稀有衰变上限测量入门

基础概念、贝叶斯方法与 CL_s 实操

以 $J/\psi \to \mu^+\mu^-$ 的模拟质量谱为例

上限的概念　　贝叶斯实操　　CL_s 实操

课堂上的问题

没有看到显著的信号峰，怎样把结果写成一个上限?
先从质量谱认识信号与本底，再用两种方法限制平均信号产额。

闫文鹏

东南大学

Winter School 2026 for Belle II China Group

课堂讲义

目录

一 上限的基本概念 **3**

为什么要测上限? . . . 3

观测数、拟合产额与上限分别是什么? . . . 4

质量谱分析需要先确定什么? . . . 5

二 贝叶斯方法与轮廓似然积分实操 **6**

在信号模拟中拟合双高斯 . . . 6

固定峰形，同时拟合信号和本底产额 . . . 7

固定信号产额，逐点重拟合本底 . . . 8

从零开始积分，读出 8.94 个事例 . . . 9

把产额上限画回质量谱 . . . 10

三 CLs 实操 **11**

CL_s 实操流程与对应脚本 . . . 11

同一张质量谱，怎样用 CL_s 给上限? . . . 12

已经有真实数据，为什么还要做 toy MC? . . . 13

为什么分别生成两组 toy MC? . . . 14

固定 5 个信号，用同一规则检验 . . . 15

CL_s 是谁除以谁? . . . 16

怎么读 CL_s 这个数? . . . 17

在 toy 分布图上找到这两个比例 . . . 18

扫描信号产额，读出 CL_s 上限 . . . 19

两条黑线回答两个问题 . . . 20

绿黄带画的是仅本底下的统计波动 . . . 21

伪实验扫描与渐近扫描差多少? . . . 22

toy 数改变的是曲线的抖动，不是趋势 . . . 23

CL_s 的误差按 $1/\sqrt{N}$ 缩小 . . . 24

上限对物理结论有什么用? . . . 25

同一张质量谱，两种上限 . . . 26

参考文献 . . . 27

延伸阅读 . . . 27

阅读和实操顺序

先讲为什么要给上限，再用同一份双缪子质量谱做轮廓似然积分和 CL_s 扫描。ROOT 命令按演示顺序排列，每次运行后可在 ROOT 提示符下输入 `.q` 回到终端。

第一部分

上限的基本概念

为什么要测上限?

这一节要回答的问题

如果质量谱上没有看到一个清楚的峰，这次实验还能告诉我们什么?

在稀有衰变搜索中，我们常常先选出一批符合条件的候选事例，再画出它们的不变质量。如果目标过程存在，它可能在某个已知质量附近形成一个峰。其他物理过程也可能通过选择，出现在同一张谱上，这些贡献统称为本底。

即使只有本底，峰附近的候选事例数也会随机起伏。一次小的突起可能来自信号，也可能来自本底涨落。如果数据无法给出足够强的信号证据，可以继续问：**假定存在这样的信号峰，峰大到什么程度就与这份数据难以相容?** 上限把这个问题的答案变成一个可报告的数值。

先讲直觉

把一个已知形状的信号峰加到本底上，逐渐调高它。很小的峰可能被统计涨落淹没，峰足够大时，数据对这种产额的支持就会变弱。具体在哪个产额给出上限，由选定的统计方法和水平决定。

上限有三种常见用途。第一，把「没有显著信号」变成定量结果。第二，检验理论预测的分支比或截面是否超过实验可以排除的范围。第三，为不同数据量、事例选择或探测器配置提供灵敏度基准。比较灵敏度时，尤其要看同一统计定义下的**仅本底预期上限**。频率学区间、贝叶斯区间与 CL_s 这几套做法的系统综述，可以看 Particle Data Group 的《Statistics》一章 [6]。

观测数、拟合产额与上限分别是什么?

量	本例	含义
观测候选数	47	通过选择后进入这份模拟质量谱的候选事例数。
最佳拟合信号产额	约 2.53	在当前模型下，最贴合数据的平均重建信号数。
信号产额上限	例如 8.94	按指定统计规则推得的平均信号产额上限。

候选事例数是实际数到的整数。拟合中的信号产额是模型的**期望值**，可以是小数。即使搜索样本完全由本底生成，拟合也可能分配少量事例给信号峰，因为本底恰好在峰附近向上波动。

