Real, False and Missed Discoveries in High Energy Physics

Luc Demortier
The Rockefeller University

Terascale Statistics Tools School
DESY, Wednesday October 1, 2008

Outline

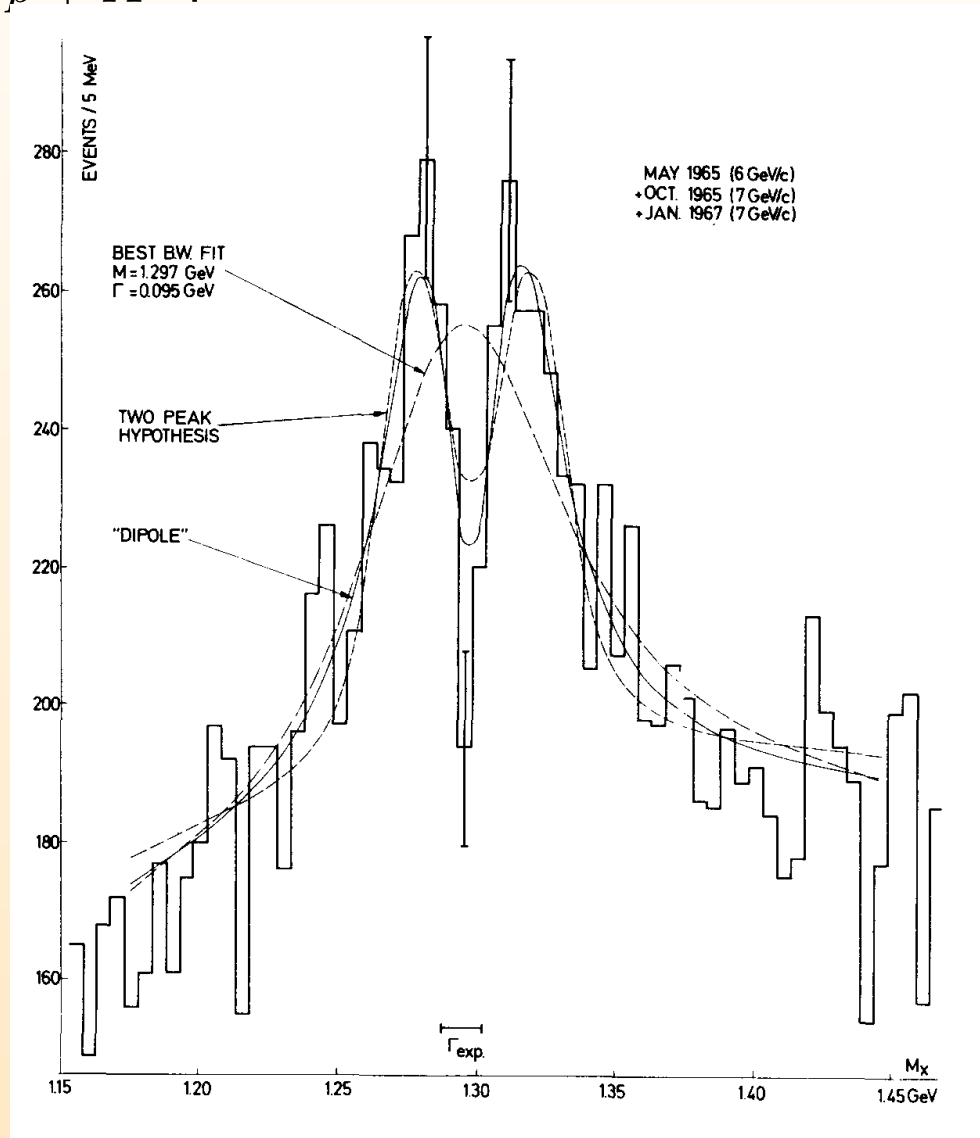
1. Examples of discoveries. . .
2. The choice of 5σ
3. P value pathologies
4. Pathological science?
5. Blind analyses
6. Summary

Disclaimer: I am neither a historian of science nor a philosopher. . .

EXAMPLES OF DISCOVERY AND NON-DISCOVERY CLAIMS AND NON-CLAIMS, AND LESSONS LEARNED AND NOT LEARNED

The A_2 meson splitting (1)

In 1967 the CERN missing mass spectrometer group reports the observation of a narrow dip of six standard deviations in the center of the A_2 peak obtained in the reaction $\pi^- + p \to p + X^-$:



The A_2 meson splitting (2)

Significance of this observation

Several fits to the data are tried:

1. Fit to a single Breit-Wigner peak yields a χ^2 p value of 0.1% ($\sim 3.1\sigma$);
2. Fit to two incoherent Breit-Wigner peaks yields $p = 15\%$;
3. Fit to two coherent Breit-Wigner peaks with equal masses, equal widths, and destructive phase yields $p = 70\%$.

If either fit 2 or 3 is correct, the resulting object will not fit easily in the quark-antiquark model of mesons:

1. If it is two particles with the same quantum numbers, this adds an uninvited tenth guest to the A_2's SU(3) nonet.
2. If it is a single object, why don't the other members of the same SU(3) nonet exhibit this behavior?

The A_2 meson splitting (3)

Historical development

1. Several other experiments observe a split A_2 in various reactions, but with much smaller significance ($\leq 3\sigma$).
2. Finally, as more data are taken, the split disappears and the A_2 remains a single, undivided particle.

Lessons learned

In his opening address to the International Conference on Hadron Spectroscopy in Maryland in 1985, S. J. Lindenbaum writes:

> "The only reasonable explanation I can think of (remembering discussions at the time) is that overzealous searching for the effect led the original investigators into inadvertent data selection which selected those statistical fluctuations in a Breit-Wigner which showed a split and rejected the other data samples."

The A_2 meson splitting (4)

Lindenbaum then goes on to describe what he calls the bandwagon effect:

- After the first report of a new effect, many investigators look at their data, and those with marginal significance ($\sim 3\sigma$) pointing to the same effect report confirmation of the original work.

- Those showing no effect assume that they are likely experiencing a statistical fluctuation or a systematic effect, and generally do not report their measurement.

The net effect is to amplify the perceived significance of the initial observation.

The lesson is that when one is checking an effect, one should make the decision to publish the result independently of preconceived notions about the outcome.

Neutral Currents (1)

Serious interest in searching for neutral currents began after 't Hooft proved in 1971 that the electroweak theory of Weinberg and Salam was renormalizable. Two experiments were looking for neutral currents, although not at the highest priority:

1. Gargamelle, a heavy liquid bubble chamber at the CERN PS;
2. Experiment 1A at Fermilab, consisting of a calorimeter that was at the same time a target, followed by a muon spectrometer.

In some sense these experiments were complementary: the bubble chamber produced beautiful event pictures from which detailed information could be obtained, but locating events of interest required vast amounts of film and time. On the other hand, spark chambers in experiment 1A could be triggered on, thereby increasing the rate of usable information.

At the International Conference on Electron and Photon Interactions at High Energies in Bonn in August 1973, both experiments reported observations suggesting neutral currents, as well as measurements of the NC:CC ratio.

Neutral Currents (2)

The background problem:

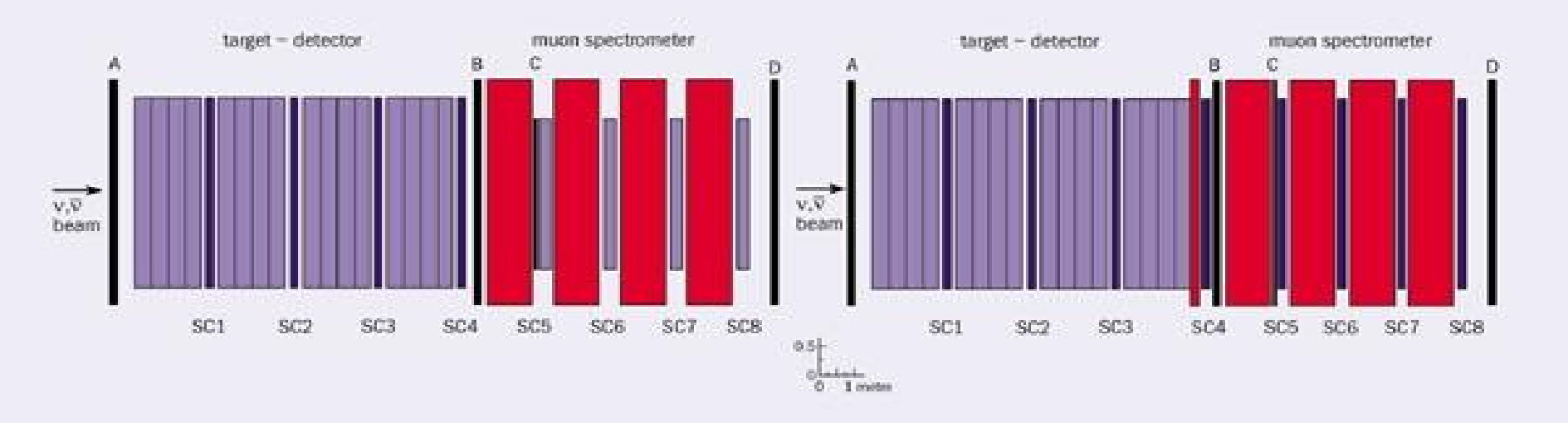
The results presented by both experiments in Bonn were on searches for *hadronic* neutral currents interactions of a $\nu_\mu/\bar{\nu}_\mu$ beam. In this case the final states of hadronic NC and CC both contain a hadronic shower, but CC also contains a muon. Therefore, if you think you observed a NC event, you must make sure you did not miss a muon somewhere, otherwise the event is really a CC! Hence the background problem:

- For Gargamelle, a neutrino could interact via CC somewhere outside the chamber, the muon could escape detection, and a neutron from the hadron shower could enter the chamber, hit a proton or neutron somewhere inside, and create a hadron shower that mimics a NC event.

