EPJ Web of Conferences **95**, 02003 (2015)
DOI: 10.1051/epjconf/ 20159502003

Extraordinary claims: the 0.000029% solution

Tommaso Dorigo[1,a]

[1] *INFN, Sezione di Padova*

Abstract. The five-standard-deviation threshold is an established standard for discovery claims in experimental particle physics; however, the criterion is an ad-hoc recipe with no solid foundations. In this report I discuss its origins and the issues it was designed to address, pointing out its shortcomings and the need for a more flexible approach to decide when a new observed effect should be taken seriously.

1 Introduction

Driven by the search for the Higgs boson and the media hype that preceded and followed the successful observation of the 125 GeV particle in July 2012, in the course of the last few years science popularization magazines and other outreach agents have been very busy explaining to the public the idea that a scientific discovery in fundamental physics requires that an effect be found with a statistical significance exceeding five standard deviations.

In accordance with Oscar Wilde's assessment that *"The only thing worse than being talked about is not being talked about"*, one might argue that science outreach brings positive effects to society regardless of the level of scientific accuracy of the distributed knowledge. While I agree to that general concept, I regret the wide exposure that the five-sigma criterion has received. It appears that we have successfully explained and popularized a convention which is entirely arbitrary and field-specific, and should be used with caution or substituted with something more scientifically motivated and suited to the specificities of the effects under study.

It is the purpose of this article to recall where the five-sigma criterion comes from, what it was designed to address, and to consider its limitations and the need for good judgement whenever a decision has to be taken on what scientific claim one may make based on the significance of the observation. In Section 2 I provide a brief introduction and a few important definitions of the essential ingredients. In Section 3 I discuss how the 5σ criterion became an established standard in particle physics searches. Section 4 reviews the merits and the limits of the criterion. In Section 5 I discuss how one could settle for different discovery thresholds depending on the characteristics of the phenomenon that is being sought. Finally, I offer some conclusions in Section 6.

[a] e-mail: dorigo@pd.infn.it

2 What is statistical significance?

Statistical significance is a number used to report the probability that an experiment obtains data at least as discrepant as those actually observed, under a given "null hypothesis", H_0. In physics H_0 usually describes the currently accepted and established theory, although there are exceptions. [1]

One usually starts with the p-value, which can be defined as the probability of obtaining a test statistic (a function of the data) at least as "extreme" as the one observed, if the null hypothesis H_0 is true. The p-value can 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 distribution equals p:

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

According to this recipe, a 15.9% probability corresponds to a one-standard-deviation effect; a 0.135% probability corresponds to a three-standard-deviation effect; and a 0.0000287% probability corresponds to five standard deviations - "five sigma" for insiders.

The alert observer will no doubt notice a few facts. First of all, the convention is to use a "one-tailed" Gaussian: we do not consider departures of x from the mean in the uninteresting direction. Hence *negative* significances (*i.e.* ones derived from p-values above 0.5) are mathematically well defined, yet not interesting as far as discovery claims are concerned; they may, if large, indicate that something is wrong with one's prediction for the behaviour expected if the null hypothesis holds.

Second, the conversion of p into σ is fixed and independent of experimental detail. As such, using σ rather than p is just a shortcut to avoid handling numbers with many digits: we prefer to say 5σ rather than "0.00000029", just as we prefer to say "a nanometer" instead of "0.000000001 meters", or "a Gigabyte" instead of "1000000000 bytes".

Third, the validity of this conversion recipe rests on a proper definition of the p-value. Any shortcoming of the properties of p (*e.g.* a tiny non-flatness of its probability density function (PDF) under the null hypothesis H_0) totally invalidates the meaning of the derived number of σ. In particular, using σ units does in no way means we are espousing some kind of Gaussian approximation for our test statistic or for other ingredients of the problem. This subtle point cannot be stressed enough, as many overlook it and are led into confusion, mis-usage, or constructing plainly wrong claims.

Fourth, an important point to make is that the "probability of the data" has no bearing on the concept of significance from a Frequentist point of view, and is not used at all. What is used is rather the probability of a *subset* of the *possible outcomes* of the experiment, defined by the outcome actually observed: ones as much or more extreme than the observed outcome.

2.1 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 actually true. In the common situation of testing a simple null hypothesis versus a composite alternative, such as when one determines at significance level α whether a signal with strength $\mu > 0$ is supported by the data or is non-existent ($\mu = 0$), the question being tested is dual to asking whether 0 is in the confidence interval for μ at confidence level $1 - \alpha$.
Strictly connected to the type-I error rate α is the concept of *power*, which is measured by $(1-\beta)$ where β is the type-II error rate. The latter is defined as the probability of accepting the null hypothesis H_0,

[1] Most notably, the discovery of the Higgs boson is one such special case, as the null hypothesis corresponded to a standard model with no Higgs boson, which is not an acceptable physical theory.

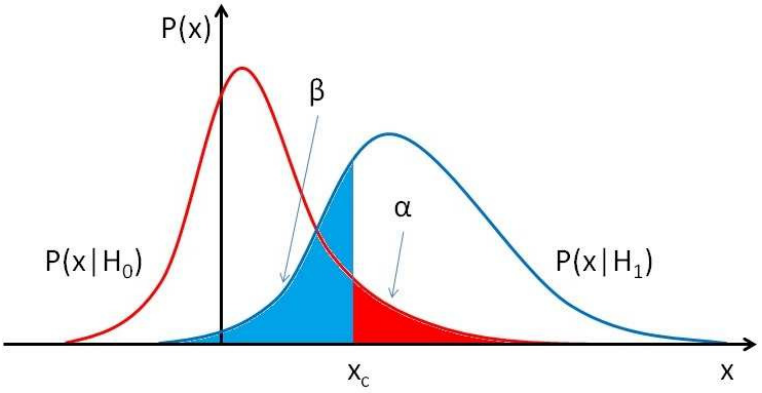

Figure 1. Meaning of α and β in the test of two simple hypotheses H_0 and H_1, here described by a parameter x. The critical region is $x > x_c$.

even if the alternative H_1 (or any of the alternatives, in the case of simple-versus-composite tests) is instead true.

