Non chiederci la parola che squadri da ogni lato
l'animo nostro informe, e a lettere di fuoco
lo dichiari e risplenda come un croco
perduto in mezzo a un polveroso prato.
...
Codesto solo oggi possiamo dirti,
cio' che non siamo, cio' che non vogliamo.

E.M.

Statistics?

Nino

May 21, 2013

Statistics can be used to support or undercut almost any argument. M. vos Savant
Facts are stubborn, but statistics are more pliable. Mark Twain

The average human has one breast and one testicle. Des McHale

Introduction

Statistics should make experiments understandable. Experiments have measured properties of Nature (the charge of electron etc.), the existence of new particles and their interactions, just remaining in the field we know better: the elementary particles.

We will start with a few important experimental results, to describe how much statistics has helped.

Statistics role is also extremely important in other fields (medicine, biology...).
Statistics started with the needs of states to base policy on demographic and economic data (hence the name). Early studies dates back to the 14^{th} century: *the Nuova Cronica*, an history of Florence by the Florentine banker Giovanni Villani that includes much statistical information on population, ordinances, commerce etc.

A very nice experiment: Anderson 1933

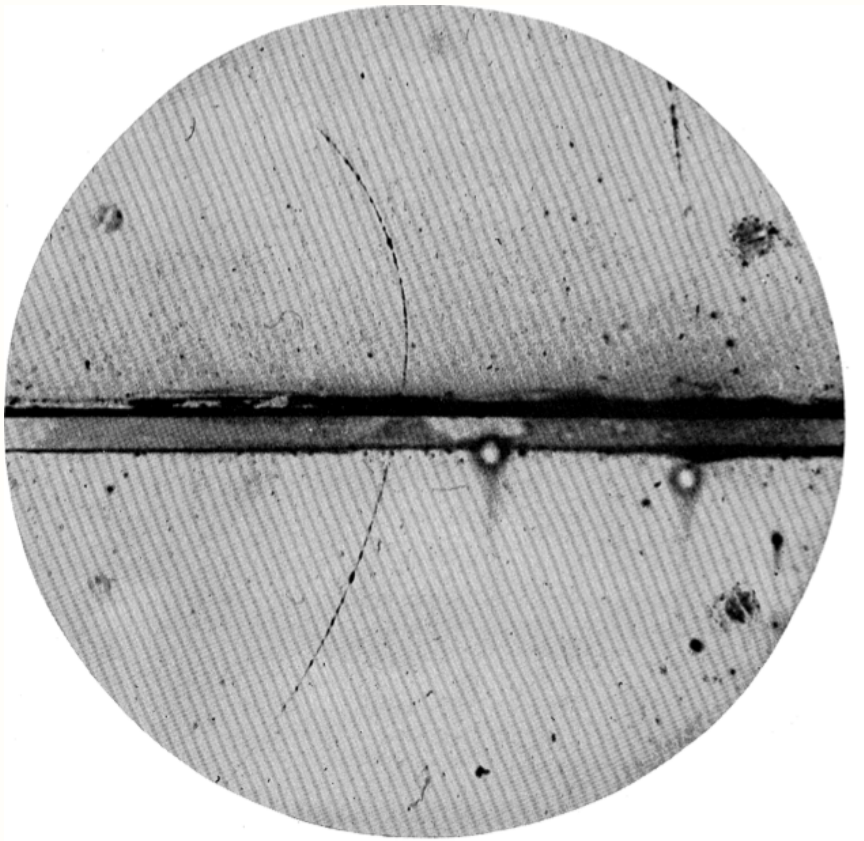


FIG. 1. A 63 million volt positron ($H\rho = 2.1 \times 10^5$ gauss-cm) passing through a 6 mm lead plate and emerging as a 23 million volt positron ($H\rho = 7.5 \times 10^4$ gauss-cm). The length of this latter path is at least ten times greater than the possible length of a proton path of this curvature.

The positive electron.
C. D. Anderson.

A well planned and lucky experiment!
There is no need of statistics: the track is coming from below (after crossing the lead plate has lost energy) it is positive (curvature) and cannot be a proton (since the range is too large).

An almost discovery: BNL 1972

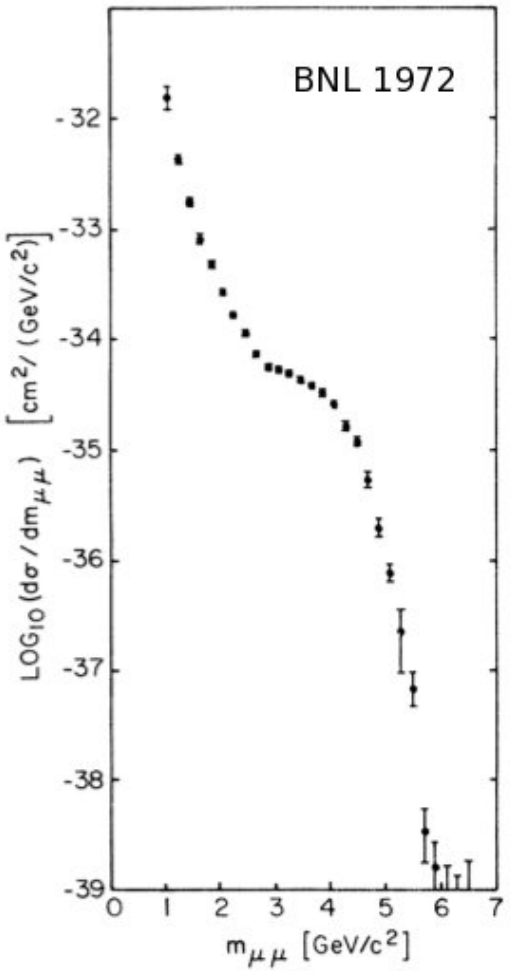


FIG. 10. $d\sigma/dm$. Weighted average of standard and "wide angle" events. Proton energy = 29.5 GeV.

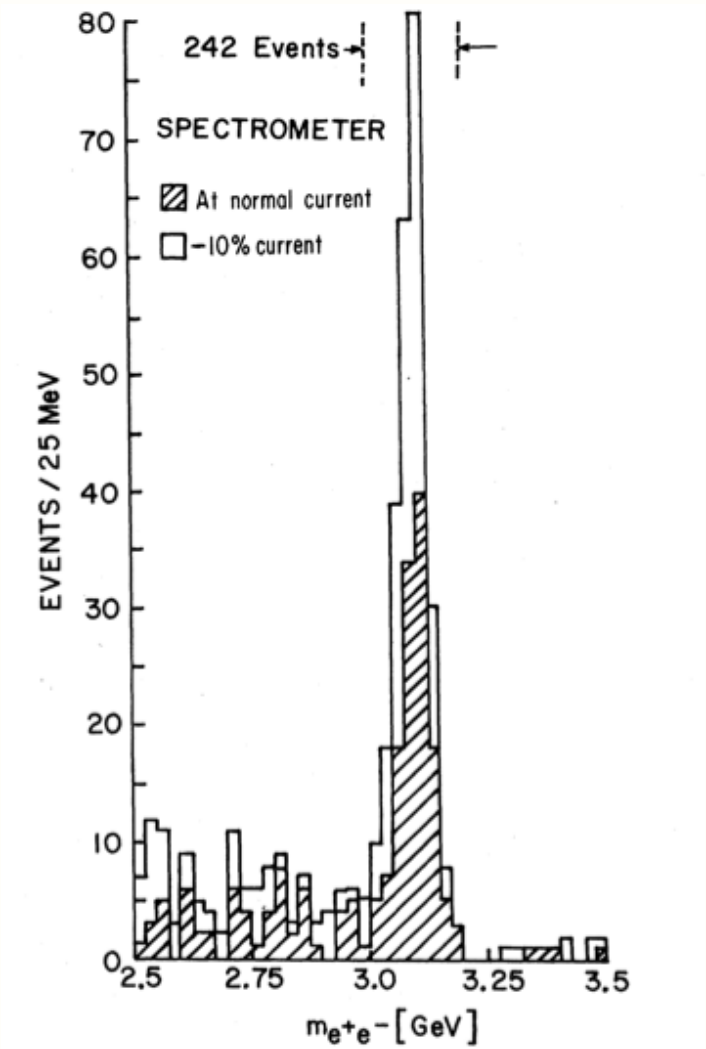
Observation of Muon Pairs in High-Energy Hadron Collisions.

J. H. Christenson, G. S. Hicks, L. M. Lederman, P. J. Limon, and B. G. Pope *Columbia University and Brookhaven National Laboratory*

E. Zavattini *CERN*

That was a case where statistics could not help (please, no unfolding!). A better experiment was needed!