- For experiment 1A, a neutrino could interact via CC inside the calorimeter target, the muon could escape at wide angle with respect to the detector axis, thereby avoiding detection in the muon spectrometer.

Neutral Currents (3)

Shortly after the Bonn conference, the Fermilab experiment withdrew its result. What had happened is that they had reconfigured their detector to reduce the correction for wide angle muons in CC events:



This resulted in a catastrophic increase in hadronic punch-through, so that genuine NC events ended up being classified as CC events, and the NC signal disappeared. It took a while to figure this out (strong interactions were not well understood at that time), and eventually the NC signal reappeared and was confirmed by experiments at Argonne, BNL, and Fermilab.

Neutral Currents (4)

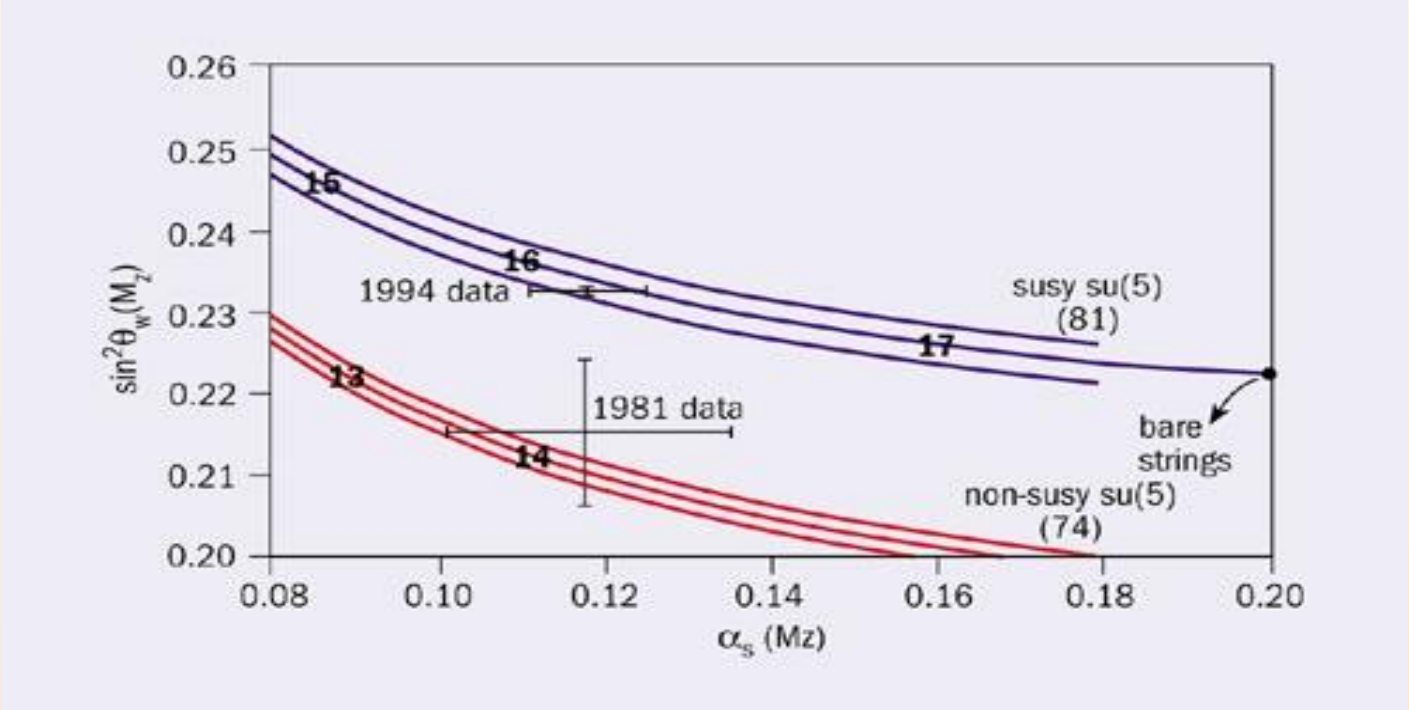
Lessons learned

1. If you do not understand the background, and especially if you do not know you do not understand the background, you are in big trouble, whether you report a positive result or a negative one.

2. Of course, the same is true for signal. In the early sixties Lederman, Schwartz, and Steinberger performed a Nobel-prize winning experiment that proved the existence of muon-neutrinos. Since they needed to find muons, they were not interested in some odd events that they had observed and that did not have a muon in the final state. These events were privately christened "crappers". Today we would call them neutral currents. . .

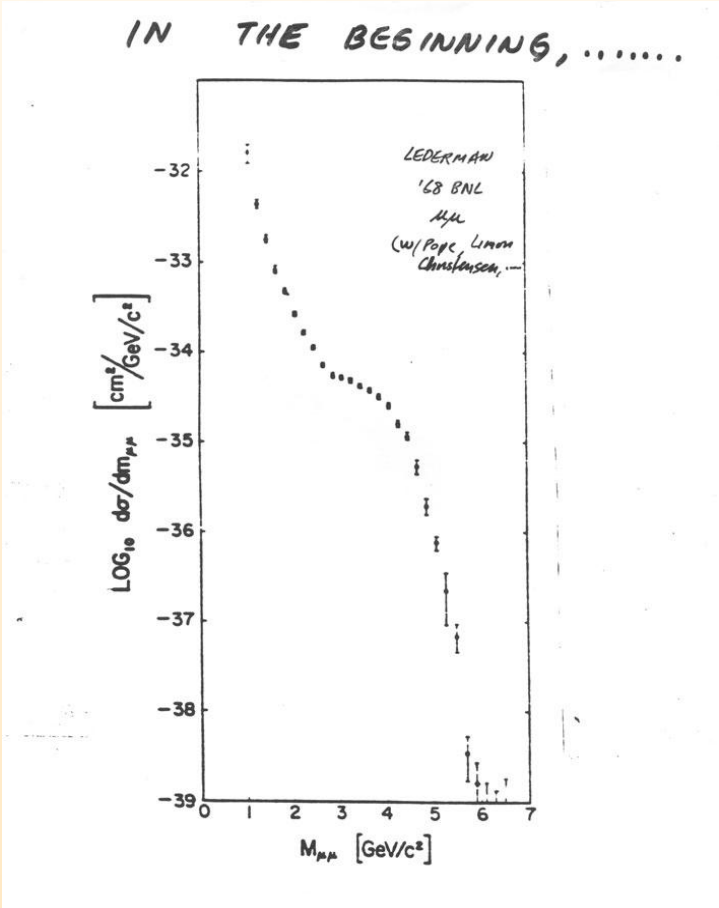
The measurement of $\sin^2\theta_W$

In a June 2003 CERN courier article about neutral currents, Don Perkins points out that wrong results can lead to very significant and positive consequences. As a case in point, he comments on the history of CERN measurements of $\sin^2\theta_W$ in neutrino experiments. In the 1970's the BEBC found an anomalously low value of $\sin^2\theta_W$, bringing the world average down to 0.21, in agreement with predictions of non-SUSY SU(5). To test this model of grand unification, underground searches for proton decay were started on three continents. A major background in these searches was atmospheric neutrino interactions, whose analysis eventually led to evidence for neutrino oscillations and masses. . .



Leon Lederman's $\mathcal{J}/\psi$

In 1968 Lederman and collaborators were studying dimuons at BNL. By measuring the range and direction of each muon, they could reconstruct the mass of the dimuon, albeit with a mass resolution of about 1 GeV at a mass of 3 GeV. The surprising observation of a large rate of direct dimuon production led to the famous Drell-Yan paper.