Once the test statistic is defined, the choice of α for an experiment (*e.g.* to decide a criterion for a discovery claim, or to set a confidence interval) automatically implies a corresponding choice of β (or $\beta(\theta)$ in a simple-versus-composite test, when θ is the parameter describing H_1). In general there is no formal recipe to guide that decision. As exemplified in Fig. 1, the choice of a smaller value of α (*i.e.* a smaller type-I error rate), performed by moving the boundary of the critical region x_c to larger values, implies a higher chance of accepting a false null hypothesis (a larger type-II error rate β, and a smaller power $1 - \beta$).

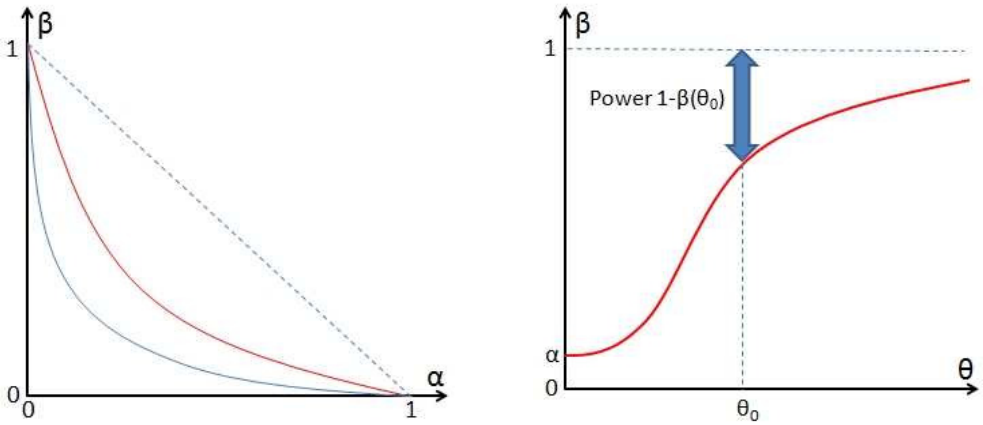

Figure 2. Left: Curves of β versus α for different tests in a simple-versus-simple setup. Curves closer to the origin have higher power $(1 - \beta)$ for a given test size α. Right: a power curve as a function of a parameter θ in a simple versus composite test..

The choices of α and β are thus conflicting: where to stay in the curve in the α versus β graph corresponding to one's analysis method highly depends on existing habits in one's research field. What makes a difference is the test statistic.

One may also study the power $1-\beta$ as a function of the parameter of interest, with graphs like the one shown in Fig. 2 (right). As the data size increases, the power curve becomes gradually closer to a step function; this corresponds to the two distributions shown in Fig. 1 becoming narrower and thus reducing their mutual overlap.

3 How 5-sigma became an established criterion in HEP

3.1 Bump searches in the sixties

In 1968 Arthur Rosenfeld wrote a paper titled "Are There Any Far-out Mesons or Baryons?" [1]. In the jargon of HEP in the sixties "far-out hadrons" indicated hypothetical hadrons not fitting 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. In the paper Rosenfeld demonstrated that the number of claims of discovery of such exotic particles published in scientific magazines in the sixties agreed reasonably well with the number of statistical fluctuations that one could expect to observe in the analyzed datasets. He examined the literature and pointed his finger at large "trials factors" (multiplicative factors affecting the p-value due to the multiple ways that an effect can manifest itself) coming into play due to the massive use of combinations of observed particles to derive mass spectra containing potential discoveries:

> *"[...] 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)."*

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

We will get back *infra* to the last point in the quote, *i.e.* the involuntary enhancement of spurious bumps. Rosenfeld used his argument to produce a ballpark estimate of the number of suggestive mass bumps that one could expect to arise in the data:

> *"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 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."*

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

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

Obviously particle physicists in the sixties were more "bump-happy" than we are today. The proposal to raise to 5σ 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.

It is instructive even for a hard-boiled sceptical physicist raised in the years of standard model precision-tests boredom to play with *Game*. In Fig. 3 are shown a few histograms, each selected by an automated procedure as the one containing *the most striking* peak among a set of 100 drawn from a smooth distribution. Histograms contain on average 1000 entries[2] distributed in 40 bins. The best histogram in each set of 100 is defined as the one with the most populated adjacent pair of bins (in the first two graphs) or triplets of bins (in the second set of two graphs). You are asked to consider what you would tell your students if they came to your office with one such histogram, claiming it is the result of an optimized selection for some doubly charmed baryon, say, that they have been looking for in their research project.

Each of the histograms shown in Fig. 3 is the best one in a set of a hundred; yet the isolated signals have p-values corresponding to roughly $3.5 - 4\sigma$ effects. In fact, some of the 2-bin bumps contain about 80 events, while the expectation is of 2*1000/40=50, and $p_{Poisson}(\mu = 50; N \geq 80) = 5.66 * 10^{-5}$, which corresponds to a significance of 3.86σ. Why do such large fluctuations arise by chance? Because the bump can appear *anywhere* in the spectrum, as we did not specify beforehand where we would look. This causes the probability of one fluctuation to increase by a trials factor of 39. Another reason is that we admit 2- as well as 3-bin bumps as "interesting", and we could extend the search to wider bumps: as there is no way to ensure that these choices are made *a priori*, they contribute to the trials factor.

One should also ponder on the often overlooked fact (but correctly identified as a source of trouble by Rosenfeld in the quotes *supra*) that researchers finding a promising bump will usually modify the selection *a posteriori*, voluntarily or involuntarily enhancing it. This makes the trials factor quite hard to estimate meaningfully.

[2]We sample the total number of entries N from a Poissonian distribution $P(N|\mu = 1000)$ to mimic a typical experimental situation.

