Data Analysis Basics

Probability Distributions

- **Poisson distribution**
- **Gaussian distribution**
- **Central Limit Theorem**
- **Propagation of errors**
- **Averaging with proper weights**

Statistics

- **Estimates of average, sigma, and errors on the estimates**

Confronting Data and Theory: Best Estimates of Theory Parameters

- **Max Likelihood Method**
- **Min χ^2 Method and dealing with χ^2 values**

Signal in Presence of background:

- **Statistical Significance of an observed signal**
- **Enhancing signal over background**
- **Confidence Levels when a signal is not seen**

Systematic Errors

Cross-checks

Traps of Wishful Thinking

Examples of low statistic "discoveries"

Probability Distributions

Poisson distribution: random (independent of each other) events occurring at rate ν.
Therefore, during time Δt, one should be expecting to detect (on *average*) $n = \nu \cdot \Delta t$ events.
However, the actually detected number of events, k, in a concrete experiment may be different:

Probability of detecting k events $P_k(n)$: $P_k = \frac{n^k}{k!} e^{-n}$

Average $\langle k \rangle = n$

Variance, Dispersion $\sigma^2 = \left\langle (k-n)^2 \right\rangle = n$

RMS (root of mean squared, or root-mean-squared) = $\sqrt{\left\langle (k-n)^2 \right\rangle} = \sqrt{n}$

Gaussian distribution is a good approximation for many typical measurement errors. Its importance is largely derived from the central limit theorem (see below).

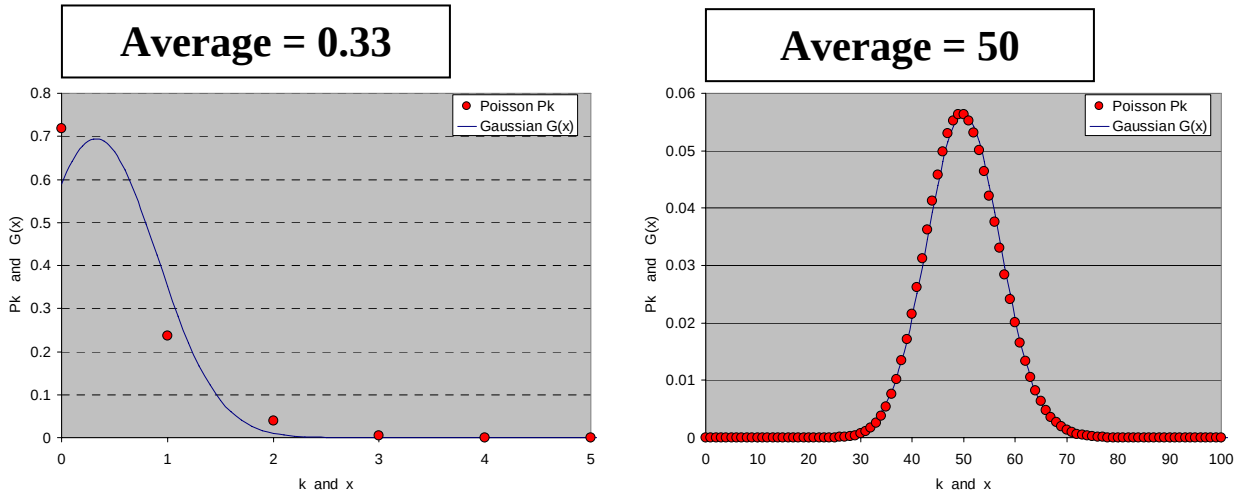
Probability of measuring x within the range from x_1 and x_2 is $P = \int_{x_1}^{x_2} p(x)dx$

Where p(x) is probability density: $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-x_0)^2}{2\sigma^2}}$

Probability to be within ±1σ is 68%
Probability to be within ±2σ is 95%
Probability to be within ±3σ is 99.7%

Central Limit Theorem: if one has n independent variables $x_1, \ldots, x_n$ having probability distribution functions of any shape (but with finite means μ_i and variances σ_i^2), the sum $X = \sum x_i$ at $n \to \infty$ will have the Gaussian distribution with the mean equal sum of μ_i and the variance equal to sum of σ_i^2.

Poisson distribution of large n (n>>1) is very close to Gaussian with x_0=n, σ^2=n.



Propagation of errors:

m=f(x):
if x has a small uncertainty σ_x,
one can estimate $\sigma_m = f_x \cdot \sigma_x$

m=f(x, y):
if x and y have small uncertainties σ_x and σ_y and *no correlations*,
$\sigma_m^2 = (f_x \cdot \sigma_x)^2 + (f_y \cdot \sigma_y)^2$

Averaging:

Assume that there are
two measurements of x (x_1 and x_2) that have estimated or known errors σ_1 and σ_2.