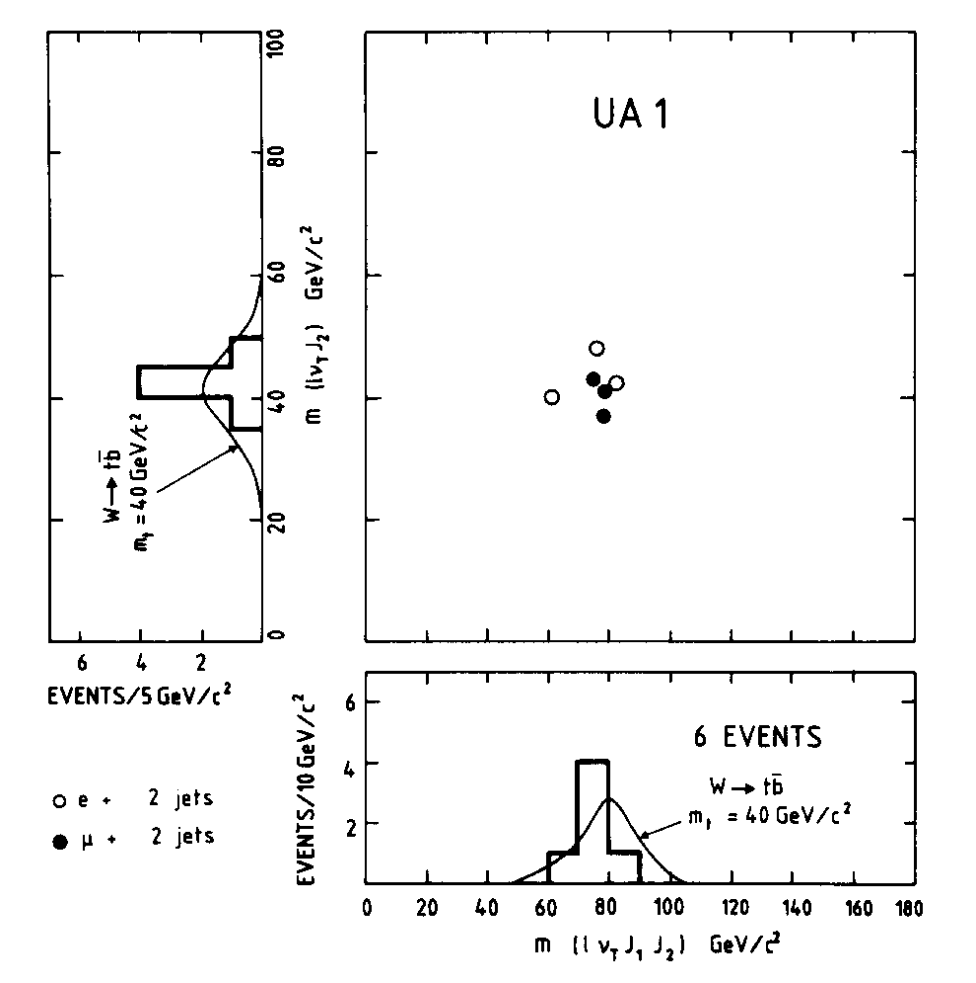


What is this bump?

Some theorists claimed that it was just another ρ' resonance, and others that it could be explained without resorting to resonances.

The Evidence for the Top Quark (1)

In 1984 UA1 published some evidence for top quark production, at the 3.4σ level ($p = 3 \times 10^{-4}$). If the six observed candidates were indeed top quarks, the top mass would be 40 ± 10 GeV/c^2 [G. Arnison *et al.*, Phys. Lett. B**147**, 493 (1984)].



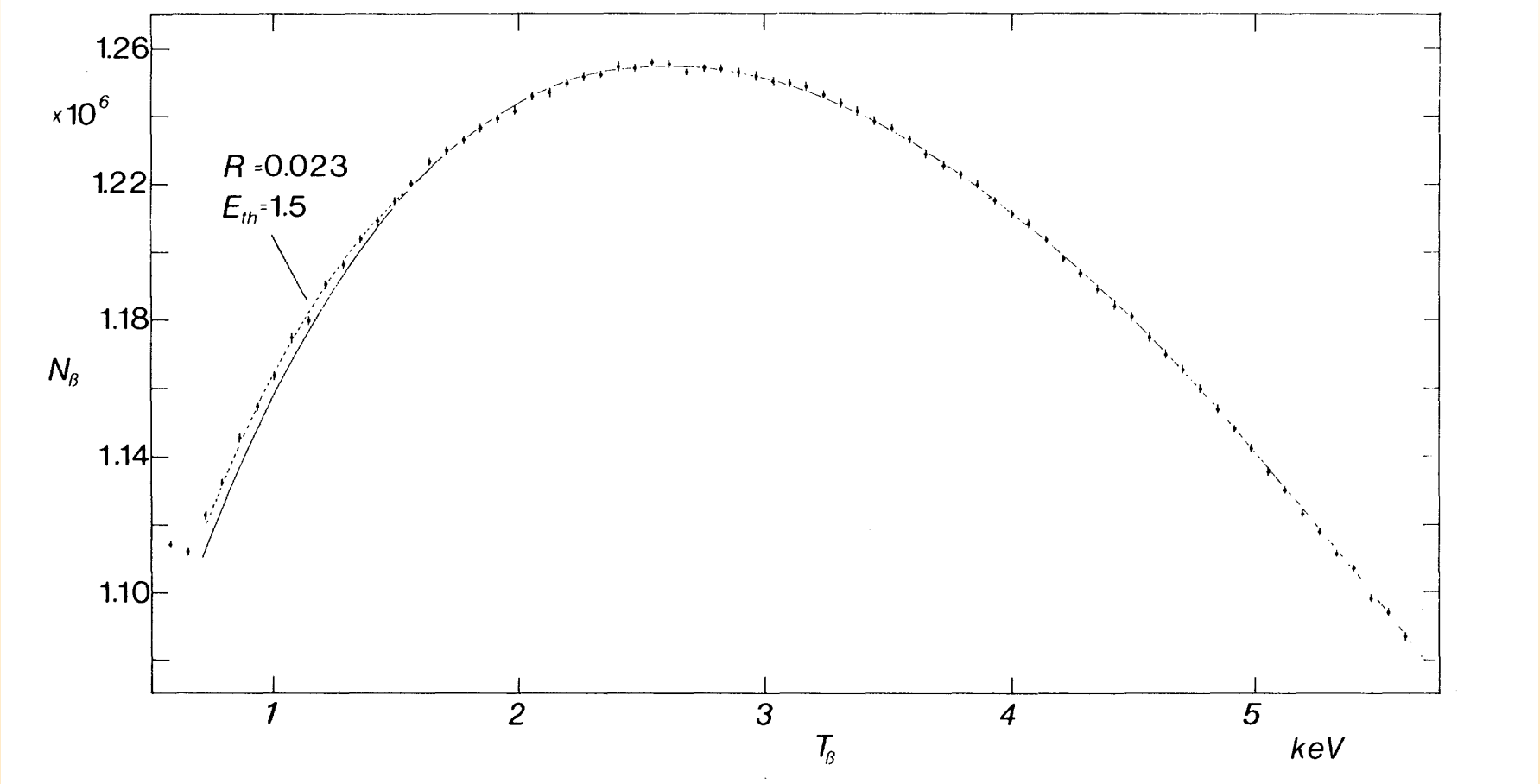
The Evidence for the Top Quark (2)

Unfortunately UA2 could not confirm this evidence, and it took until 1995 for the CDF and DØ experiments at the Tevatron to announce the discovery of the top quark, each with close to 5σ significance. The mass of the top quark turns out to be 172.4 ± 1.2 GeV/c^2 (July 2008 Tevatron combination).

An interesting point is that, although everybody places the top quark discovery in 1995, we are just now beginning to accumulate large enough samples to measure such elementary properties as the charge of the top quark. There is even a theory that claims that the quark discovered in 1995 is not the top, with charge $+2/3$ and decaying into W^+b, but an exotic quark with charge $+4/3$ and decaying into $W^+\bar{b}$. Furthermore, consistency of this exotic model with the rest of the electroweak dataset can be maintained provided the real top quark has a mass around 270 GeV/c^2. At present the *only* way to settle this question is to measure the top quark charge. The evidence for the top quark is not complete yet. . .

The 17 keV neutrino (1)

In 1985 John Simpson announces the discovery of a new species of neutrino's, with a mass of 17 keV, based on a measurement of the energy spectrum of β particles emitted during the decay of tritium (Phys. Rev. Lett. **54**, 1891 (1985)):



The 17 keV neutrino (2)

Significance of this observation (with hindsight)

1. Standard Model: A 17 keV neutrino must be "exotic" in order to remain consistent with the LEP results (e.g. coupling to Z^0 very weak);
2. Astrophysics: 17 keV is far above the maximum mass consistent with dark matter theories and astronomical observations.