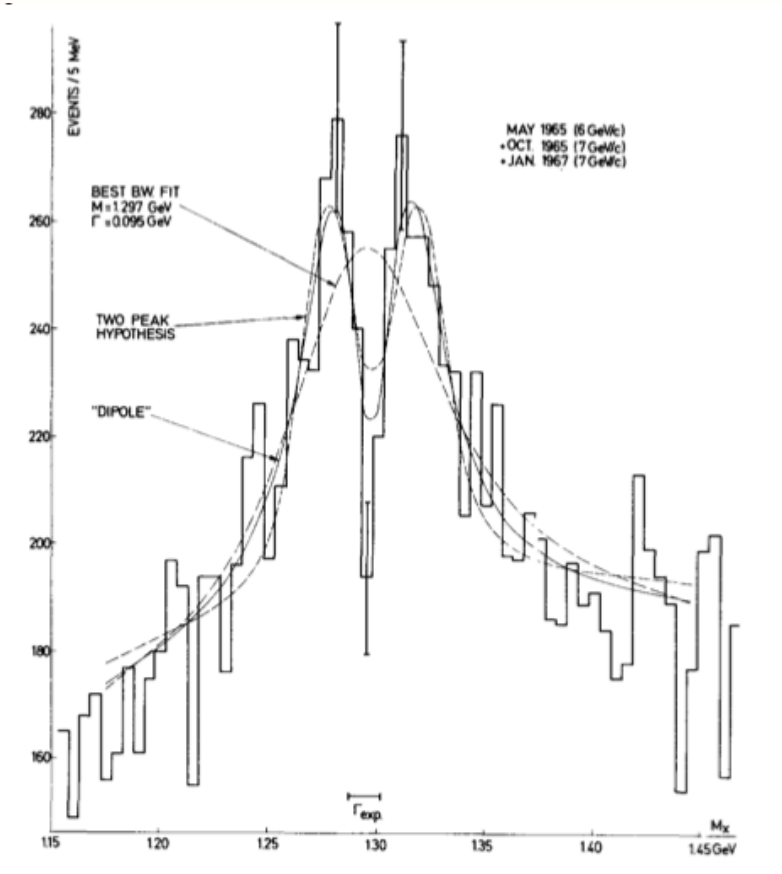
A classical discovery: J/ψ discovery



S. Ting 1974

In this case there is no need of statistics to claim for a discovery. The J/ψ stands clearly above a small and flat background.
Still Ting waited for a confirmation of his discovery from Richter, before publishing it.

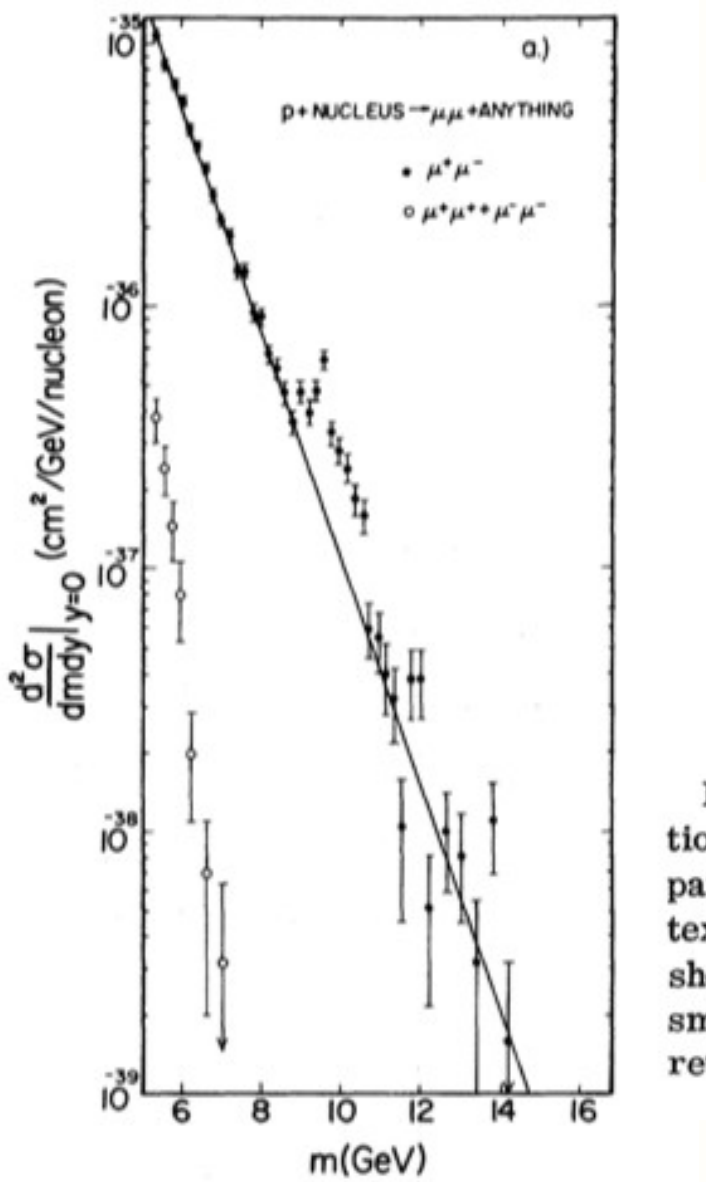
Another classical discovery: the A_2 – *split* effect



In 1970 two CERN experiments, MMS and CBS, claimed that the structure around 1300 MeV, believed to be the 2+ A_2 meson produced in pion proton collisions had a mass spectrum split into two peaks (six sigma effect). Other experiments confirmed this finding.

Finally, the removal of a suspect cut and as more data are taken, the split disappeared and the A_2 became a normal, single, undivided particle.

Another classical discovery: the Y discovery



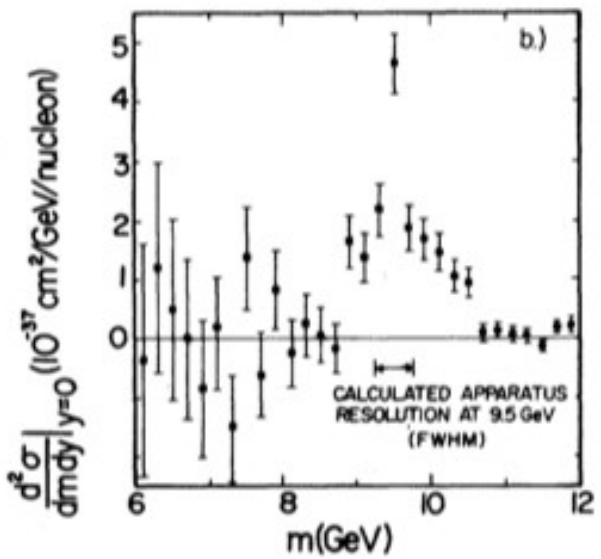


FIG. 3. (a) Measured dimuon production cross sections as a function of the invariant mass of the muon pair. The solid line is the continuum fit outlined in the text. The equal-sign–dimuon cross section is also shown. (b) The same cross sections as in (a) with the smooth exponential continuum fit subtracted in order to reveal the 9–10-GeV region in more detail.

The pentaquark case, (can statistics become a killer?)

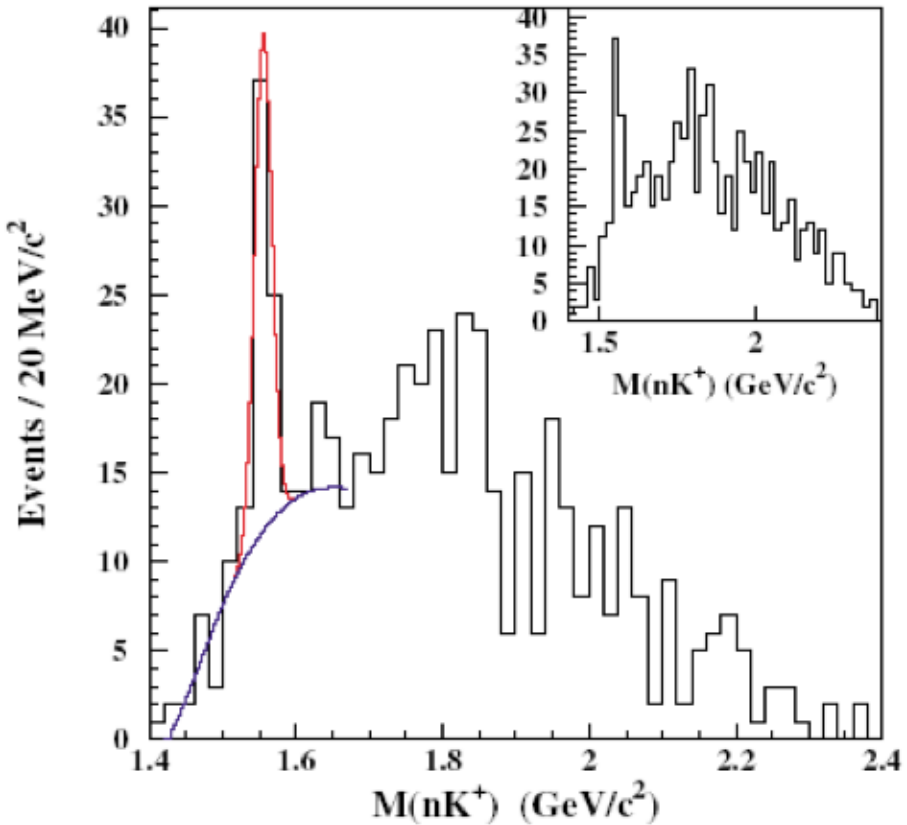


FIG. 4 (color online). The nK^+ invariant mass spectrum in the reaction $\gamma p \to \pi^+ K^- K^+(n)$ with the cut $\cos\theta^*_{\pi^+} > 0.8$ and

CLAS collaboration.

