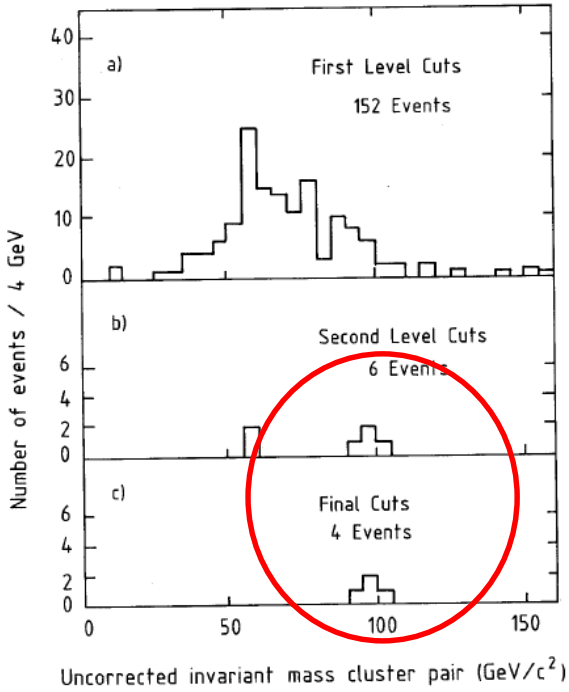
The W and Z Bosons

The 1983 W discovery was announced based on 6 electron events with missing energy and no jets.

- No statistical analysis is discussed in the **discovery paper[7]**, 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[8]**, except notes that backgrounds were negligible





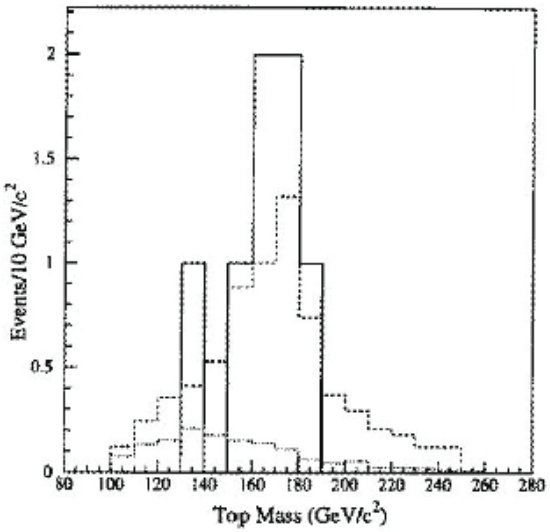
The Top Quark Discovery

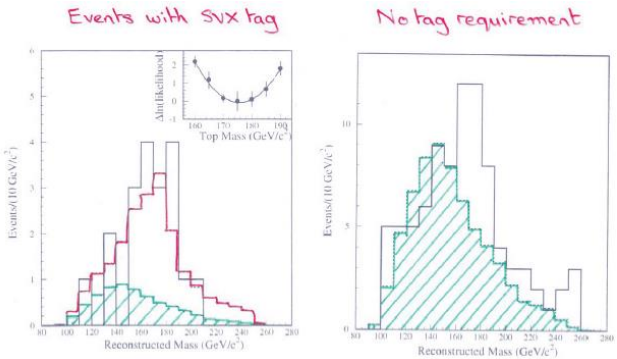
- In 1994 the CDF experiment had a serious counting excess (2.7σ) in b-tagged single-lepton and dilepton datasets, plus a mass peak at a value compatible with theory predictions
  - the mass peak, or corresponding kinematic evidence, was over 3σ by itself; **M = 174 +- 10^{+13}_{-12} GeV**

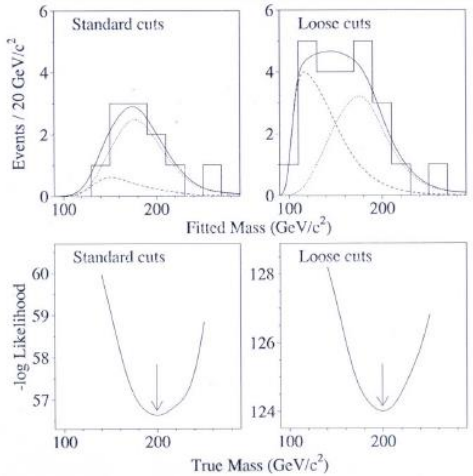
  The paper describing the analysis (120-pages long) spoke of "**evidence**" for top quark production[9]