Historical development

1. Several groups repeat the measurement and find zero effect. Critics point out that atomic effects are important for tritium in the low-energy region below 1.5 keV and could cause the observed kink.
2. Simpson reduces the mixing probability of the new neutrino from 3 to 1%. He also points out that the null experiments are conducted in a magnetic field, which requires the spectrum shape to be corrected; this could conceal the presence of a kink. He also claims that some null experiments do in fact show a kink.

The 17 keV neutrino (3)

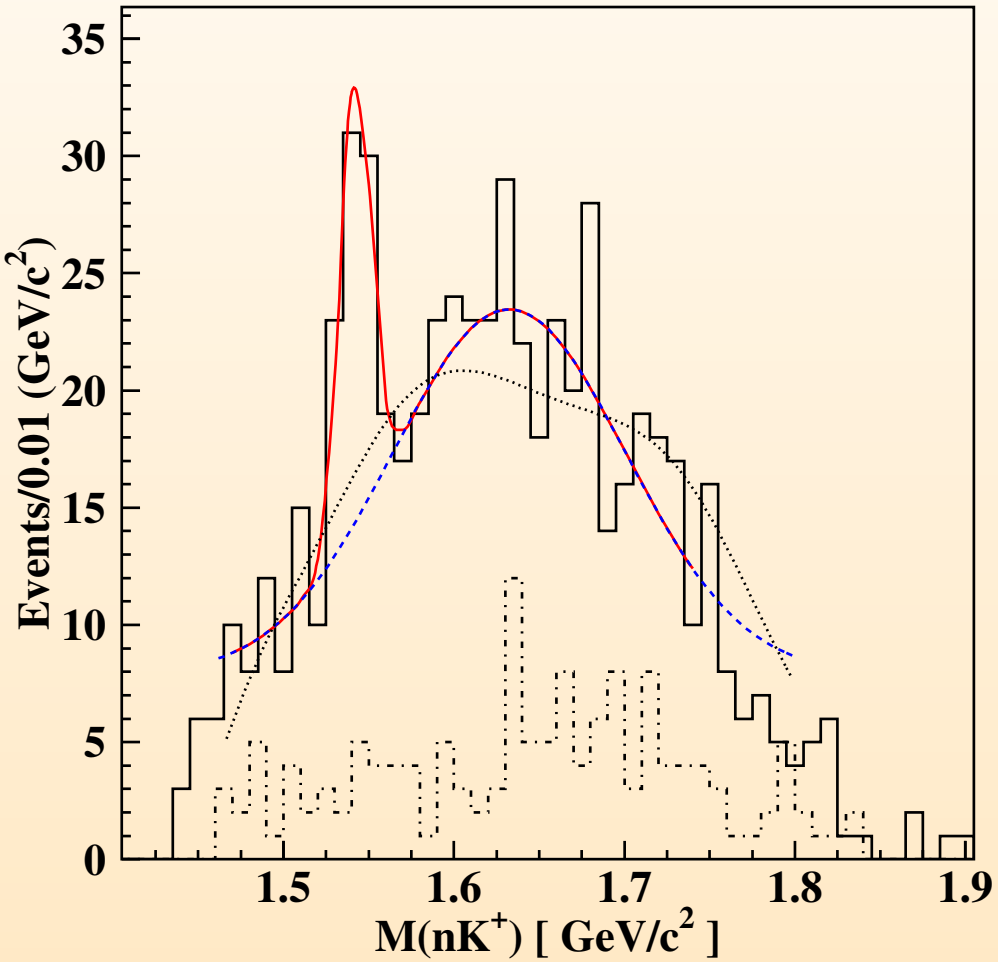
3. In 1989 Simpson and Hime perform a new experiment on the decay of both tritium and ^{35}S, whose β spectrum has an endpoint of 165 keV. This puts the expected kink at 150 keV, far from the region sensitive to atomic effects. Kinks are found, with a mixing ratio of about 1%.
4. It is later found that backscattering of electrons from the target could be a problem, and collimators are introduced to try to deal with this effect.
5. In the early 90's ingenious new experiments are performed, which answer all objections to the previous null experiments. They find zero effect. Various exclusion limits on the 17 keV neutrino are published.

Lessons learned

1. Over-enthusiasm and selection of only positive experiments by some theorists caused this story to have more impact than it deserved.
2. Considerable ingenuity was needed to design experiments that avoided the objections made against the initial generation of null experiments.

The Pentaquark Sightings (1)

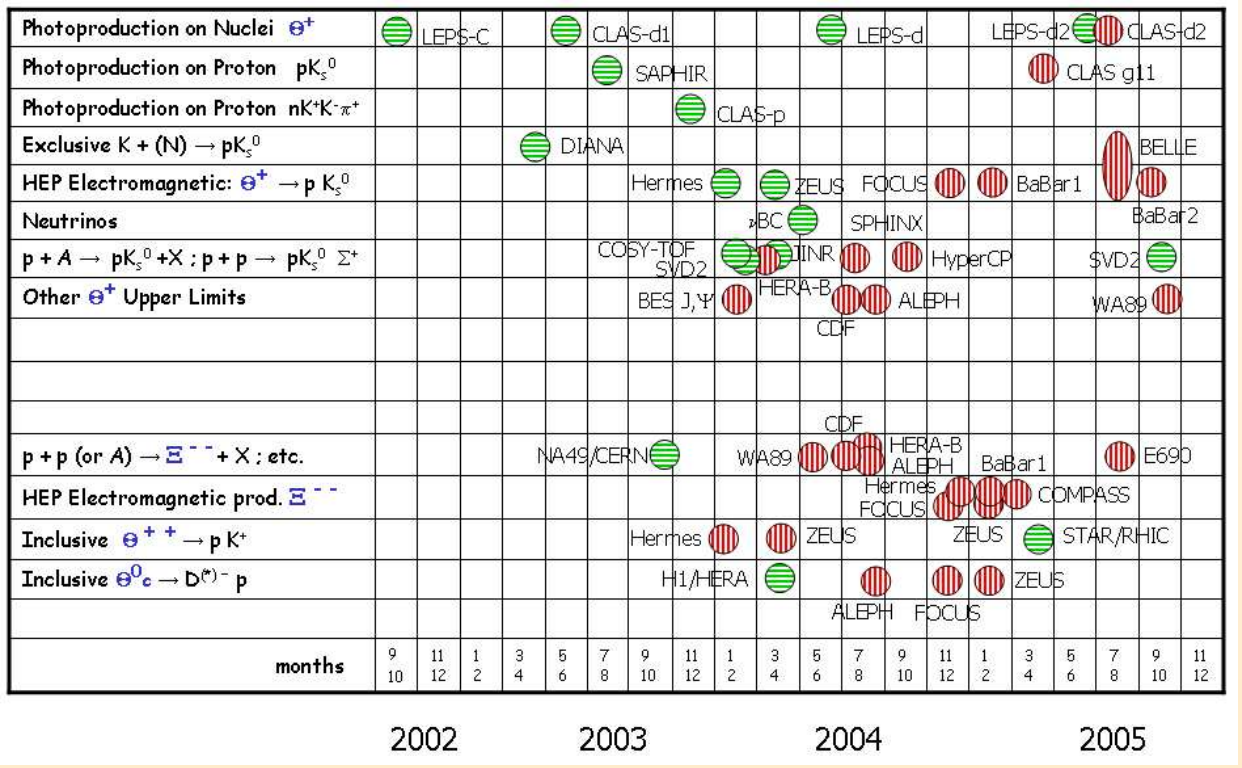
In 2003, the CLAS collaboration at Jefferson Lab announced the observation of an exotic $S = +1$ baryon (the Θ^+ pentaquark) in exclusive photoproduction from the deuteron ($\gamma d \to K^+K^-pn$) [S. Stepanyan *et al.*, Phys. Rev. Lett. **91**, 252001 (2003)]. The significance of the peak is $(5.2 \pm 0.6)\sigma$:



The Pentaquark Sightings (2)