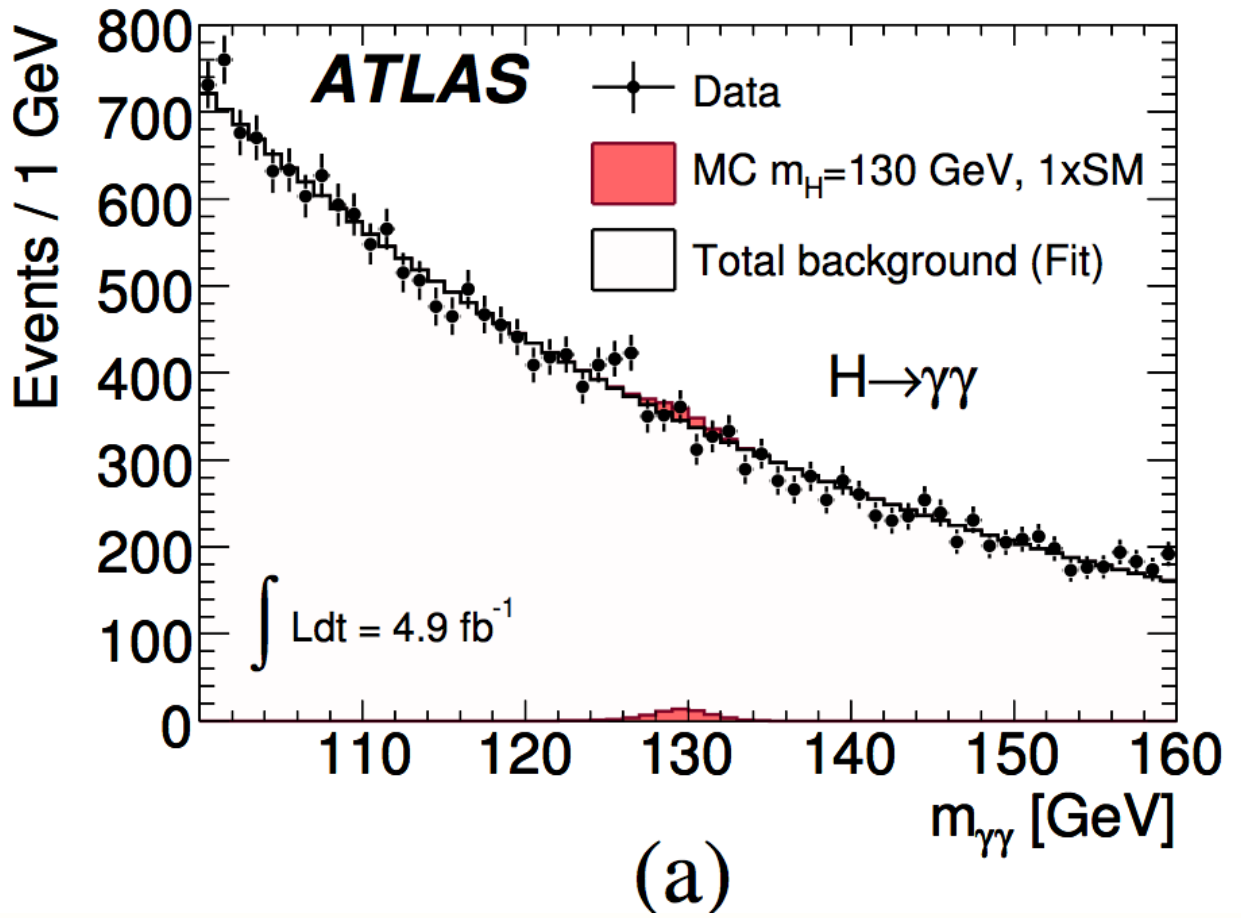
Claimed evidence:

$$\frac{N_S}{\sqrt{N_B}} = 7.8 \pm 1$$

there is ... overwhelming evidence that the claimed pentaquarks do not exist...
The whole story - the discoveries themselves, the tidal wave of papers by theorists and phenomenologists that followed, and the eventual "undiscovery" is a curious episode in the history of science. (C. Amser et al. (Particle Data Group) (2008)).

Modern "discovery" (performed using statistical methods)

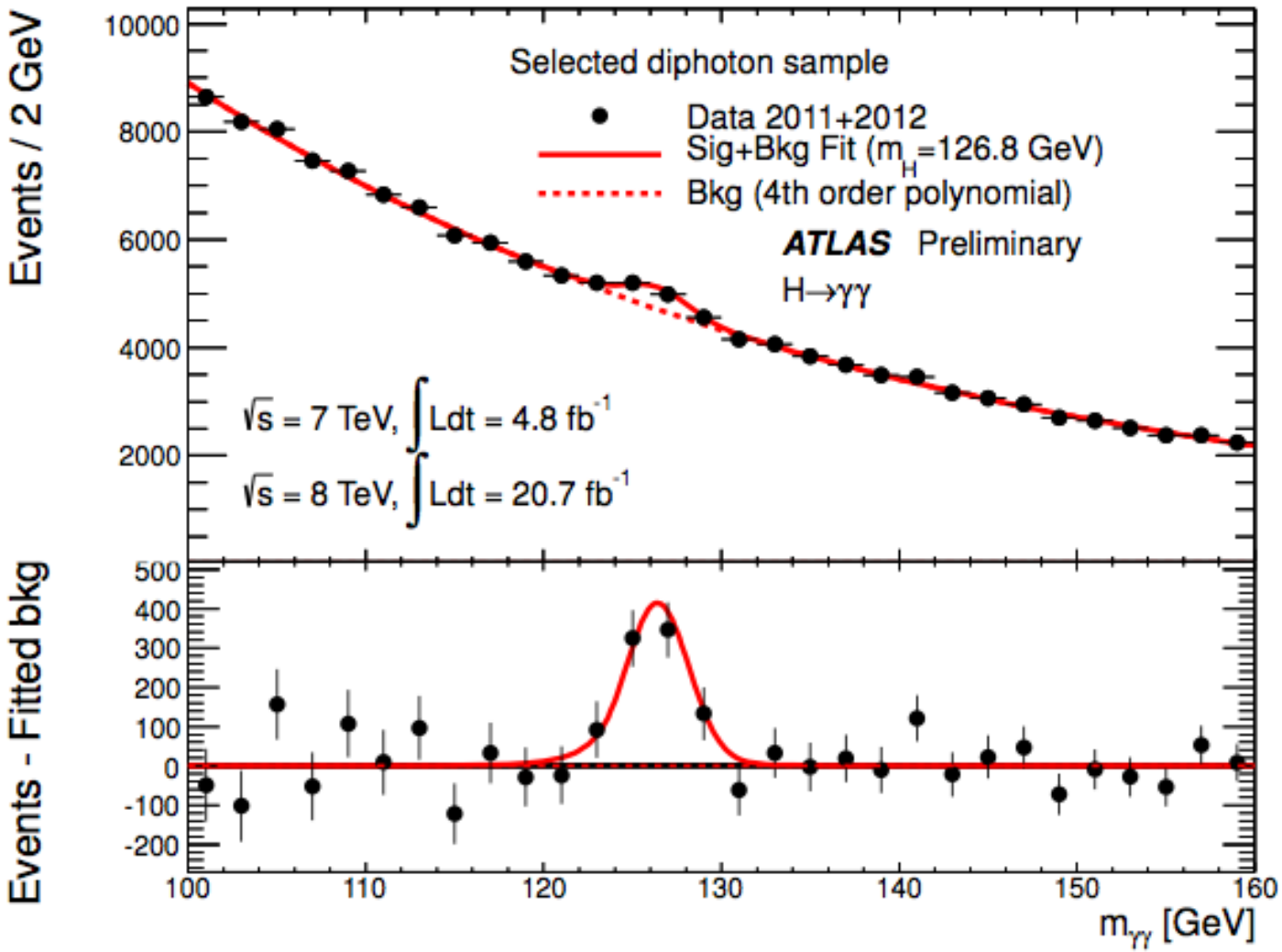
The quest for the Higgs is a much larger challenge:



Statistics and its magic is all needed, hoping that it will also provide some confidence (in the psychological sense) on the Higs discovery.