One can easily calculate that the best estimate of the value of x and the error on this estimate are:

$$x_m = w_1 x_1 + w_2 x_2, \quad \text{where } w_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \quad \text{and} \quad w_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$$

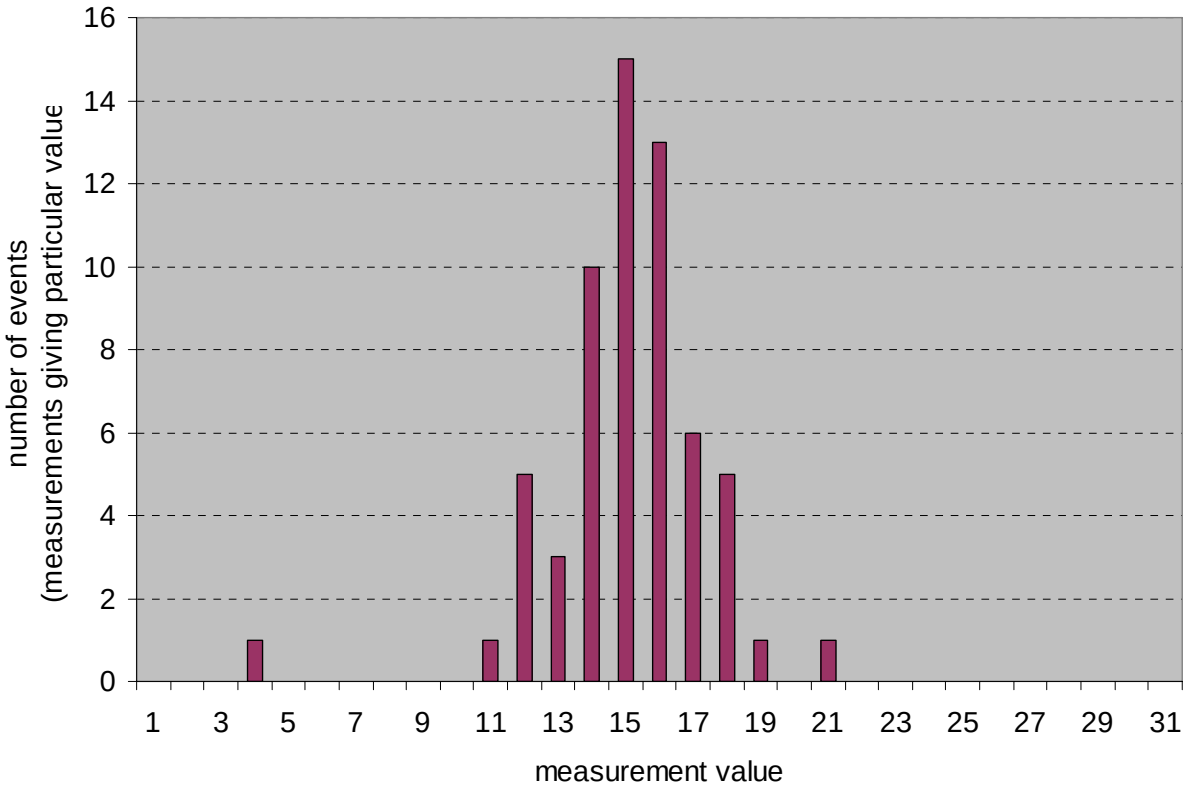
$$\sigma_m^2 = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

Trivial consequences:

- a lousy measurement can be ignored, it hardly adds any weight for the estimate and does not improve the error on the estimate
- two equally good/bad measurements should be counted with equal weighs, and the error from two measurements is $1/\sqrt{2}$ better than from a single measurements

Statistics:

Given the finite number of measurements,
a) estimate probability distribution function parameters (e.g., mean, width, …) and
b) evaluate errors on the estimations



Assume that the true probability distribution has mean x_0 and dispersion $D=\sigma_0^2$

Best estimate of mean: $x_m = \frac{1}{N}\sum_{i=1}^{N} x_i$

Best estimate of dispersion $\sigma_m^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - x_m)^2$

Estimate on error in x_m: $\delta x_m = \frac{\sigma_m}{\sqrt{N}}$

Estimate on error in σ_m: $\delta\sigma_m = \frac{\sigma_m}{\sqrt{2N}}$ (for Gaussian distribution and large N)

Confronting Data and Theory: Best Estimates of Theory Parameters

The primary questions one must answer are:

a) is the theory consistent with data?
b) what are the best estimates on theoretical parameters?
c) what are the errors on the estimates?
d) are there any indications that experimental data are not self-consistent?

Max Likelihood Method

Generic Example:

- Data: a set of y_i measurements at x_i points with
  - known $f_i(y_i|y)$ error distribution functions: probability of measuring y_i when the true value is y
  - and no correlations between points
- Theory with parameter(s) a: $y=F(x, a)$

Probability to get a particular set of measurements y_i for a given choice of parameter(s) a:

$$dP = \prod_i dp_i = \prod_i f_i(y_i \mid F(x_i,a))dy_i = \prod_i f_i(y_i \mid F(x_i,a))\prod_i dy_i = L(y_i \mid a)\prod_i dy_i$$

$L(y_i \mid a)$—Likelihood function.

We will choose the best possible theoretical parameter by maximizing the probability dP, or equivalently, the Likelihood function.
Note, it is often more convenient to maximize the log of L, $\ln\left(L(y_i \mid a)\right)$, instead of L—the answer would be the same as the log-function is monotonous.