怎样报告一个上限

说清限制的是哪个物理量、采用了哪种方法、对应什么水平。例如，可以说「该模拟样本的平均重建信号产额的 90% 轮廓似然积分上限为 8.94 个事例」。

直觉：把置信水平想成撒网

把要限制的参数想成一条**待着不动的鱼**，把每次实验的结果想成**一次撒网**。置信水平描述的是这张网：长期看，大约 90% 的撒网能把鱼罩住。它并不是说「这条鱼有 90% 的概率在网里」——鱼待着不动，有随机性的是我们算出来的那个上限。所以「90% 上限为 8.14」的正确读法是：如果真值就是 8.14，本方法有 10% 的机会给出更小的上限，把它错误地排除。

贝叶斯可信上限对应给定模型和先验下的后验累计概率。频率学派的 CL_s 上限由排除规则确定。本讲义的第一种实操对轮廓似然积分，第二种实操用 CL_s 扫描。两边都用 90% 水平，但它们赋予这个数字的统计含义不同。

质量谱分析需要先确定什么?

- **信号峰形**。用信号模拟确定峰的位置、宽度和尾部。搜寻数据时通常沿用这个形状，拟合峰的产额。
- **本底模型**。峰附近可能有连续本底或其他物理过程形成的峰。可借助质量边带、控制样本或模拟来约束这些贡献。本实操用平坦本底，便于观察基本逻辑。
- **验证与系统误差**。实际分析可在查看目标区域前确定选择、模型和统计步骤。效率、归一化和本底模型的不确定性通常需要进入最终的上限计算。

先区分峰的形状和峰的大小

信号模拟回答「如果有信号，它在质量谱上长什么样」。搜索样本回答「眼前这份谱需要放多大的信号峰」。知道第一件事，才能用第二件事约束信号产额。

第二部分

贝叶斯方法与轮廓似然积分实操

在信号模拟中拟合双高斯

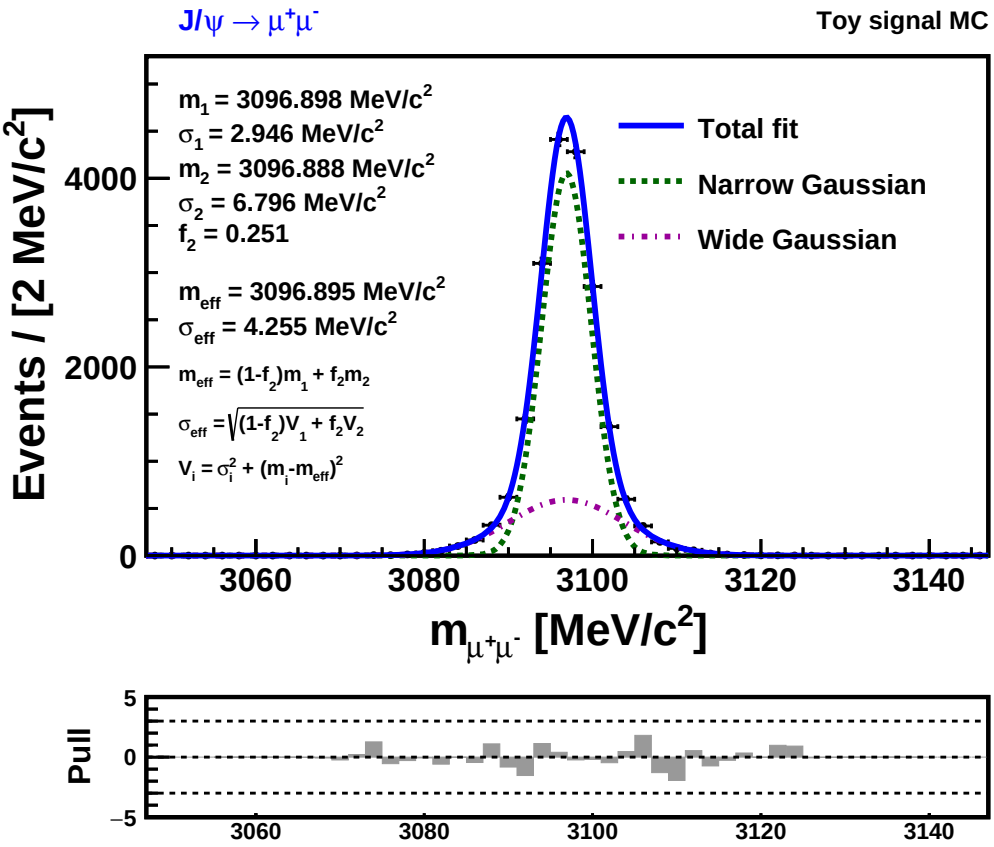


图 1 双缪子纯信号模拟样本的双高斯拟合。下方灰柱是 pull，用来看各质量 bin 中数据和拟合的偏差。

例题以 $J/\psi \to \mu^+\mu^-$ 为背景，模拟中心约为 $3096.9\,\mathrm{MeV}/c^2$。窄、宽分量的生成宽度分别设为 3 和 $7\,\mathrm{MeV}/c^2$，用来模拟两种质量分辨率。双高斯拟合得到两个均值、两个宽度以及宽分量比例。这些宽度描述本例的重建峰形。

```
cd /gpfs/group/belle/users/wenpeng/winSchool_2026/UL/method1
root -l steps/fit_signal_mc.C
```

