Discovery and Significance

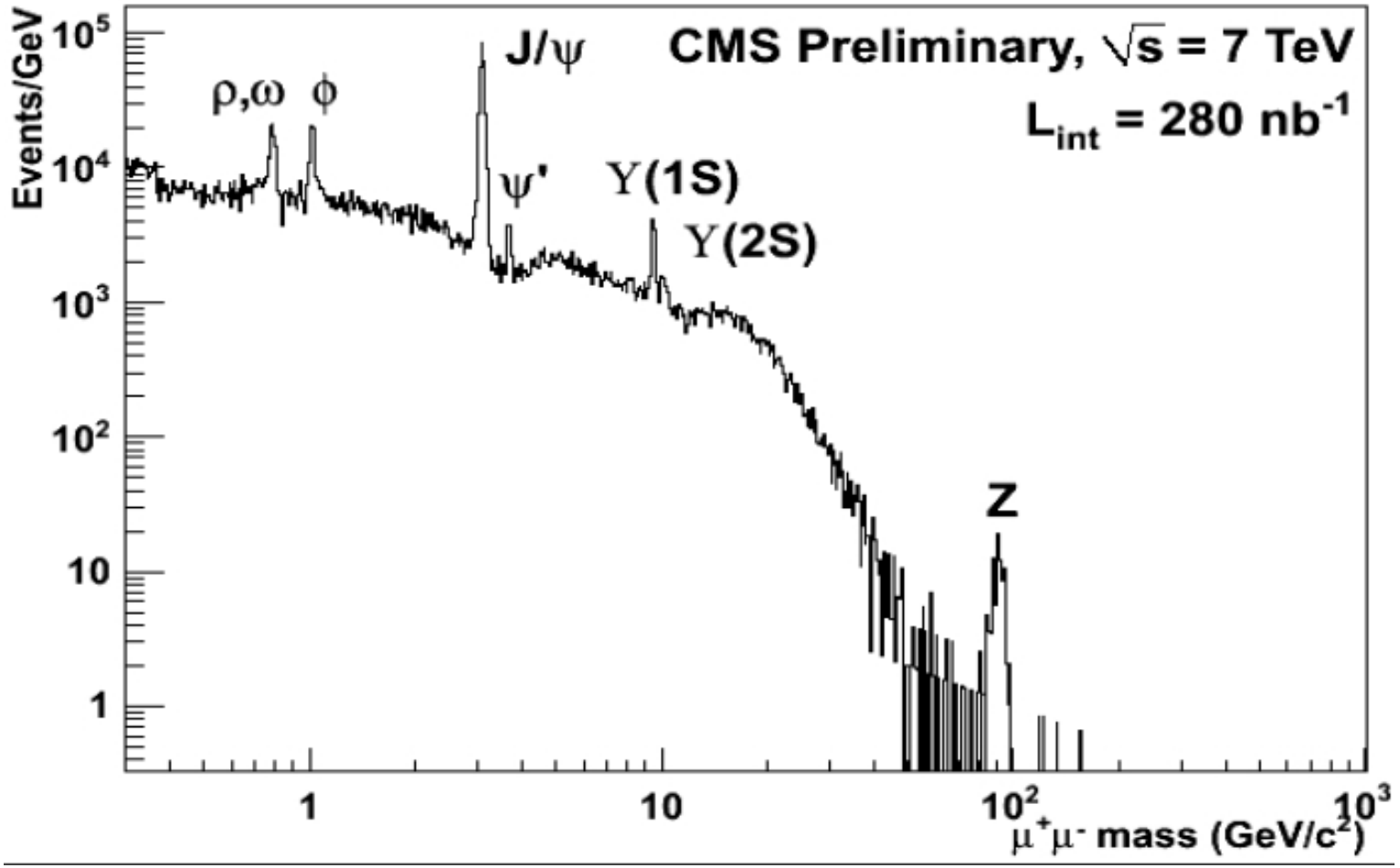
M. Witherell

5/10/12

Discovering particles

- Much of what we know about physics at the most fundamental scale comes from discovering particles.
- We discovered these particles by finding bumps in invariant mass plots.
  - $\rho^0 \to \pi^+\pi^-$
  - $\psi \to \mu^+\mu^-$
  - $Z \to \mu^+\mu^-$
- When is a particle discovered?
- What can go wrong?

The history of particle physics, as told by CMS



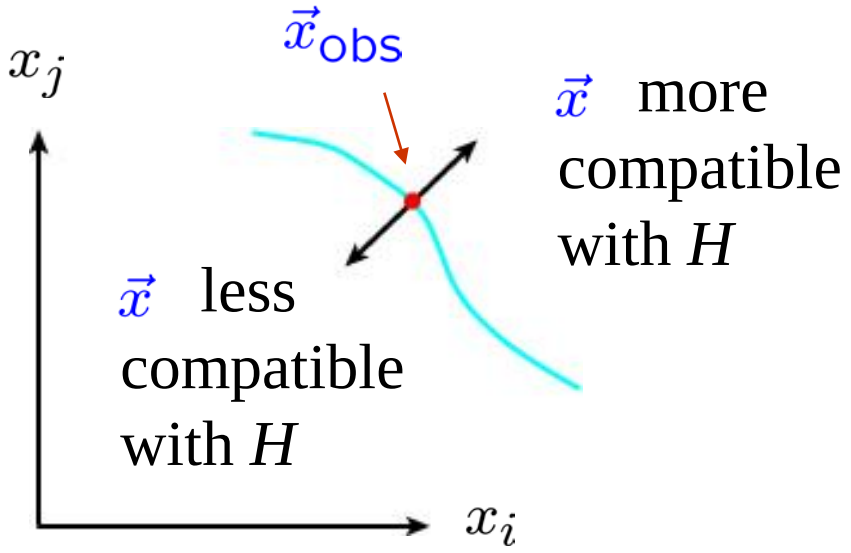
Testing significance / goodness-of-fit

Suppose hypothesis *H* predicts pdf $f(\vec{x}|H)$ for a set of observations $\vec{x} = (x_1, \ldots, x_n)$.

We observe a single point in this space: $\vec{x}_{\text{obs}}$

What can we say about the validity of *H* in light of the data?

Decide what part of the data space represents less compatibility with *H* than does the point $\vec{x}_{\text{obs}}$. (Not unique!)



p-values

Express 'goodness-of-fit' by giving the *p*-value for *H*:

p = probability, under assumption of *H*, to observe data with equal or lesser compatibility with *H* relative to the data we got.

This is not the probability that *H* is true!

In frequentist statistics we don't talk about $P(H)$ (unless *H* represents a repeatable observation). In Bayesian statistics we do; use Bayes' theorem to obtain

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

where $\pi(H)$ is the prior probability for *H*.