Case of Gaussian errors:
Maximum Likelihood method is equivalent to the Minimum χ^2 method:

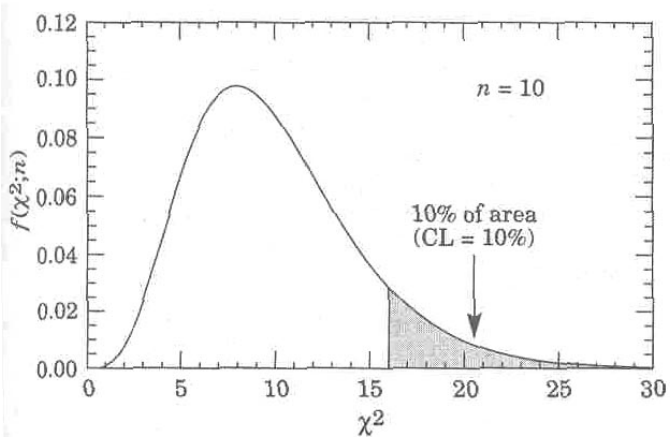
$$\ln L(y_i \mid a) = \ln \prod_i f_i(y_i \mid F(x_i,a)) = \sum_i \ln f_i(y_i \mid F(x_i,a))$$

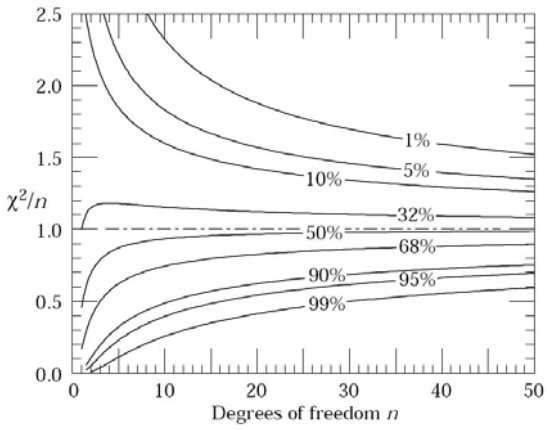
$$= \sum_i \ln \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(y_i - F(x_i,a))^2}{2\sigma_i^2}}$$

$$= Const - \sum_i \frac{(y_i - F(x_i,a))^2}{2\sigma_i^2}$$

$$= Const - \frac{1}{2}\chi^2$$

- Statistical expectations for χ^2 and what if you get something very different





  - $\overline{\chi^2} = n_{measurements} - k_{parameters} = n.d.f.$ (number of degrees of freedom)

  - Large χ^2
    - Theory does not describe Data
    - Errors are underestimated
    - There are large "negative" correlations (systematic errors)
  - Small χ^2
    - Errors are overestimated
    - There are large "positive" correlations (systematic errors)
  - Other cross-checks for "hidden" systematic errors

- Estimation of errors on parameter estimates from χ^2
  - $a \rightarrow a \pm \sigma_a, \qquad \chi^2 \rightarrow \chi^2 + 1$

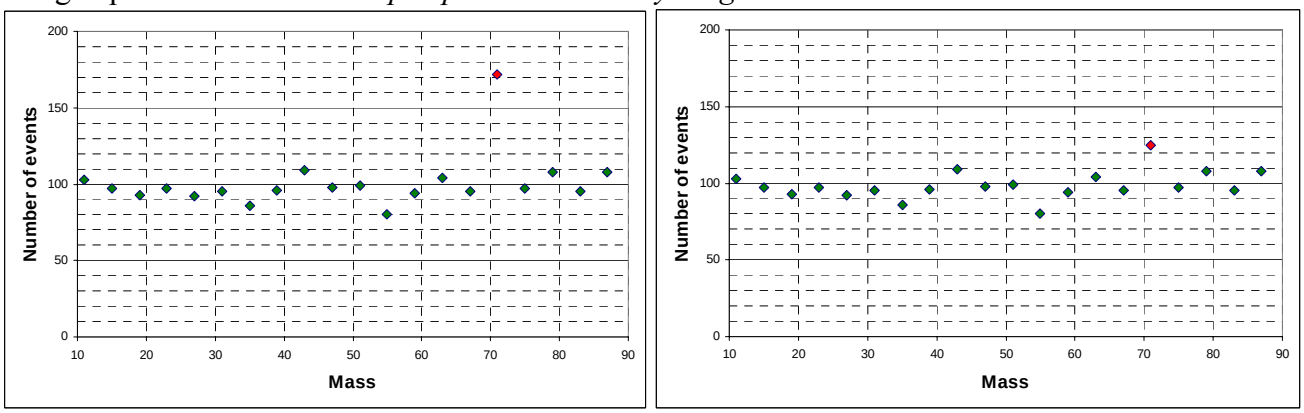
- When using the χ^2 minimization method is wrong:
  - Errors are not Gaussian, e.g.:
    - Gaussian with long tails
    - Small statistics (must use Poisson errors)
    - Flat error distribution for digitized signal (bin width >> noise)
  - Errors have correlations:
    - Both Max Likelihood and Min χ^2 Methods can be appropriately modified

Signal in presence of background: statistical significance of signal presence

You expect b events (background) and observe n_0 events and n_0 is greater than b. What is the significance of this observation? Have you discovered a new process that would account for the observed access of events? Or, maybe, this excess is a plain statistical fluke? Significance S is introduced to quantify the probability of a statistical fluctuation to observe n_0 events or more when you expect only b events. It maps a probability of a statistical fluctuation into a "number of Gaussian sigmas":

$$p(n \geq n_0) = \sum_{k=n_0}^{\infty} p(k) = \int_S^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx$$

Two plots below show histograms of reconstructed invariant masses for positive-negative charged particles in reactions $p + p \rightarrow e^+ + e^- + anything$



What is significance of the excess in the bin at Mass=70 in the left- and right-hand histograms?

The answer will depend strongly on whether you know a priori the mass of this resonance.

