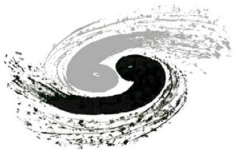
粒子物理中的统计分析简介

陈明水



中国科学院高能物理研究所

iSTEP2019，华南师范大学

What is statistics?

- **Statistics** is a branch of mathematics dealing with the collection, organization, analysis, interpretation, and presentation of data. Wikipedia

AI is actually statistics

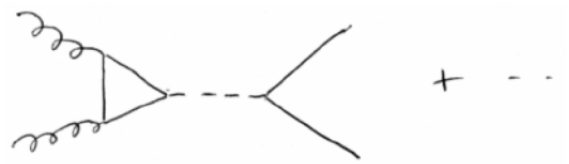
- “Artificial intelligence is actually statistics, but with a very gorgeous phrase, in fact, is statistics. Many of the formulas are very old, but we say that all artificial intelligence uses statistics to solve problems. ”

  ---- Thomas J. Sargent, winner of the 2011 Nobel Prize in Economics

Theory ↔ Statistics ↔ Experiment

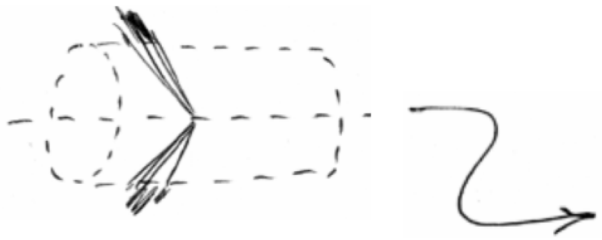
Theory (model, hypothesis):

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + i \bar{\psi} \not{D} \psi + \cdots$$

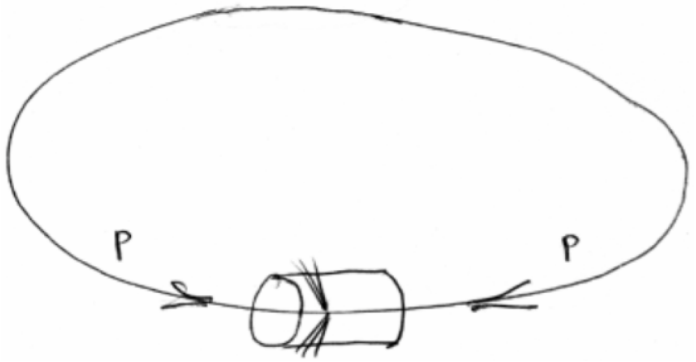


$$\sigma = \frac{G_F \alpha_s^2 m_H^2}{288 \sqrt{2\pi}} \times \sim$$

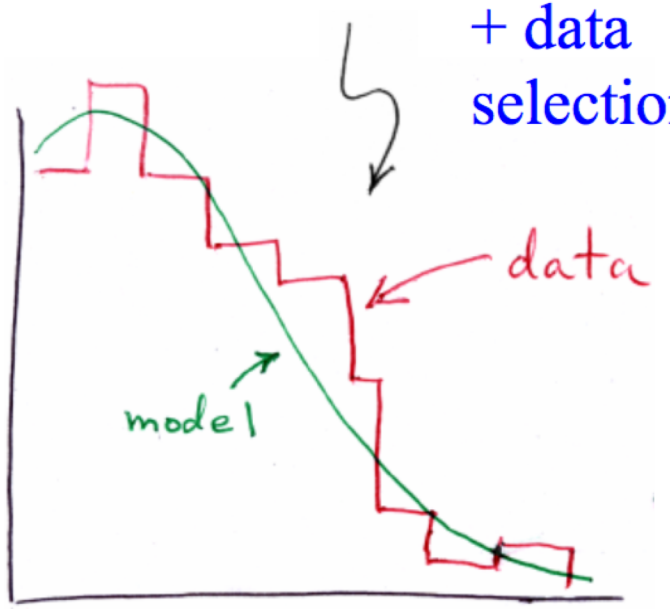
+ simulation of detector and cuts



Experiment:



+ data selection



Outline

- Probability distribution function
- Parameter estimate
- Significance and limit
- Systematic uncertainties and examples

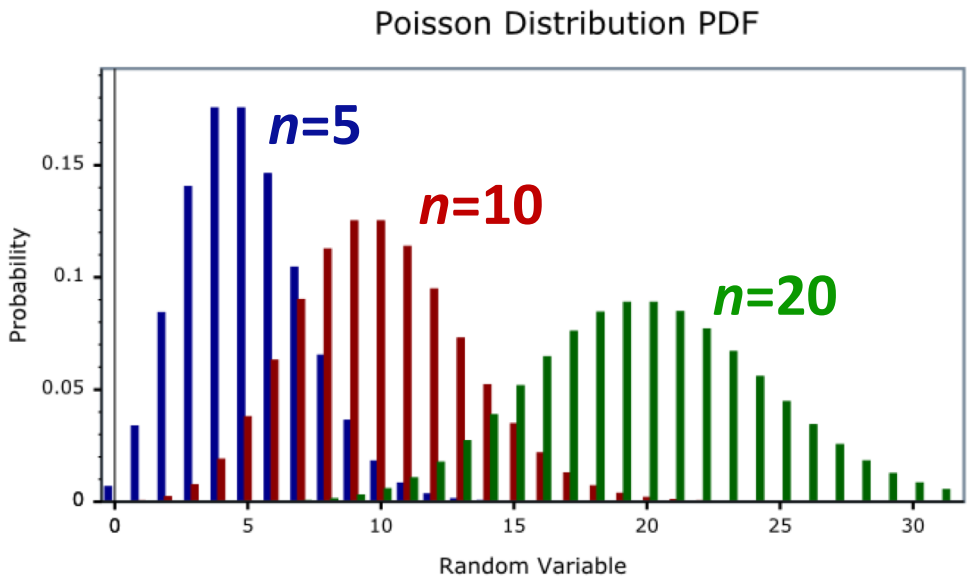
一些参考书

- 朱永生, 实验数据分析(上、下册), 科学出版社, 2012
- G. Cowan, *Statistical Data Analysis*, Clarendon, Oxford, 1998
- Ilya Narsky and Frank C. Porter, *Statistical Analysis Techniques in Particle Physics*, Wiley, 2014
- K.A. Olive et al. (Particle Data Group), *Review of Particle Physics*, Chin. Phys. C, 38, 090001 (2014).; see also **pdg.lbl.gov** sections on probability, statistics, Monte Carlo

泊松分布 (Poisson distribution)

- 泊松分布: 描述随机(互相独立)事件发生的频率(ν)的概率分布。因此单位时间(Δt)内，事件发生的平均次数的**期望值**为$n=\nu\cdot\Delta t$.
- 实际上每一次实验中探测到的事例数 k 是不一样的:
  - 探测到k个事例的概率为$P_k(n)$:

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



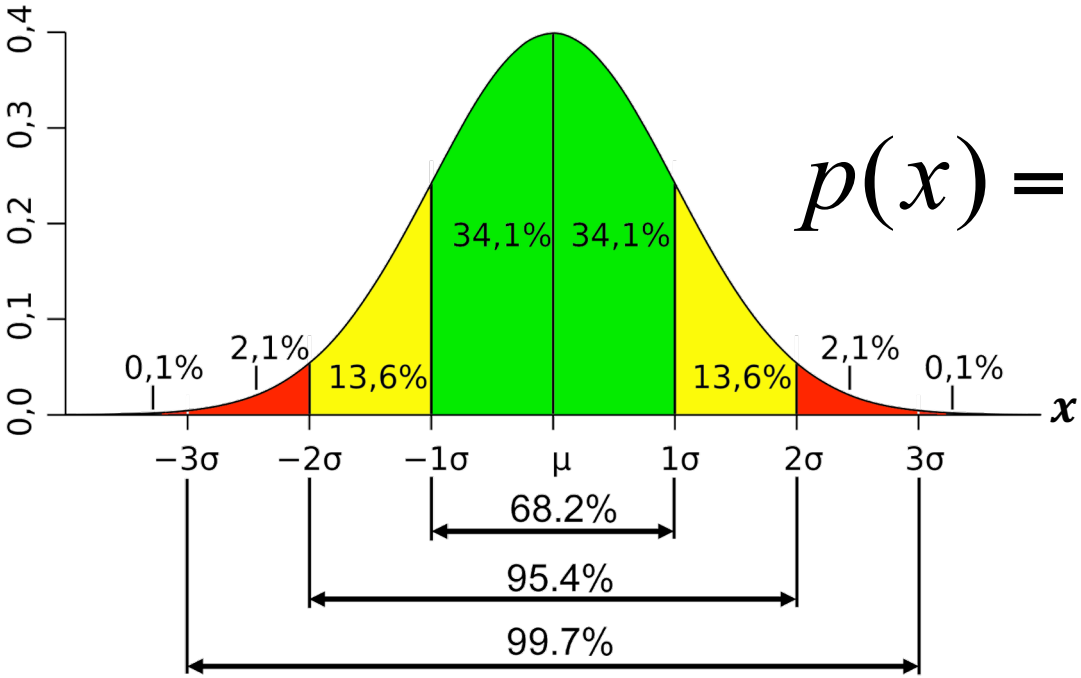
方差:

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

高斯分布 (Gaussian distribution)

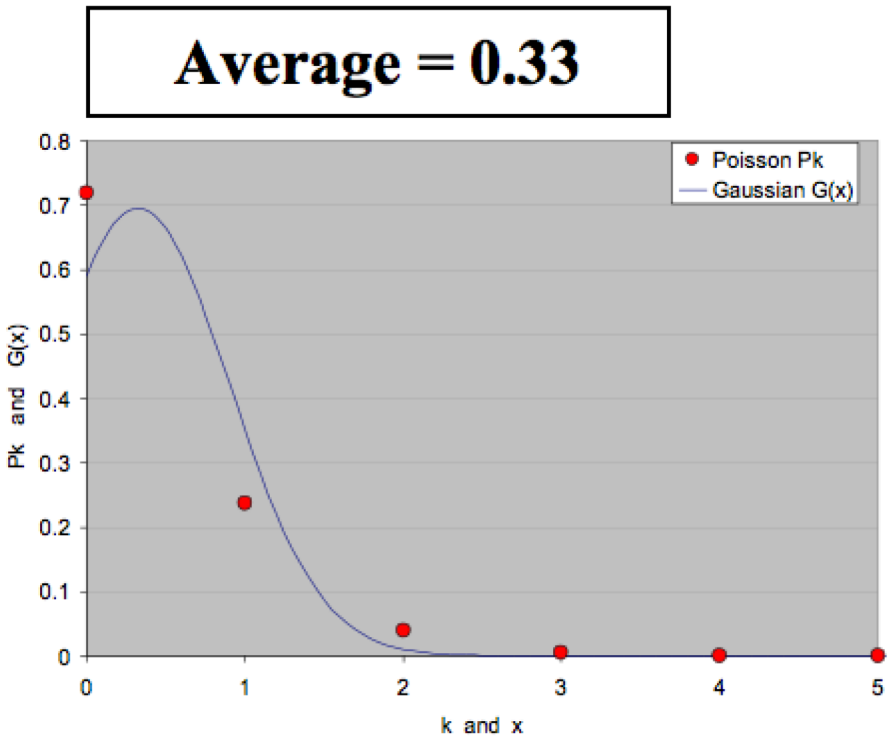
- 对于许多典型的测量误差，高斯分布是较好的近似。
- 随机量x的测量值落于[x_1, x_2]的几率 $P = \int_{x_1}^{x_2} p(x)dx$