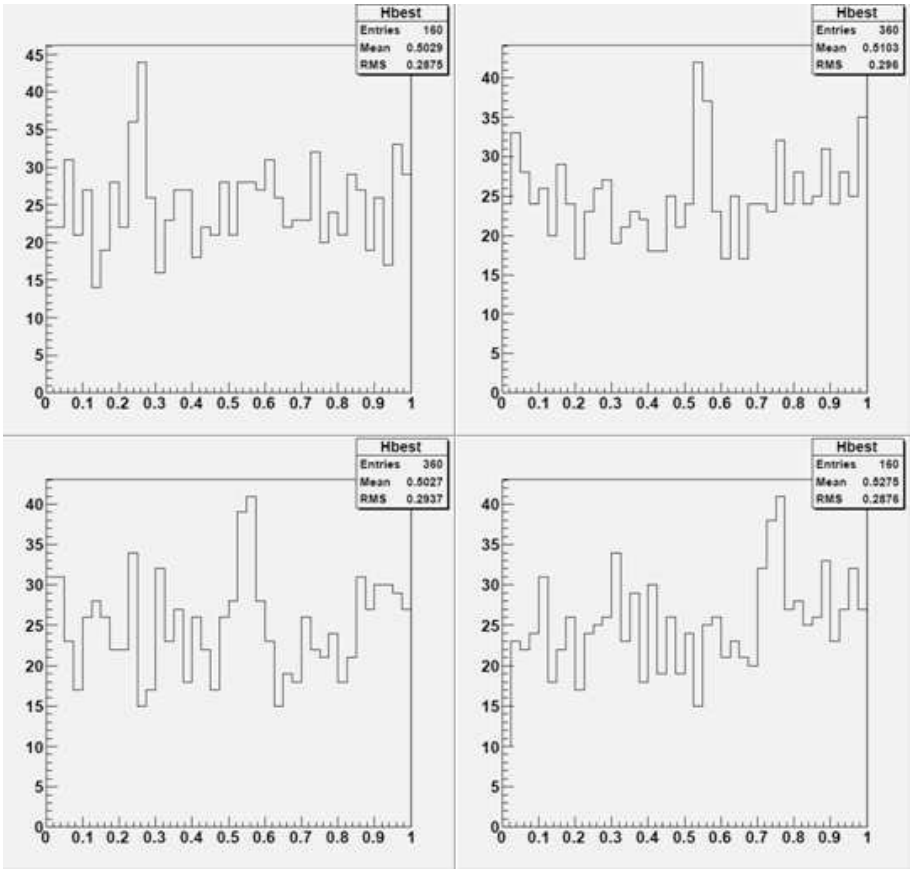

Figure 3. Example of histograms of random data drawn from a uniform distribution, selected by the procedure described in the text. Top: selected two-bin bumps. Bottom: selected three-bin bumps.

3.2 What 5-sigma may do for you

Setting the bar at 5σ for a discovery claim undoubtedly removes the large majority of spurious signals that originate due to statistical fluctuations. The trials factor required to reach 10^{-7} probabilities is of course very large, and yet the large number of searches being performed in today's experiments can still make up for that. Nowadays we call this *LEE*, for "look-elsewhere effect". 46 years after Rosenfeld published his study we do not need to compute the trials factor by hand: we can estimate a *global* as well as a *local* p-value using brute force computing, or an useful approximation which will be mentioned *infra* (see Sec. 4).

The other reason at the roots of the establishment of a high threshold for significance was the ubiquitous presence in HEP measurements of unknown, or ill-modeled, systematic uncertainties. To some extent, a 5σ threshold in fact protects systematics-dominated results from being hastily published as discoveries. Protection from large trials factors and unknown or ill-modeled systematics constitute the rationale behind the 5σ criterion. It is however worth stressing that 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, and other sciences.

3.3 How 5σ became a standard

A lot has happened in high-energy physics since 1968. In the seventies, the gradual consolidation of the standard electroweak model shifted the focus of particle hunts from random bump hunting to more targeted searches. It is useful to have a look at a few important searches for new particles or phenomena, in order to understand how the 5σ criterion gradually became a standard.

We may start with the November revolution. When the J/ψ discovery was announced in November 1974, statistical significance was not mentioned by the Brookhaven and Stanford groups that jointly claimed the new find: the observed effects were too big for anybody to bother fiddling with statistical tests. One year later, in the long arguments about whether a new lepton had been discovered by Martin Perl and collaborators in the $e\mu$ final state of electron-positron collisions at the Stanford Linear Collider, there still was no question on the significance of the observed excess; rather, a very long debate on hadron backgrounds ensued which lasted for at least a couple of years. Eventually Perl earned recognition for his discovery, and a well-deserved Nobel prize in 1995.

1976 was the year of the Oops-Leon, a potential resonance which was spotted in the mass distribution of pairs of muons by the team led by veteran bump hunter Leon Lederman. The authors explain:

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

And 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%."*

The latter estimate above must include a trials factor of the order of 20, as the Poisson probability to observe 12 or more events when 4 are expected is 0.09%. This hypothesis was confirmed to me by one of the authors[3] during a coffee break of the ICNFP 2014 conference.

For the real Υ discovery in 1977 the E-288 scientists were more careful. Burned by the "Oops-Leon" fiasco, they waited patiently for more data after seeing a promising 3σ peak at a mass of about 9.5 GeV; the only bold step forward I am aware of was the one of post-doctoral researcher John Yoh, who put in the refrigerator a bottle of Champagne with the tentative resonance mass written on its label with a marker. The E-288 team also performed several statistical tests to account for the trials factor (comparing MC probability to Poisson probability); and even after obtaining a peak with very large significance, they continued to investigate systematic effects. Their final announcement claims a discovery but does not quote a number of σ, noting however that the signal is indeed "statistically significant" [3].

Six years had to pass after the Υ observation before HEP got another major discovery. The W boson was announced on January 25^{th} 1983 by Carlo Rubbia on behalf of the UA1 experiment, based on finding six electron events featuring missing energy and no jets. No statistical analysis is discussed in the discovery paper [4], which however tidily and systematically rules out all possible backgrounds

[3] Daniel Kaplan, private communication.

as sources of the effect. It is worth noting that in the W search there was no trials factor to account for, as the signature was unique and predetermined; further, the theory prediction for the particle mass (82 ± 2 GeV) was matched well by the actual UA1 measurement (81 ± 5 GeV). The Z boson was discovered shortly thereafter, with an official CERN announcement made in May 1983 which was based on observing 4 events. Also for the Z no trials factor was applicable, due to the unicity of the signature. The article describing the find makes no mention of statistical checks [5], but it notes that background sources had been estimated to be negligible.