For now stick with the frequentist approach;
result is *p*-value, regrettably easy to misinterpret as $P(H)$.

p-value example: testing whether a coin is 'fair'

Probability to observe n heads in N coin tosses is binomial:

$$P(n; p, N) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

Hypothesis H: the coin is fair ($p = 0.5$).

Suppose we toss the coin $N = 20$ times and get $n = 17$ heads.

Region of data space with equal or lesser compatibility with H relative to $n = 17$ is: $n = 17, 18, 19, 20, 0, 1, 2, 3$. Adding up the probabilities for these values gives:

$$P(n = 0, 1, 2, 3, 17, 18, 19, \text{ or } 20) = 0.0026 \; .$$

i.e. $p = 0.0026$ is the probability of obtaining such a bizarre result (or more so) 'by chance', under the assumption of H.

The significance of an observed signal

Suppose we observe n events; these can consist of:

n_b events from known processes (background)
n_s events from a new process (signal)

If n_s, n_b are Poisson r.v.s with means s, b, then $n = n_s + n_b$ is also Poisson, mean = $s + b$:

$$P(n; s, b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

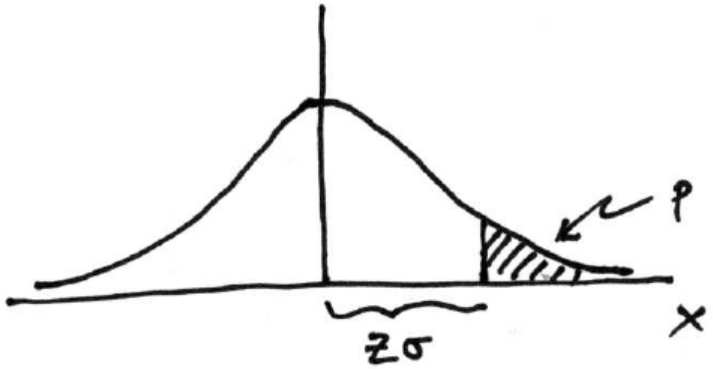
Suppose $b = 0.5$, and we observe $n_{obs} = 5$. Should we claim evidence for a new discovery?

Give p-value for hypothesis $s = 0$:

$$p\text{-value} = P(n \geq 5; b = 0.5, s = 0)$$
$$= 1.7 \times 10^{-4} \neq P(s = 0)!$$

Significance from *p*-value

Often define significance Z as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same *p*-value.

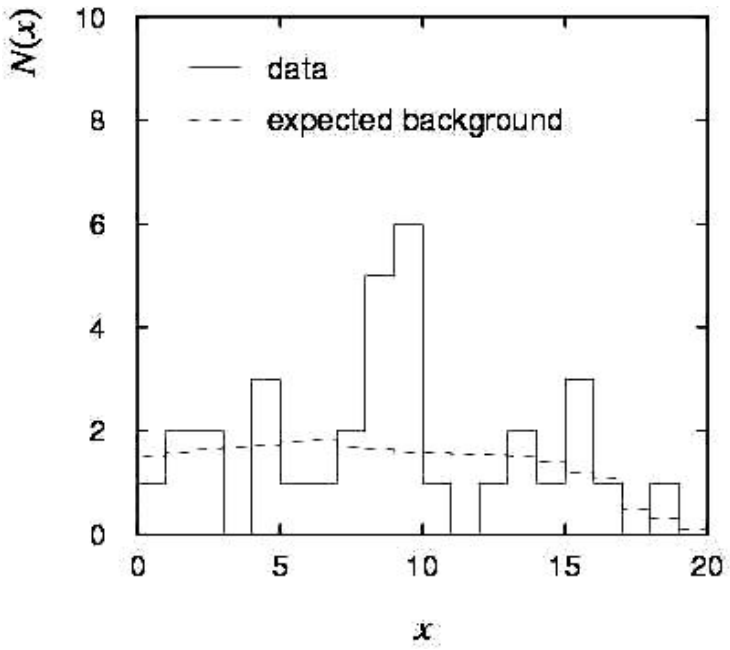


$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z)$$ `1 - TMath::Freq`

$$Z = \Phi^{-1}(1-p)$$ `TMath::NormQuantile`

The significance of a peak

Suppose we measure a value x for each event and find:



Each bin (observed) is a Poisson r.v., means are given by dashed lines.

In the two bins with the peak, 11 entries found with $b = 3.2$.
The p-value for the $s = 0$ hypothesis is:

$$P(n \geq 11; b = 3.2, s = 0) = 5.0 \times 10^{-4}$$

The significance of a peak (2)

But... did we know where to look for the peak?

→ give $P(n \geq 11)$ in any 2 adjacent bins

Is the observed width consistent with the expected x resolution?

→ take x window several times the expected resolution

How many bins × distributions have we looked at?

→ look at a thousand of them, you'll find a 10^{-3} effect

Did we adjust the cuts to 'enhance' the peak?

→ freeze cuts, repeat analysis with new data

How about the bins to the sides of the peak... (too low!)

Should we publish????

When to publish

HEP folklore is to claim discovery when $p = 2.9 \times 10^{-7}$, corresponding to a significance $Z = 5$.

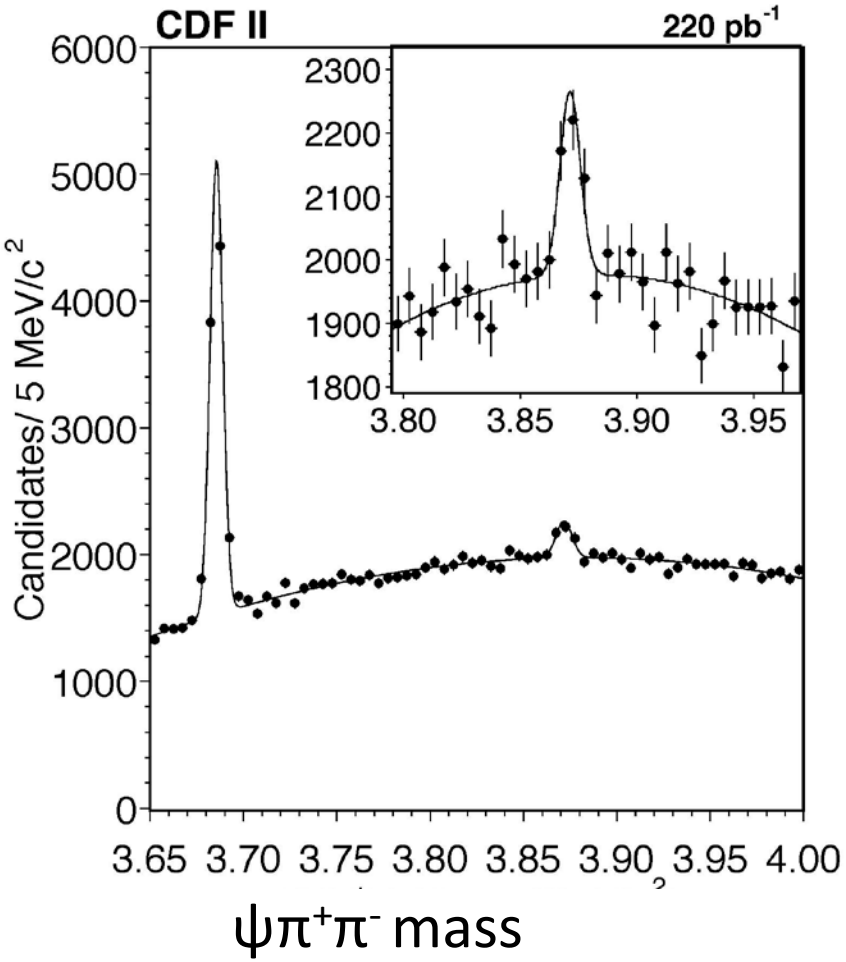
This is very subjective and really should depend on the prior probability of the phenomenon in question, e.g.,