图留在 ROOT 画布中。看完后输入 `.q`，回到终端。若重拟合后参数有变化，把新得到的五个峰形参数填入第二个宏中对应的位置。

固定峰形，同时拟合信号和本底产额

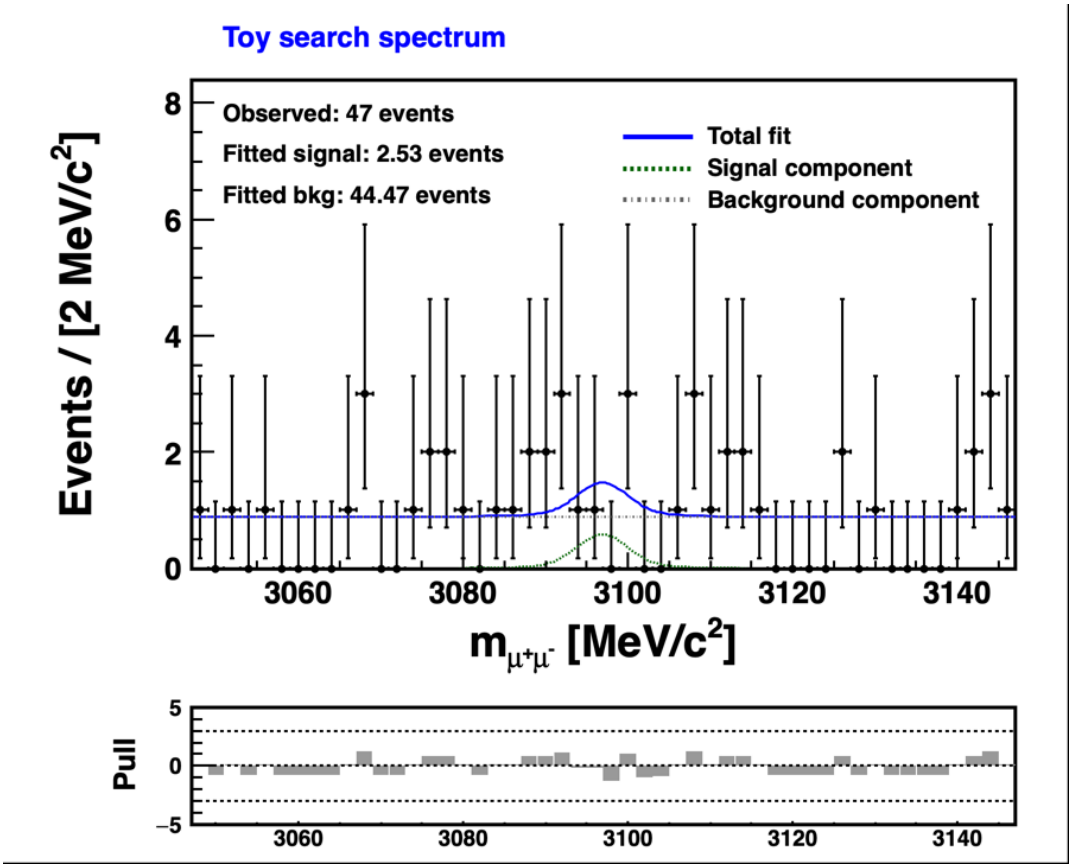


图 2 模拟搜索样本的拟合。两种成分分别为固定峰形的双高斯信号和连续平坦本底。

搜索样本由纯本底生成，这次抽到 47 个候选事例。拟合时仍然允许模型同时调整信号和本底的产额。自由拟合得到约 2.53 个信号和 44.47 个本底。这个小的正信号产额展示了本底涨落也能在固定峰形上产生一些拟合贡献。

```
root -l steps/fit_search_sample.C
```

该宏把数据和模型保存到 `steps/search_fit.root`，供轮廓似然扫描和下一部分的 CL_s 分析读取。

固定信号产额，逐点重拟合本底

假设我们想知道平均重建 5 个信号事例与数据的相容程度，就把信号产额固定为 5，再让本底产额在这个条件下找最佳值。在一系列测试产额上重复这两步，就得到轮廓似然。

测试信号产额	重拟合本底产额	ΔNLL	相对似然
0.00	46.92	0.321	0.725
2.55	44.45	0.000	1.000
5.00	42.83	0.221	0.802