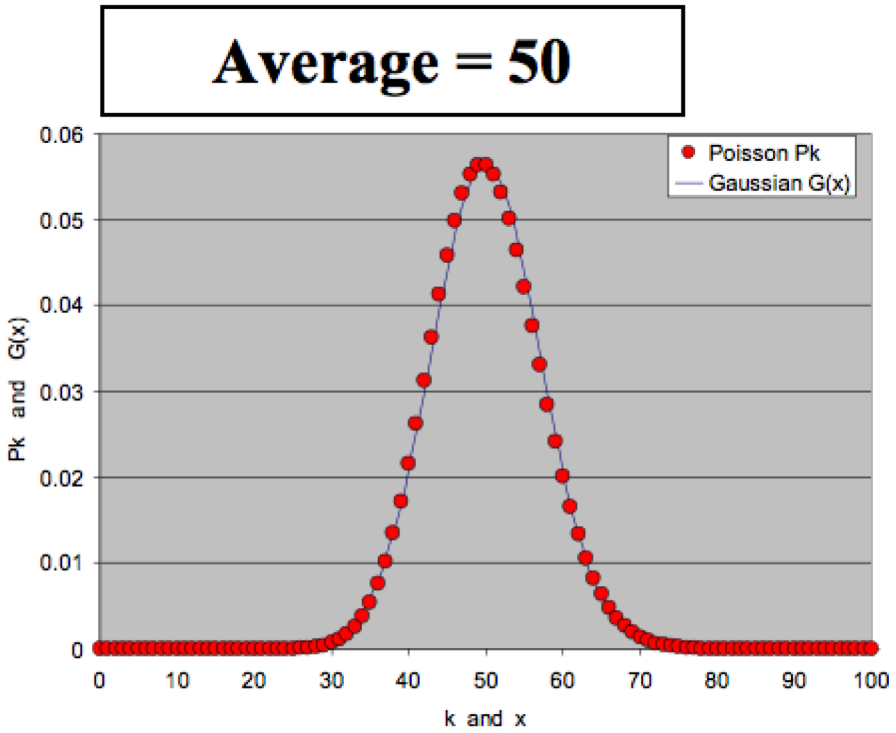
  其中p(x) 是概率密度函数:



$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Poisson -> Gaussian distribution





期望值为n(远大于1时)的泊松分布趋近于μ=n, σ^2=n的高斯分布

中心极限定理(Central Limit Theorem)

- 中心极限定理:对于 n 个互相独立的符合任意的概率分布函数的变量 $x_1,\ldots\ldots, x_n$ (要求各自的平均值 μ_i 和方差 σ_i^2 是有限的),它们的和 $X=\sum x_i$ 在 $n\to\infty$ 时符合平均值为 $\sum\mu_i$，方差为 $\sum\sigma_i^2$ 的高斯分布。
  - 在适当的条件下，大量相互独立随机变量的均值经适当标准化后依分布收敛于正态分布。这组定理是数理统计学和误差分析的理论基础，指出了大量随机变量之和近似服从正态分布的条件

误差传递

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***,

$$\sigma_m^2 = (f_x \cdot \sigma_x)^2 + (f_y \cdot \sigma_y)^2$$

加权平均值

- 假定自由变量x的两个测量值 x_1 and x_2 的误差分别是 σ_1 and σ_2。 那么通过加权平均我们可以得到x的最佳估计值及其误差:

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

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

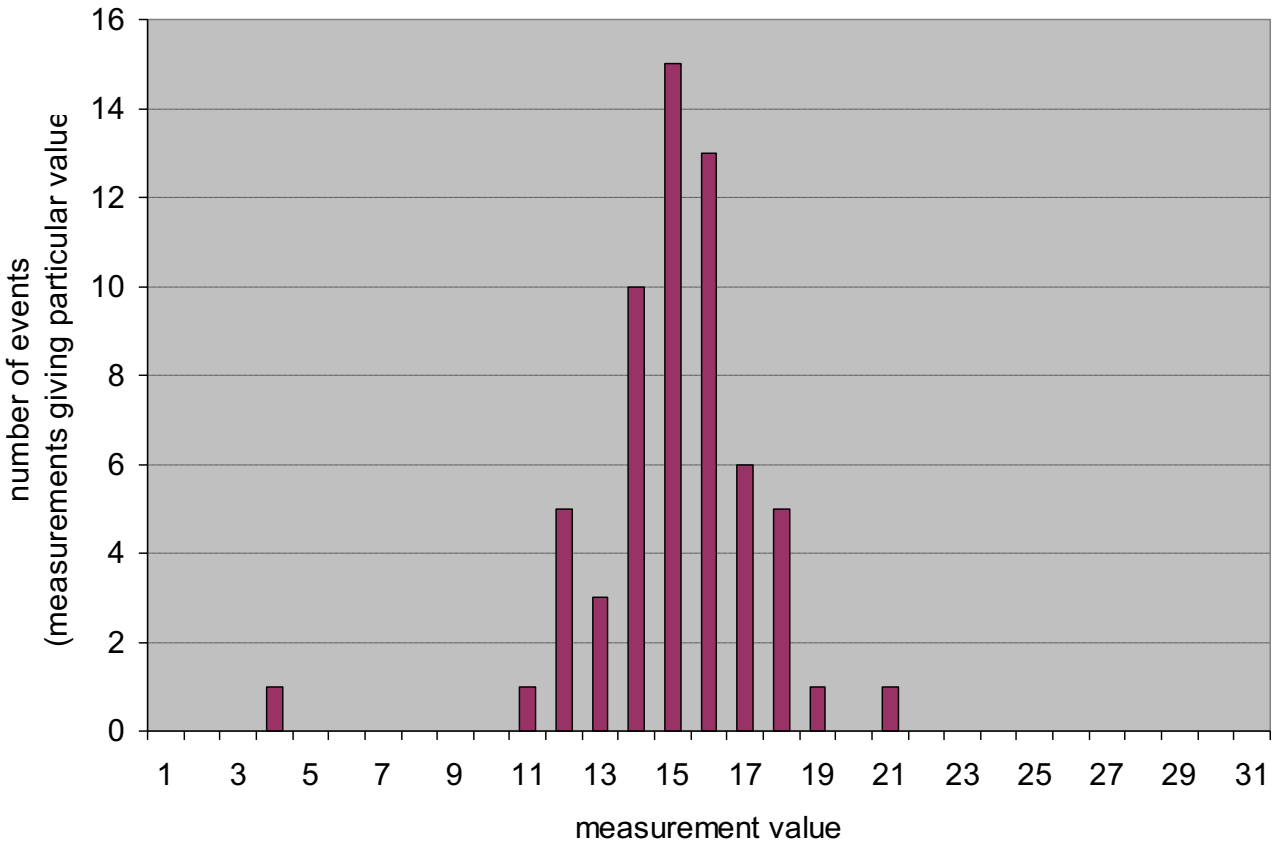
→ 误差相对很大的测量可以被忽略，因为它权重很小，很难对最终估计值产生影响，也不会改善测量精度(误差)

→ 两个一样精度的测量权重一样，加权合并后的误差是单个测量误差的1/√2

参数估计

参数估计

- 给定一组有限的测量值，
  - →估计概率密度分布函数的参数(e.g., mean, width, …)，
  - →同时评估计算这些参数的误差



参数估计

- 假设真实的概率分布的mean=x_0，方差D=σ_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 Gauss distribution and large N)

数据和理论：理论参数的最佳估计

需要回答的一些问题：

- 理论是否与数据相符？
- 理论参数的最佳估计是什么？
- 这些参数估计值的误差是多少？
- 有没有任何显示实验数据不自洽的迹象？

最大似然法(Max Likelihood Method)

一般的例子

- 数据：在x_i位置的测量值y_i的集合，以及
  - 已知$f_i(y_i|y)$分布函数:
    即当真值为y时，测量值为y_i的概率
  - 并且各位置点之间**没有关联**
- 理论：参数为a的理论预测$y=F(x, a)$

→对于给定参数值a，实验上获得一套特定的测量值y_I的概率为：

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

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

$$= \boxed{L(y_i \mid a)} \prod_i dy_i$$

$L(y_i|a)$ - 似然函数
通过最大化似然函数数值来确定最有可能的理论参数值

最小二乘法(Minimum χ^2 Method)

测量值为高斯随机变量时

- Maximum Likelihood method 与 Minimum χ^2 method等效:

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

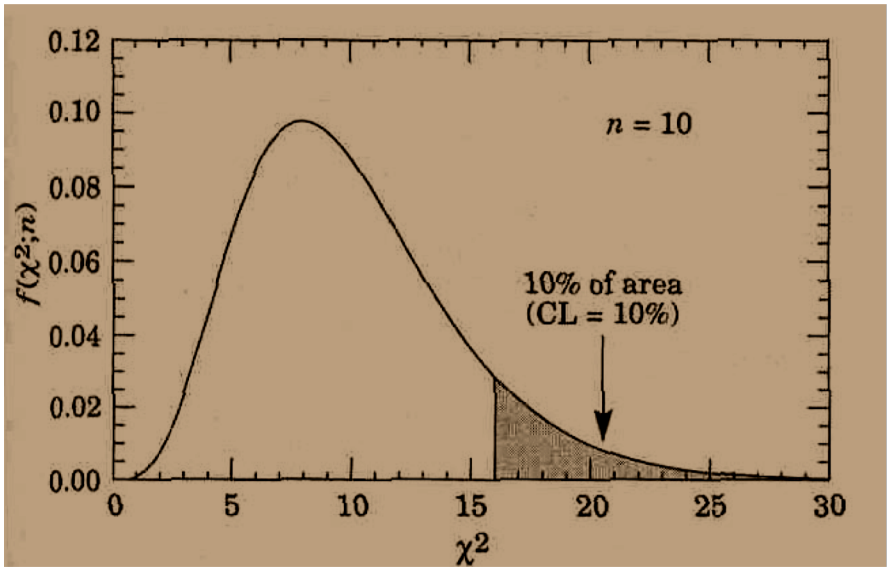
$$= \sum_i \ln \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(y_i - F(x_i,a))^2}{2\sigma_i^2}}$$

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

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

Statistical expectations for χ^2

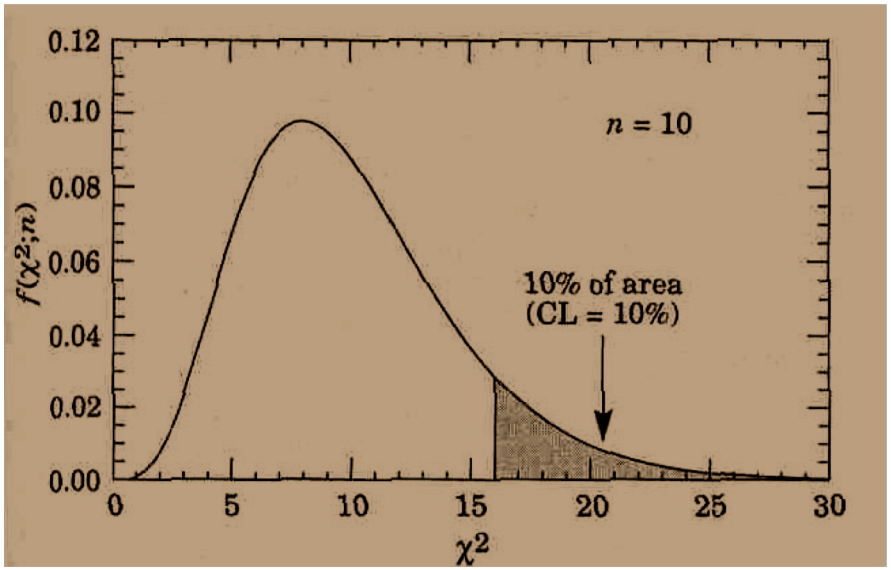
$$\overline{\chi^2} = n_{measurements} - k_{parameters} = n.d.f. \qquad \text{(number of degrees of freedom)}$$