phenomenon	reasonable *p*-value for discovery
D^0D^0 mixing	~0.05
Higgs	~ 10^{-7} (?)
Life on Mars	~10^{-10}
Astrology	~10^{-20}

One should also consider the degree to which the data are compatible with the new phenomenon, not only the level of disagreement with the null hypothesis; *p*-value is only first step!

Basic method of peak fitting. Fit a smooth sideband.

- In 2006 BELLE discovered an unknown charmonium state with a mass of 3.872 GeV, decaying into $\psi\pi^+\pi^-$.
- CDF immediately looked at their data, shown at right.
  - 6200 events in 3 bins.
  - Background of ~ 5600 events.
  - Excess of ~600 ± 80.
- Fit using a quadratic background.
  - 730 ± 90 in peak at 3872.
  - Gaussian σ=4.9±0.7 MeV, consistent with resolution.
  - Discovery confirmed with 8σ.



Discoveries that turned out to be false

- Several discoveries that were 5,6,8 σ turned out to be false discoveries. What happened?
  - Uncertainties in background were dismissed.
  - Uncertainty in shape of signal was mishandled.
  - Many cuts were tried, but few were chosen.
  - Many plots were looked at, and one had a fluctuation.
- 5 σ is a reliable discovery only if all these other possibilities are removed or accounted for.

Background uncertainties

- Statistical errors
  - measure how much the result would fluctuate if the experiment were repeated.
  - do not reflect uncertainties in background or theoretical models.
- Systematic errors
  - do not vary upon repeated experiments, or increased data sample.
  - include uncertainties in theory, calibrations, background shape and other sources of error.
- Systematic uncertainties can be taken into account via "nuisance parameters."
- For example, we may have a measurement of the background level b which has an error σ_b, or an error in an energy calibration. Such errors can be included in the fit.

Upsilon discovery – take 1

In 1976, Leon Lederman's group looked at the mass spectrum of e^+e^- pairs produced in pN collisions with 400 GeV protons at Fermilab.

This was just two years after the revolutionary discovery of the ψ at Brookhaven and SLAC.

They wanted to see if the bottom quark could be out there just beyond charm.

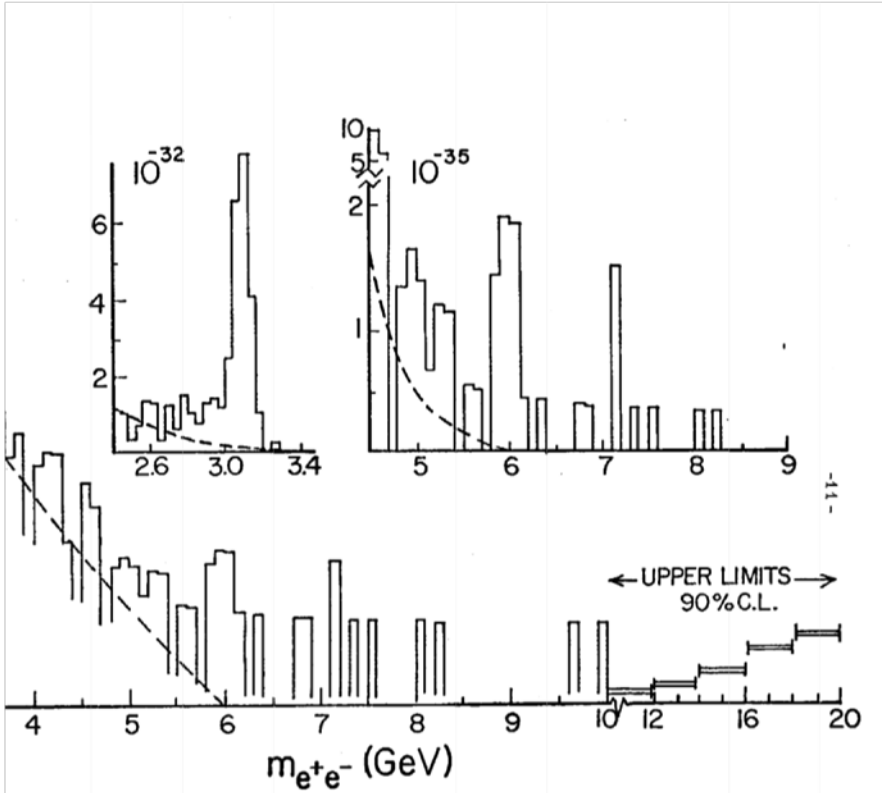
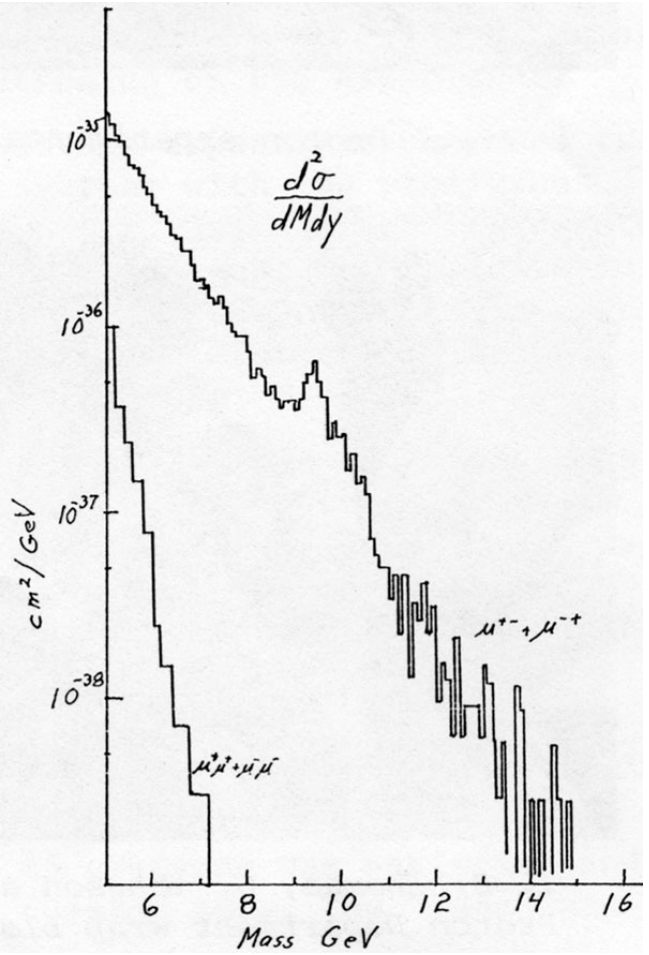


Fig. 2: Electron-Positron Mass Spectrum: dσ/dm per nucleon versus the effective mass. A linear A-dependence is assumed. Note bin width changes.