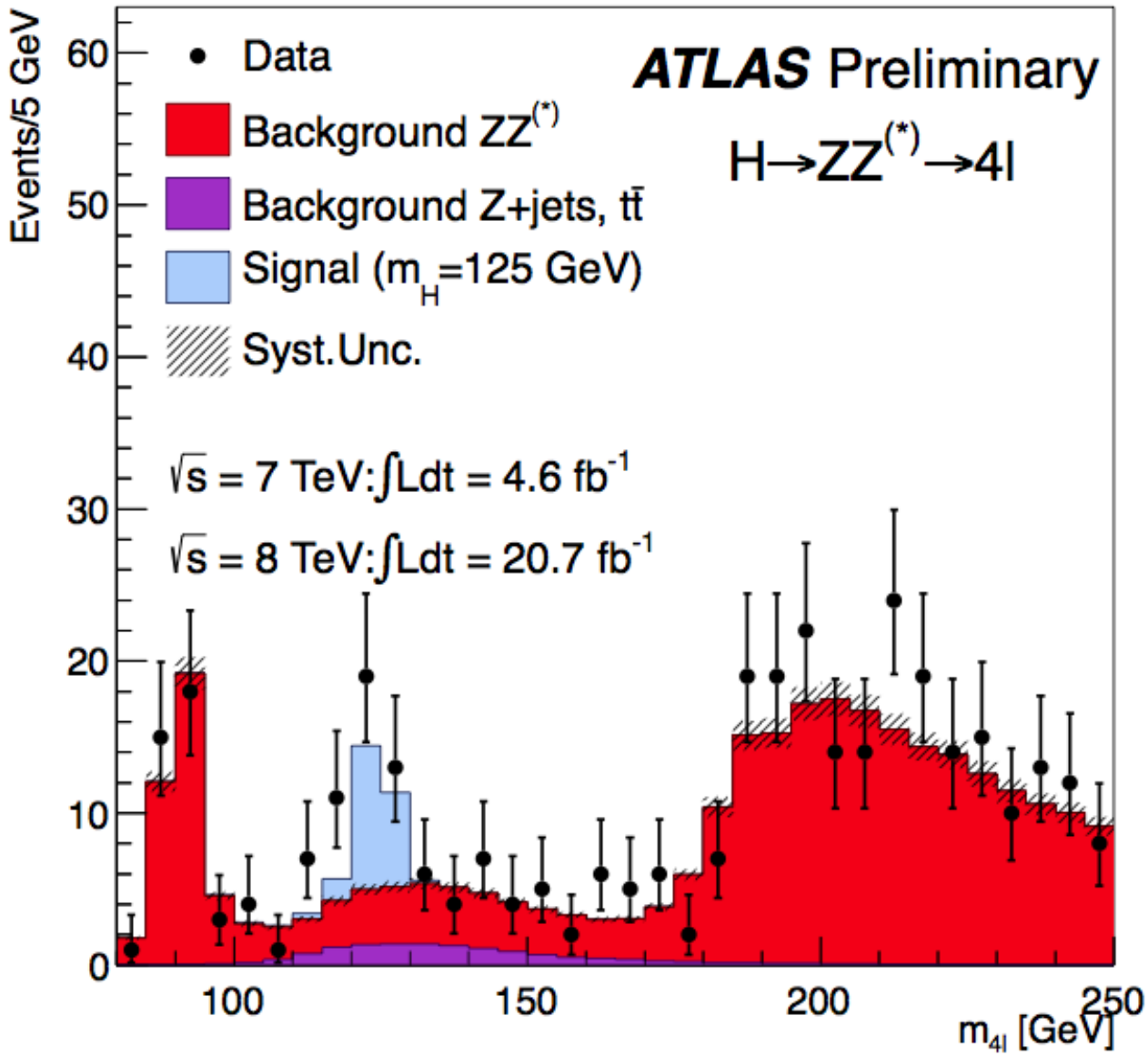
Modern "discovery"

Updates in Higgs hunting, recent plots with $L \approx 25 fb^{-}1$.



Modern "discovery"

Updates in Higgs hunting, recent plots with $L \approx 25 fb^{-}1$.



The first steps to understand

There is a lot of technique in doing statistics. We will try to be very light.

Some definitions an explications are unfortunately needed. Thus we will start by revisiting, very shortly, the standard statistical tools:

1. Hypothesis test ← we will discuss only of this
2. Determine parameters
3. Confidence Intervals
4. Coverage

And then some more advanced statistical tools designed for our search.

Definition of probability: the frequency school

The probability is a property of the system under study hence can be measured by:

$$P(E) = \lim_{N_{tot} \to \infty} \frac{N_E}{N_{tot}}$$

- $P(E)$ is a property of the system.
- exists only for repeatable experiments;
- the theory relies on two concepts: the *random event* and the possibility of performing *long run of experiments* in uniform conditions, if not in practice, at least in principle.
- since $P(E)$ exists we can repeat the samplings. Even virtual samplings (without making the experiment) are allowed and compared with the actual measurement.
- This probability is the one used in QM:

$$P(x \in S) = \int_S | \psi(x) |^2 \, dV$$

Definition of probability: the Bayes school

This probability $P(E)$ expresses one's opinion on the proposition (E) and depends on the information available to the observer.

- To any proposition, on which there is no certainty, we associate a numerical value, the probability.
- The support of the frequency interpretation to probability is lost. The probability cannot be measured. It is assigned to the event/proposition by the observer. "The probability does not exist" (de Finetti).
- There is NO random event, NO repeated samples.
- It is used very often in everyday life (*...he is probably right... ...m_H is probably less than 200 GeV... etc. etc.*)
- It is subjective and cannot be falsified.
- Bayes rule mixes the prior believes to the experiment:

$$P(\theta \mid data) \propto P(data \mid \theta) \cdot Prior(\theta)$$

Definition of probability: the Bayes school

There is nothing like the sample mean:

$$\bar{x} = \frac{\sum_i x_i}{n}$$

to estimate the true value of the parameter since there is NO real value of the parameter... The mean is computed as:

$$\bar{\theta} = \int P(\theta \mid data) d\,\theta$$

All information is stored in the likelihood $L(\theta, measured\ data)$. Data are NOT random variables but fixed constant, the random variable is θ. All inference MUST be done on L only. Thus repeated samples are prohibited as concepts like $E(X)$ which relies on sampling on the whole sample space are also meaningless.

Fisher significance test (frequency)

H_0 and no alternatives

Example:

In a counting experiment, we measure a rate n_0. The background rate is b.

H_0: is n_0 compatible with the background only? To quantify the answer we use the *Pvalue*: the probability to observe results even more extreme than what is predicted by H_0:

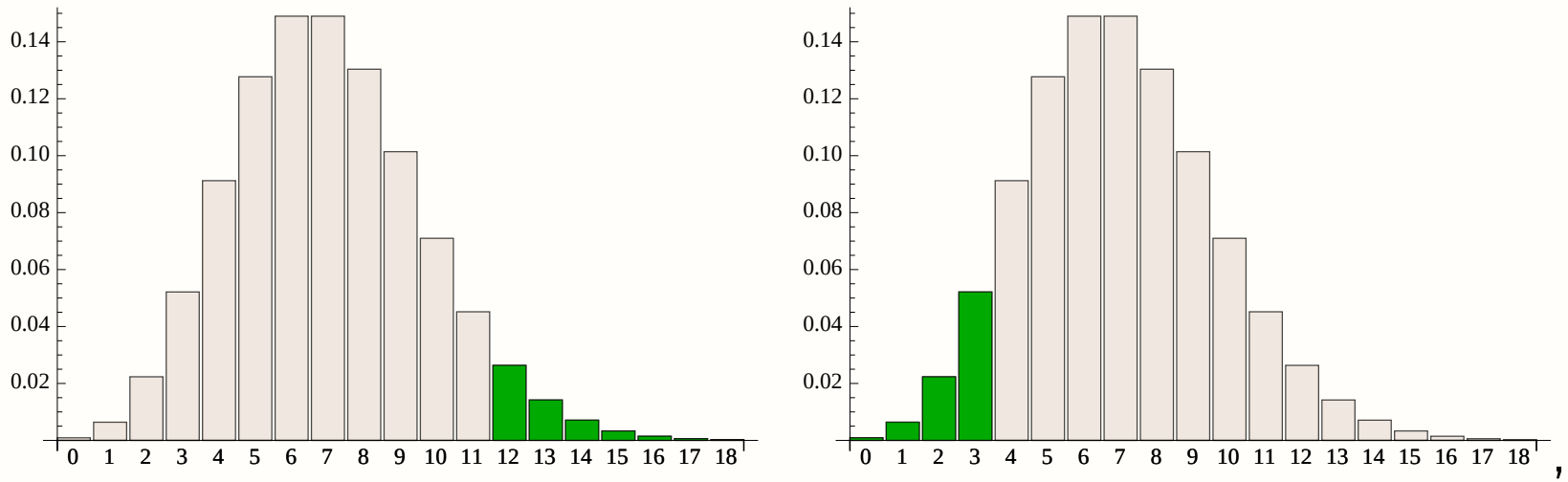
$$Pvalue = P(n \geq n_0 \mid H_0) \quad \left(\text{ or } \quad P(n \leq n_0 \mid H_0) \right)$$

The argument is: if the Pvalue is too small either H_0 is wrong or we got a rare result.