Assuming you knew that the resonance mass was predicted to be exactly M=71 and it would be very narrow, much narrower than the bins used in these histograms ΔM=4. Then, using bins other than the one centered at M=71, one can estimate background rate to be B=100 counts. Assuming that the background in bin at M=71 is the same as in the other bins, it is expected to fluctuate with σ=√(100)=√(B)=10. The excess of events in the resonance-containing bin in the first case is S=172-100=72, or 7.2σ, which can be written as $S/\sqrt{B} = 7.2\sigma$. The second histogram gives 25 excess events, or $S/\sqrt{B} = 2.5\sigma$. Probabilities p of such upward fluctuations are $<10^{-12}$ and 0.6%. Both numbers are very small and one can feel confident enough to claim the discovery of the predicted resonance.

If one did not know at what mass the resonance might show up, the significance of the peaks would be very different. Now we need to take into account that there are 20 bins and chances that at least one of them would fluctuate upward as measured would be larger that the probability of a particular a priori predetermined bin. Probabilities of none of the bin with flat background fluctuating upward as shown is $(1-p)^{20}$. Therefore, probability of at least one bin fluctuating upward is $1-(1-p)^{20}$, which gives $\sim 10^{-11}$ and 12%. One can see that the statistical significance of the discovery in the second case is not as striking and one would have to collect more data.

Enhancing Signal over Background:

Collecting more data. Collecting more data implies a reduction in relative statistical errors resulting in a cleaner signal identification.

- same histogram
- assuming that signal was real in the second histogram, collect 10 times more data.
- the background would be B=100×10=1000 events,
- the excess would also grow 10-fold, S=25×10=250 events
- Then, signal significance per bin would be $S/\sqrt{B}=250/\sqrt{1000}=7.9\sigma$.

Data cuts (offline selection/cuts). One can enhance signal significance by using some special criteria that allow one to suppress background by a large factor while leaving the signal events relatively intact. For example, if background charged tracks are mostly pions, one can use electron/pion separation criteria (e.g. electromagnetic calorimeter). Let's assume that such criteria allow to cut pions by a factor of f=10, while remain ε=90% efficient to electrons/positrons. So statistics will be reduced, but with very different factors for background and signal.

- same histogram and assuming that signal was real
- the background would be $B_{new}= B_{old}\times f$=10 events,
- the excess would also decrease, $S_{new}=S_{old}\times\varepsilon$=22 events
- Then, signal significance per bin would be $S_{new}/\sqrt{B_{new}}= (S_{new}/\sqrt{B_{new}})\times(\varepsilon/\sqrt{f})=7\sigma$.

Note: once statistics becomes very small, one must not use $\sigma=\sqrt{N}$….

Trigger (online selection/cuts).

Often one is limited not by a number events that can be produced, but by the number of events one can record. Then, online selection/cuts (trigger conditions) can be applied to enhance the statistical significance of the signal being looked for. For instance, identification of electrons discussed above can and is often done online.

Signal in presence of background: Confidence Levels

One of the most popular ways of estimating confidence levels for observing or not observing a signal is based on so-called Bayes' theorem:

$$p(a \mid y) = \frac{L(y \mid a) \cdot \pi(a)}{\int L(y \mid a) \cdot \pi(a)\, da}, \text{ where}$$

$p(a|y)$—probability that theory's parameter is a, given we have a set of measurements y_i;
$L(y|a)$—Likelihood function of getting a set of measurements y_i, if the theory's parameter is a;

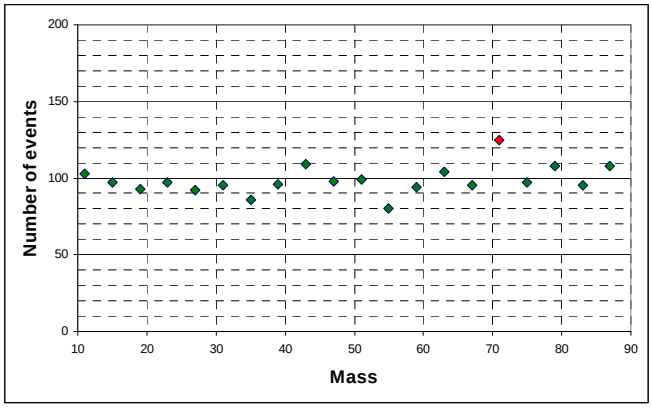
$\pi(a)$—a priory probability distribution function for the theoretical parameter a, which might be based on theoretical reasoning, practical considerations, or plain common sense… At the end, it always boils down to some a priori believes… For example, an a priori probability distribution function for signal rate can be naturally assumed to be the step-function: zero for negative values and uniformly distributed for positive values. However, what is flat in one parameterization, may not be flat in an another (e.g., one can assume that it is the matrix element that must have flat distribution; in this case the rate will be zero for negative values and NOT flat for positive values). Bayes' theorem shows this arbitrariness explicitly.