Upsilon discovery – take 2

- The next phase of the experiment in 1977 used dimuons instead of dielectrons. The sensitivity was 1000x greater.
- The peak at 9.5 GeV is the real upsilon, and represents the discovery of the b quark.
- See anything at 6 GeV?

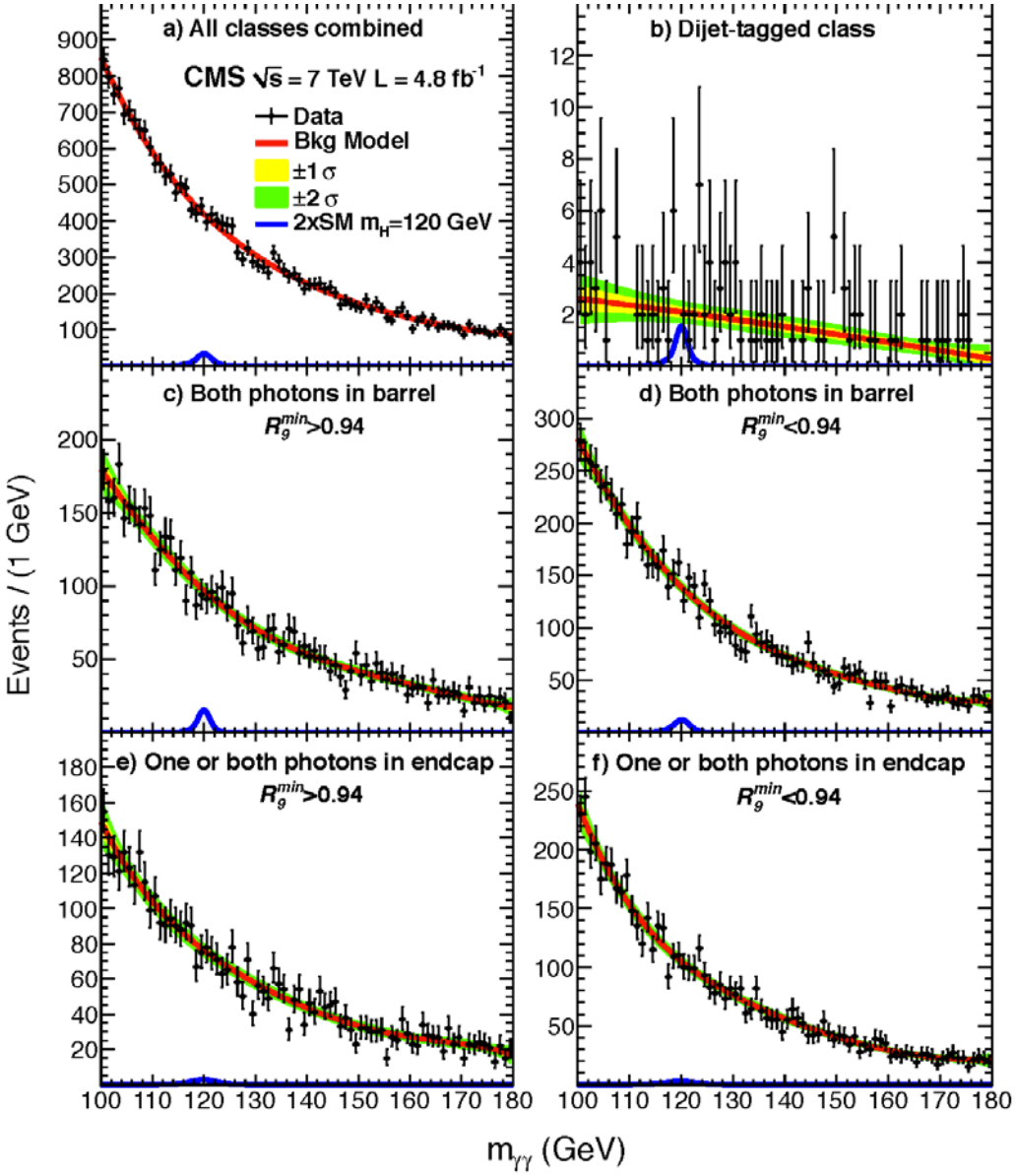


Oops-Leon or Upsilon

- What went wrong?
- Experimenter's hope for a peak led them to assume a background shape that fell off too sharply.
- How should this have been handled?
  - Get best fit with no peak and a plausible background first.
  - Then allow a Gaussian peak on top, and see how much the fit improves.
  - Change binning, or do unbinned maximum likelihood fit.

Higgs -> $\gamma\gamma$ is much harder.

- We need to know the background shape to a couple percent, or it will swamp the statistical error.
- Also, one needs to know the resolution function.
- Finally, since the Higgs mass is not known, one needs to refit this for every value that the Higgs mass could take.