The decision to discard H_0 is finally left to the observer.

Pvalue: an aside comment

We have to consider the probability of all those results still more extreme w.r.t. H_0 expectations than what we actually measured .
Thus for instance the plots below show, in green the p-value in case of two measurements (n=12 and n=3) of a Poisson process with $\mu = 7$.



H_0 has to be rejected if n is too low or too large.
The reason for considering only the right tail is that H_0 is the background only hypothesis. The signal, if exist will produce counts on the right tail.

The Pvalue is a random variable

The Pvalue is function of a random variable thus is a RV itself.

It can be shown that the Pvalue distribution is uniform $U(0, 1)$.

The Pvalue is a tool to inform the experimentalist on the agreement between H_0 and data.

If the Pvalue is too low, we call the result significant.

The Pvalue is a piece of information used to help the decision whether reject H_0.
The automatic rejection if the Pvalue is too low has to be avoided.

Significance, an example

The experiment consists in measuring a Poisson variable: $n \sim Poiss(n \mid s + b)$ s is the hypothetical signal, b is the background known to be: $b = 1.2$ (exactly).

The experiment has measured $n = 7$

H_0: s=0

$$Pvalue = 1 - \sum_{i=0}^{i=6} Poiss(i \mid b = 1.2) = 2.5\ 10^{-4}$$

It is a significant result, $s = 0$ has probably to be rejected.

The argument is that a worse result (n equal or larger than 7) will occur, in average only once in 5000 experiments.

The Neyman decision making test

We have two hypotheses:

H_0: the null hypothesis and H_A: the alternative hypothesis.

The measurements ($\vec{X} = X_1 \cdots X_n$) are distributed either as $f_0(X)$ or $f_A(X)$. We shall define a region of sample space, w_α in such a way that:

$$P(\vec{X} \in w_\alpha \mid H_0) = \alpha \quad \alpha \text{ is called the size of the test}$$

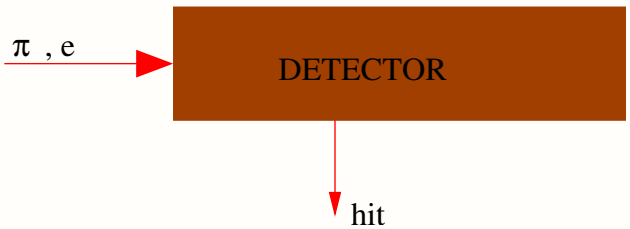
If the measurements fall in w_α we reject H_0 and accept H_A. α is the probability of wrongly reject H_0, (type I errors). This will happen in a fraction α of repeated experiments if H_0 is true.

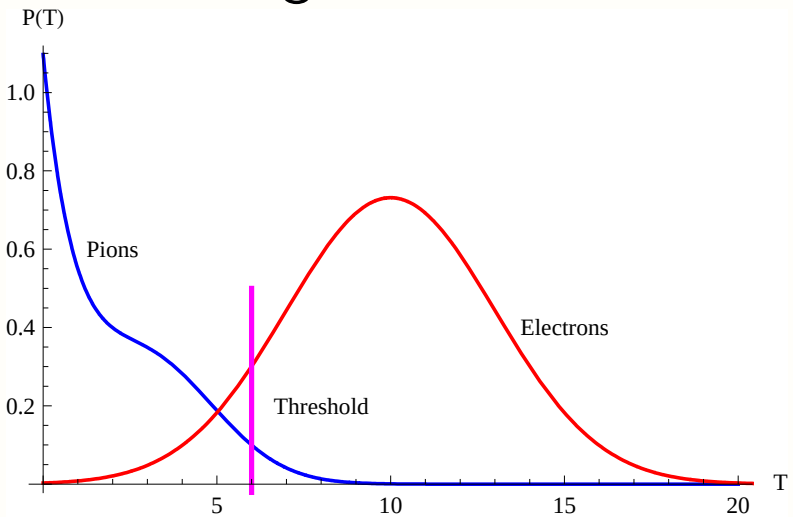
$$1 - \beta = P(\vec{X} \in w_\alpha \mid H_A) \quad \text{is called the power of the test}$$

A well designed test has a large power for a fixed size. A large power means that the hypotheses an be safely separated.

Example of Neyman test

In a beam of e, π we have a detector to tag electrons.





H_0: the hit is an electron,
H_A: the hit is a pion
Type I Error: $P(T \in w_\alpha \mid H_0) = \alpha$ is the loss of electrons,
Type II error: $P(T \in \bar{w}_\alpha \mid H_A) = \beta$ is the contamination of pions in the electron sample.

The test is designed in order to keep the losses under control (not larger than α) and the contamination is minimized.

The difficult part is finding w_α

In case the two hypotheses are fully specified, Neyman showed that the most powerful test is based on the likelihood ratio:

$$\lambda(\vec{X}) = \frac{L(\vec{X} \mid H_A)}{L(\vec{X} \mid H_0)}$$

If λ is large then H_A is preferred. Thus the test is:

$$\lambda(data) > \lambda_\alpha \quad \text{we reject } H_0 \text{ and accept } H_A$$

but, at the same time:

$$P(\lambda(\vec{X}) > \lambda_\alpha \mid H_0) \leq \alpha$$

This defines λ_α.

The Neyman-Pearson lemma

The test is designed in such a way that the fraction of losses (fraction of experiments rejected even id H_0 is true) is not larger than α.
The power of the test is then optimized.

In decision theory the plane $\{\alpha, \beta\}$ is also called ROC plane and is used to optimize the tradeoff between false positive and true positive rates.

Neyman hypothesis test: example

The experiment consists in measuring a Poisson variable: $n \sim Poiss(n \mid s + b)$ s is the hypothetical signal, b is the background known to be: $b = 1.2$ (exactly).

The experiment has measured $n = 7$.

H_0: s=0 $\quad$ H_A : $s > 0$

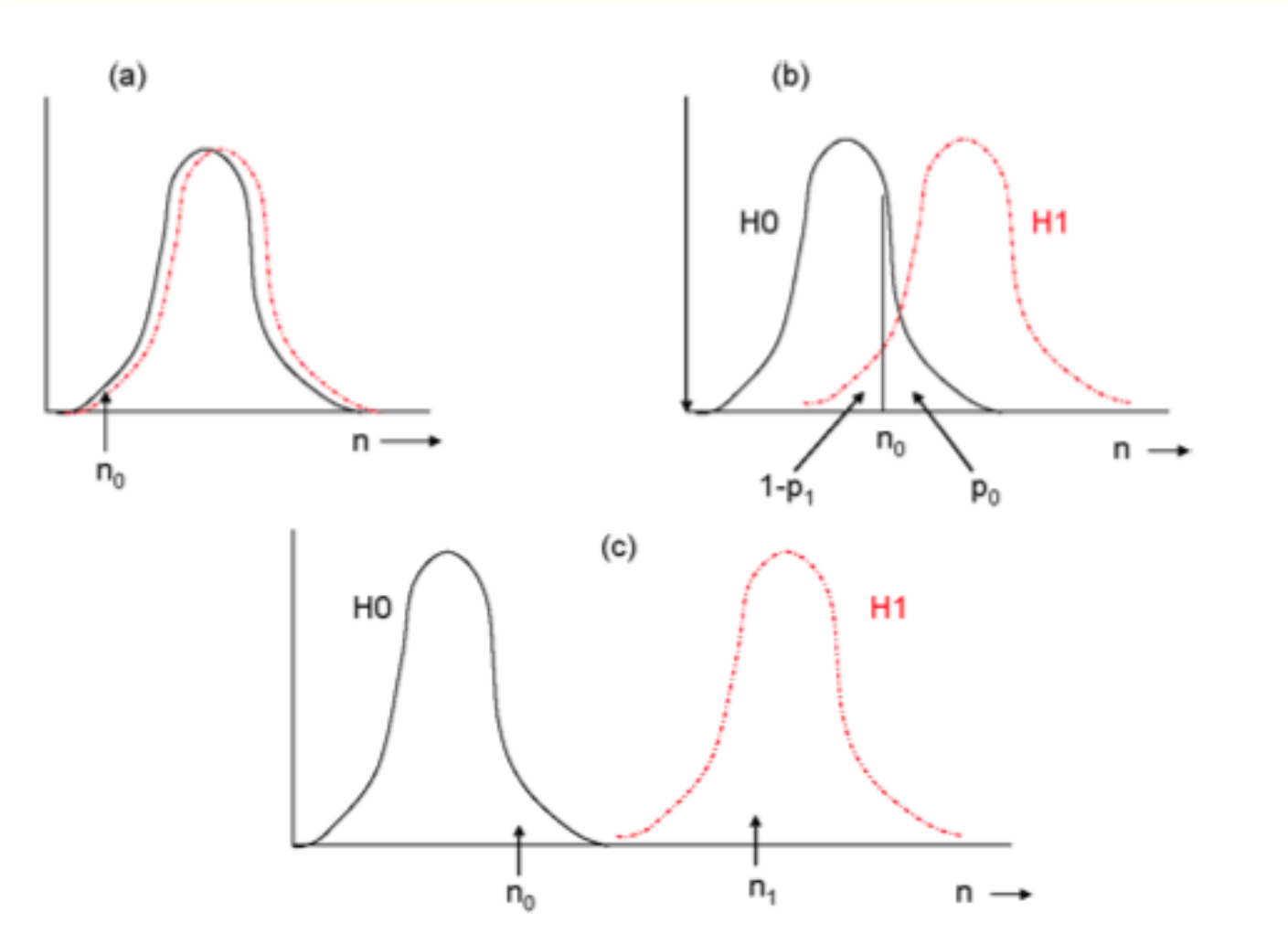
Find a region w_α that has probability $P(n \in w_\alpha \mid H_0) \geq \alpha$.