- 服从χ^2分布的随机变量的期望值等于自由度的数目n (即数据点的数目减去独立参数的数目)
- 通常也用χ^2除以自由度的数目n来衡量拟合的好坏
- 显著水平(p-值)表示当前假设将导致相比于实际数据更差(即更大的)χ^2值的概率

What if you get something very different

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



- 如果χ^2/n远小于1，则在给定的测量误差条件下，拟合好于预期，虽然不见得有问题，但是这时通常需要仔细检查误差是否被高估，或者是否存在正关联现象(系统误差)
- 如果χ^2/n远大于1，则有理由怀疑假设的正确性，通常先检查是否低估误差，或者是否存在很大的负关联(系统误差)
- 通过其它cross-checks找出“隐藏”的系统误差

参数值的误差估计

- 利用χ^2估计参数值的误差 （自由度为1时）
  - $a \rightarrow a \pm \sigma_a$ 使得 $\chi^2 \rightarrow \chi^2+1$

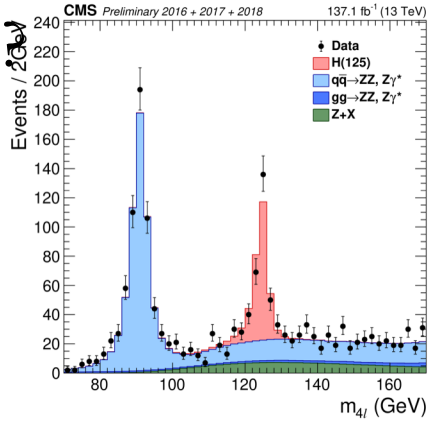
不适用最小二乘法的情况

- 下面这些情况最小二乘法不适用:
  - 随机变量不符合高斯分布, e.g.:
    - 带有长尾巴的高斯分布
    - 小统计量(必须使用Poisson errors)
  - 各测量点之间有关联时:
    - 当然可以适当修改Max Likelihood and Min χ^2 Methods

Significance and limit

有本底情况下的信号统计显著性

- 本底事例期望值为 b，观测到 n_0 个事例,且 $n_0 > b$
  - 这个实验观测的显著性(significance)是多少？
  - 是否发现了一个新效应(导致上述超出)?
  - 或者这个超出仅仅是本底的统计涨落?



- 显著性 S：假设仅期待本底事例数为 b 的条件下，由于统计涨落观测到不少于当前观测事例数 n_0 的概率。

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

通常也把统计涨落的概率与高斯分布的标准偏差对应起来：

significance	1	2	3	4	5
probability (p-value)	16%	2.3%	0.14%	3×10^{-5}	3×10^{-7}

估计信号显著性的简单方法：S_1

- 对于大统计量N:

$$S_1 = \frac{signal}{\sqrt{bkgd}} = \frac{n_{observed} - b}{\sqrt{b}} = \frac{s}{\sqrt{b}}$$

  - 很常用的估计方式，但是对于小统计量的情况，所得结果与正确结果相差较大
  - 比如 b<100时， 这个方法算出的结果太大(高估信号显著性)

估计信号显著性的简单方法：S_{cL}

- 最好的简单方法

$$S_{cL} = \sqrt{2n_0 \ln(1 + s/b) - 2s}$$

- 源于比较分别基于s+b和b-only两种假设情况下观测到 n_0 事例数的概率，即似然比(likelihood ratio):

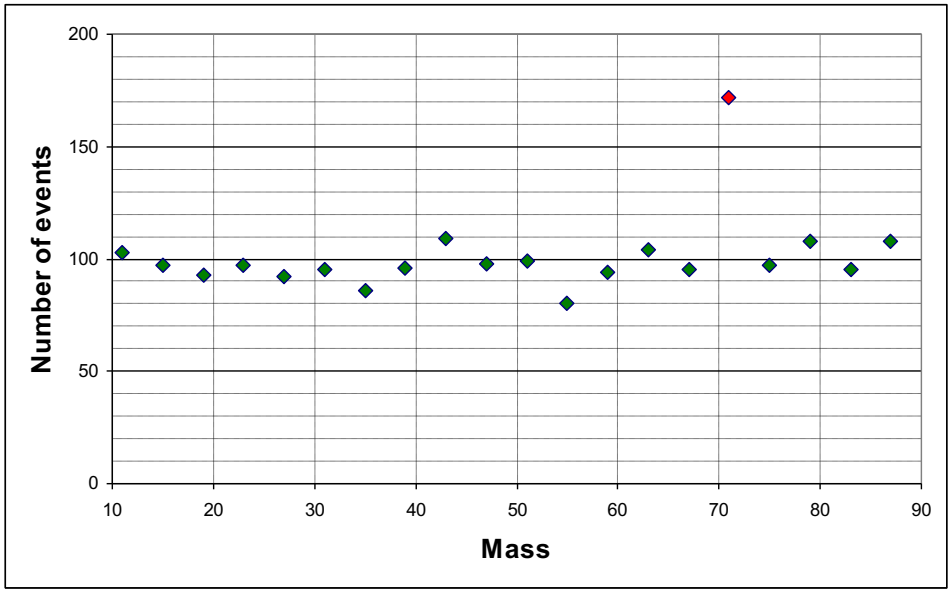
$$S_{cL} = \sqrt{2\ln Q}, \text{ where } Q = \frac{p(n_0 \mid s+b)}{p(n_0 \mid b)}$$

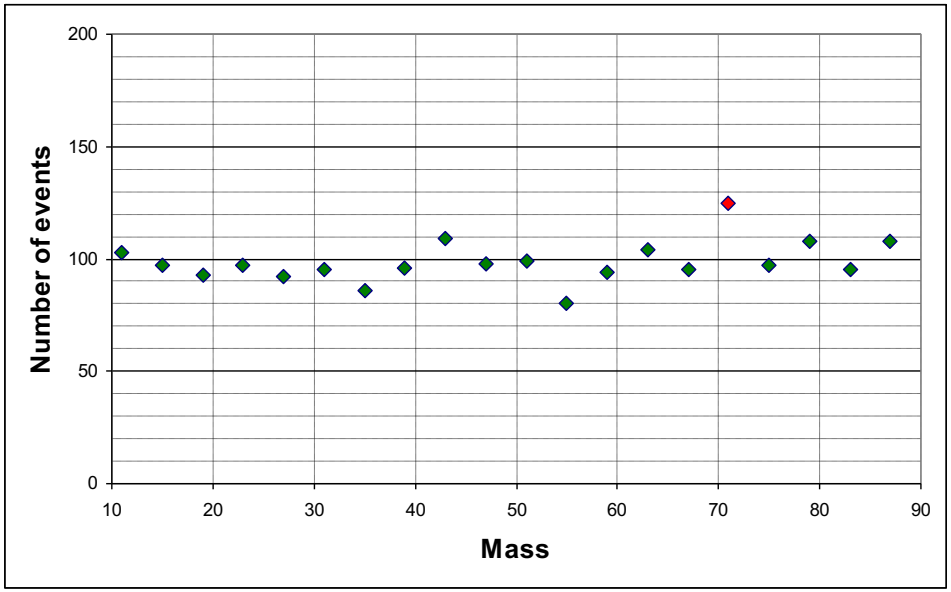
$$Q = \frac{p(n \mid s+b)}{p(n \mid b)} = \frac{\frac{(s+b)^n}{n!}e^{-(s+b)}}{\frac{(b)^n}{n!}e^{-b}} = \frac{(s+b)^n}{(b)^n}e^{-s}$$

$$\ln Q = \ln[\frac{(s+b)^n}{(b)^n}e^{-s}] = n\ln\frac{s+b}{b} - s = n\ln(1+s/b) - s$$

- 所得结果通常与真实显著性非常接近，即使是非常小的统计量，结果也基本不会偏离超过0.2σ

例子





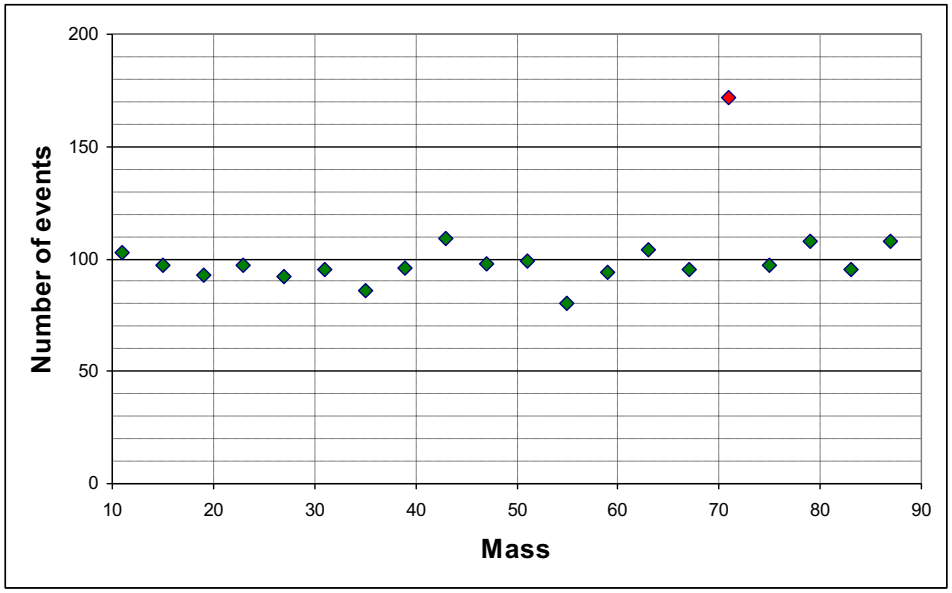
pp对撞过程中正负电子对的不变质量谱

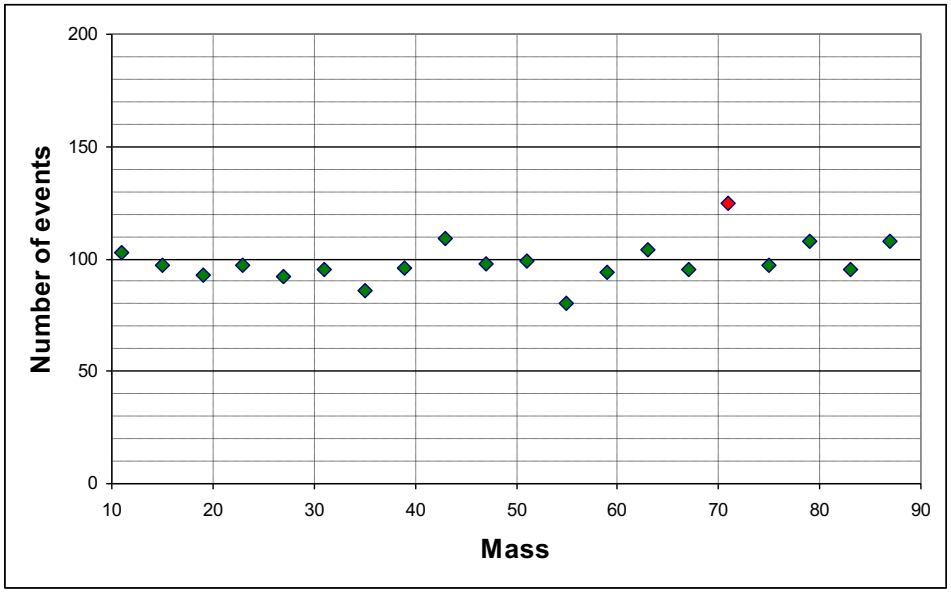
$(p + p \rightarrow e^+ + e^- + anything)$