- One year later CDF and DZERO[10] both presented 5σ significances based on their counting experiments, obtained by analyzing 3x more data

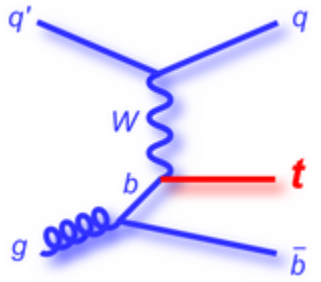
  The top quark was thus the first particle discovered by a willful application of the "5σ" criterion





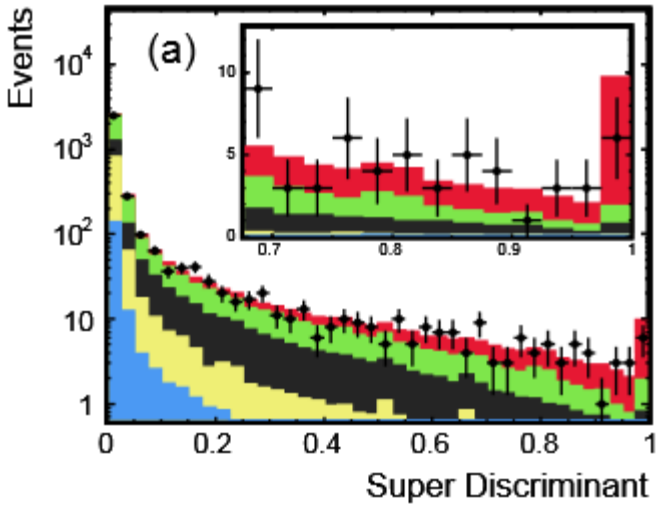


Following the Top Quark...

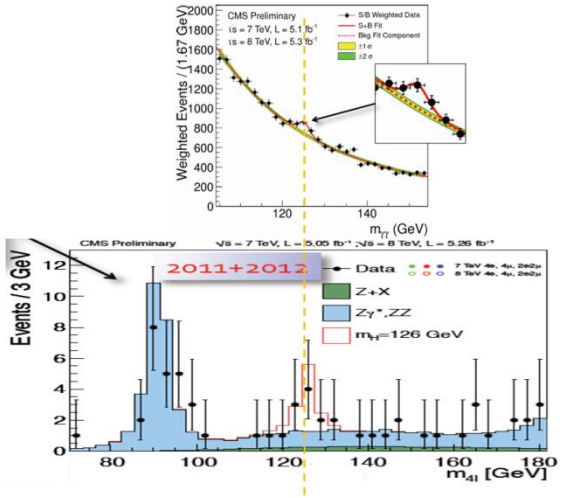


- Since 1995, the requirement of a p-value below $3*10^{-7}$ slowly but steadily became a standard.
- Striking examples of searches that diligently waited for a 5-sigma effect before claiming discovery:

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 **[11]**, over clear 5-sigma effects



2) In 2012 the **Higgs boson** was claimed by **ATLAS and CMS [12]**. 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 search for the Higgs what brought the five-sigma criterion to the attention of media

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, a.k.a. LEE

- The LEE was a concern 50 years ago, but nowadays we have enormously more CPU power. Still, the complexity of our analyses has also grown considerably
  - Take the Higgs discovery as an example: hundreds of nuisances, many final states
    → we still occasionally cannot compute the trials factor by brute force!
  - A further complication is that in reality **the trials factor also depends on the significance of the local fluctuation,** adding dimensionality to the problem.
- A study by **E. Gross and O. Vitells[13]** demonstrated in 2010 how it is possible to estimate the trials factor in most experimental situations, without resorting to simulations

Trials Factors

In Statistics **trials factors** arise in a hypothesis test when a nuisance parameter is present only under the alternative hypothesis (a simple-vs-composite test).

Let us consider a particle search when the mass **x** is unknown.

The null hypothesis is that the data follow the background-only model **b(x)**, and the alternative hypothesis is that they follow the model **b(x)+ μs(x|m_H)**, with **μ** a signal strength parameter and **m_H** the particle's true mass (the nuisance parameter!)

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

One then defines a test statistic summarizing all possible mass values,

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

This could e.g. be the maximum of the likelihood ratio b/w models **b(x)** and **b(x)+μs(x|m)**. The problem is assigning a p-value to the maximum of q(m_H) **given the wide search range**.