History of the pentaquark

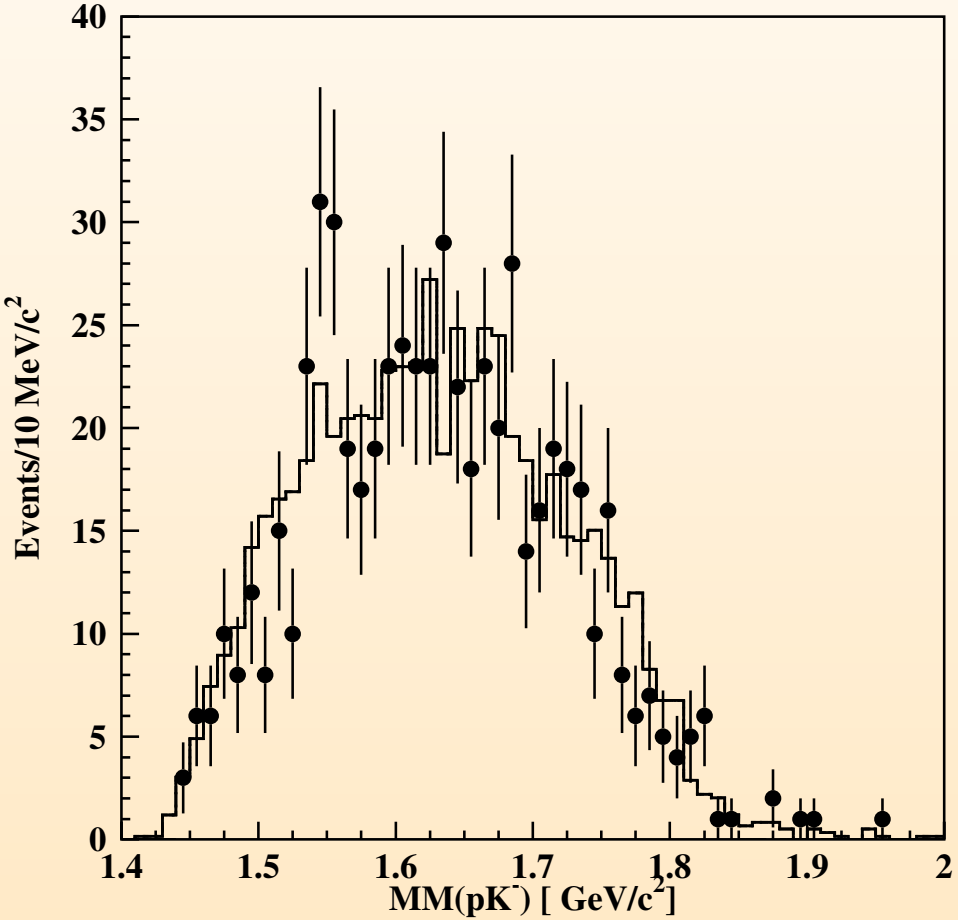
In 1997 some theorists had proposed the existence of a low-mass anti-decuplet of pentaquark baryons with spin $1/2$ and even parity, and predicted the Θ^+, a $uudd\bar{s}$ quark combination with a mass of about 1530 MeV and a width of 15 MeV or less. The first "observation" was made at LEPS in Japan in 2003, from an analysis of $\gamma n \to nK^+K^-$, and quoted a significance of 4.6σ. Many other reports followed:



[Reinhard A. Schumacher, "The Rise and Fall of Pentaquarks in Experiments," arXiv:nucl-ex/0512042v1, 27 Dec 2005.]

The Pentaquark Sightings (3)

In 2006, the CLAS collaboration withdrew its 2003 discovery claim after a higher statistics study failed to reproduce the effect [B. McKinnon *et al.*, Phys. Rev. Lett. **96**, 212001 (2006)]. In addition, an improved estimate of the background led to a downward revision of the significance of the original observation, from $(5.2 \pm 0.6)\sigma$ to 3.1σ:



The Pentaquark Sightings (4)

It is clear that the latest trend among experiments searching for pentaquarks is a string of negative results. In its 2006 Review of Particle Properties, the Particle Data Group points out that among the confirming experiments, there was a "large variation in the locations of the observed peaks (~ 30 MeV) for what had to be a very narrow resonance; thus, the various experiments were not truly confirming one another." Furthermore, the background uncertainties may not have been adequately taken into account in the significance calculations.

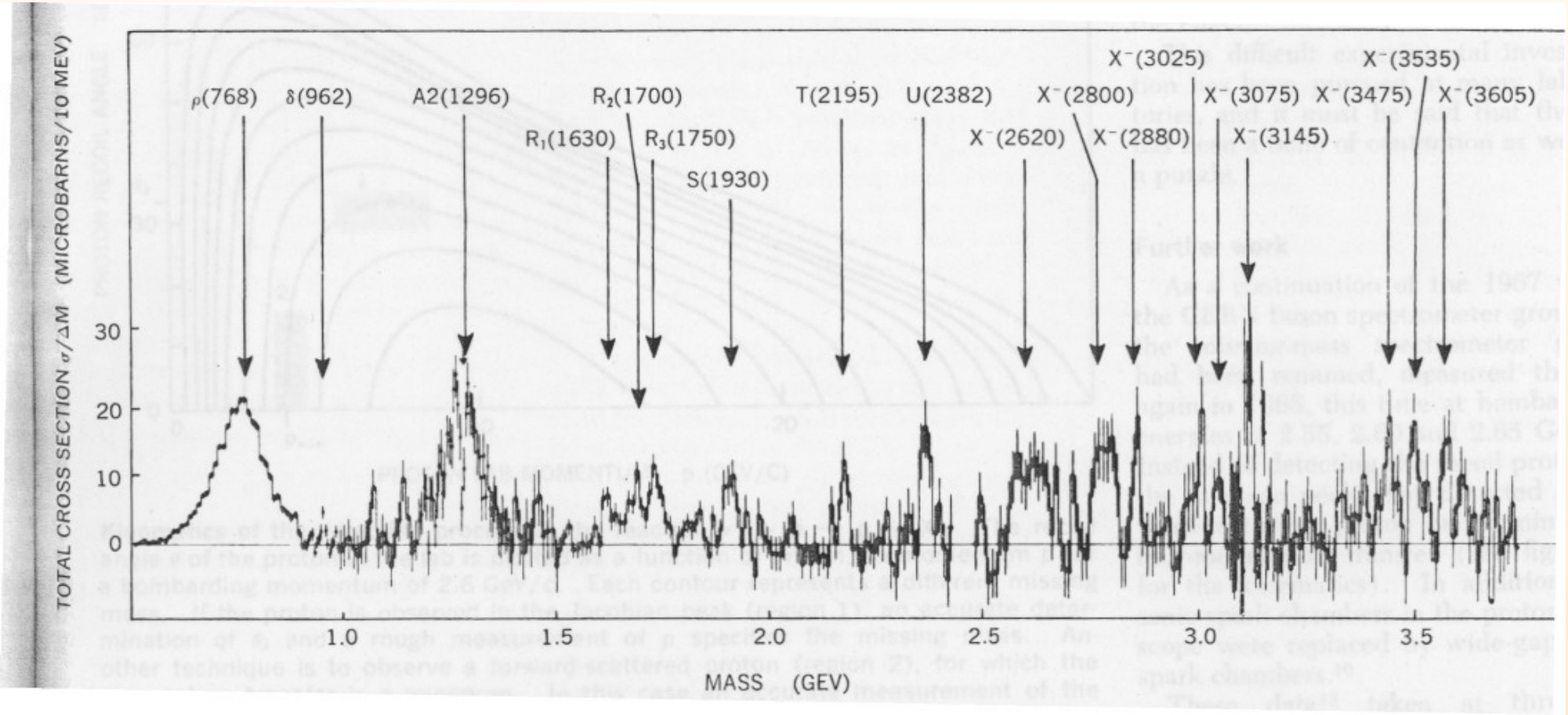
Lessons learned

1. The lesson of the split A_2 affair has not been learned, or has been forgotten.
2. The bandwagon effect is alive and well.
3. The PDG insists that "the burden of proof for the confirmation of an important new result should be about as high as for the original claim of discovery. Only then can one hope to separate the influence of the original discovery from the supposedly independent results of the confirming papers and convince oneself that the confirmation adds significantly to the confidence in the discovery."

THE CHOICE OF 5σ

The Choice of 5σ: Effect of the Trials Factor

Consider this missing-mass spectrum of non-strange boson resonances discovered in the reaction $\pi^- + p \to p + X^-$ between 1965 and 1970 at CERN:



So many resonances on this plot! How can we be sure that each one is real?

The Choice of 5σ: Rosenfeld's 1968 Calculation

At the April 1968 Conference on Meson Spectroscopy in Philadelphia, Arthur Rosenfeld claims that one should expect *several* 4σ fluctuations per year. He argues that the number of "potential resonances" equals the number of histograms looked at times the number of possible deceptive fluctuations per histogram:

1. Number of histograms looked at

   In 1967, approximately two million events with four outgoing prongs were measured over the whole world. Depending on the number of neutrals assumed in each event, there are from 10 to 25 mass combinations that can be formed; let's say 17 on average. This yields 35 million masses. Furthermore, a typical mass histogram has about 2,500 entries, so that

   the total number of histograms looked at is $\sim 15,000$.

2. Number of deceptive fluctuations per histogram

   This is much more difficult to estimate, because most physicists do exercise some judgment and restraint. . . As a conservative estimate, Rosenfeld takes

   ~ 10 deceptive fluctuations per histogram.