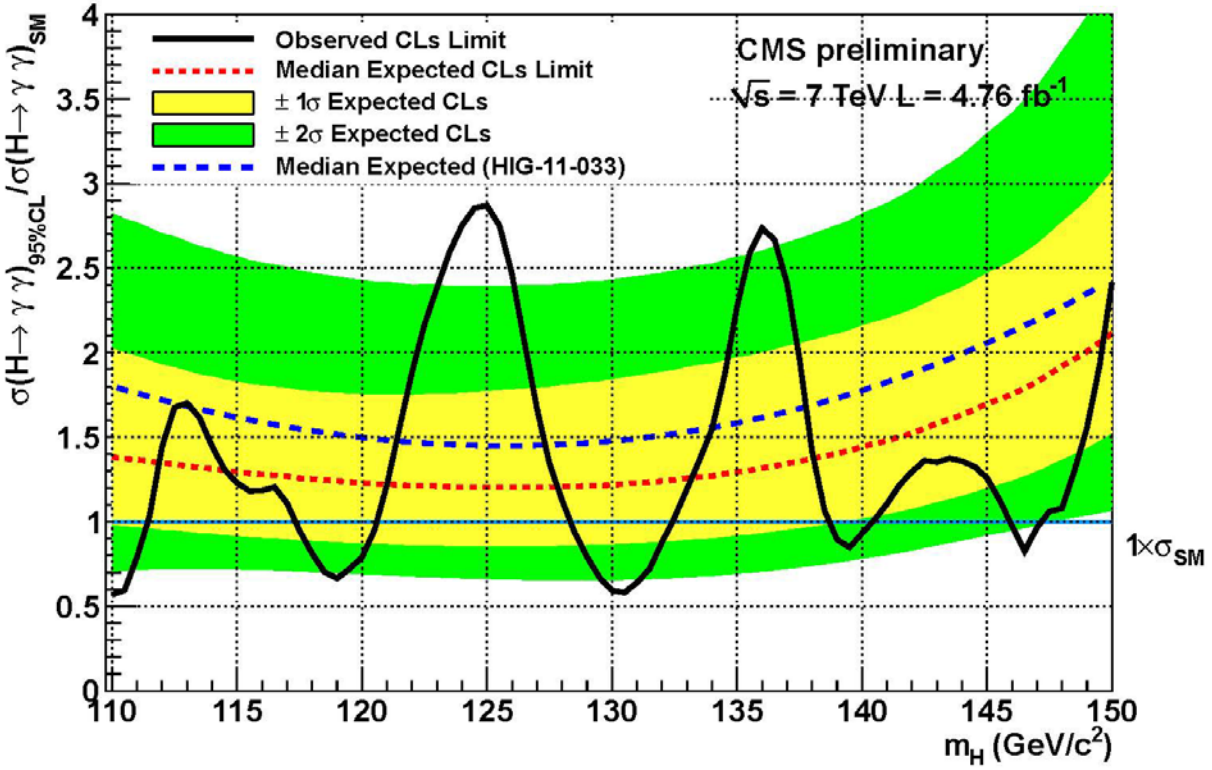


Higgs Searches at the LHC

This is not a mass plot.

Rather, it is a plot of the significance and expected significance as a function of a theory parameter, the Higgs mass.

One needs to keep in mind the mass plots behind it, however.

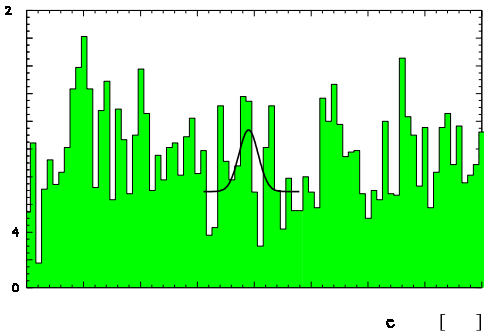


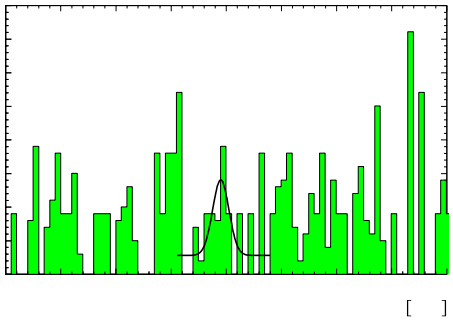
Search for neutrinoless double beta decay

- For 30 years physicists have been looking for neutrinoless double beta decay.

$$^{76}Ge \rightarrow {}^{76}Se + e^- + e^-$$

- Observation of this decay would establish that the electron neutrino is a Majorana particle and measure its mass.
- The signal would be a peak in the energy spectrum at a known value, 2.04 MeV.
- In 2002, Klapdor et al. claimed 2-3 σ evidence, after selecting 3 of 5 detectors.





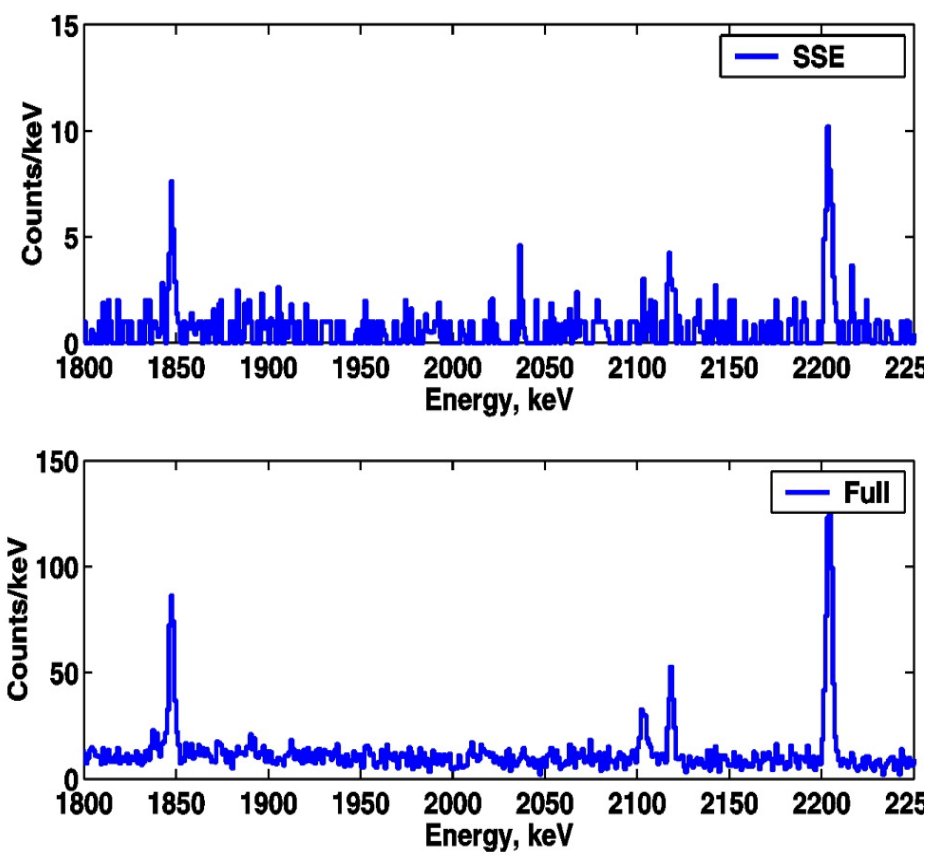
Comment on "Evidence for Neutrinoless Double Beta Decay"