左右两图中在质量为71的位置上红色数据点的统计显著性分别是多少？

首先进行本底估计：利用不包含M=71的其它bins来拟合估计本底事例率为b=100，假设M=71的bin的本底跟其它bin一样,统计误差σ=√(100)=√(B)=10

局域显著性(Local Significance)





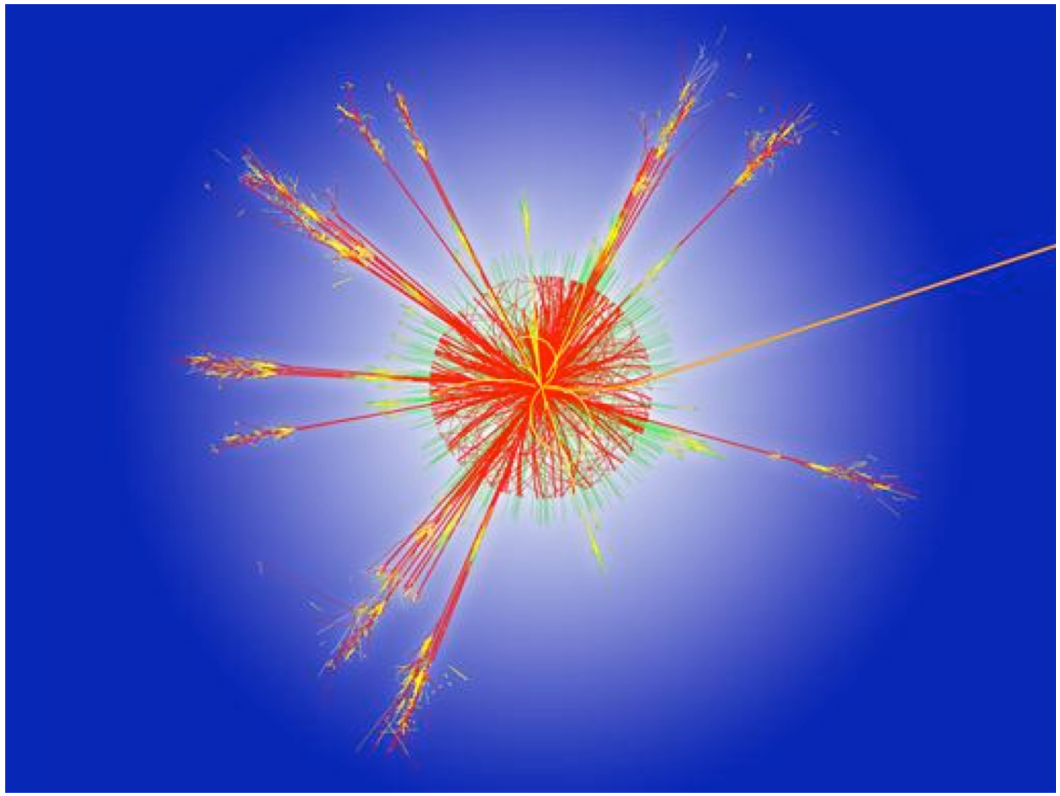
- 假定我们预先知道共振态质量的预测值为71，并且衰变宽度很窄远小于直方图中的bin宽=4。
- 左图超出是S=172-100=72, S/√(B)=7.2σ；右图超出是S=125-100=25, S/√(B)=2.5σ。
- 由统计涨落造成such upward fluctuations的概率分别是<10^{-12} 和 0.6%，二者都很小，有足够把握声明发现了所预言的共振态。

全局显著性(Global Significance)

- 如果实验寻找前我们不知道共振态的质量,那么这超出的显著性会很不一样
- 需要考虑总共有互相独立的20个bin，其中至少有一个bin向上波动至所观测的峰值的概率要大于预先指定位置的概率
  - Probabilities of none of the bin with flat background fluctuating upward as shown is $(1-p)^{20}$
  - Therefor, probability of at least one bin fluctuating upward is $1-(1-p)^{20}$, which gives ~10^{-11} and 12%.
- 对应右图的信号统计显著性就没那么强烈了

信号排除限(Exclusion Limits)

- 我们在LHC上寻找黑洞，努力了好几年结果什么都没找到。这种情况下应该怎么总结寻找的结果呢?



信号排除限(Exclusion Limits)

- 我们在LHC上寻找黑洞，努力了好几年结果什么都没找到。这种情况下应该怎么总结寻找的结果呢?
  - “找黑洞失败了”感觉有点负面。
  - 一般换种说法：基于实验数据，我们有99%的信心认为如下结论是对的：如果LHC上可以产生黑洞，它的产生截面小于XXX fb 。
  - 99% 置信度(confidence level) 意味着允许1%的概率上述结论是错的。

信号排除限(Exclusion Limits)

- 类似于显著性的定义，我们也可以构造一个概率计算公式(probability of observing no more than n_0 events, in assumption that the signal was s):

$$p(n \le n_0 \,|\, b+s) = \sum_{0}^{n_0} p(k \,|\, b+s)$$

- “对于给定的信号期望值s，如果这个概率小于α，那么大于等于这样强度的信号可以被排除。”这个定义有个逻辑缺陷：如果我们不幸观测到比本底期望值b更少的事例数，就可能排除s=0的情况。

信号排除限：CLs

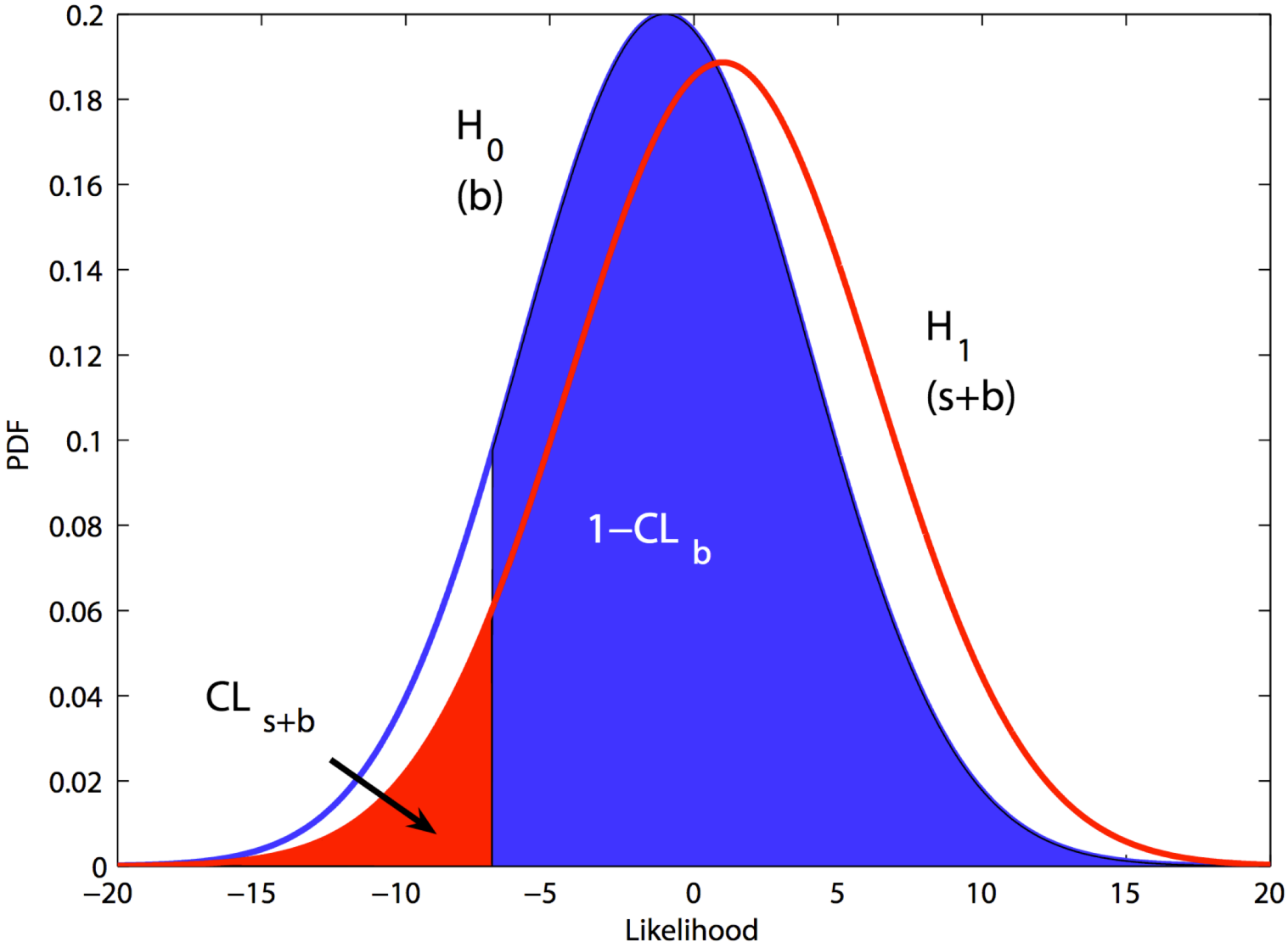
- 实验上通常是构造两个概率的比值，即有信号假设和零假设情况下，分别观测到不多于 n_0 事例数的概率比值：

$$r = \frac{\sum_{n=0}^{n_0} p(n|b+s)}{\sum_{n=0}^{n_0} p(n|b)} = \frac{\sum_{n=0}^{n_0} \frac{(b+s)^{n_0}}{n_0!} e^{-(b+s)}}{\sum_{n=0}^{n_0} \frac{b^{n_0}}{n_0!} e^{-b}}$$

- 其中 $CL_{b+s} = \sum_{n=0}^{n_0} p(n|b+s), \quad CL_b = \sum_{n=0}^{n_0} p(n|b), \quad CL_s = \frac{CL_{b+s}}{CL_b} = r$

信号 s≥0 时，CL_s值范围为 (0 , 1].