Hence the total number of potential resonances is 150,000. Given that a 4σ upwards fluctuation happens once every 32,000 potential bumps, one could expect *4 or 5 claims per year at the 4σ level, and hundreds of claims at the 3σ level.*

The Choice of 5σ: the Posterior Error Rate (1)

[See Branco Sorić, "Statistical 'Discoveries' and Effect-Size Estimation," J. Amer. Statist. Assoc. **84**, 608 (1989).]

Define:

- $N =$ the number of independent tests or measurements made;
- $n_t =$ the number of true null hypotheses;
- $n_c =$ the number of discovery claims;
- $\alpha =$ the rejection threshold.

Then for large N, the proportion of false discoveries in the set of discovery claims will be:

$$Q = \frac{\alpha\, n_t}{n_c}.$$

Do not confuse Q with α! If n_t is large, Q can be much larger than α.

The Choice of 5σ: the Posterior Error Rate (2)

Unfortunately we do not know the true value of n_t. However, it is possible to calculate an upper bound on Q. Indeed, note that the number of experiments without discovery claim is at least as large as the number of true null hypotheses that were accepted (the difference being the number of false null hypotheses that were accepted). Thus:

$$N - n_c \;\geq\; (1-\alpha)\, n_t, \qquad \text{so that} \qquad n_t \;\leq\; \frac{N-n_c}{1-\alpha},$$

and therefore:

$$Q \;=\; \frac{\alpha}{n_c}\, n_t \;\leq\; \frac{\alpha}{n_c}\, \frac{N-n_c}{1-\alpha} \;=\; \frac{N/n_c - 1}{1/\alpha - 1} \;\equiv\; Q_{\max}.$$

If the number of false null hypotheses that were accepted is small, i.e. if the average power is high, then $Q_{\max} \approx Q$.

As an example, suppose that in the course of its lifetime, one of the LHC experiments makes 1,000 searches, out of which it claims only 10 discoveries. Then, for a 3σ threshold $Q_{\max} \sim 13\%$, whereas for a 5σ threshold $Q_{\max} \sim 2.8 \times 10^{-5}$.

The Choice of 5σ: Summary

Rosenfeld's calculation is probably the first one to mention 5σ as a standard in HEP. His recommendation is more nuanced:

- For the experimental group:

  Go ahead and publish your tantalizing bump: you worked and paid for it. However, realize that anything less than 5σ calls for a repeat of the experiment. With twice the sample size, the number of standard deviations should increase by $\sqrt{2}$ to confirm the original effect.

- For the theorist:

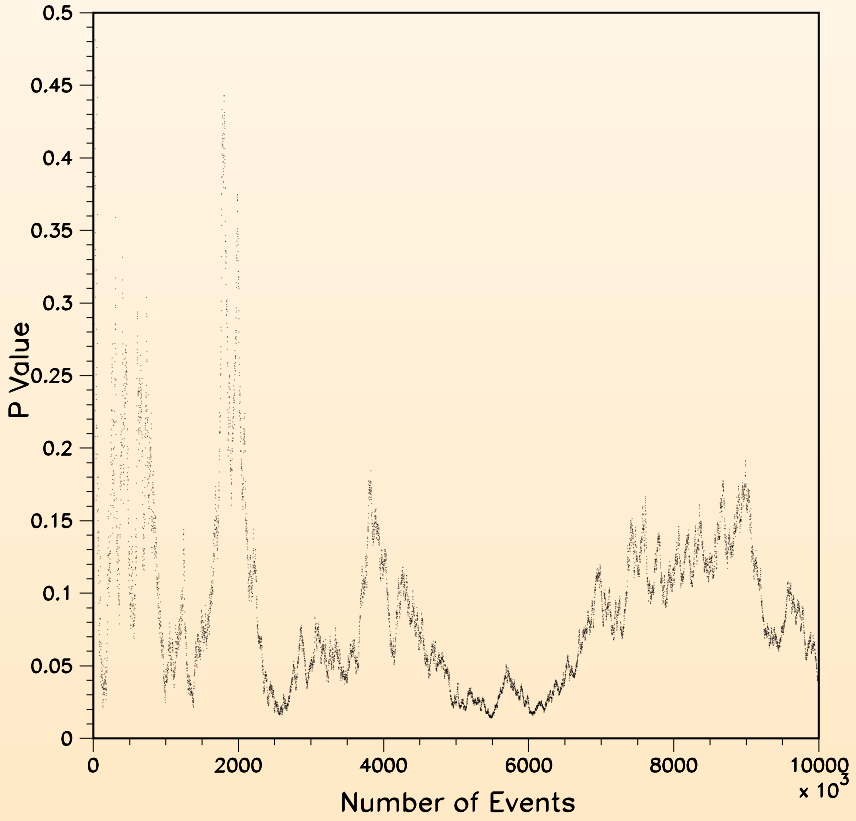
  Wait for nearly 5σ effects.

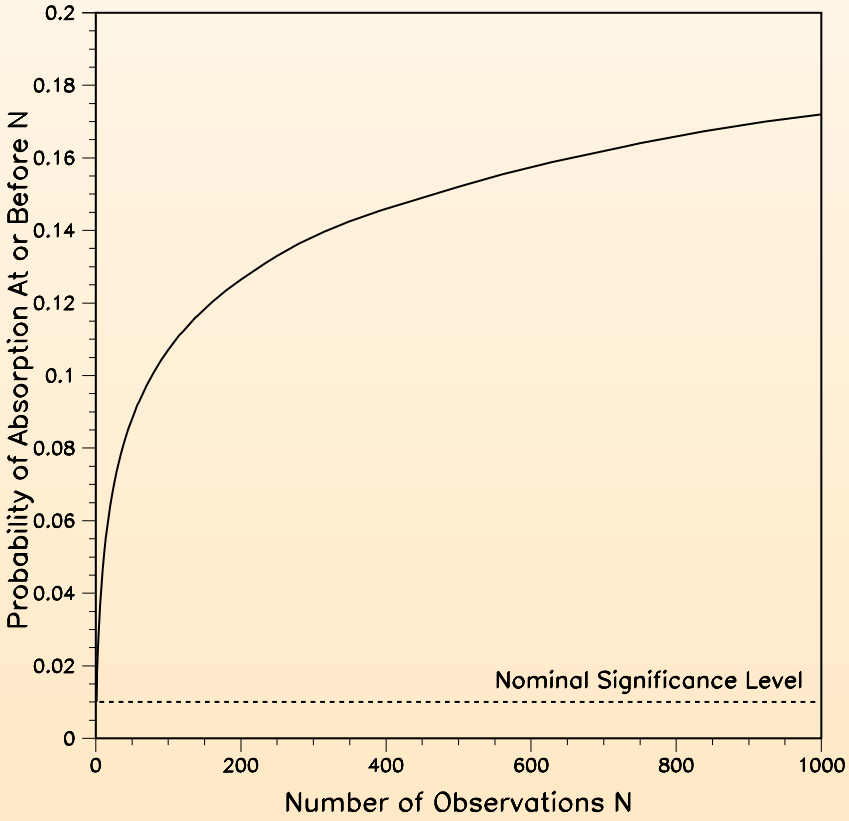
Rosenfeld also notes that the U.S. measurement rate seems to double every two years, "so things will get worse." This was back in 1968, and we haven't adjusted the 5σ threshold since. Should we? Perhaps not; there are things one can do to keep the posterior error rate under control: blinding the data analysis, specifying the alternative hypothesis before looking at the data, and keeping track of the probability of being wrong, whether via Bayesian or frequentist methods.

P Value Pathologies

Random Walk Effects

Test statistics perform a random walk as the sample size increases. Therefore, the true Type I error rate depends on how one decides to stop the experiment...
Furthermore, if one decides not to stop until the null hypothesis is rejected, one will eventually stop, with probability one, even if the null hypothesis is true (a consequence of the Law of the Iterated Logarithm).





Jeffreys' Paradox (1)

Suppose you wish to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$, where μ is the mean of a Gaussian distribution with known width σ.