Assume $\alpha = 10^{-3}$ and solve for n_α:

$$\sum_{i=0}^{i=n_\alpha} Poiss(i \mid b = 1.2) \geq \alpha \rightarrow \quad n_\alpha = 5, \ (w_\alpha : \{5, \infty\})$$

H_0 is rejected ($n \in w_\alpha$) and H_A is accepted with significance α.

The CL_s method or what to do if the power is low



Expected distributions for a data statistic n: H_0,(background only) (black), and H_1 (dashed, red curve), where there is signal, resulting in larger n.

The CL_s method or what to do if the power is low

The signal production is very small in (a), marginal in (b), while in (c) the signal is abundantly produced.

n_0 events are detected (see (b)). The tail areas of H_0 above n_0 and of H_1 below n_0 correspond to probabilities p_0 and $1 - p_1$.

(c) shows a situation where H_0 and H_1 are well separated. Thus, n_0 would result in H_1 being excluded, while n_1 would be taken as evidence in favor of new physics.

(b) the signal is weak, and H_0, H_1 curves largely overlap. If the sensitivity is small (small power) it is preferred NOT to exclude the model.

CL_s penalizes the Pvalue by an amount that increases with decreasing sensitivity.

CL_s

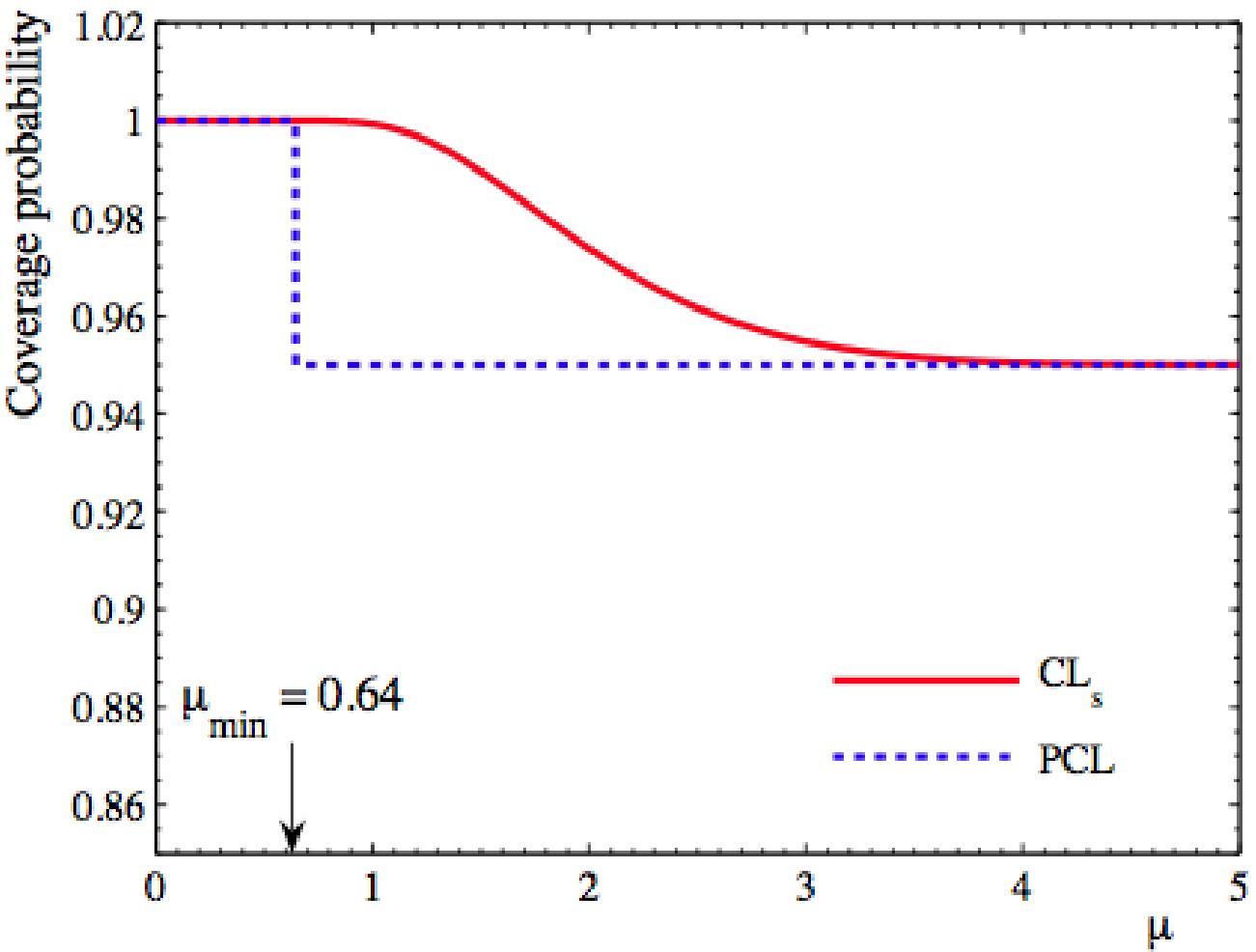
$$CL_s = \frac{p_\mu}{1 - p_0}$$

A value of signal intensity μ is excluded if $CL_s(\mu) < \alpha$.

Since $1 - p_0 < 1$:

- CL_s is more conservative,
- CL_s coverage is $> 1 - \alpha$,
- CL_s is used for exclusion.

Coverage CL_s



All this started long ago with Zech (1988)

In a count experiment with s and b the mean rate for signal and background, the distribution of counts n is:

$$P(n; s+b) = \sum_{n_b=0}^{n} \sum_{n_s=0}^{n-n_b} P(n_b, b) P(n_s, s) = Poiss(n, s+b)$$

Zech noticed that since N events have been observed, than $P(n_b, b)$ no longer corresponds to our improved knowledge of the background distributions. Since n_b can only take the numbers $n_b < N$, its distribution has to be renormalized to the new range.

$$P^{'}(n, s+b) = \frac{P(n, s+b)}{P(n_b \leq N)}$$

and

$$CL = \frac{P(n \leq N, s+b)}{P(n_b \leq N, b)} = \frac{CL_{s+b}}{CL_b}$$

All this started long ago with Zech (1988)

In the previous formula CL is fixed (95%) and the upper limit for the signal s (if b is well known) is obtained by solving the equation.

The procedure looks formally frequentistic but in fact has Bayesian flavor.

How Atlas searched for Higgs?

In the experiment each event fills an histogram. In the signal sample we have:

$$\{n_i, i = 1 \cdots n\}, \quad E(n_i) = \mu s_i + b_i$$

s_i, b_i are the mean of signal and background events, μ is the strength of the signal ($\mu = 0$ is background only) In a control sample we monitor the background:

$$\{m_i, i = 1 \cdots m\}, \qquad E(m_i) = u(\Theta) \quad \Theta \text{ are nuisance}$$

The experiment consider two hypotheses:

$$H_0 \quad : \quad \mu = 0$$
$$H_1 \quad : \quad \mu \geq 0 \qquad 1 \text{ could be SM predictions}$$

The statistics used in the analysis is the LR:

$$\lambda = \frac{L(\mu, \hat{\hat{\Theta}})}{L(\hat{\mu}, \hat{\Theta})} \qquad \text{double hat means conditioned to } \mu$$

Atlas search method

The idea is: if we manage to reject H_0, we are on the way to discover new signals.

The statistics to be used is:

$$\boxed{q_0 = -2Ln\ \lambda(0) \qquad \text{if } \hat{\mu} \geq 0, \text{ else } 0}$$

(data disagree from the model only if $\hat{\mu} \geq 0$.)