信号排除限：CLs


0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
PDF
−20
−15
−10
−5
0
5
10
15
20
Likelihood
H0
(b)
H1
(s+b)
1−CLb
CLs+b

信号排除限：Bayesian

- 另外一种方法是基于贝叶斯定理

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

  - ***p(a|y)* 是验后概率**：基于实验观测结果 *y*，理论参数是 *a* (例如黑洞产生截面)的概率
  - ***L(y|a)*** 是理论参数为 *a* 时，得到实验观测结果 *y* 的**似然函数**
  - ***π(a)*** 是理论参数 *a* 的**验前概率密度函数**
- 按照预先指定犯错概率(α)可以排除参数空间的区域

$$\int_{a_x}^{+\infty} p(a \mid y)\ da\ =\ \alpha$$

信号排除限：Bayesian

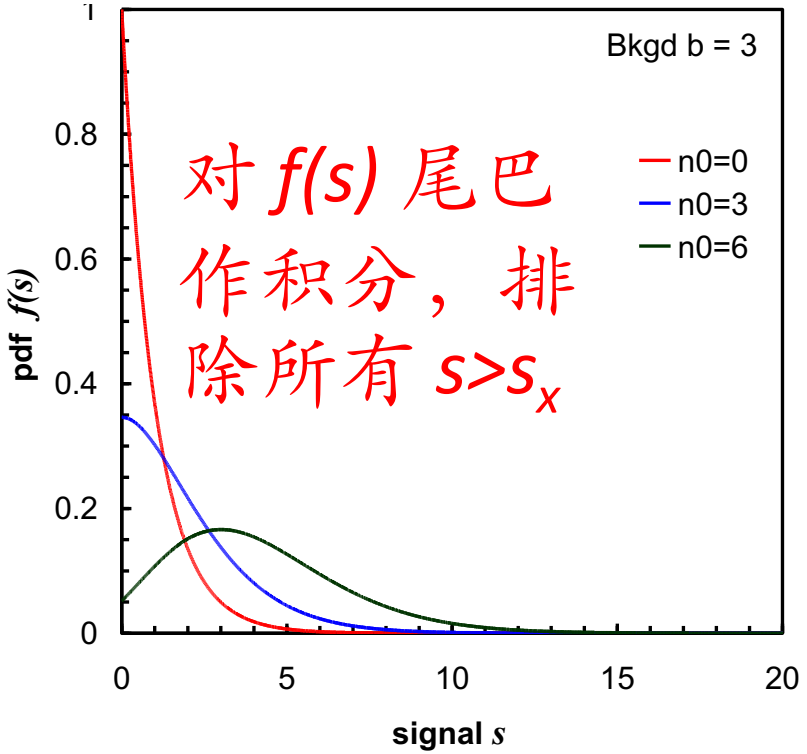
- 假设信号强度 s 的验前概率分布是负值为0、正值为均匀分布的step-function

$$f(s) = p(s \mid b, n_0) = \frac{p(n_0 \mid b+s) \cdot \pi(s)}{\int_0^{+\infty} p(n_0 \mid b+s) \cdot \pi(s) \cdot ds}$$

$$= \frac{p(n_0 \mid b+s)}{\int_0^{+\infty} p(n_0 \mid b+s) \cdot ds} = \frac{\frac{(b+s)^{n_0}}{n_0!} e^{-(b+s)}}{\sum_{n=0}^{n_0} \frac{b^n}{n!} e^{-b}}$$

信号排除限：Bayesian

- 假设信号强度 s 的验前概率分布是负值为0、正值为均匀分布的step-function



$$\int_{s_x}^{+\infty} f(s)\,ds = \frac{\sum_{n=0}^{n_0} \frac{(b+s_x)^n}{n!} e^{-(b+s_x)}}{\sum_{n=0}^{n_0} \frac{b^n}{n!} e^{-b}} = \alpha$$

由于信号强度大于 s_x 的概率 α 很小 (通常选1% or 5%)，因此可以以 1-α 的置信水平 (99% or 95%)排除 $s>s_x$ 的信号可能性。

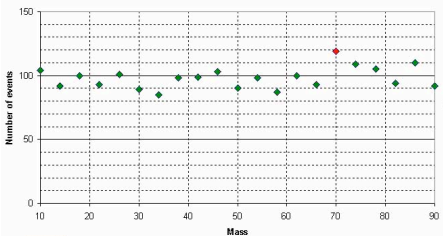
Systematic uncertainties and examples

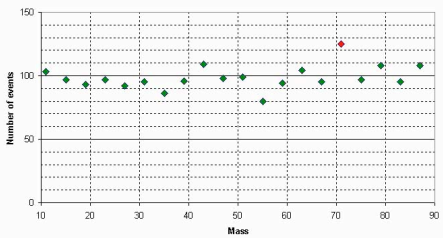
系统误差(Systematic errors, estimation of biases)

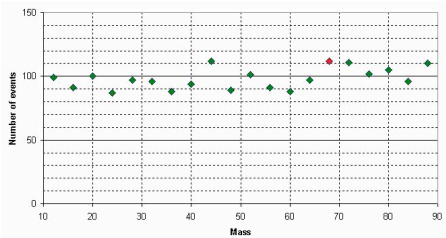
- Biases due to theory (background level and/or shape, signal shape)
- Biases due to event selection/cuts (either at trigger or offline levels)
- Biases due to reconstruction and corrections (apparatus effects, why error function tails are so dangerous in new physics searches)
- Biases due to the analysis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

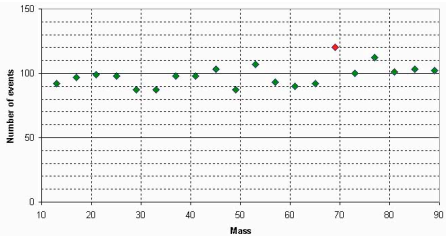
一些注意事项(posteriori adjustments)

- **直方图分区(Histogram Binning):** 通常是基于预期的事例数和探测器分辨率等，但是没有严格的规则。并且可以自由向左或右移动分区。只要是先验(a priori)的，大部分可能的分区选择都是同等有效的。
- **问题在于有些人发现验后调整分区可以“增强”信号的表面的统计显著性，特别是在事例数很少和显著性不大的情况下。**
- 下面四张ppt展示四个直方图，它们分区宽度是一样，只是起始点offsets不同。四张图使用完全相同的一套数据，按照每个单位质量内25个事例的密度随机产生。