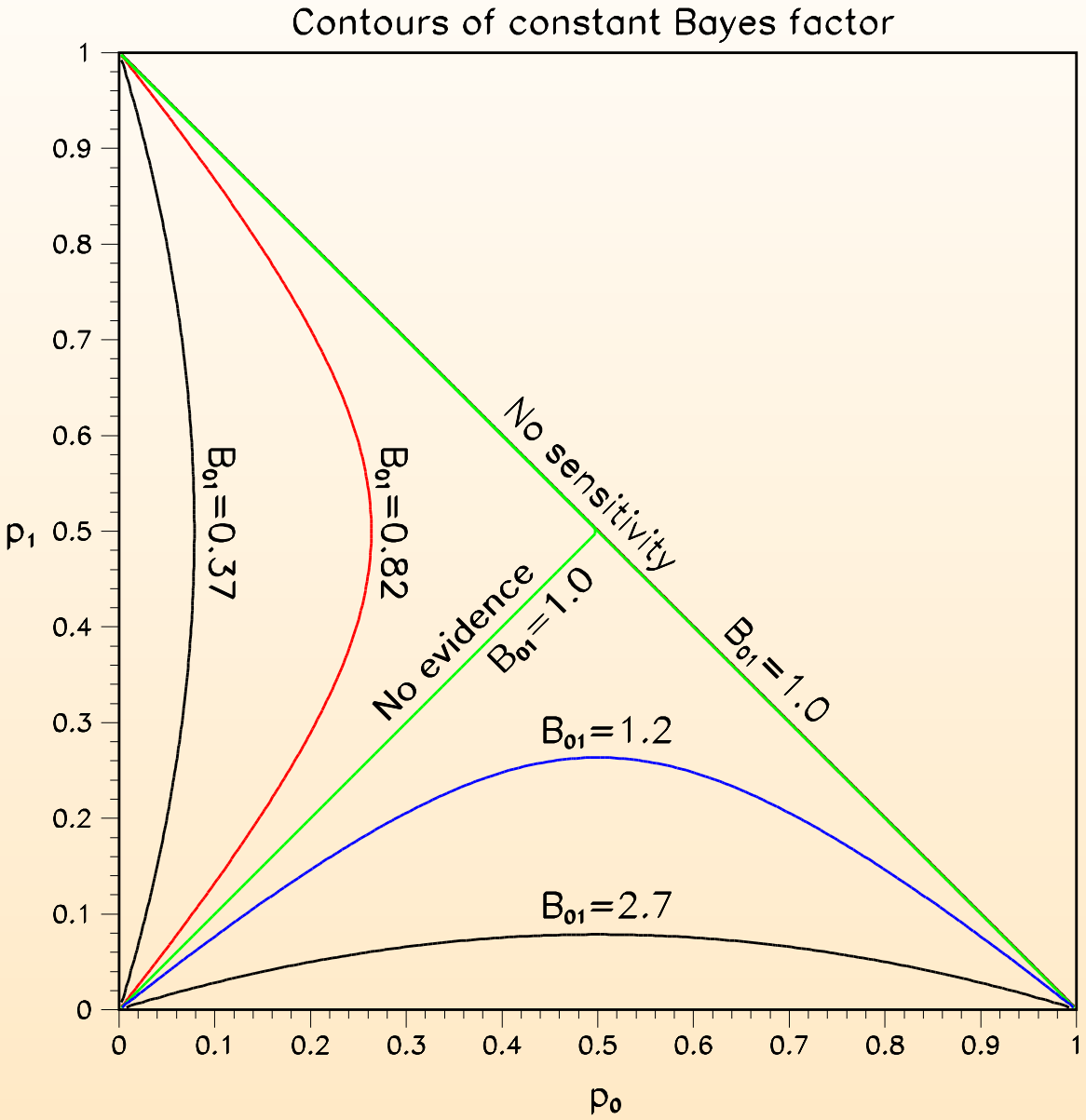
Suppose also that the sample size available for the test keeps increasing at regular time intervals, and that the p value against H_0, p_0, stays constant at some value below the rejection threshold α.

Thus, from a frequentist point of view, we can reject H_0 and this conclusion does not change as the sample size increases.

Interestingly however, it can be shown that, under the given assumptions, the Bayes factor B_{01} in favor of H_0 will eventually exceed 1 (note that B_{01} is a simple likelihood ratio in this problem!). Thus, a Bayesian will eventually accept H_0 even though a frequentist will keep rejecting it!

The reason for this paradox is that, as the sample size increases, for constant α the probability β of incorrectly accepting H_0 decreases. Eventually it will become more advantageous to accept H_0 than to reject it. The Bayesian procedure takes this automatically into account. The frequentist one does not, unless α is made to decrease at some appropriate rate with sample size.

Jeffreys' Paradox (2)



P Values versus Bayesian Measures of Evidence

A popular misunderstanding of p values is that they somehow represent the probability of H_0. What can we actually say about the relationship between p and $\mathbb{P}\mathrm{r}(H_0 \,|\, \mathbf{x}_{\mathrm{obs}})$? Unfortunately the answer depends on the choice of prior.

Idea: Compare p to the smallest $\mathbb{P}\mathrm{r}(H_0 \,|\, \mathbf{x}_{\mathrm{obs}})$ obtained by varying the prior within some large, plausible class of distributions.

It is useful to study separately two cases:

1. $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$;

2. $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

See G. Casella and R. Berger, JASA **82**, 106 (1987); J. Berger and T. Sellke, JASA **82**, 112 (1987).

P versus Bayes: the One-Sided Case

Casella and Berger consider the test $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$, based on observing $X = x$, where X has a location density $f(x - \theta)$. f is assumed to be symmetric about zero and to have monotone likelihood ratio. The following classes of priors are used:

- $\Gamma_S = \{\text{all distributions symmetric about } 0\}$;
- $\Gamma_{US} = \{\text{all unimodal distributions symmetric about } 0\}$;
- $\Gamma^\sigma(g) = \{\pi_\sigma : \pi_\sigma(\theta) = g(\theta/\sigma)/\sigma, \sigma > 0,\ g(\theta) \text{ bounded, symm., unimodal}\}$.

The following theorems are then proved (all assume $x > 0$):

$$\inf_{\pi \in \Gamma_{US}} \mathbb{P}\mathrm{r}(H_0 \,|\, x_{\text{obs}}) = p(x) \tag{1}$$

$$\inf_{\pi_\sigma \in \Gamma^\sigma(g)} \mathbb{P}\mathrm{r}(H_0 \,|\, x_{\text{obs}}) = p(x) \tag{2}$$

$$\inf_{\pi \in \Gamma_S} \mathbb{P}\mathrm{r}(H_0 \,|\, x_{\text{obs}}) \leq p(x) \tag{3}$$

P versus Bayes: the Two-Sided Case

Berger and Sellke consider the test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, based on observing $\mathbf{X} = (X_1, \ldots, X_n)$, where the X_i are iid $\mathcal{N}(\theta, \sigma^2)$, σ^2 known; the usual test statistic is $T(\mathbf{X}) = \sqrt{n}|\bar{X} - \theta_0|/\sigma$.

The prior is of the form $\pi(\theta) = \pi_0$ if $\theta = \theta_0$, and $\pi(\theta) = (1 - \pi_0)\, g(\theta)$ if $\theta \neq \theta_0$, where $g(\theta)$ belongs to one of the classes:

- $G_A = \{\text{all distributions}\}$;
- $G_S = \{\text{all distributions symmetric about } \theta_0\}$;
- $G_{US} = \{\text{all unimodal distributions symmetric about } \theta_0\}$.

The following theorems are then proved:

$$\text{For } t_{\text{obs}} > 1.68 \text{ and } \pi_0 = \frac{1}{2}: \quad \inf_{g \in G_A} \frac{\mathbb{P}\text{r}(H_0 \,|\, \mathbf{x}_{\text{obs}})}{p\, t_{\text{obs}}} > \sqrt{\frac{\pi}{2}} \cong 1.253 \tag{4}$$

$$\text{For } t_{\text{obs}} > 2.28 \text{ and } \pi_0 = \frac{1}{2}: \quad \inf_{g \in G_S} \frac{\mathbb{P}\text{r}(H_0 \,|\, \mathbf{x}_{\text{obs}})}{p\, t_{\text{obs}}} > \sqrt{2\,\pi} \cong 2.507 \tag{5}$$

$$\text{For } t_{\text{obs}} > 0 \text{ and } \pi_0 = \frac{1}{2}: \quad \inf_{g \in G_{US}} \frac{\mathbb{P}\text{r}(H_0 \,|\, \mathbf{x}_{\text{obs}})}{p\, t_{\text{obs}}^2} > 1 \tag{6}$$

P Values as Measures of Support (1)

If we wish to use p values as measures of support, there are some properties we will need them to have. Think of the simple problem of testing the mean of a normal density by using the average of several measurements. Then:

1. The farther the hypothesis is from the observed data, the smaller the p value should be.
2. If H implies H', then anything that supports H should *a fortiori* support H'. This property is known as coherence.

It is easy to see that p values satisfy the first of these requirements. However, they do not always satisfy the second. For example, consider the following two test situations:

$$H_1 : \mu = \mu_0 \quad \text{versus} \quad A_1 : \mu \neq \mu_0$$
$$H_2 : \mu \leq \mu_0 \quad \text{versus} \quad A_2 : \mu > \mu_0$$

Note that H_1 implies H_2; hence, coherence of p values as measures of support requires that $p_{H_2} \geq p_{H_1}$.
Suppose however that we observe $\bar{x} > \mu_0$, but with relatively large p values under both H_1 and H_2. Then $p_{H_2} = 0.5\, p_{H_1} < p_{H_1}(\bar{x})$, which is *in*coherent.