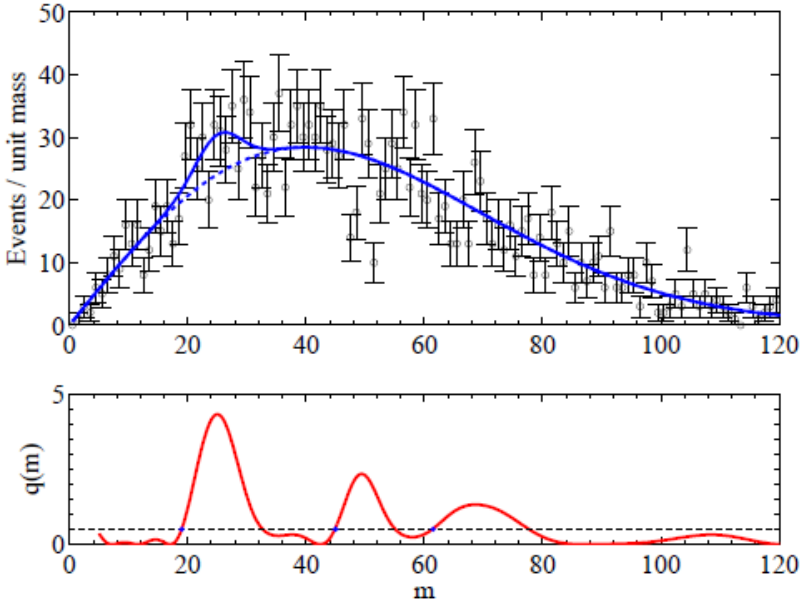
One can use an asymptotic "regularity" of the distribution of the above q to get a global p-value by using the technique of Gross and Vitells.

Local Minima and Upcrossings

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

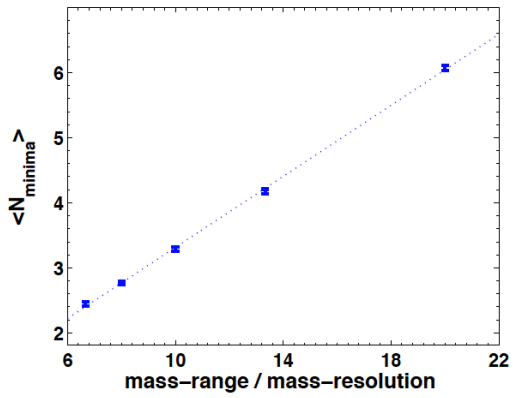
The number of times that the test statistic (below, the likelihood ratio between H_1 and H_0) crosses some reference line can be used to estimate the trials factor:

$$p_b^{global} = P(q_0(\hat{m}_H) > u) \leq \langle N_u \rangle + \frac{1}{2} P_{\chi^2_1}(u)$$



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. have I tried different selections before settling on the one I actually ended up with?
4. Have I been 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 ?

→ **There is ambiguity on the LEE depending who you are (grad student, experiment spokesperson, lab director...)**

In fact, Rosenfeld considered the whole world's database of bubble chamber images in deriving a trials factor

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

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 variation of an observable due to the imprecise knowledge of a nuisance parameter **at the 1-sigma level**; then one assumes that the probability density function of the nuisance be Gaussian.

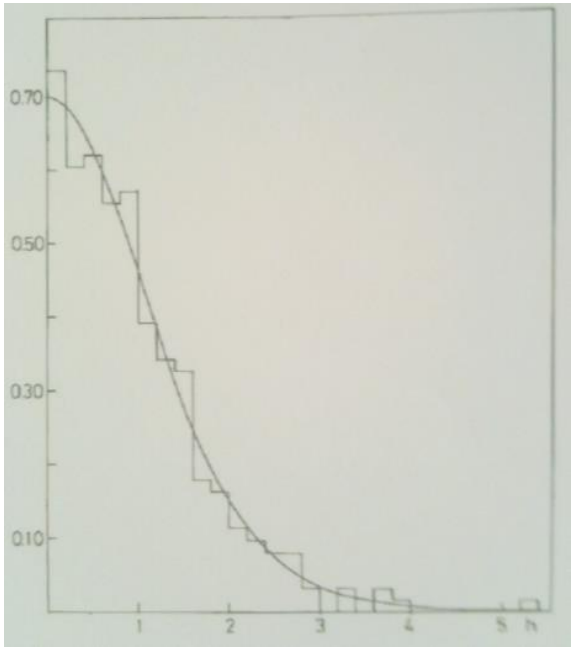
→ if however the PDF has larger tails, it makes the odd large bias much more frequent than estimated

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

- One quick example: if a 5σ effect has uncertainty dominated by systematics, and the latter are underestimated by a factor of 2, the 5σ effect is actually a 2.5σ one (a p=0.006 effect): in p-value terms this means that the size of the effect is **overestimated by a factor 20,000**!