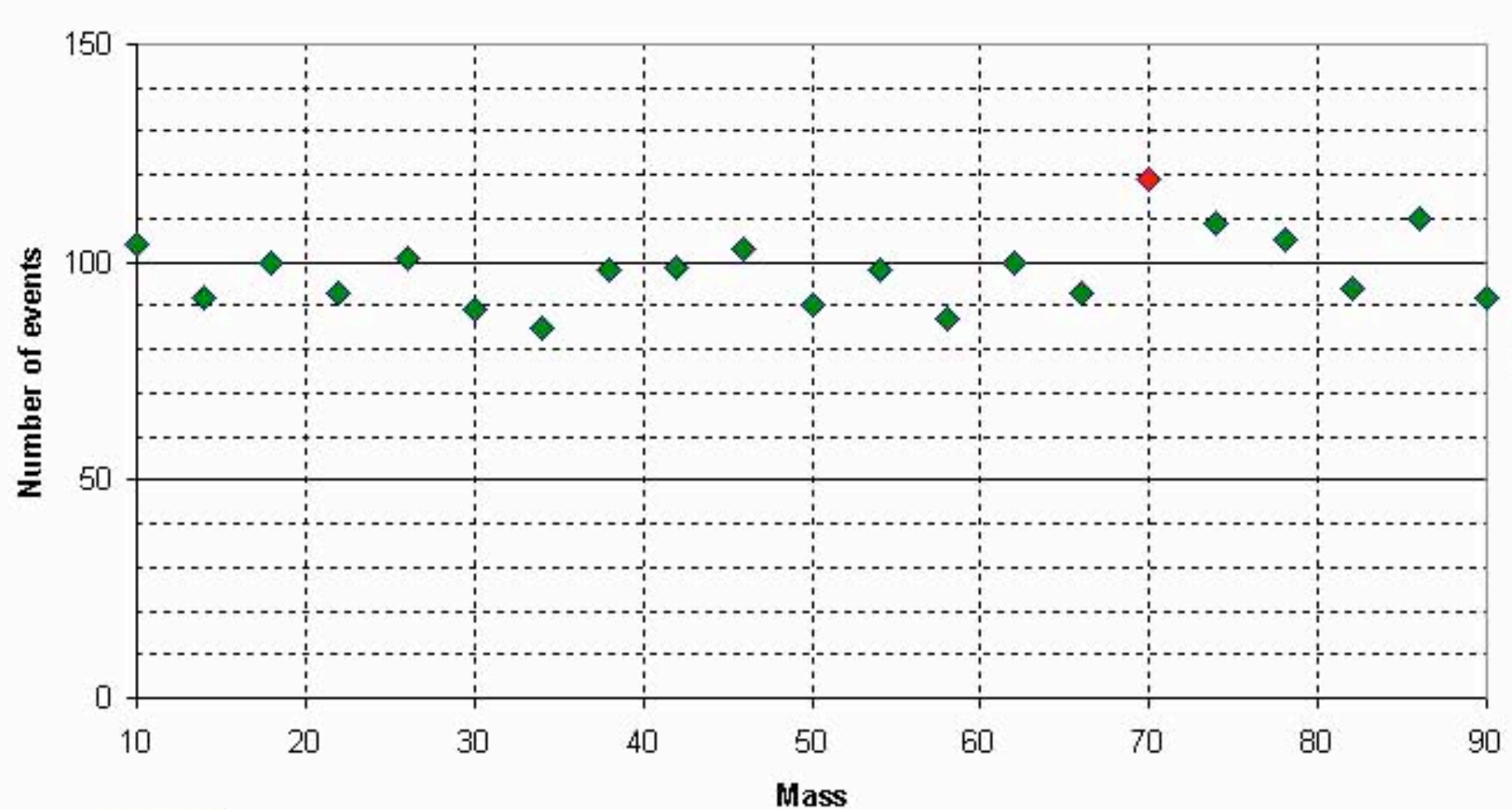




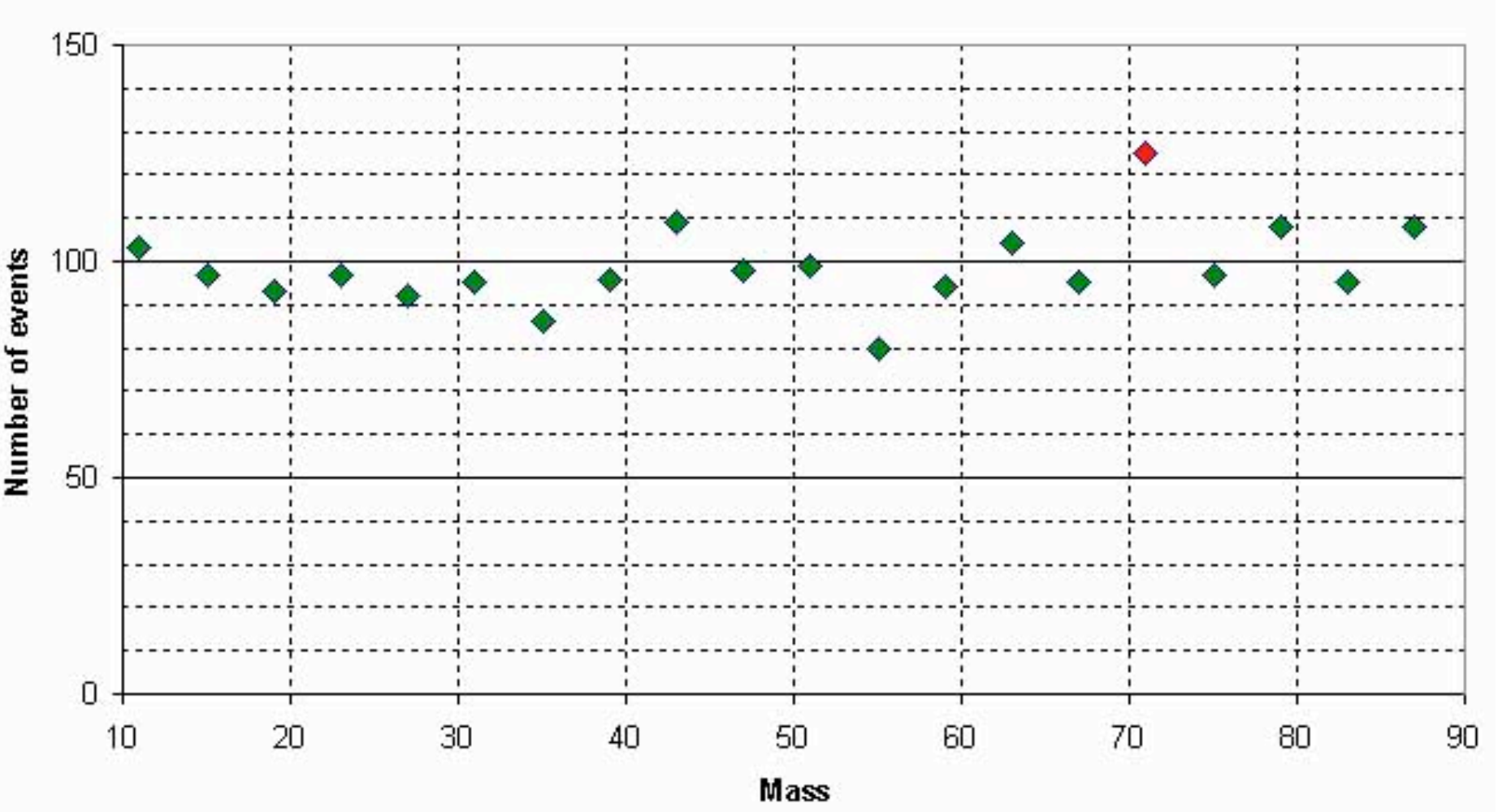




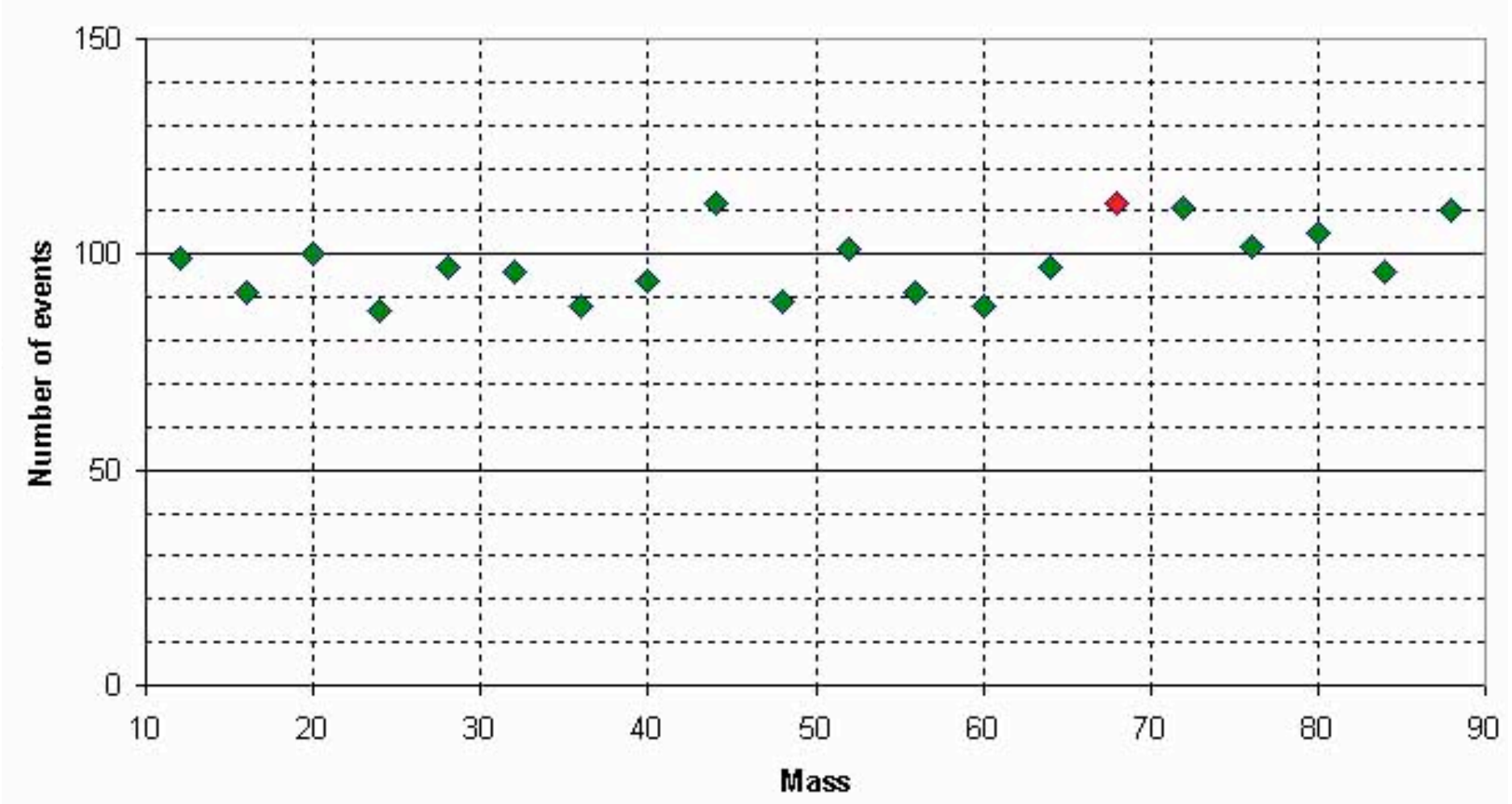
直方图1：bins 0-4-8-12-...



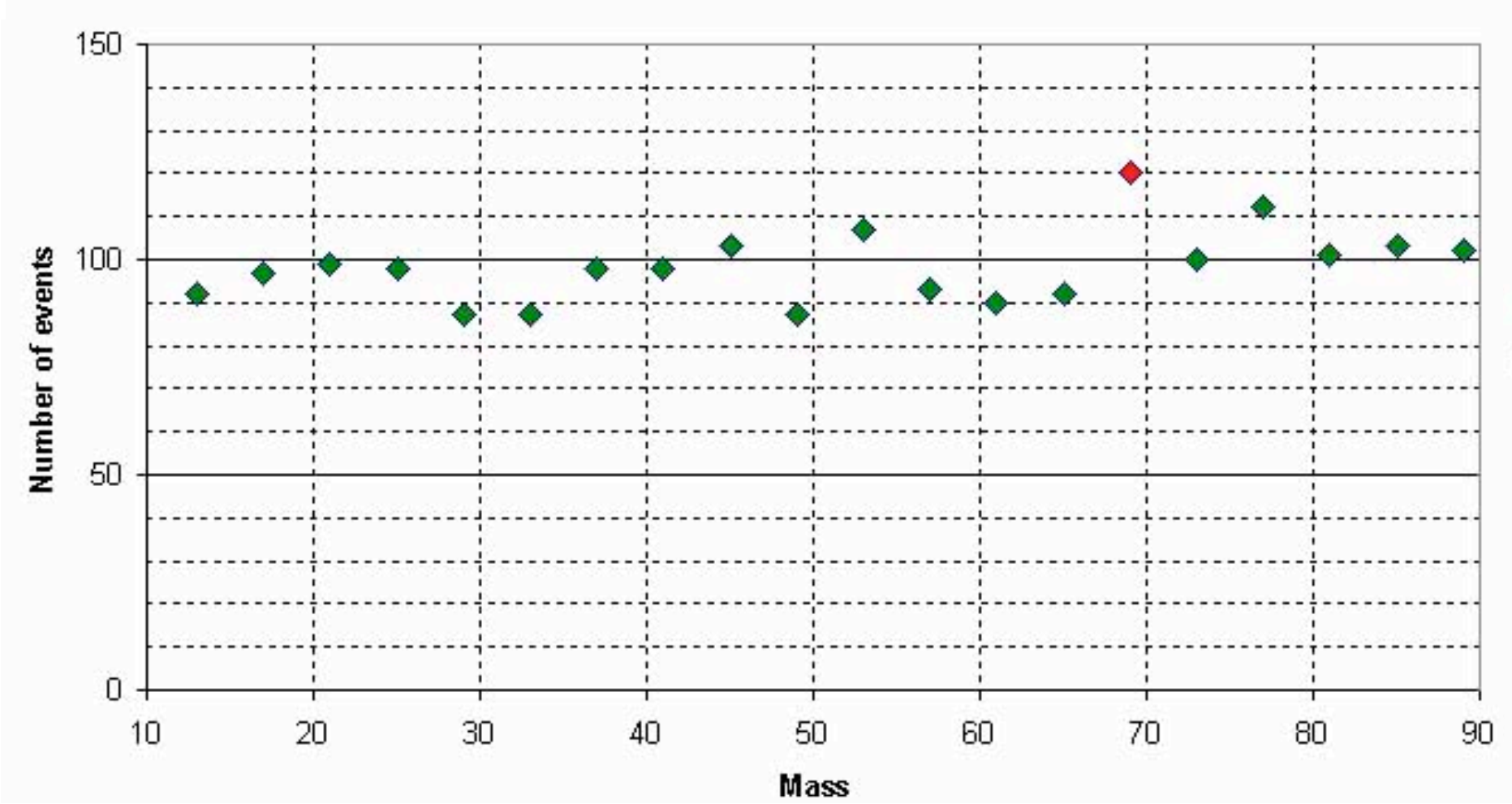
直方图2：bins 1-5-9-13-...



直方图3：bins 2-6-10-14-...



直方图4：bins 3-7-11-15-...



一些注意事项(posteriori adjustments)

- 大家可以观察到通过分区起始点的左右移动，M=70左右的accidental “peak”的事例数可以在[12,25]范围内[illegible]（平均[illegible]为100），对应S/√B是[1.2σ, 2.5σ]。
- 另一个**“优化”点是[illegible]多少个分区(sideband)来估计本底**。前面[illegible]方图2中如果只使用M=70峰值附件[illegible]正负4个[illegible]区来估计本底，就会得益于M=55附近刚好向下波动的有利情况，这样平均本底将只有97.5，后果就是“peak”的显著性 S/√B=(125-97.5)/√97.5=2.8σ。

一些注意事项(posteriori adjustments)

- **事例挑选Cuts**：类似的，如果事例挑选cuts“优化”是基于验后促进信号显著性的角度而不是出于先验的物理考虑，也会“增强”想要的信号。
- **丢弃“不好”的数据**：有时候会发现去掉特定集合的数据，例如每周一采集的数据(或者特定晶体触发的数据，再或者每一个数据采集run的初期数据等等)可以使信号更加显著。这种情况下通常大家会去琢磨有什么在周一那天出错了导致“不好”的数据，而不会去想有什么在周二至周五出错了导致“过好”的数据。这两个方面的错误类型都可以发生，但是人们思考问题的角度偏见导致第一种情况更经常出现，而且有时候找到的解释较勉强。显然这会导致朝向“discovery”的偏见。
- **这个新物理寻找只是非常多新物理分析中一个**：这是Look Elsewhere Effect的一种。LHC上的ATLAS、CMS实验每年都各有上百个新物理分析，如果我们只选择99%置信度作为发现新物理信号的标准，那么每年出来一些个“突破”也不足为奇了

相应解决方案

- **分区间(Binning)**：当处于有风险时(典型的情况是期望的事例数很少、小统计量情形)，使用不分区间(unbinned)的数据分析方案...
- **显著性计算**：小统计量时不要使用 $S/\sqrt{B}$，而是采用泊松概率分布函数。通常要尽力找出正确的误差概率分布函数，因为系统误差的存在经常会显著影响计算结果。
- **事例挑选Cuts**: cuts条件的优化要基于**先验**的考虑，蒙特卡洛样本事例，确实需要时也可以利用一小部分数据(比如20%)；把优化后的cuts应用到剩余的数据，之后不允许利用剩余数据作进一步的cuts调整。
- **丢弃"不好"的数据**：没有recipe... 要当心...