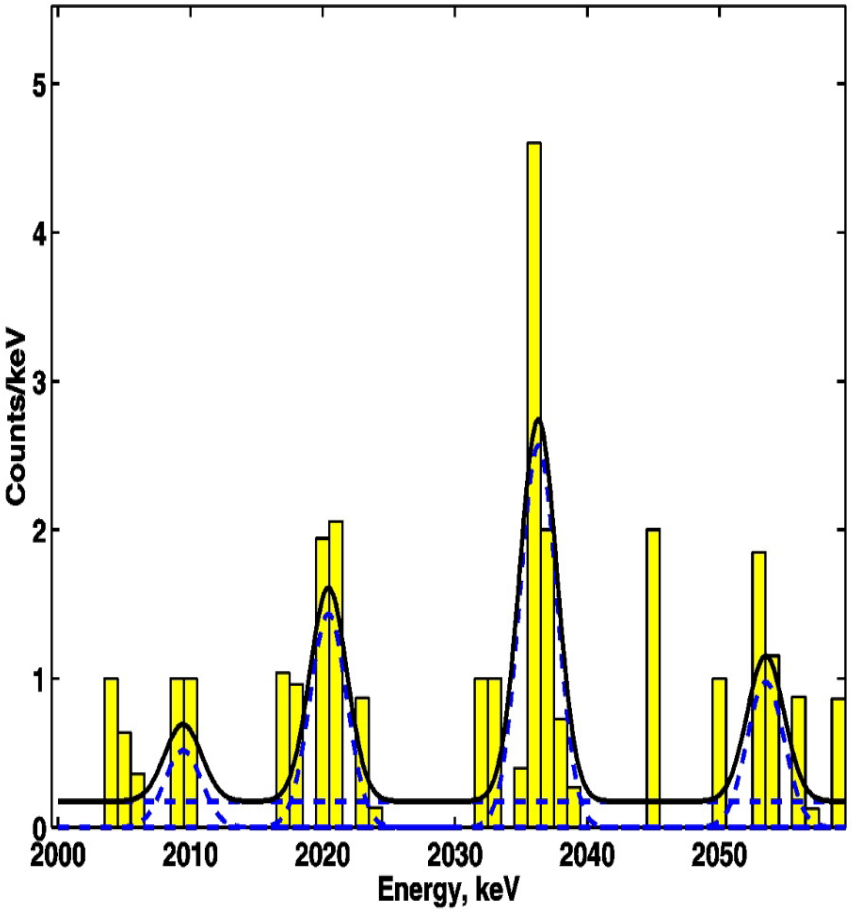
- "However, the analysis in KDHK makes an extraordinary claim, and therefore requires very solid substantiation. In this letter, we outline our concerns for the claim of evidence. Unfortunately, a large number of issues were not addressed in KDHK. Some of these are:
  - There is no null hypothesis analysis demonstrating that the data require a peak. Furthermore, no simulation has been presented to demonstrate that the analysis correctly finds true peaks or that it would find no peaks if none existed. Monte Carlo simulations of spectra containing different numbers of peaks are needed to confirm the significance of any found peaks.
  - There are three unidentified peaks in the region of analysis that have greater significance than the 2039-keV peak. There is no discussion of the origin of these peaks. "

More Comments

- "However, KDHK found numerous peaks in the 2000-2080 keV region in their search for a peak. Next, they constrained their double-beta decay (ββ) analysis to a small region that excluded these peaks. An analysis only within that limited region is used to claim a 2039-keV peak at the 2-3 σ level. The conclusion in KDHK must depend on the choice of window and on the number of peaks in the region near the window. "

In 2004 Klapdor et al. raise the stakes by refining cuts on the same data. The peak is up to 4 σ





The Split A_2

1967 The CERN Missing-mass spectrometer saw a broad meson resonance with a narrow dip in

$$\pi^- + p \rightarrow X^- + p$$

This was a very startling result. There was no prediction for such a structure.

1968 The same group, with a different apparatus, confirmed the deep dip.

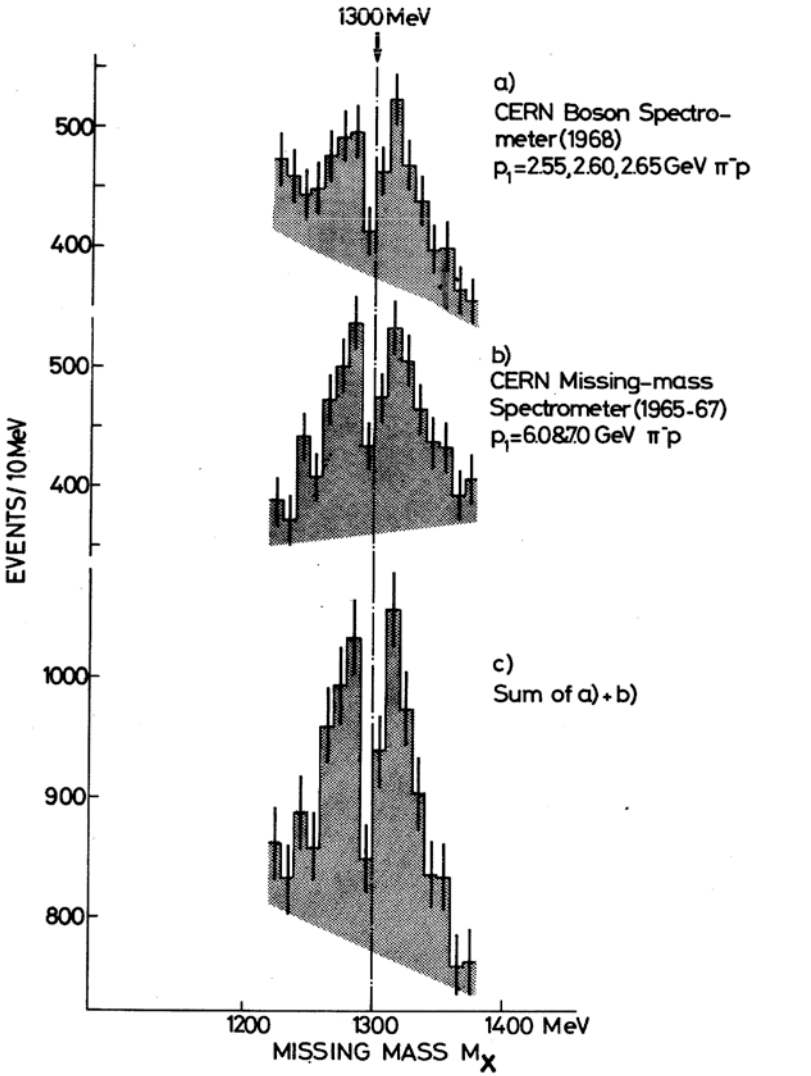


FIG. 4 Compilation of the total A2 data from CERN Boson Spectrometer (0° method) 1968, and CERN Missing-mass Spectrometer (Jacobian peak method) 1965-67.

The Split A_2

The probability of a fit to a single Breit-Wigner is <<0.1%. Combined, this was a 7σ dip.

Best fit is with a dipole field, but no plausible mechanism for such.