A Study of Residuals

The distribution of residuals of 306 measurements in [14]



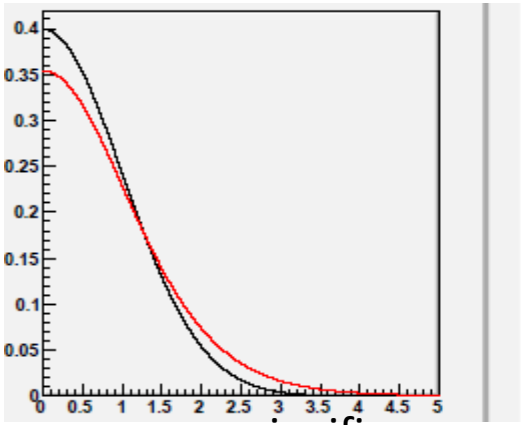
A study of the residuals of particle properties in the RPP in 1975 revealed that they were **not Gaussian**. **Matts Roos *et al.* [14]** considered residuals in kaon and hyperon mean life and mass measurements, and concluded that these are well described by a Student distribution $S_{10}(h/1.11)$:

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

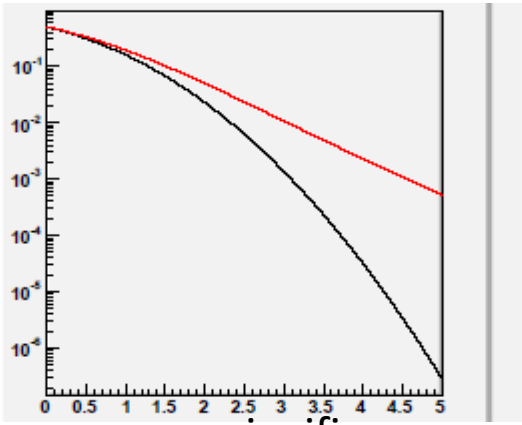
One should not extrapolate to 5-sigma the behaviour found by Roos and collaborators in the bulk of the distribution; yet it is evidence that the uncertainties evaluated in experimental HEP may have a significant non-Gaussian component

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

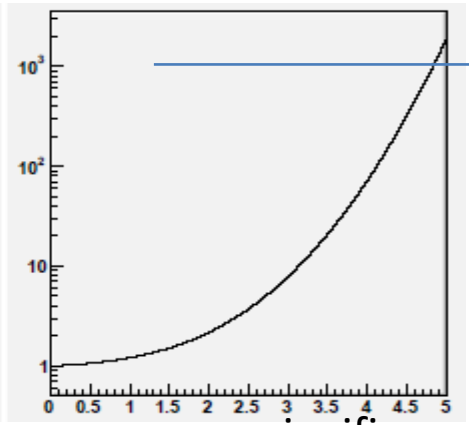
Left: 1-integral distributions of the two functions. Right: ratio of the 1-integral values as a function of z