In 1994 the CDF experiment published a counting excess (amounting to 2.7σ) in b-tagged single-lepton and dilepton data, plus a towering mass peak at a value not far from the one predicted by indirect electroweak constraints. The mass peak, corroborated by some additional kinematic evidence, coresponded to an effect of over 3σ by itself. The unusually long article described the analysis in great detail, and spoke of "evidence" for top quark production [6]. One year later CDF and DZERO both presented [7] 5σ significances based on their counting experiments, obtained by analyzing three times more collision data. The top quark was thus the first particle discovered by a willful, disciplined application of the 5σ criterion.

Since the top discovery, the requirement of a p-value below $3 * 10^{-7}$ slowly but steadily became a standard. Two striking examples of searches that diligently waited for a 5-sigma effect before claiming discovery are the ones of single (*i.e.* electroweak-mediated) top quark production at the Tevatron and of the Higgs boson at the LHC. Single top quarks produced by electroweak processes in hadron-hadron collisions are harder to detect than top-antitop pairs produced by strong interactions, due to the less distinctive final state of the former process: it took 14 more years to conclusively observe it. The CDF and DZERO collaborations competed for almost a decade in the attempt to claim discovery of single top production, obtaining 2σ, then 3σ and 4σ effects, and only resolving to call their find "observation" in 2009 [8], when clear 5σ effects were observed by both experiments. Then, three years later it was the LHC's turn to be conservative: in 2012 the Higgs boson was claimed by ATLAS and CMS [9] only after obtaining 5σ significances each. The two experiments had mass-coincident $> 3\sigma$ evidence in their data already 6 months earlier, but the 5σ recipe was followed strictly, as already noted.

3.4 Discoveries that petered out

In April 1995 CDF collected an event that fired four distinct "alarm bells" of the online trigger monitoring program, Physmon. The event featured two clean energetic electrons, two clean photons, large missing transverse energy, and nothing else. It raised huge interest, as no standard model process appeared to come even close to explain its presence in the data: possible standard model expected rates were estimated to lay well below 10^{-7}. That was therefore a close to 6σ find. The observation [10] caused a whole institution to dive in a 10-year-long campaign to find "cousins" of the anomalous event and the search for an exotic explanation; it also caused dozens of theoretical papers as well as the revamping or development of SUSY models. In the Tevatron Run 2 no similar events were found; the competitor experiment DZERO did not see anything similar in its datasets, either.

In 1996 CDF found a clear resonance structure of b-quark jet pairs at 110 GeV, produced in association with photons. The signal [11] corresponded to an almost 4σ effect, and looked quite good –but there was no compelling theoretical support for the state, nor any additional evidence in orthogonal samples. The significance estimate did not pass the threshold for a discovery claim; after fiddling with it for a while, the researchers archived it. Yet 1996 was a prolific year for particle ghosts in the 100-110 GeV region: ALEPH also observed a 4σ-ish excess of Higgs-like events at 105 GeV in the 4-jet final state of electron-positron collisions at 130-136 GeV. They published the search [12], which found 9 events in a narrow mass region with a background of 0.7, and estimated the probability

of the effect at the 0.01% level. The ALEPH paper reports a large number of different statistical tests based on the event numbers and their characteristics. Of course one should note that a sort of Look-Elsewhere Effect is at work also when one makes many different tests.

Two years later CDF observed 13 "superjet" events in its Run 1 sample of W boson candidates featuring two or three additional hadronic jets; the superjet was a very energetic jet containing both a secondary vertex b-tag and a high-p_T electron or muon embedded in the jet core. A 3σ excess from background expectations could be estimated, but the most surprising aspect of that handful of events was their weird kinematics. A hypothesis test using a set of variables completely describing the event kinematics yielded a significance in the 6σ ballpark. The analysis was published [13] only after a fierce, three-year-long fight within the collaboration; the article did not claim any observation, but pointed out the weird observation. No similar effects were seen in the 100-times larger statistics of Run 2, so the explanation of that really anomalous find must lay in a mixture of researcher's bias (see Sec. 3.1) and a-posterioriness: to some extent the 13 anomalous events were singled out because of their weirdness, rather than based on an *a priori* selection strategy.

A more recent example of spurious signals is the one which appeared in 2004, when the H1 collaboration published a claim of having observed a pentaquark signal at 6σ significance [14]. Their prominent peak at 3.1 GeV was indeed suggestive; however it was not confirmed by later searches. In their paper the H1 researchers explain that

> *"From the change in maximum log-likelihood when the full distribution is fitted under the null and signal hypotheses [...], the statistical significance is estimated to be p=6.2σ".*

It is to be noted that H1 worded their result as an "evidence" in the title. That was a wise departure from the blind application of the 5σ rule, and one along the same line of reasoning offered *infra* (Sec. 5).

Claim	Significance				Verified or Spurius
Top quark evidence	3				Verified
Top quark observation			5		Verified
CDF $bb\gamma$ signal		4			Spurious
CDF ee$\gamma\gamma ME_T$ event				6	Spurious
CDF superjets				6	Spurious
B_s oscillations			5		Verified
Single top observation			5		Verified
HERA pentaquark				6	Spurious
ALEPH 4-jets		4			Spurious
LHC Higgs evidence	3				Verified
LHC Higgs observation			5		Verified
OPERA v>c neutrinos				6	Spurious
CDF Wjj bump		4			Spurious

Table 1. List of several observations of physics effects, claimed significance, and real nature of the effect. See the text for details.