1971 Two high-statistics experiments saw no dip, with very high statistics.

1972 My thesis experiment saw no dip, with high statistics.

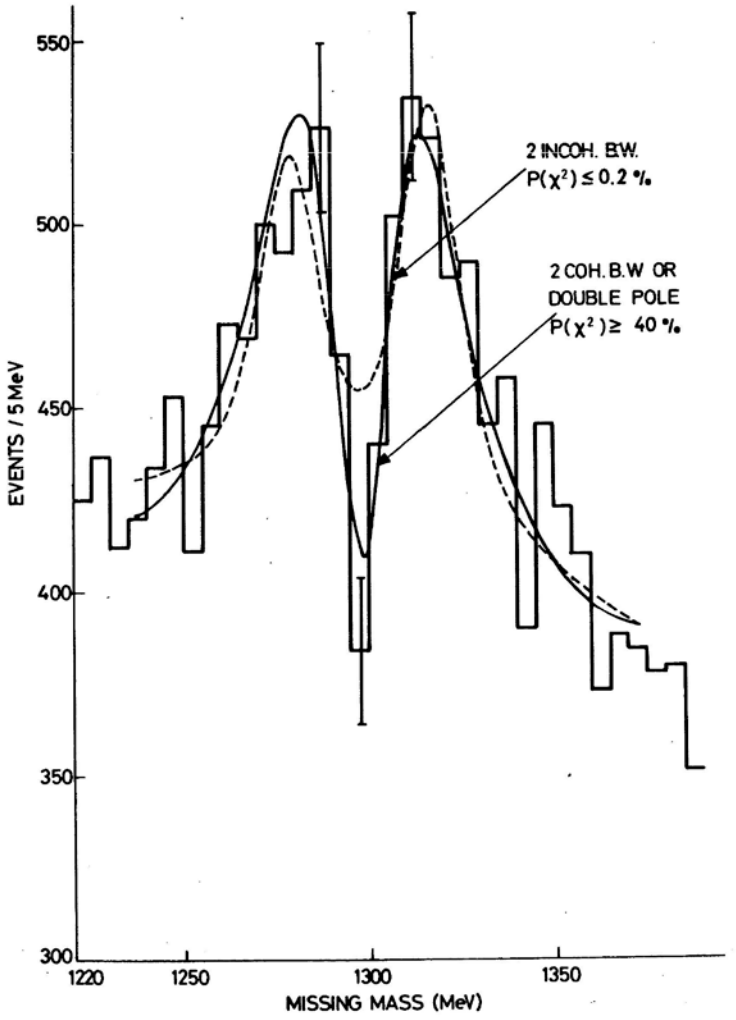


FIG. 5 FITS TO THE TOTAL(MMS+CBS)A2 DATA

Statistics gone wrong

What can go wrong?

- Selection criteria favor the peak (or dip).
  - This is surprisingly hard to get right.
  - The problem is usually self-deception, not conscious bias.

For the A_2, what went wrong?

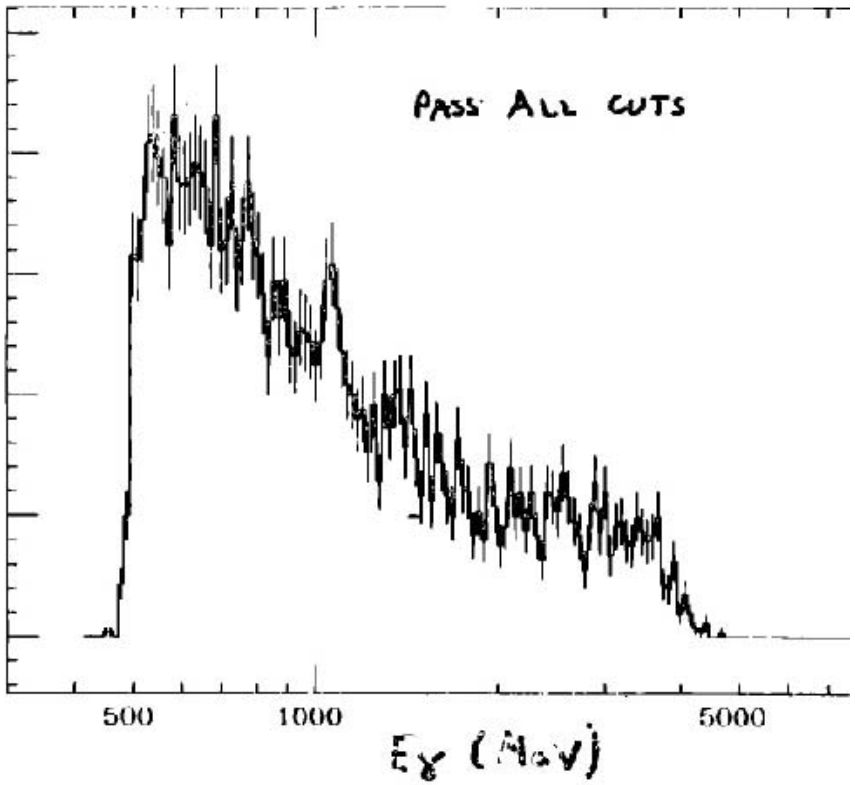
- The first dip was probably a fluctuation.
- After that, people had an unconscious bias toward it.
- "Whenever we didn't see the dip...we checked the apparatus and found something wrong."

An early Higgs candidate

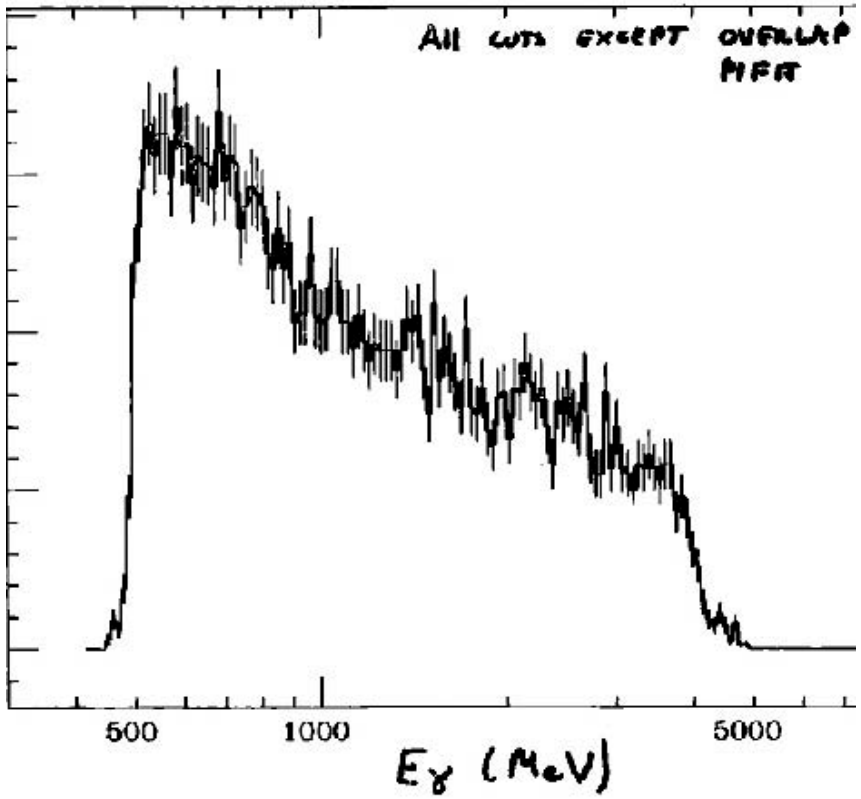
- In 1983 the Crystal Ball was operating at the DORIS collider at DESY, looking at $e^+e^- \to \Upsilon(9.5\text{GeV})$. They saw a narrow peak in the energy spectrum of photons at E = 1.0 GeV, implying the existence of a long-lived particle (zeta) with M = 8.32 GeV.
- "What most excites the high-energy community is the very real possibility that the zeta may be teh long-sought-after Higgs particle" (Physics Today, October 1984).
- They saw the peak after applying a cut that no hadron be within 30° of the photon.
- The problem was that the cuts were chosen while looking at the data. They got an additional year of running, used the same cuts, and saw no peak.