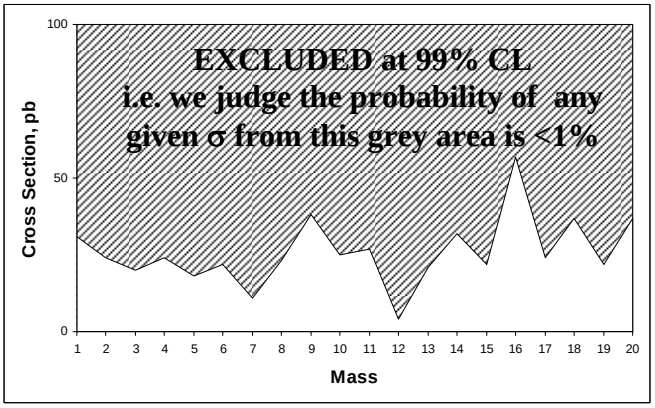
Using this way obtained distribution $p(a|y)$, one can exclude regions of parameter space with some predefined confidence level:

$$CL = 1 - \alpha, \quad \text{where} \quad \alpha = \frac{p(n \le n_0 \mid s+b)}{p(n \le n_0 \mid b)}$$

Popular confidence levels are CL=95% and 99%

Example:





The plot on the left shows a histogram of reconstructed invariant masses for positive-negative charged particles in reactions $p + p \rightarrow e^+ + e^- + anything$. Assume that experimental setup was such that, if resonances were to be produced at all, one would record on average 1 electron-positron pair per each 1 pb of the resonance production cross section.

The plot on the right shows the CL-contour (line in this case) of signal cross section being higher than the line. For calculating these limits, I used $\pi(\sigma)$=const for all values, including negative ones. Note that the line is the function of mass and the wiggling results from the actual numbers of observed counts.

Systematic errors (estimation of biases)

- biases due to theory (background level and/or shape, signal shape)
- biases due to event selection/cuts (either at trigger or offline levels)
- biases due to reconstruction and corrections (apparatus effects, why error function tails are so dangerous in new physics searches)
- biases due to the analysis methodology (e.g. ignoring correlations between errors)

Cross-checks:

A good data analysis presents a large number of crosschecks and auxiliary measurements to show that an experimenter understands what he/she is doing

Traps of wishful thinking (posteriori adjustments)

Histogram Binning: The choice of in width is usually based on the expected statistics of events and detector resolution; however, there are no strict rules. And there is always a freedom of shifting bins left and right. Although, a priori many of possible choices are equally valid, one finds that by tweaking them posteriori one can "enhance" the apparent statistical significance of a signal, especially in the cases of small number of events and marginal significance. Below are four histograms with the same bin width, but with different offsets. The data used are exactly the same set of points, generated to be randomly distributed with the density of 25 events per unit of Mass. One can see that, by shifting bins left-right, the accidental "peak" around *Mass*=70 can be tuned to vary from 25 to 12 over the average background of 100 ($S/\sqrt{B}$ is 1.2 to 2.5σ). Another "optimization" can be can be done by choosing how many bins are to be used for estimating the background. By using ±4 bins around the "peak" at M=70 in the second histogram, one can take advantage of statistical downward fluctuation around M=55. This choice would give the average Background=97.5, and, consequently, "peak" significance $S/\sqrt{B}=(125-97.5)/\sqrt{97.5}=2.8\sigma$. One can play the game further and pick the "optimal" bin width…

Selection Cuts: Similarly, "optimization" of event selection cuts will "enhance" the desired signal, if the optimization is based on promoting the significance of the signal posteriori, rather than on a priori physics considerations.

Dismissing "bad" data: Another trap: one can notice that removing a particular subset of data, say, taken on Mondays (or with crystal sample 1, or at the beginning of each data collection run, or anything else) makes "signal" more prominent. Typically, this prompts one to think what may have gone wrong on Mondays that lead to "bad" data, rather than to think what may have gone wrong on Tuesday-Fridays that lead to "too good" data. The errors of both types do happen, but such biases in thinking lead to finding real errors of the first kind more often. Sometimes, explanations may end up being merely plausible. Obviously, this may lead to biases toward "discovery".

That search is just one of many: There are many on-going searches, 100s, all proceeding at the same time and coming up for publications every year. If one chooses 99% CL of observing signal as a sufficiently convincing criterion, then, he/she should not be surprised to see a few “breakthroughs” every year…

Solutions (if you do analysis)

- Binning: When at risk (typically, when you expect to have or actually have small statistics), use methods devised for unbinned data analysis…

- Significance evaluation: Never use $S/\sqrt{B}$ for small statistics—use Poisson probabilities. In general, make the best effort to find the correct error distribution functions. Presence of systematic errors may drastically effect the CL calculations.

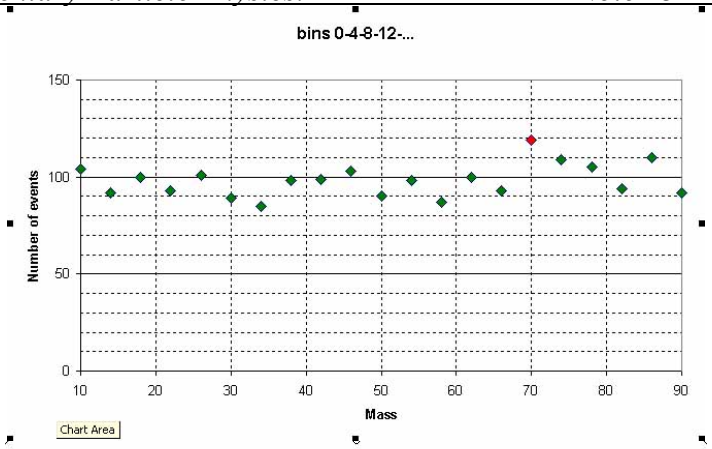
- Selection Cuts: To optimize the cuts, use a priori considerations, Monte Carlo generated events and, if absolutely needed, only a small fraction of data (e.g., 20%); apply the optimized cuts to the rest of the data (no further tuning of cuts is allowed after opening the "box" with the remaining data); the results should include the fraction of data used for cut optimization.