significance



significance

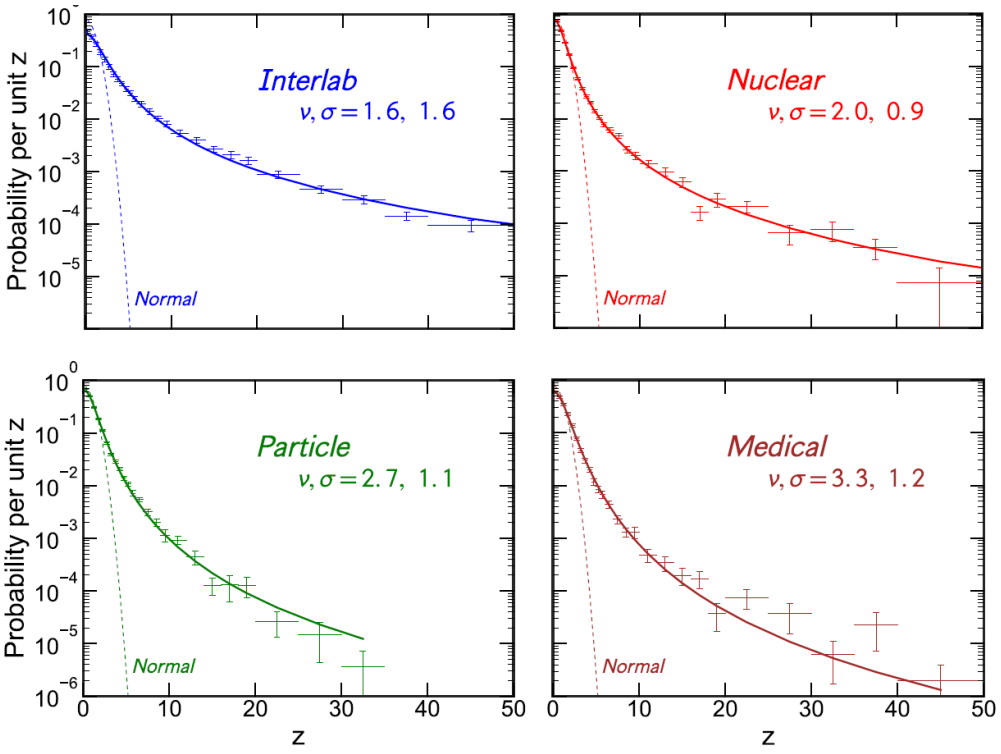


x1000!

significance

A Bigger, Meaner Study of Residuals

- David Bailey (U. Toronto) recently published an **article[15]** 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 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[16]** 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

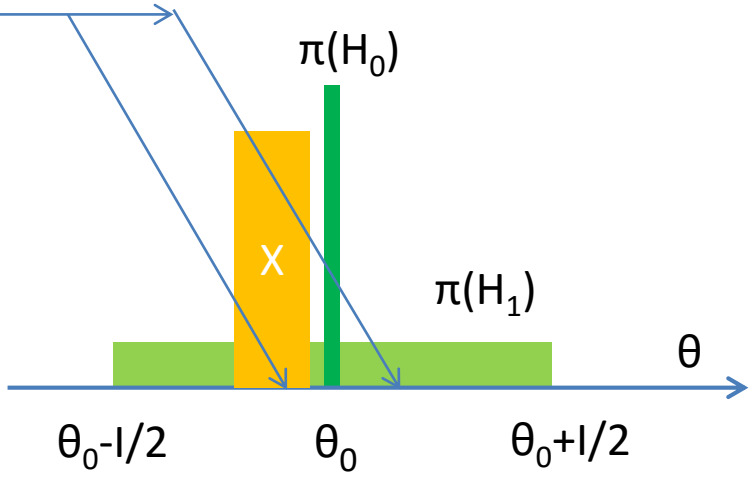
Take $X_1 \ldots X_n$ i.i.d. as $X_i|\theta \sim N(\theta,\sigma^2)$, and a prior belief on θ constituted by a mixture of a point mass **p** at θ_0 and **(1-p)** uniformly distributed in $[\theta_0\text{-}I/2, \theta_0\text{+}I/2]$.

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

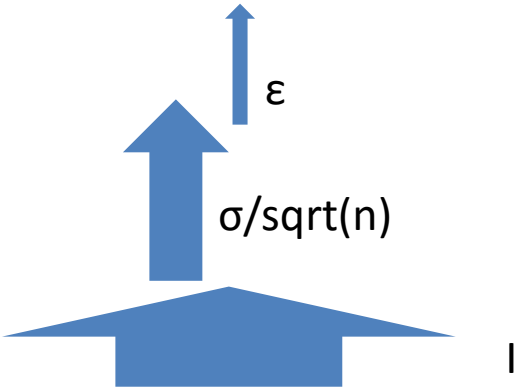
$$\bar{X} = \theta_0 \pm (\sigma/\sqrt{n}) z_{\alpha/2}$$

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[17]**, the paradox arises if there are <u>three independent scales in the problem</u>, **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!!