The Zeta data

with special cuts



with those cuts removed



Feynman on Experimenter's Bias

- "We have learned a lot from experience about how to handle some of the ways we fool ourselves. One example: Millikan measured the charge on an electron by an experiment with falling oil drops, and go an answer which we now know not to be quite right. It's a little bit off because he had the incorrect value for the viscosity of air."
- "It's interesting to look at the history of measurements of the charge of an electron, after Millikan. If you plot them as a function of time, you find that one is a little bit bigger than Millikan's, and the next one's a little bit bigger than that, and the next one's a little bit bigger than than, until finally they settle down to a number which is higher."

Feynman on Experimenter's Bias

- "It’s a thing that scientists are ashamed of—this history—because it’s apparent that people did things like this: When they got a number that was too high above Millikan’s, they thought something must be wrong—and they would look for and find a reason why something might be wrong. When they got a number closer to Millikan’s value they didn’t look so hard. . . "
- **"The first principle is that you must not fool yourself—and you are the easiest person to fool."**

Statistics in particle physics

- The reason that we now use blind analyses and more formal methodologies of statistics are to avoid the historical pattern of experimenter's bias.
- There is a danger in looking at refined statistical analyses, however. The implicit assumptions made in a given analysis have proven to be the weak spot that leads to false discoveries.

Pearson's χ^2 statistic

Test statistic for comparing observed data $\vec{n} = (n_1, \ldots, n_N)$ (n_i independent) to predicted mean values $\vec{\nu} = (\nu_1, \ldots, \nu_N)$:

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\sigma_i^2} \,, \quad \text{where } \sigma_i^2 = V[n_i] \,.$$

(Pearson's χ^2 statistic)

χ^2 = sum of squares of the deviations of the ith measurement from the ith prediction, using σ_i as the 'yardstick' for the comparison.

For $n_i \sim$ Poisson(ν_i) we have $V[n_i] = \nu_i$, so this becomes

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} \,.$$

Pearson's χ^2 test

If n_i are Gaussian with mean ν_i and std. dev. σ_i, i.e., $n_i \sim \mathrm{N}(\nu_i\,,\,\sigma_i^2)$, then Pearson's χ^2 will follow the χ^2 pdf (here for $\chi^2 = z$):

$$f_{\chi^2}(z;N) = \frac{1}{2^{N/2}\Gamma(N/2)} z^{N/2-1} e^{-z/2}$$

If the n_i are Poisson with $\nu_i >> 1$ (in practice OK for $\nu_i > 5$) then the Poisson dist. becomes Gaussian and therefore Pearson's χ^2 statistic here as well follows the χ^2 pdf.

The χ^2 value obtained from the data then gives the p-value:

$$p = \int_{\chi^2}^{\infty} f_{\chi^2}(z;N)\,dz\ .$$

The 'χ^2 per degree of freedom'

Recall that for the chi-square pdf for N degrees of freedom,

$$E[z] = N\,, \quad V[z] = 2N\,.$$

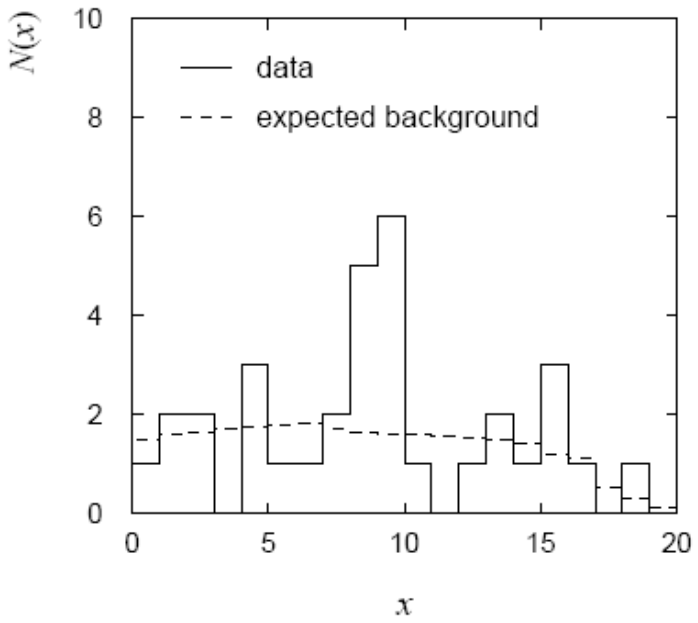
This makes sense: if the hypothesized ν_i are right, the rms deviation of n_i from ν_i is σ_i, so each term in the sum contributes ~ 1.

One often sees χ^2/N reported as a measure of goodness-of-fit. But... better to give $\chi^2$and N separately. Consider, e.g.,

$$\chi^2 = 15,\ N = 10 \rightarrow p - \text{value} = 0.13\,,$$
$$\chi^2 = 150,\ N = 100 \rightarrow p - \text{value} = 9.0 \times 10^{-4}\,.$$

i.e. for N large, even a χ^2 per dof only a bit greater than one can imply a small p-value, i.e., poor goodness-of-fit.

Example of a χ^2 test



← This gives

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} = 29.8$$

for $N = 20$ dof.

Now need to find p-value, but... many bins have few (or no) entries, so here we do not expect χ^2 to follow the chi-square pdf.

Using MC to find distribution of χ^2 statistic

The Pearson χ^2 statistic still reflects the level of agreement between data and prediction, i.e., it is still a 'valid' test statistic.

To find its sampling distribution, simulate the data with a Monte Carlo program: $n_i \sim \text{Poisson}(\nu_i)\,,\ i = 1, N\,.$

Here data sample simulated 10^6 times. The fraction of times we find $\chi^2 > 29.8$ gives the p-value:

$p = 0.11$

If we had used the chi-square pdf we would find $p = 0.073$.