The disagreement is quantified by the Pvalue:

$$Pvalue = P(q_0 \geq q_0^{obs}) = \int_{q_0^{obs}}^{\infty} f(q_0 \mid 0) dq_0$$

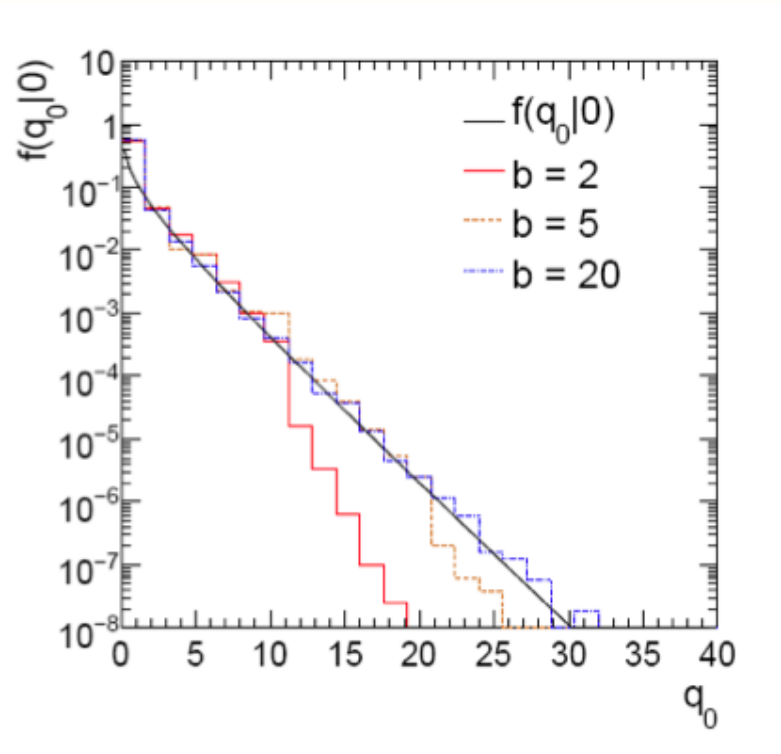
Here $f(q_0 \mid 0)$ is the distribution of $q_{\mu=0}$ computed assuming that the true strength of the data is also 0.

Atlas search method

The distribution $f(q_0 \mid 0)$ cal be computed analytically using the Wilks asymptotic approximation:

$$f(q_0 \mid 0) = \frac{1}{2}\delta(q_0) + \chi^2_1$$

and verified using toy MC.



Atlas Search methods

It is often useful to quantify the sensitivity of an experiment, to report the expected significance.

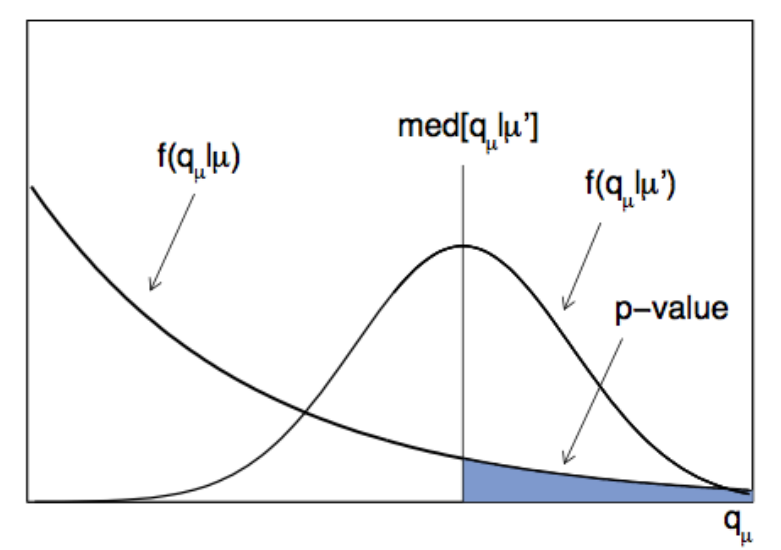
For instance we could characterize the sensitivity to discover a signal (μ) by the mean value of the Pvalue, to reject H_1 by testing H_0.

In fact it is more convenient to use the median instead of the mean because of its invariance under transformations of variables.

If we place limits on a coupling constant or a cross section (which is usually proportional to a coupling constant squared), then the median limit on one corresponds to the median limit on the other, while an average will be pulled to one side by the transformation.

Atlas Search methods

The sensitivity is illustrated in the next figure:



The distribution $f(q_\mu \mid \mu)$ is shown as a decreasing line (the curve is a χ^2_1).
In the same plot is shown $f(q_\mu \mid \mu^1)$, the curve is a *non central* χ^2.
The sensitivity is computed as shown:

$$Pvalue = P(q_\mu \geq Median(f(q_\mu \mid \mu^1)))$$

Comparing frequency and Baysian methods

Frequency statistics and Bayes statistics use often the same words to mean different concepts.

Let us see how they compare on the problem of hypothesis test!

Pvalue in Frequency and Bayes. Berger (PhyStat 2008).

Counting experiment: background and (may be) signal:

$$\begin{cases} H_0: & s = 0 \\ H_A: & s \geq 0 \end{cases} \qquad X \sim \frac{(s+b)^x e^{-(s+b)}}{x!} \qquad b \text{ exactly known}$$

We consider two cases:

x	b	pvalue	the pvalue is (Fisher)
7	1.2	$2.5\ 10^{-4}$	$pv = P(X \geq x \mid b, s = 0)$
6	2.2	$2.5\ 10^{-2}$	

Two comments:

... a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. (Jeffreys)

... a small p-value is only the first step in the interpretation of data. (R.A. Fisher)

Pvalue in Frequency and Bayes. Berger (PhyStat 2008).

There are a few other remarks to be done:

- Two ways of testing: based on the p-value or for fixed size:

  Fisher The p-value is used for rejecting H_0. In case 1 we reject at the level 10^{-4}

  Neyman Fix the size $\alpha = 10^{-3}$ thus reject H_0 if $x \geq n_\alpha$, n_α the largest x such that $P(x \geq n_\alpha \mid H_0) \geq \alpha$.

  In Neyman test we should quote an evidence of 10^{-3}. Fisher test suggest a factor ten smaller.

- Sequential experimentation. At LHC each month data are collected and analyzed. Each month there is space for discovery the Higgs or rejecting a previous discovery... How can we be sure that the probability of errors of type 1 is still α?

  In clinical experimentation the experiment stops after a discovery (for ethical necessity).

Pvalue in Frequency and Bayes. Berger (PhyStat 2008). I

Bayesian analysis starts from posterior distribution: H_i is the model or hypothesis. In our case Bayes rule is:

$$P(H_i \mid data) = \frac{P(data \mid H_i)P(H_i)}{P(data)} \qquad \text{i=1,2}$$

If the model depends on unknown parameters θ: $P(data \mid H, \theta)$, with a subjective prior $\pi(\theta)$ (subjective since it reflects the experimenter's believes), then it is *marginalized*:

$$P(data \mid H_i) = \int P(data \mid H_i, \theta)\pi(\theta)d\theta \qquad \text{then:}$$

$$\boxed{\frac{P(H_0 \mid data)}{P(H_1 \mid data)} = B_{01}\frac{P(H_0)}{P(H_1)}} \qquad B_{01} \text{ is the Bayes factor}$$

$$\text{Posterior Odds} = \text{Bayes Factor} \times \text{Prior Odds}$$

Pvalue in Frequency and Bayes. Berger (PhyStat 2008). II

In our case:

$$B_{01} = \frac{Poisson(x \mid b+0)}{\int_0^\infty Poisson(x \mid b+s)\pi(x)ds} = \frac{b^{x-1}e^{-b}}{\Gamma(x-1,b)}$$

Where we have used $\pi(s) = b(s+b)^{-2}$ as subjective prior.
The *objective prior* for the hypotheses is taken as:
$P(H_0) = P(H_1) = 0.5$. Since $P(H_0 \mid x) + P(H_1 \mid x) = 1$ we get:

$$P(H_0 \mid data) = \frac{B_{01}}{1+B_{01}} \qquad \text{evidence}$$

Finally the comparison between Bayes evidence and p-values is:

x	b	pvalue	Bayes evidence $P(H_0 \mid x)$
7	1.2	$2.5\ 10^{-4}$	$7.5\ 10^{-3}$
6	2.2	$2.5\ 10^{-2}$	$21\ 10^{-2}$

How to interpret B-factors? I

A value of $B_{01} \geq 1$ means that H_0 is more strongly supported by the data than H_A.
Harold Jeffreys gave a scale for interpretation of B:

B_{01}	Strength of Evidence (H_0)
$\leq 1:1$	Negative (supports H_A)
1:1 to 3:1	Barely worth mentioning
3:1 to 10:1	Substantial
10:1 to 30:1	Strong
30:1 to 100:1	Very strong
100:1 and larger	Decisive

Comparison to p-value is impossible. P-value considers only H_0 and built evidence against it. Here, instead, we have to compare two hypotheses and make a choice.
Significance test is impossible in Bayes statistics.

Pvalue in Frequency and Bayes. Berger (PhyStat 2008). I

Berger: *The Bayesian error probabilities given in the previous section differ from the corresponding p-values by factors of 30 and 10 in the two cases, respectively. What explains this?*
..... a serious discrepancy remains even when the prior is eliminated. This can be traced to the fact that the p-value is based on the probability of the tail area of the distribution, rather than the probability of the actual observed data.