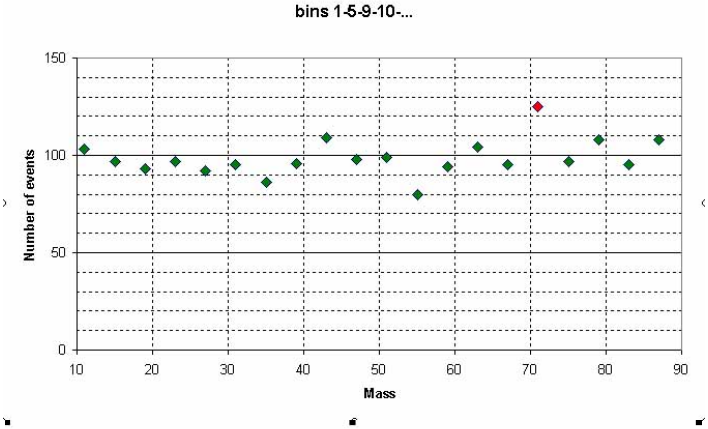
- Dismissing "bad" data: No recipe… Be aware…

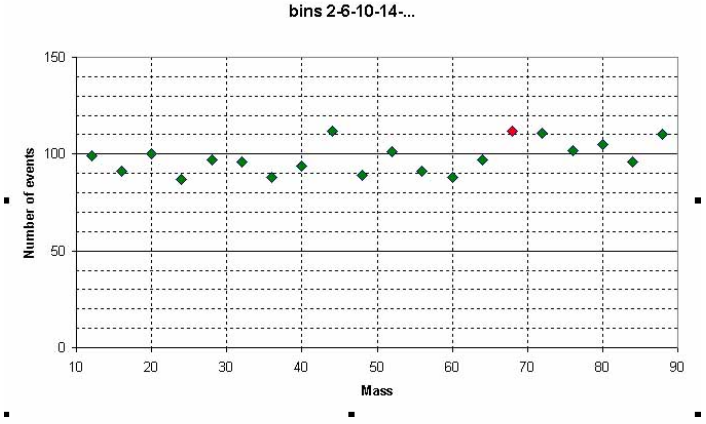
Rules of thumb:

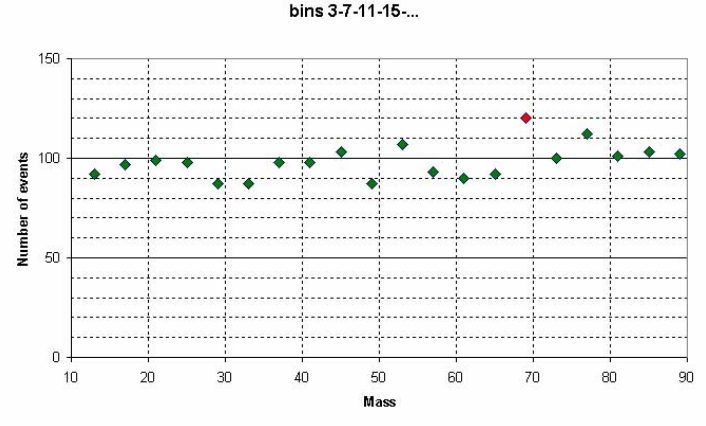
3σ--might be a real thing or might be a statistical fluke, worth publishing, do NOT claim a discovery, more data and/or independent experiments are needed…

5σ--time to get serious, independent experiments are needed…


bins 0-4-8-12-...
Number of events
Mass
Chart Area


bins 1-5-9-10-...
Number of events
Mass


bins 2-6-10-14-...
Number of events
Mass


bins 3-7-11-15-...
Number of events
Mass

Examples of low statistic "discoveries":
"Double-β neutrinoless decay" by Heidelberg-Moscow Experiment

$^{76}Ge \rightarrow {}^{76}Se + 2e^-$
This implies that neutrino is its own antiparticle, a la photon…
Energy of two electrons is known: Q = M(^{76}Ge) - M(^{76}Se) = 2039.00±0.05 keV

Paper of January 2001 claimed the discovery of neutrinoless double-β decay…
A good fraction of the collaboration did not sign the paper…