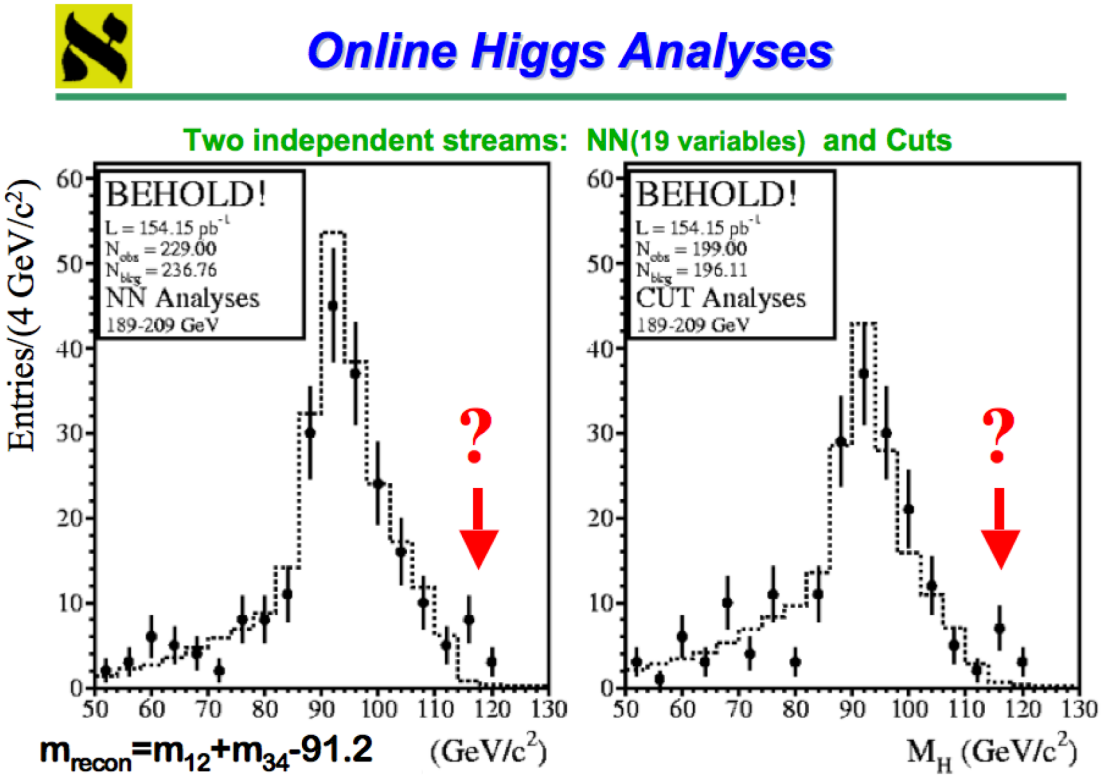
经验法则

- 3σ--可能是真东西，也可能是本底统计涨落；虽然值得发表，但是**不要宣称是个发现**；需要更多的数据以及/或者其它独立实验验证...

- 5σ--是时候严肃起来，需要独立的实验验证...

Higgs at LEP?

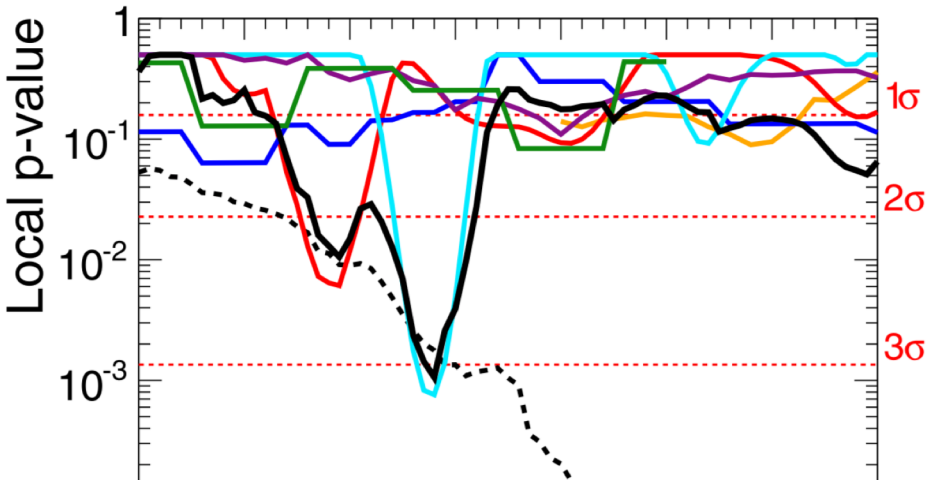
- 2000: ALEPH: ~4σ for Higgs signal present @ ~**115** GeV
  All four collaborations combined: ~2.9σ
- 2002: More thorough re-analysis of the same data:
  ALEPH: ~3σ
  All four collaborations: <2σ