Two comments:

In spite of the efforts by Berger and others to merge Bayes and frequency methods it seems to me that the best strategy is to keep distinct the two procedures, keeping well clear in mind that the meaning of concepts is often quite different even if the name is the same.

Pvalue in Frequency and Bayes. Berger (PhyStat 2008). II

Some problems can only be solved by Bayes methods, other by both. My personal suggestion is to stick to frequency methods, whenever possible and not get too confuse by a different answer from Bayes methods. Frequency methods look preferable in scientific research, for several reasons.

We will consider next the problem of dependence of CI from the amount of background where Bayes seem to provide a superior answer.

Blind analysis

Now a few comments on a recent trend in data analysis...

Blind analysis

The method of data analysis that is becoming more and more popular in HEP is to blind the data or part of the data to physicists that actually do the analysis, to avoid personal bias that could distort the final results.

The data will be completely available only after the analysis method (cuts, corrections etc.) are completely defined.

There are many ways to blind the analysis:

- Adding fake data to the experiment data set (that are removed at the end of the analysis),
- Use MonteCarlo only to set up the procedures,
- Hide the signal region,
- Keep visible only part of data... and many others.

There is a question that arises: after unblinding are we permitted to modify the cuts?

Blind analysis, example of bias

We will describe now a few results where bias could be suspected.

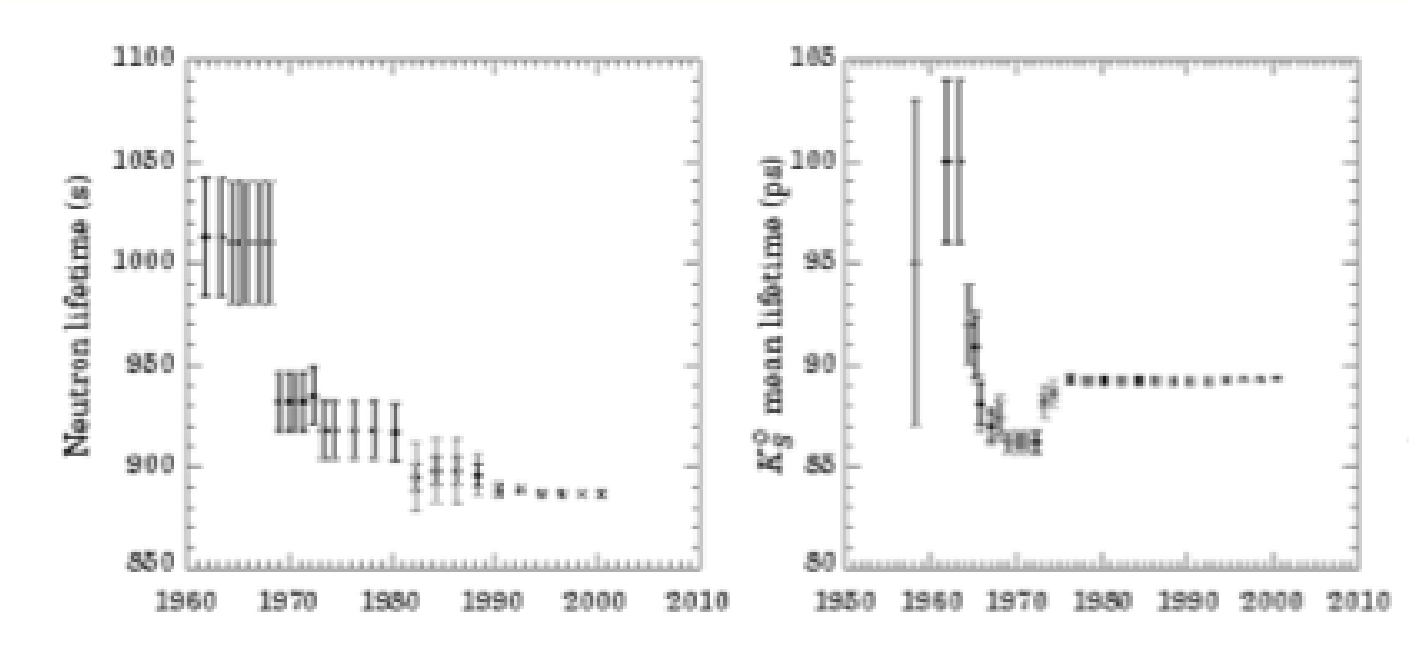
Could a blind analysis procedure have avoided these biases?

I am not sure that all the blind analyses did not have some smart student that, to avoid bad surprises did not peer in the hidden data...

How much do we pay for this (excess) of rigor?

Blind analysis, example of bias

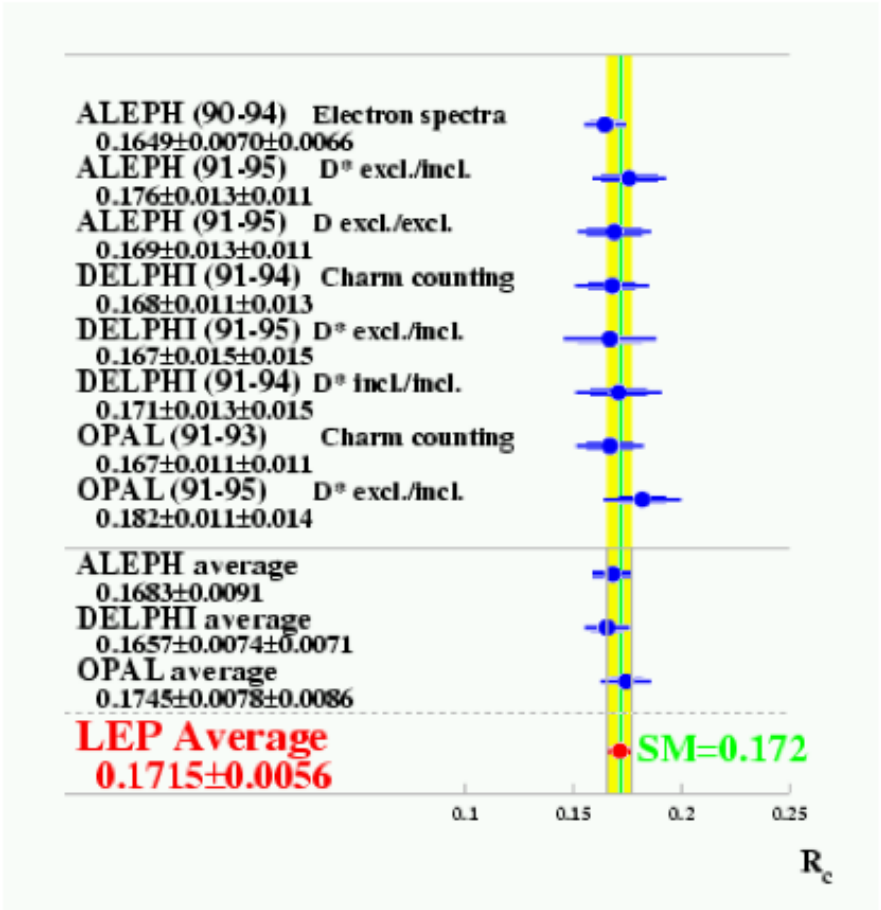
Here are examples of the measurement of neutron and K_s^0 lifetime, as a function of the year of the measurement.



Clearly there is a trend in the sequence of measurements; there is the suspect that the measurement are not independent of previous results.

Blind analysis, example of bias

This is a summary of LEP results on $R_c = \frac{Z \to c\bar{c}}{Z \to q\bar{q}}$:



Here the problem is that data do not fluctuate enough and any test statistics would fail.

Blind analysis

Mendel in his long and careful study of the genetic laws, made many hybridization experiments on garden peas (Pisum sativum). Mendel studied in particular the color of peas that do not blend upon cross pollination.

F1 was pure line of yellow peas. Two traits were identified: Y (dominant yellow), g(recessive green). When F1 plants breed, each has an equal chance of passing on either Y or g units to each offspring (F2 generation).

The possible combinations are gY, Yg, YY and gg. Three will give yellow offsprings since have at least one Y dominant unit. Only one combination will give green offsprings since it has two recessive traits. This is the origin of the 3:1 law.

Fisher analyzed the results of the F2 ratio and found the ratio Y-to-g to be implausibly close to the expected ratio of 3 to 1 and boosted against Mendel accusing him of fraud. (Mendel was dead since many years.)

Blind analysis

Reproduction of his experiments has demonstrated the validity of his hypothesis and correctness of the results.
It is possible that this is as an example of confirmation bias. This might arise if he detected an approximate 3 to 1 ratio early in his experiments with a small sample size, and continued collecting more data until the results conformed more nearly to an exact ratio.
To be noted that a blind analysis in this experiment would be impossible. Each plant of pea has to be inspected carefully and eventually rejected.