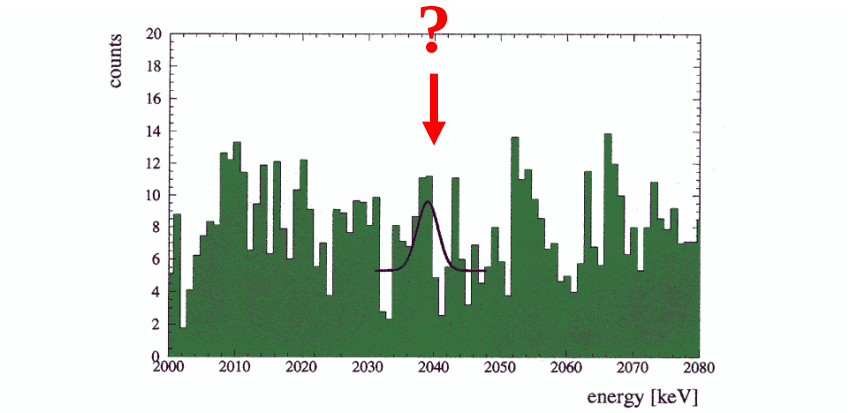
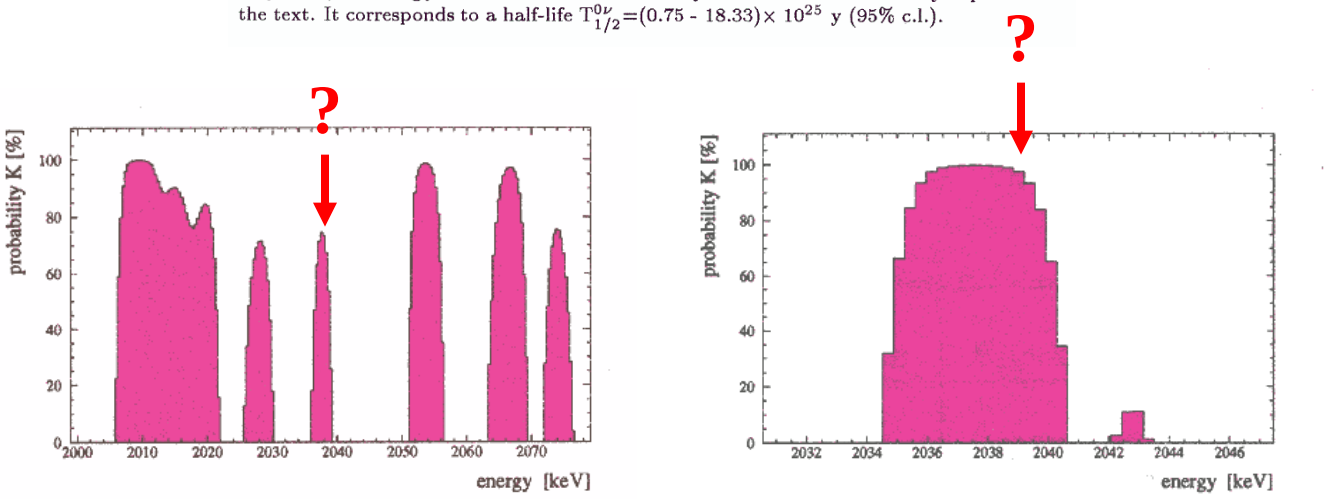
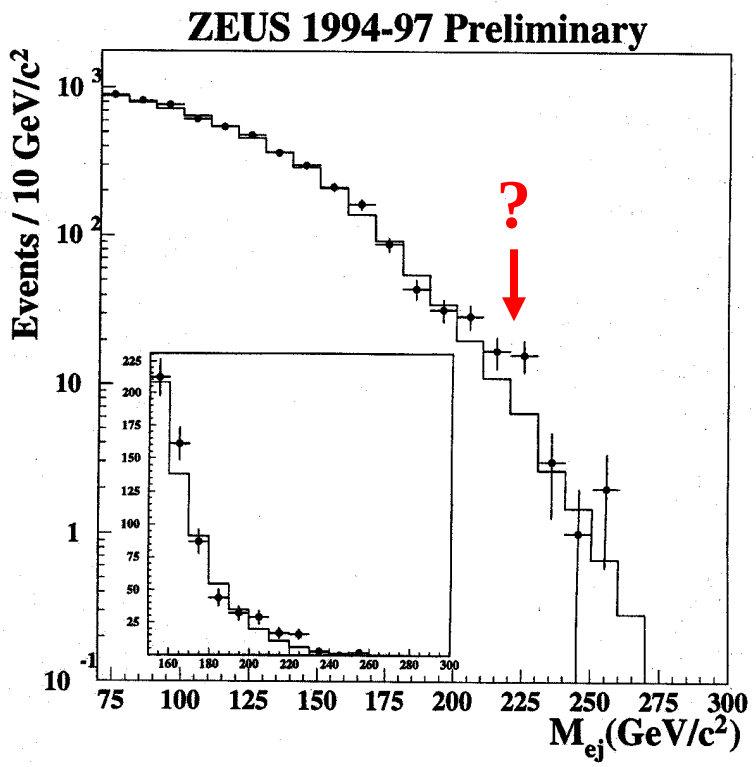


Figure 2. Sum spectrum of the ^{76}Ge detectors Nr. 1,2,3,5 over the period August 1990 to May 2000, 46.502 kg y. The curve results from Bayesian inference in the way explained in the text. It corresponds to a half-life $T^{0\nu}_{1/2}=(0.75 - 18.33)\times 10^{25}$ y (95% c.l.).



Large window scan gives CL~70% for observing a non-zero signal at Q=2039 keV
Smaller ("optimized"?) window scan gives CL~97% for observing a non-zero signal

Examples of low statistic "discoveries":
Signs of "lepto-quark" at HERA?