Higgs at LHC—real thing that started from ~3σ

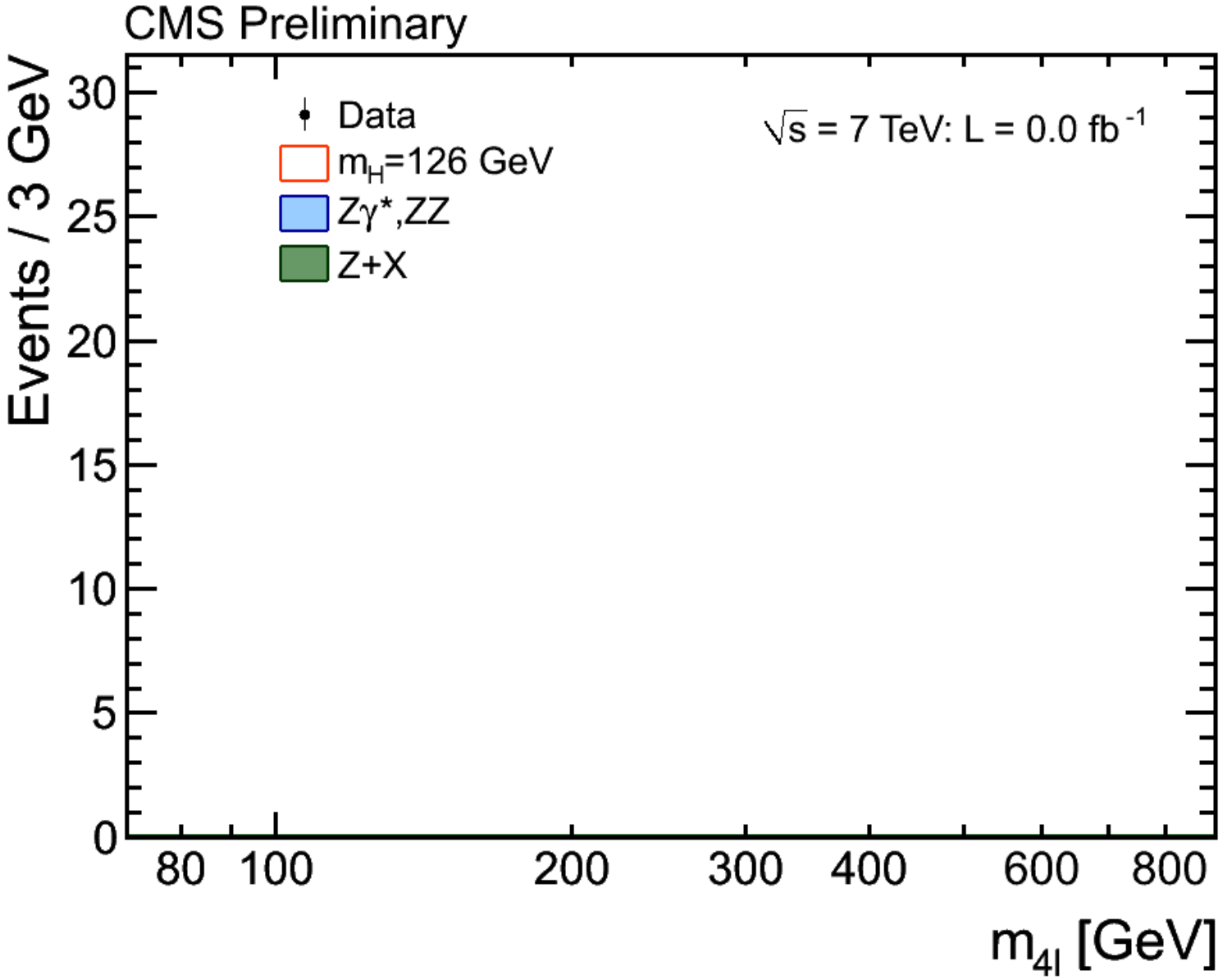
2011 December, CERN Council Seminar

Based on ~5/fb 7 TeV data, only **tantalizing** evidence, in both ATLAS and CMS

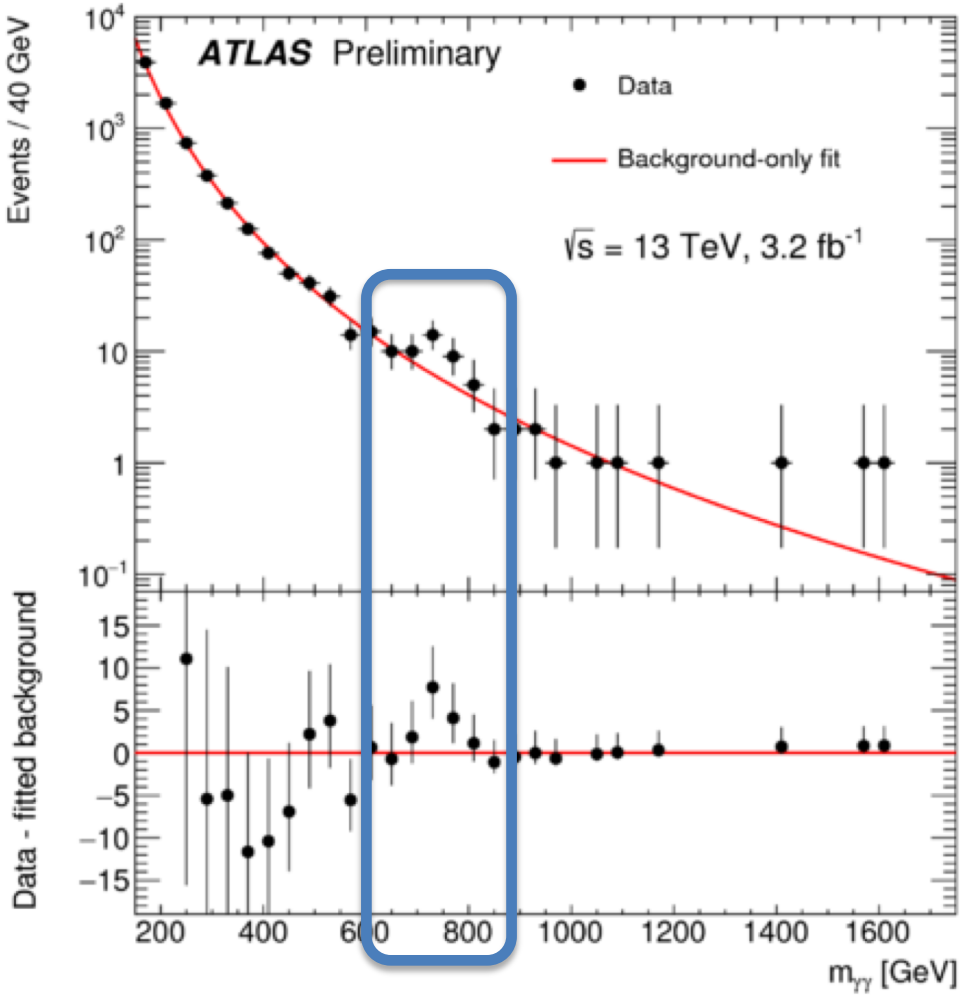


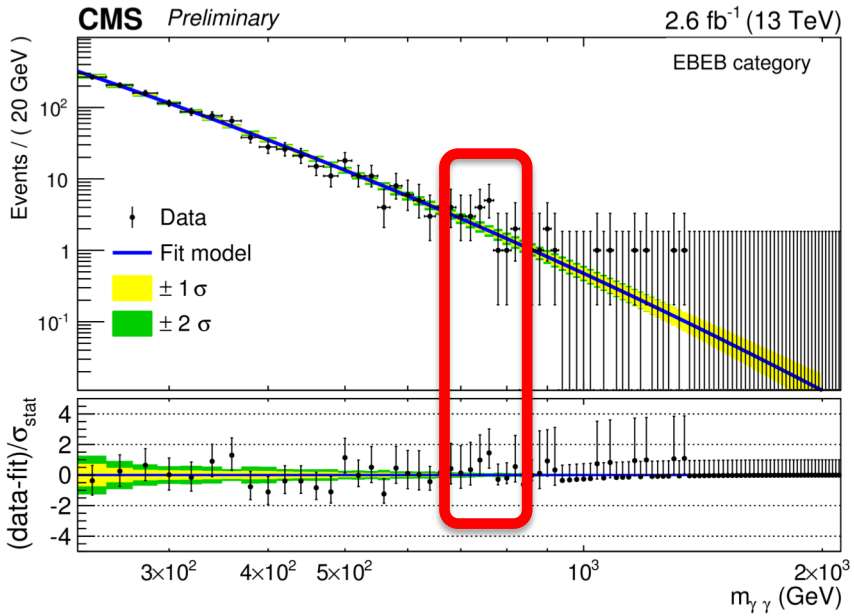
- **CMS paper title** was "Combined results of searches for the standard model Higgs boson in pp collisions at √*s* = 7 TeV"
- ***Abstract:* …**The largest excess, with a local significance of 3.1σ, is observed for a Higgs boson mass hypothesis of **124 GeV**. The global significance of observing an excess with a local significance >3.1σ anywhere in the search range 110–600 (110–145) GeV is estimated to be 1.5σ (2.1σ). More data are required to ascertain the origin of the observed excess.
- Subsequent papers based on much larger statistics confirmed the signal and were titled *"Observation of…"*

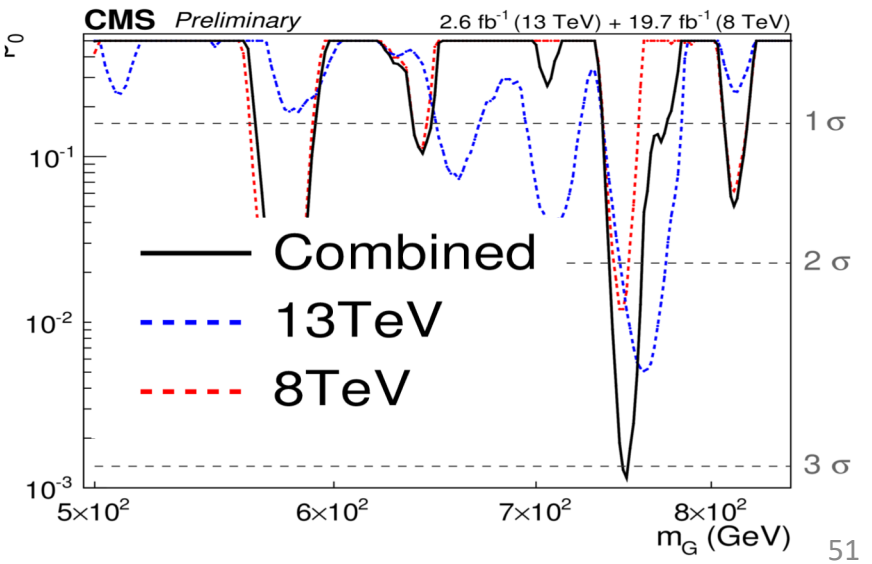
Growth of Higgs



750 GeV diphoton excess @ December 2015



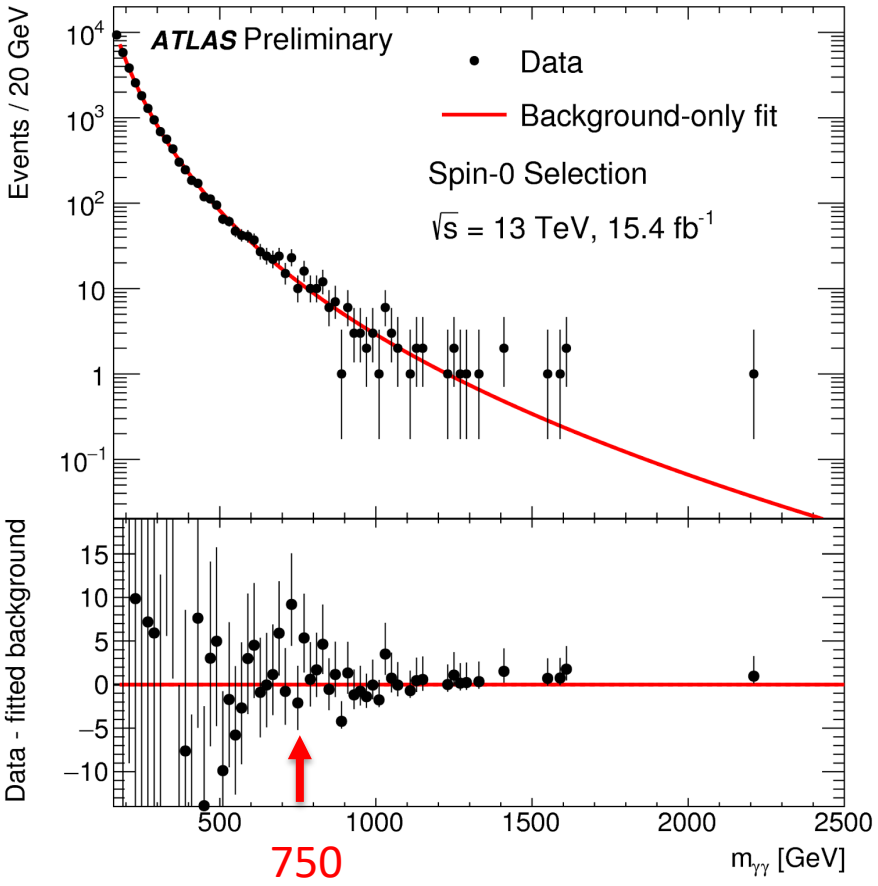


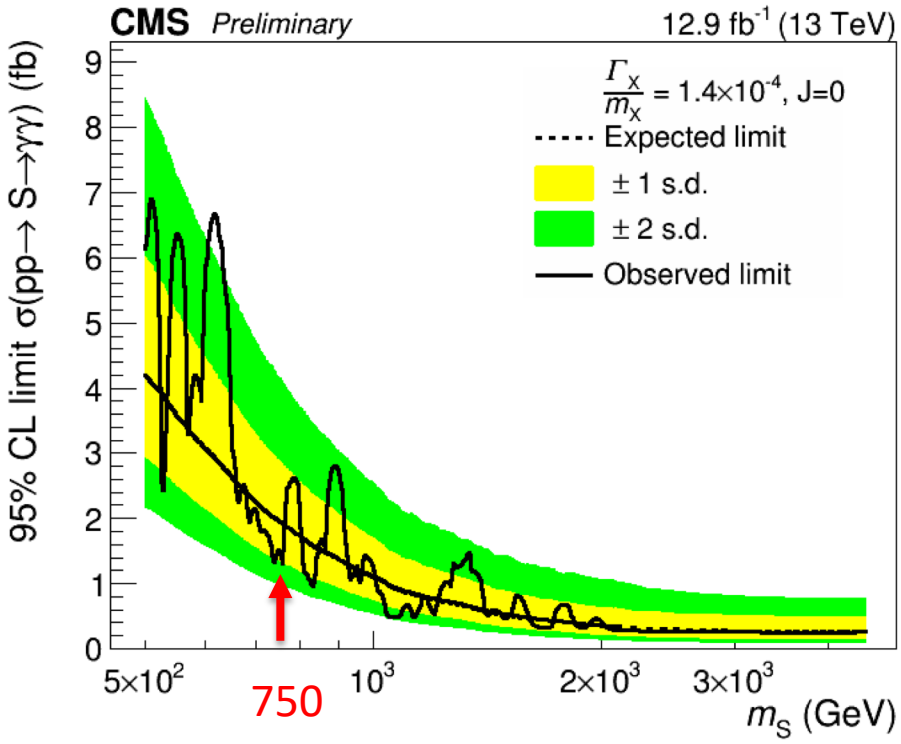


Hundreds of theory papers produced

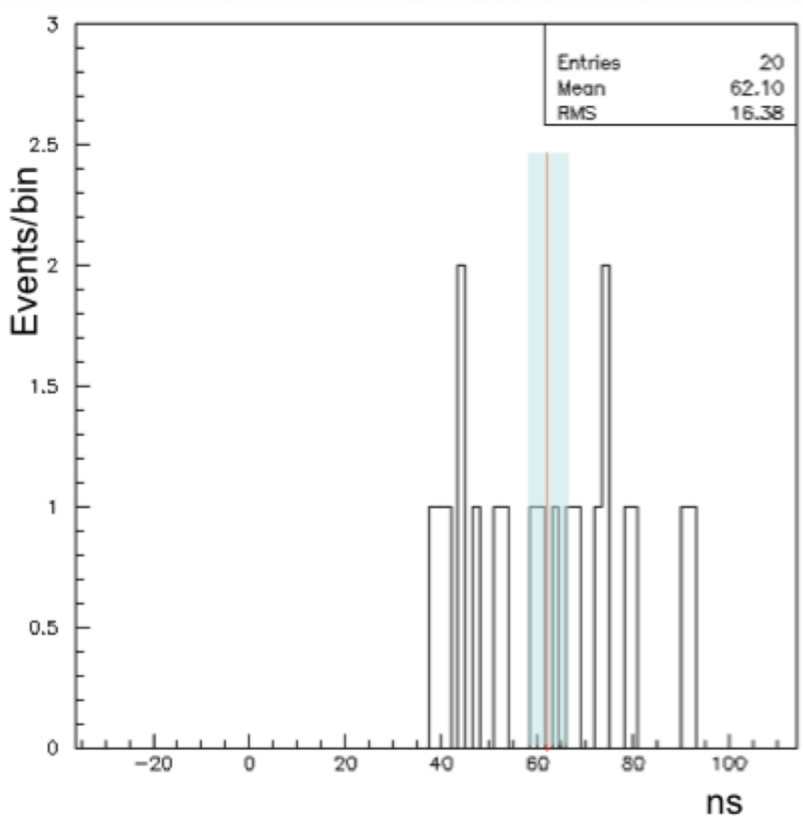
The loss of 750 GeV diphoton excess

At 2016 July ICHEP conference, with ~5 times more data analyzed, it's gone





Faster-than-light neutrino from OPERA 2011



- A loose fiber optic cable had introduced a delay in their timing system that explained the effect.
- Its spokesperson and physics coordinator resigned their leadership positions.

The end

实验物理分析要多做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