A mention has also to be made of two recent, striking examples. In 2011 the OPERA collaboration produced a measurement of neutrino travel times from the CERN CNGS beam target to Gran Sasso which appeared smaller by 60ns than the travel time of light in vacuum [15]. The effect spurred lively

debates, got enormous media coverage, and triggered independent measurements by the neighbor ICARUS experiment; dedicated beam runs were performed to collect data less affected by jitter effects in the timing structure of the beam. After several months of investigations the effect was finally understood to be due to a single large source of systematic uncertainty, which had not been accounted for: the delay was produced by a loose signal cable [16]. In the same year the CDF collaboration showed a large, 4σ signal at 145 GeV in the mass distribution of jet pairs produced in association with leptonic W boson decays in the Tevatron 1.96 TeV proton-antiproton collisions [17]. The effect grew with data size and was clearly systematical in nature; the collaboration investigated it for over two years before finally understanding it as due to the combination of background contaminations and energy response differences in quark and gluon jets [18].

In light of the above information, one might be tempted to see an intriguing pattern in the correspondence between the parity of claimed significances of the effects and their genuinity, as shown in Table 1. More seriously, one feels bound to look a bit more into the causes that bring about large-significance effects later proven spurious, namely a large trials factor or unaccounted-for systematics (or non-Gaussian tails of accounted ones).

4 LEE, systematics, and other factors

From the quotes reported in the previous Section it transpires that a compelling reason for enforcing a very small test size α as a prerequisite for discovery claims is the presence of large trials factors. In principle, the LEE was a concern 50 years ago, but nowadays we have at our disposal an enormously greater CPU power. On the other hand, the complexity of our analyses has also grown considerably.

We can take the Higgs discovery as a classical example, if not a typical one: in order to reach the maximum possible sensitivity from their data the ATLAS and CMS collaborations combined together dozens of final states, with hundreds of systematic uncertainties. The latter were modeled as "nuisance parameters" (parameters not of interest but affecting the extraction of the measurement), some of them treated as partly correlated with others, or partly constrained by external datasets and ancillary measurements; often the assumed density of those nuisances was non-Gaussian. In such complex cases, despite the large computing power available today we still have trouble computing the trials factor satisfactorily by brute force.

A further complication which was not understood until recently is that in reality the effective trials factor of a search also depends on the significance of the local fluctuation, adding dimensionality to the problem. A study by E. Gross and O. Vitells [19] demonstrates how it is possible to produce a reasonable estimate of the trials factor with the data themselves in most experimental situations.

It is important to note that even if we can compute the trials factor using a large number of pseudo-datasets produced by toy simulations, or estimate it with approximate methods, there is always a degree of uncertainty in how to define it. In the classical case of mass bump searches, for instance, one may consider the multiplicity arising from the location parameter $\hat{m}$ and its freedom to lay anywhere in the considered mass spectrum; from the possibly unknown width of the signal peak; or from the fact that the final data selection may be the result of a non-blind investigation of several different selection cuts. Then one also needs to take into account the fact that one might have been searching for the signal in several possible final states. Further, one's colleagues in the experiment have probably been performing similar searches in different datasets. Overall, there is an ambiguity on the size of the LEE which depends on who you are: a graduate student, an experiment spokesperson, or a laboratory director. The bottomline is that while we can always compute a local significance for our search, it may not always be clear what we should quote as the true, "LEE-corrected", global significance.

4.1 Systematic uncertainties

Systematic uncertainties affect any physical measurement and it is sometimes quite hard to correctly assess their impact. Often one sizes up the typical range of variation of an observable due to the imprecise knowledge of a nuisance parameter at the 1-sigma level; then one stops there and assumes that the PDF of the nuisance be Gaussian. This is a reasonable assumption in the majority of cases. However, when the PDF of the nuisance parameter has wider tails than a Gaussian distribution, it makes the odd large bias much more frequent than estimated, such that large significances become increasingly meaningless. Furthermore, one should consider the possibility that additional non-considered sources of systematic uncertainty are present.

The potential harm of large non-Gaussian tails of accounted systematic effects or totally ignored ones can be seen as a reason for sticking to the very strict 5σ significance level even when we can somehow cope with the LEE. However, the safety margin that the criterion provides to avoid incorrect discovery claims is not always sufficient, as suggested by *e.g.* the OPERA neutrino speed measurement. One quick example to further stress the point is the following: if a 5σ effect has its uncertainty dominated by systematic sources, and the latter are underestimated by a factor of two, 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 has been overestimated by a factor 20,000.

A study of the distribution of residuals in measurements of particle properties was undertaken in 1975 using the large database collected in the Review of Particle Properties. The study revealed that the residuals were in fact not Gaussian. Matts Roos et al. [20] considered residuals in kaon and hyperon mean life and mass measurements, and concluded that they seemed to all have a similar shape, well described by a Student distribution $S_{10}(x/1.11)$:

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

Of course, one cannot extrapolate to 5σ the behaviour observed by Roos and collaborators in the bulk of the distribution; if one did, one would find that 5σ residuals are 1000 times more frequent than the simple Gaussian approximation would imply (see Fig. 4, right). One may consider this as evidence that the uncertainties evaluated in experimental HEP may have a significant non-Gaussian component. Detection and measurement techniques have however changed significantly in the course of the past forty years, so the effect can only be taken as a qualitative indication that caution is required when applying Gaussian approximations to nuisance parameters.