JLP Example: Charge Bias of a Tracker

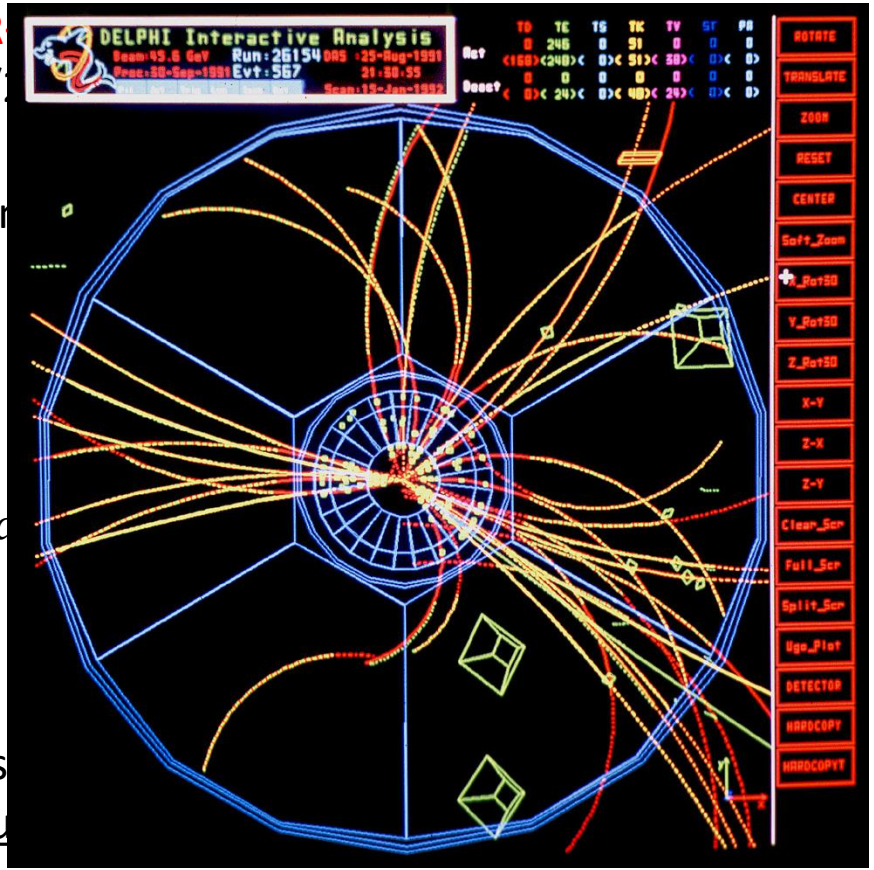
Imagine you want to investigate whether your detector has a bias in reconstructing positive versus negative curvature, say at a lepton collider (e^+e^-). You take a unbiased set of collisions, and count positives and negatives in a set of n=1,000,000.

- 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 !

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***" **[18]**

- 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 (*e.g.* particle searches), others one does not (*e.g*. OPERA's ν speed)
- The cost of a wrong claim, as backfiring of media hype, can vary dramatically
- Some claims are intrinsically less likely to be true, and deep within we have a subconscious Bayes factor at work.

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

Conclusions

- **50 years** after the first suggestion of a 5-sigma threshold for discovery claims, and **25 years** after the start of its consistent application, the criterion appears inadequate
  - It does not protect from steep claims that later peter out
  - It delays acceptance of uncontroversial finds
  - It is arbitrary and illogical in many aspects
- Bayesian hypothesis testing does not offer a robust replacement, due to hard-to-circumvent prior dependence of conclusions
- A single number never summarizes the situation of a measurement
  - experiments have started to publish their likelihoods, so combinations and interpretation get easier
- My suggestion is that for each considered (relevant) search the community should seek a consensus on what could be an acceptable significance level for a media-hitting claim
- For searches of unknown effects and fishing expeditions, the **global** p-value is the only real weapon – but in most cases the trials factor is hard to quantify
- Probably 5-sigma are insufficient for unpredicted effects, as large experiments look at thousands of distributions, multiple times, and the experiment-wide trials factor is extremely high

Expect some spurious 5-sigma effect from the LHC soon!

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] D. Cox, "*Some Problems Connected with Statistical Inferenc*e", Ann. Math. Stat. 29 (1958) no. 2, 357-372.

[3] 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).

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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

[13] E. Gross and O. Vitells, "*Trials factors for the Look-Elsewhere Effect in High-Energy Physics*", Eur. Phys. J. C70 (2010) 525-530.

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

[15] D. Bailey, "*Not Normal: the uncertainties of scientific measurements*", ArXiv:1612.00778 (2016).

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

[17] R. D. Cousins, "*The Jeffreys-Lindley Paradox and Discovery Criteria in High-Energy Physics*", arxiv:1310.3791v4 (2014).

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