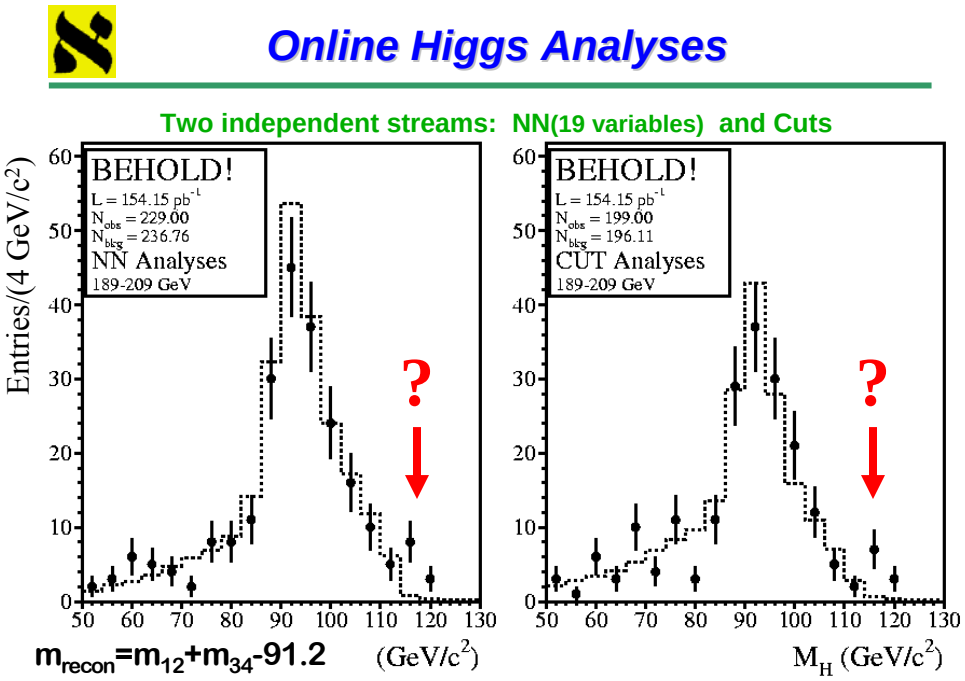


Examples of low statistic "discoveries": Higgs at LEP?

One does not know where it is, but, if it is there to be observed, it must be at the very tail (otherwise, it would have been seen before)…

2000: **ALEPH: ~4σ for Higgs signal present**
All four collaborations combined: ~2.9σ

2002: **More thorough re-analysis of the same data:**
ALEPH: ~3σ
All four collaborations: <2σ,
or, by inverting logic, there is no Higgs with M_H<114 GeV at 95% CL.



Examples of low statistic "discoveries":
Top quark at Tevatron—real thing that started from 2.8σ (99.7% CL)

Debate at CDF over the title key word: Discovery? Study? Evidence? Observation? Search for? The jokes were: Evidence for Study… Observation of Search for…

The final compromise was
"Evidence for top quark production in p-bar p collisions at sqrt(s) =1.8 TeV"

Abstract: …*The probability that the measured yield is consistent with the background is 0.26%. Though the statistics are too limited to establish firmly the existence of the top quark, a natural interpretation of the excess is that it is due to tt-bar production….*

Subsequent papers based on much larger statistics confirmed the signal and were titled
"Discovery of..."

Appendix: Deriving Poisson distribution and its parameters

- Assume average rate of μ events per second.
- Large time interval T: expect average **n=μT**
- Calculate probabilities to get none, one, two, three, etc. events:
  - The time interval can be broken in M=T/dt small intervals
  - probability to get one event during dt: $p_1=\mu dt=\mu T/M$
  - probability of getting more than 1 is vanishing in comparison to p_1 at $M\to\infty$
  - probability to get no events during very short time dt: $p_0=1-\mu dt$
  - probability to get no events during T: $P_0=p_0^M=(1-\mu dt)^M=(1-\mu T/M)^M \to e^{-\mu T}= e^{-n}$
  - probability to get 1 event: $P_1=C(M,1)\cdot p_1\cdot p_0^{(M-1)}= M\cdot(\mu T/M)\cdot p_0^M/p_0 \to ne^{-n}$
  - probability to get k event: $P_k=C(M,k)\cdot(p_1)^k\cdot p_0^{(M-k)}= \ldots. \to$

$$P_k = \frac{n^k}{k!}e^{-n}$$

  - Cross-check: Average $= 0\cdot P_0 + 1\cdot P_1 + \ldots + k\cdot P_k + \ldots$
    $= \Sigma\, k\cdot(n^k/k!)\cdot e^{-n} = n\cdot e^{-n}\Sigma\, n^{k-1}/(k-1)! = n\cdot e^{-n}e^{n}$
    $= n$

- RMS (root of mean squared) $= \sqrt{\left\langle (k-n)^2 \right\rangle} = \sqrt{n}$

$$\left\langle (k-n)^2 \right\rangle = \left\langle k^2 - 2kn + n^2 \right\rangle = \left\langle k^2 \right\rangle - 2n\left\langle k \right\rangle + n^2 = \left\langle k^2 \right\rangle - n^2$$

$$\left\langle k^2 \right\rangle = \sum_{k=0}^{\infty} k^2 P_k = \sum_{k=0}^{\infty} k^2 \frac{n^k}{k!}e^{-n} = \sum_{k=1}^{\infty} k \frac{n^{k-1}}{(k-1)!}ne^{-n}$$

$$= ne^{-n}\sum_{k=1}^{\infty}(k-1+1)\frac{n^{k-1}}{(k-1)!} = ne^{-n}\sum_{k=1}^{\infty}(k-1)\frac{n^{k-1}}{(k-1)!} + ne^{-n}\sum_{k=1}^{\infty}\frac{n^{k-1}}{(k-1)!}$$

$$= ne^{-n}\sum_{k=2}^{\infty} n\frac{n^{k-2}}{(k-2)!} + ne^{-n}\sum_{k=1}^{\infty}\frac{n^{k-1}}{(k-1)!}$$

$$= n^2e^{-n}\sum_{k=0}^{\infty}\frac{n^k}{k!} + ne^{-n}\sum_{k=0}^{\infty}\frac{n^k}{k!} = n^2 + n$$

$$\left\langle (k-n)^2 \right\rangle = \left\langle k^2 \right\rangle - n^2 = n^2 + n - n^2 = n$$