4.2 The "subconscious Bayes factor"

The term “Bayes factor” indicates the ratio of posterior to prior odds of the alternative hypothesis H_1 in a two-hypothesis test. Louis Lyons [21] named “subconscious Bayes factor” the ratio of prior probabilities we subconsciously assign to the two hypotheses under test. When comparing a background-only H_0 hypothesis with a background+signal H_1 one, one often uses the likelihood ratio $\lambda = L_1/L_0$ as a test statistic. To claim a discovery, the $p < 0.000029\%$ criterion is then applied to the distribution of λ under H_0. However, what would be more relevant to the claim would be the ratio of the probabilities:

$$\frac{P(H_1|data)}{P(H_0|data)} = \frac{p(data|H_1)}{p(data|H_0)} \times \frac{\pi_1}{\pi_0} = \lambda\frac{\pi_1}{\pi_0} \tag{3}$$

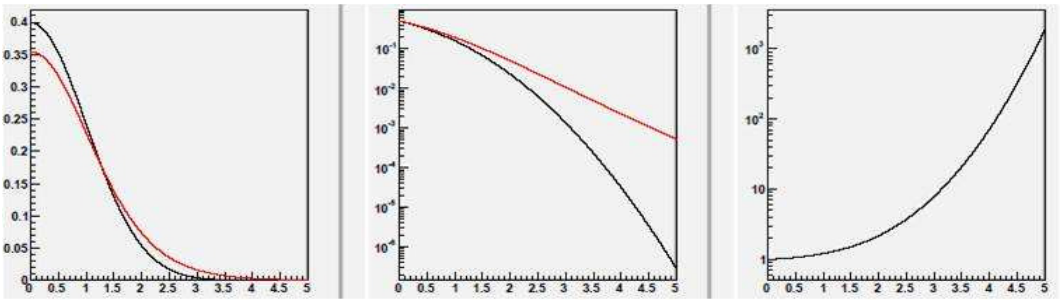

Figure 4. Left: Student S_{10} distribution (in red, curve with higher tails on the right) compared to a Gaussian. Center: distributions of the tail integral functions of the two distributions, $f_1(x) = \int_x^\infty S_{10}(t/1.11)dt$ and $f_2(x) = \int_x^\infty G(t)dt$. Right: ratio of the functions shown in the center graph as a function of x.

where $p(data|H)$ are the likelihoods, and π are the priors of the hypotheses under test. In that case, if our prior belief in the alternative hypothesis p_1 were low, we would still favor the null hypothesis even in presence of a large evidence λ against it.

The one described is a legitimate application of Bayes' theorem, yet the majority of HEP physicists prefer to remain in Frequentist territory and avoid assigning probabilities to the hypotheses under test. Lyons however notes:

> *"This type of reasoning does and should play a role in requiring a high standard of evidence before we reject well-established theories: there is sense to the oft-quoted maxim 'extraordinary claims require extraordinary evidence'."*

4.3 The issue of the "point null" and the Jeffreys-Lindley paradox

All what we have discussed so far makes sense strictly in the context of classical Frequentist statistics; on the other hand one might well ask what is the Bayesian view of the problem. The issue revolves around the existence of a null hypothesis, H_0, on which we base a strong belief. It is quite special to physics that we do believe in our "point nulls". The standard model is a classic example, as it works only when some of its parameters have very specific values, which are known with arbitrary accuracy; such is the case *e.g.* for the mass of the photon, which is exactly zero in the standard model; or the absolute equality of proton and positron electric charges. In other sciences a true point null hardly exists.

The fact that we must often compare a simple null hypothesis (for which a parameter θ has a very specific value θ_0) to a composite alternative (where the parameter under test may take any value in a continuous range) bears on the definition of a prior belief for the parameter. Bayesians speak of a "probability mass" at $\theta = \theta_0$. The use of probability masses in the priors in a simple-vs-composite test throws a monkey wrench in the Bayesian calculation. In fact, it can be proven that no matter how large and precise is the data, Bayesian inference strongly depends on the scale over which the prior is non-null: that is, on the prior belief of the experimenter. The Jeffreys-Lindley paradox [22] which arises in that situation may bring Frequentists and Bayesians to draw opposite conclusions on some data when comparing a point null to a composite alternative. This fact bears relevance to the kind of tests we are discussing, hence it is useful to review the paradox below.

We take X_1 ...X_n as independent and identically-distributed as $X_i|\theta \sim N(\theta, \sigma^2)$, *i.e.* Normally distributed, and a prior belief on θ constituted by a mixture of a point mass probability p at $\theta = \theta_0$ and $(1 - p)$ uniformly distributed in $[\theta_0 - I/2, \theta_0 + I/2]$. Here I being the width of the interval over which we consider the parameter to have any chance of lying. In classical hypothesis testing, the "critical values" of the sample mean delimiting the rejection region of H_0: $\theta = \theta_0$ in favor of H_1: $\theta \neq \theta_0$ at significance level α are

$$\bar{X} = \theta_0 \pm (\sigma/\sqrt{n}) z_{\alpha/2} \tag{4}$$

where $z_{\alpha/2}$ is the significance corresponding to test size α for a two-tailed Normal distribution. Given the above, it can be proven that the posterior probability that H_0 is true conditional on the data in the critical region (*i.e.* excluded by a classical α-sized test) approaches 1 as the sample size becomes arbitrarily large.

As evidenced by Bob Cousins [23], the paradox arises when there are three different scales in the problem, $\epsilon < \sigma/\sqrt{(n)} < I$, *i.e.* the width of the point mass, the measurement uncertainty, and the scale I of the prior for the alternative hypothesis (see Fig. 5). The three scales are usually independent in HEP, and this makes the paradox extremely relevant there.

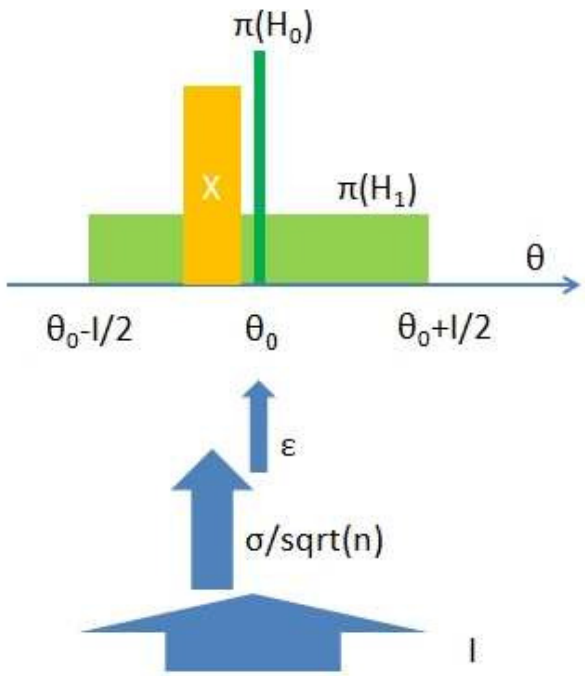

Figure 5. The figure sketches the existence of three different scales in a problem of simple versus composite hypothesis testing: the scale where the null hypothesis is non-zero (ϵ); the scale set by the measurement precision $\sigma/\sqrt{n}$ of data X; and the scale where the continuous prior under the alternative is non-null, I.