P Values as Measures of Support (2)

One would like systematic uncertainties to decrease one's confidence in the result of a test, whether it is to reject the null hypothesis H_0 or to accept it. However:

1. to decrease confidence in a rejection of H_0, p values must *increase*, whereas

2. to decrease confidence in an acceptance of H_0, p values must *decrease*.

It is generally not possible to satisfy both requirements simultaneously. In fact, most methods for incorporating systematic uncertainties in p values tend to increase them.

This is a major obstacle to using p values as measures of support.

Pathological Science?

Langmuir's List of Symptoms of Pathological Science

Irving Langmuir (1932 Nobel in chemistry) gave a talk on pathological science at the General Electric research labs in 1953. This talk is now available on the web: `http://www.cs.princeton.edu/~ken/Langmuir/langmuir.htm`. In it, Langmuir discusses several famous cases of false discovery claims (Davis-Barnes effect, N-rays, Mitogenetic rays, extrasensory perception, flying saucers, etc.) and then lists what he calls "characteristic symptoms of pathological science":

1. The maximum effect that is observed is produced by a causative agent of barely detectable intensity, and the magnitude of the effect is substantially independent of the intensity of the cause.
2. The effect is of a magnitude that remains close to the limit of detectability; or, many measurements are necessary because of the very low statistical significance of the results.
3. Claims of great accuracy.
4. Fantastic theories contrary to experience.
5. Criticisms are met by ad hoc excuses thought up on the spur of the moment.
6. Ratio of supporters to critics rises up to somewhere near 50% and then falls gradually to oblivion.

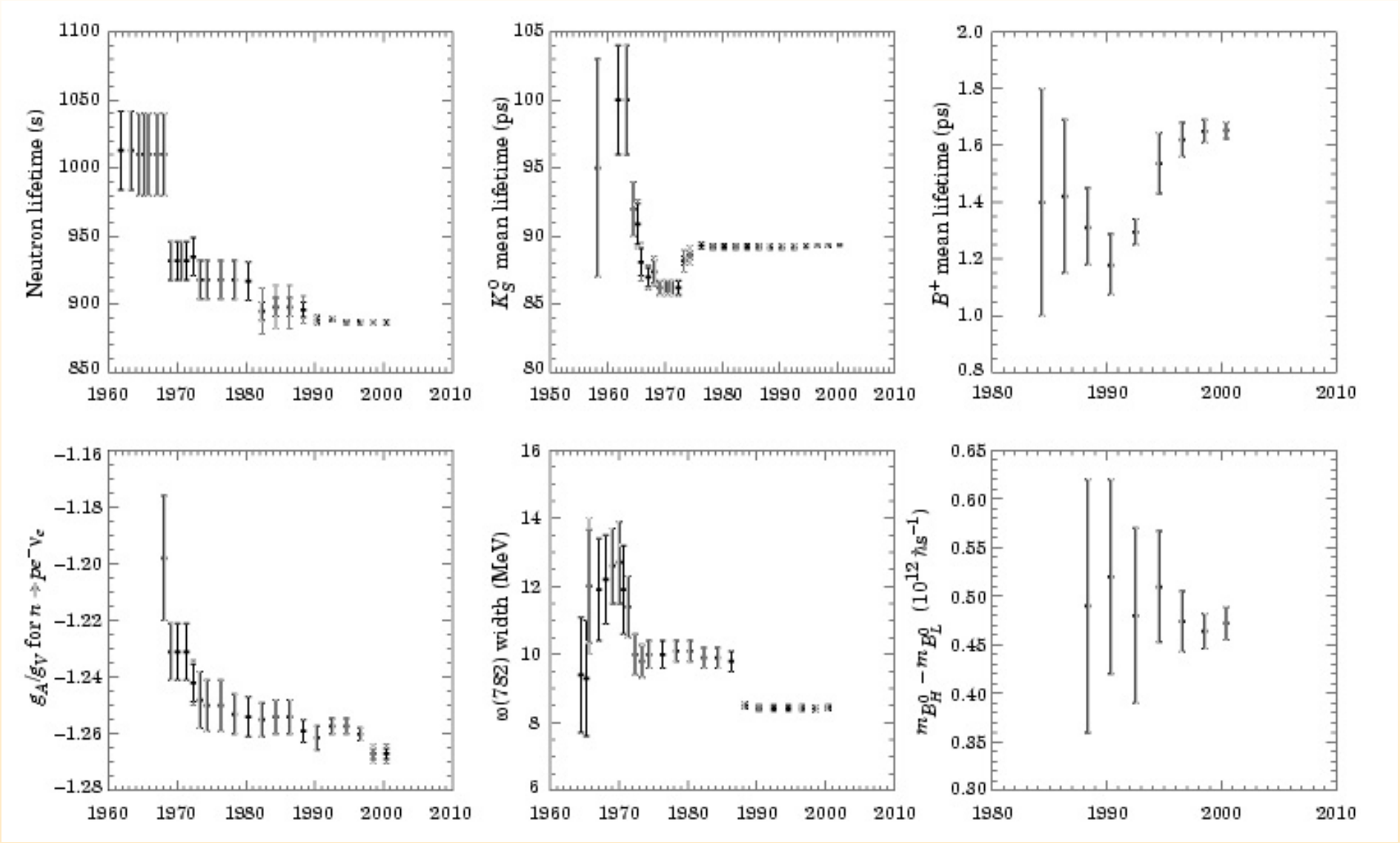
Langmuir's List of Symptoms of Pathological Science

Not everybody agrees with this characterization however: several of the symptoms can be observed in discoveries that have stood the test of time. . .

Blind Analyses

Motivation for Blind Analyses

The Particle Data Group has produced some interesting plots of the "evolution" of various measurements with time:



Definition of Blind Analyses

An analysis is blind if it is conducted in such a way that the actual data leading to the final result are somehow hidden from the analyst.

This allows one to avoid experimenter bias (remember the split A_2 analysis, where unnecessary cuts, based on "running conditions", caused runs to be discarded in which no split showed up).

Blind analyses are possible because the value of a measurement does not contain any information about its correctness; they have been used in rare decay searches at BNL, in E791, KTeV, BABAR, BELLE, CDF, etc.

See Aaron Roodman's talk at the 2003 PhyStat conference at SLAC.

Blind Analysis Methods (1)

Blinding techniques depend on the type of analysis done:

1. Hidden signal region

   This technique works well in searches for rare decays, for example $K_L^0 \to \mu^\pm e^\mp$. Events inside a box surrounding the signal region are excluded from the analysis and plots. The analysis is then optimized using data sidebands to characterize the background and Monte Carlo simulations to characterize the signal. A good procedure is to use half the data and Monte Carlo to *optimize* the analysis, and the other half to determine the background normalization and signal efficiency.

   Before unblinding, the analysis is fully documented and the analyst provides the expected number of background events in the signal box, the signal efficiency, the sensitivity, and the expected statistical uncertainty

   Only after unblinding are plots made of the signal region in the data.

Blind Analysis Methods (2)

2. Hidden offset

   When making precision measurements, it is often the case that previous measurements are already available, so that blinding should help avoid biasing the new measurement towards the old ones. Often the method involves a maximum likelihood fit, and the result of the fit can then be hidden from the analyst by having the code add a fixed, unknown random number to the result:

$$x_{\mathrm{blind}} = x + \mathcal{R}.$$

   Then, x_{blind} is returned, along with the true error and likelihood value instead of x. It is sometimes useful to run two analyses in parallel, in which case a different offset can be used for each analysis; this will prevent them from biasing each other. At some point the offsets can be made the same in order to allow a (blind!) comparison of the two analyses.

Blind Analysis Methods (3)

In some cases one may wish to hide the direction in which a result changes when the analysis is modified. This happened with the KTeV measurement of ϵ'/ϵ:

$$\left[\frac{\epsilon'}{\epsilon}\right]_{\rm blind} = c \times \frac{\epsilon'}{\epsilon} + \mathcal{R},$$

where c is a hidden random number equal to $+1$ or -1.

Other techniques exist. For example, one can mix a hidden amount of simulation data into the real data while optimizing the analysis.

Searches for new particles are difficult to conduct in an unbiased way because the location of the signal is not known a priori (the hidden signal region method does not work here). One needs separate control samples.

Summary

I hope I have convinced you that the scientific method works, that no matter how often we trip or get diverted into dead-end streets, we always end up back on the straight and narrow path to truth.

Nevertheless, please keep in mind that:

- You can fool yourself (experimenter bias).
- You can be fooled by others (bandwagon effect).
- You can be fooled by mathematics (p values).
- You can fool yourself (again).

There are some things one can do to minimize the chance of a false discovery: apply sound statistical methods, use blind analyses. . . With some care, it will not be necessary to raise the 5σ discovery threshold in the foreseeable future.