I provide a proof of Jeffreys-Lindley paradox in what follows. We wish to compute the posterior probability P that H_0 is true given data that lay in the critical region. We start by writing it using Bayes' theorem as

$$P(H_0|\bar{X} = \bar{x} = \theta_0 + (\sigma/\sqrt{n})z_{\alpha/2}) = \frac{P(H_0)P(data|H_0)}{P(H_0)P(data|H_0) + P(H_1)P(data|H_1)} \tag{5}$$

We now insert in the above expression the actual priors p and $(1-p)$ and the likelihood values in terms of the stated Normal density of the i.i.d. data X:

$$= \frac{p\frac{\sqrt{n}}{\sqrt{2\pi}\sigma}e^{(-1/2)[(\sqrt{n}/\sigma)(\hat{x}-\theta_0)]^2}}{p\frac{\sqrt{(n)}}{\sqrt{2\pi}\sigma}e^{(-1/2)[(\sqrt{n}/\sigma)(\hat{x}-\theta_0)]^2} + (1-p)\int_{\theta_0-I/2}^{\theta_0+I/2}\frac{\sqrt{n}}{\sqrt{2\pi}\sigma}e^{(-1/2)[(\sqrt{n}/\sigma)(\hat{x}-\theta_0)]^2}\frac{1}{I}d\theta} \tag{6}$$

Now we can rewrite two of the exponentials using the conditional value of the sample mean in terms of the corresponding significance z, and remove the normalization factors $\sqrt{n}/(\sqrt{2\pi}\sigma)$:

$$= \frac{pe^{-(1/2)z^2_{\alpha/2}}}{pe^{-(1/2)z^2_{\alpha/2}} + \frac{1-p}{I}\int_{\theta_0-I/2}^{\theta_0+I/2}e^{(-1/2)[(\sqrt{n}/\sigma)(\hat{x}-\theta_0)]^2}d\theta} \tag{7}$$

Finally, we maximize the expression by using the integral of the Normal distribution:

$$= \frac{pe^{-(1/2)z^2_{\alpha/2}}}{pe^{-(1/2)z^2_{\alpha/2}} + \frac{1-p}{I}\frac{\sqrt{2\pi}\sigma}{\sqrt{n}}} \to 1 \tag{8}$$

that is, P goes to 1 as $n \to \infty$, *i.e.*, as the data size grows indefinitely, the posterior of the null hypothesis becomes unity.

The paradox is often used by Bayesians to criticize the way inference is drawn by Frequentists. *E.g.* Jeffreys:

> *"What the use of [the p-value] implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred" [24].*

Alternatively, the criticism concerns the fact that no mathematical link between p and $P(H|x)$ exists in classical hypothesis testing. On the other hand, the problem with the Bayesian approach is that there is no clear substitute to the Frequentist p-value for reporting experimental results. Bayesians prefer to cast the hypothesis test problem as a Decision Theory one, where by specifying the loss function one is allowed to design a quantitative and well-specified (although subjective) recipe to choose between alternatives. Yet Bayes factors, which describe by how much prior odds are modified by the data, are not factorizing out the subjectivity of the prior belief when the Jeffreys-Lindley paradox holds: even asymptotically, they retain a dependence on the scale where the prior of the alternative hypothesis is non-null.

In their debates on the Jeffreys-Lindley paradox, Bayesian statisticians have blamed the concept of a point mass, as well as suggested n-dependent priors. There is a large body of literature on the subject; for suitable references the reader is advised to look into the cited paper by Cousins. As assigning to the null hypothesis a non-zero prior is the source of the problem, Bayesian statisticians tend to argue that "the precise null" is never true. However, physicists do believe their point nulls, especially in particle and astro-particle physics.

To come back to the issue of the choice of α for discovery claims, the Jeffreys-Lindley paradox draws attention to the fact that a fixed level of significance does not cope with a situation where the amount of data increases, which is common in particle physics. Hence the trouble of defining a test size in a classical hypothesis testing is not automatically solved by moving to Bayesian territory.

5 So what to do with 5σ ?

I believe it is useful to summarize here the points made in the previous section.

1. The LEE can be estimated analytically as well as computationally; experiments in fact now routinely produce "global" and "local" p-values and significances for the fluctuations or signals they observe in their data. Hence one might argue that there is no point in choosing a small test size to account for large trials factors, which was the original motivation of Rosenfeld as discussed in Sec. 1. Sometimes the trials factor is 1 and sometimes it is enormous; a one-size-fits-all approach is then hardly justified, and it is illogical to penalize an experiment for the LEE of others.

2. As far as systematic uncertainties are concerned, their impact varies widely from case to case. Sometimes one has control samples of data to verify the absence of unknown effects (*e.g.* in particle searches); in other cases one does not (like in the neutrino speed measurement by OPERA).

3. The cost of a wrong claim, in terms of image damage or the backfiring of media hype, can vary dramatically.

4. Some claims appear intrinsically less likely to be true because we have a subconscious Bayes factor at work. How much value you give to a significance estimate does depend on whether you are discovering a new meson or a violation of physical laws.

Given the points listed above, you could ask why we should settle on a fixed discovery threshold. One may take the attitude that any claim is subject to criticism and independent verification, and the latter is always more rigorous when the claim is steeper and/or more important; and it is good to just have a reference value for the level of significance of the data. One also often hears the argument that the 5σ criterion is a *tradition* and an useful standard. Yet the issue remains.

One suggestion to overcome the impasse comes from a recent paper by Louis Lyons [21]. He considered several known searches in particle and astro-particle physics, both past and ongoing ones, and produced a table where for each effect he listed several of the inputs we discussed *supra*: the degree of surprise of the potential discovery of the effect, its impact on the progress of science, the size of the trials factor at work in the search, and the potential impact of unknown or ill-quantifiable systematics. Lyons then derived a reasonable significance threshold which accounted for the different factors at work in each of the considered effects. Such an approach is of course only meant to provoke a discussion, and the numbers in Lyons' table are entirely debatable. The message is however clear: we should be wary of a one-size-fits-all standard. For the sake of this discussion I have slightly modified the original table to reflect my personal bias on some of the inputs. Table 2 is thus a subjective view of the situation.

Search	Surprise level	Impact	LEE	Systematics	Z-level
ν oscillations	Medium	High	Medium	Low	4
B_s oscillations	Low	Medium	Medium	Low	4
Single top prod.	Absent	Low	Absent	Low	3
$B_s \to \mu\mu$	Absent	Medium	Absent	Medium	3
Higgs boson	Medium	Very High	Medium	Medium	5
SUSY searches	High	Very High	Very High	Medium	7
Pentaquarks	High	High	High	Medium	6
G-2 anomaly	High	High	Absent	High	5
H spin>0	High	High	Absent	Low	4
4^{th} gen. fermions	High	High	High	Low	6
v>c neutrinos	Huge	Huge	Absent	Very High	THTQ
Direct DM search	Medium	High	Medium	High	5
Dark energy	High	Very High	Medium	High	6
CMB tensor modes	Medium	High	Medium	High	5
Grav. waves	Low	High	Huge	High	7

Table 2. Possible discovery-level significances (Z-level, last column) of several past and present searches for real or hypothetical phenomena, according to the personal opinion of the author. THTQ = too high to quote.

5.1 Too high to quote?

In Table 2 I voluntarily refrained to quote a proper significance level for one of the considered effects, reasoning that no single striking observation, regardless of the size of the effect, could convince me of the reality of the claimed phenomenon. The loss of meaning of very high significance levels brought in by that consideration has however another independent cause.

I recently heard the following claim from a respected astrophysicist who was giving a talk at a workshop: *"The quantity has been measured to be non-zero at* 40σ *level"*. He was referring to a measurement which had been quoted by its authors as as $x = 0.110 \pm 0.0027$. I believe that was a really silly statement, and a very improper usage of the Gaussian approximation. In fact, as the number of significance units goes above 7 or so we are rapidly losing contact with the reality of experimental situations. To claim *e.g.* a 5σ effect, one has to be reasonably sure to know the PDF of the p-value to the 10^{-7} level or below, for we have to recall that the number of sigmas is just a proxy for a small number, no less than are funny measurement units such as femtobarns or attometers. Hence before quoting blindly very large significances, we should really think about what they really mean. In the case of the astrophysicist, it is not even easy to directly make the conversion, as most of the common Gaussian-integral calculation routines break down when the lower bound of the integrated region goes above 7.5. We must resort to approximations, like the one by Karagiannidis and Lioumpas [25],

$$Q(x) = \frac{(1 - e^{-1.4x})e^{-\frac{x^2}{2}}}{1.135\sqrt{2\pi}x}, \; x > 0. \tag{9}$$

For $N = 40$ my computer still refuses to return anything larger than 0, but for $N = 38$ it gives $p = 2.5 * 10^{-316}$. It transpires that the astrophysicist quoted above was basically saying that the data had a probability of less than a part in 10^{316} of being observed if the null hypothesis held. That claim qualifies for one of the steepest claims ever made by a scientist; it is beyond ridiculous. Of course, we will never be able to know the tails of our systematic uncertainties to a level of precision even remotely similar to that.

6 Conclusions

Forty-six years after the first suggestion of a 5σ threshold for discovery claims, and 20 years after the start of its consistent application, the criterion appears inadequate to address the experimental situation in particle and astro-particle physics. In fact it did not protect us from steep claims that later petered out, while it significantly delayed acceptance of some relatively uncontroversial finds. The search for electroweak production of single top quarks at hadron colliders is a prime example of the latter shortcoming: in Run 2 at the Tevatron the DZERO and CDF collaborations competed for eight years to be the first to reach a 5σ observation, when in fact they could have used their thinning forces much better in other searches. A fixed discovery threshold is arbitrary and illogical in many aspects, as I hope this article has shown.

A solution that many advocate is to switch to Bayesian hypothesis testing. However, Bayesian hypothesis testing does not appear ready to offer a robust replacement for the procedures of experimental particle physics. The Jeffreys-Lindley paradox is still an active area of debate, and there appears to be no consensual view on how to address the problem in the professional statistics literature.

One suggestion to break the impasse is that for each considered search the community should seek a consensus on what could be an acceptable significance level of a media-hitting claim. Probably five standard deviations are insufficient to convince the community of the genuine nature of observations of unpredicted effects, and on the other hand a smaller significance would be advisable for effects that are expected and well defined.

Acknowledgements

I wish to thank the organizers of the Third International Conference on New Frontiers in Physics (ICNFP 2014) for their great work and for offering me to give an invited lecture there. I also thank Bob Cousins and Louis Lyons for providing inspiration for the present report, and for the invaluable insight they offered on the subject during a number of discussions. Finally, special thanks are due to Matthew Dearing, Michael McCracken, Alex Reinhart, and Matthew West for reviewing the draft of this article.

References

[1] A.H. Rosenfeld, “Are there any far-out mesons and baryons?”, In: Baltay, Rosenfeld (eds.), Meson Spectroscopy: A collection of articles, W.A. Benjamin, New York (1968) 455-483.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[19] E. Gross and O. Vitells, "Trials factors for the Look-Elsewhere Effect in High-Energy Physics", Eur. Phys. Journ. C 05 (2010) 70.

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

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

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

[23] R.D. Cousins, "The Jeffreys-Lindley Paradox and Discovery Criteria in High-Energy Physics", arxiv:1310.3791v4, June 28th, 2014; to appear in Synthese (2014).

[24] H. Jeffreys, "Theory of Probability", 3rd ed., Oxford University Press, Oxford (1961) 385.

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