令 $\mathrm{NLL} = -\ln L$。每个扫描点记录该点的最小 NLL，再减去扫描中的最低 NLL。

$$L_{\mathrm{rel}}(s) = \frac{L_{\mathrm{prof}}(s)}{L_{\mathrm{prof,max}}} = \exp\!\big[-\Delta\mathrm{NLL}(s)\big].$$

同一张谱下，固定更多信号时，本底产额可以调整，所以每个测试点都要重新拟合本底。归一化后的峰顶为 1；这里的相对似然用于比较不同信号产额的拟合质量。

从零开始积分，读出 8.94 个事例

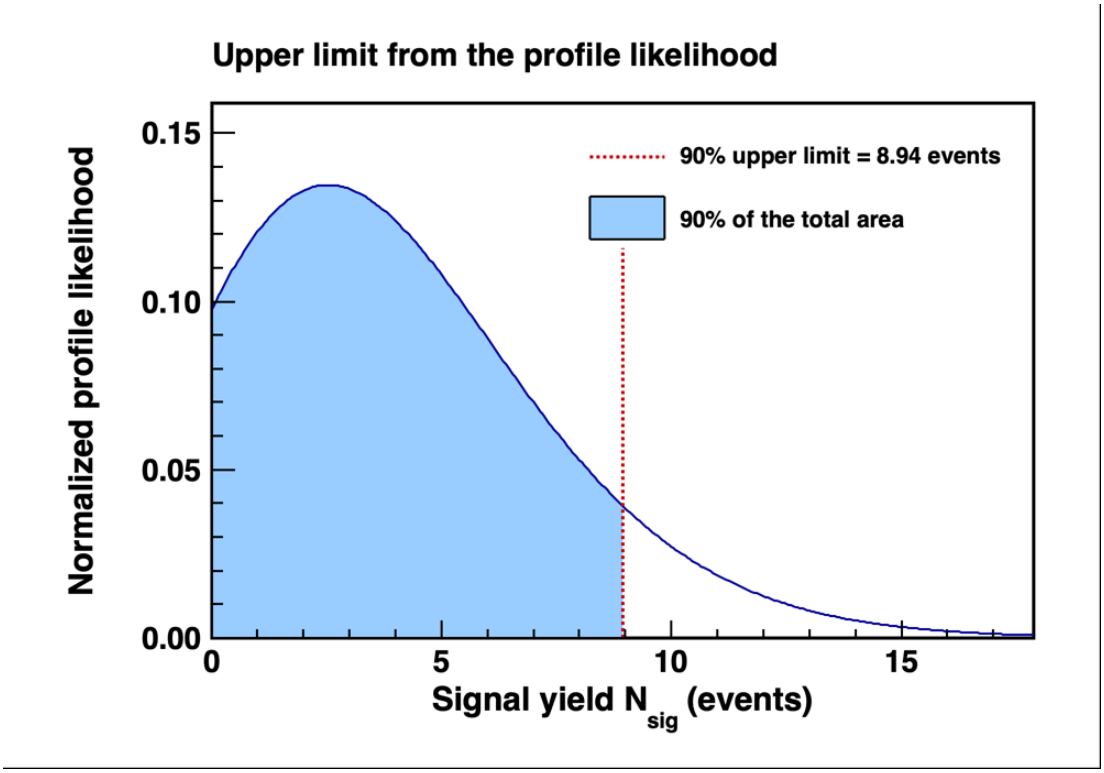


图 3 轮廓似然扫描和积分。归一化积分从零开始，累计到 90% 时给出信号产额上限。

在非负信号产额上采用平坦权重，从零往右累计面积。设 s_{UL} 为达到总面积 90% 的位置，$s_{\max}$ 为程序的扫描终点。

$$\frac{\int_0^{s_{\mathrm{UL}}} L_{\mathrm{prof}}(s)\,\mathrm{d}s}{\int_0^{s_{\max}} L_{\mathrm{prof}}(s)\,\mathrm{d}s} = 0.90.$$

```
root -l steps/scan_bayesian_ul.C
```

本例得到 $s_{\mathrm{UL}} \simeq 8.94$ 个事例。本底在每个信号产额处重新最优化，因此这里报告的是**轮廓似然积分上限**。完整的贝叶斯处理会给本底等干扰参数指定先验，并对它们做边缘积分。

把产额上限画回质量谱

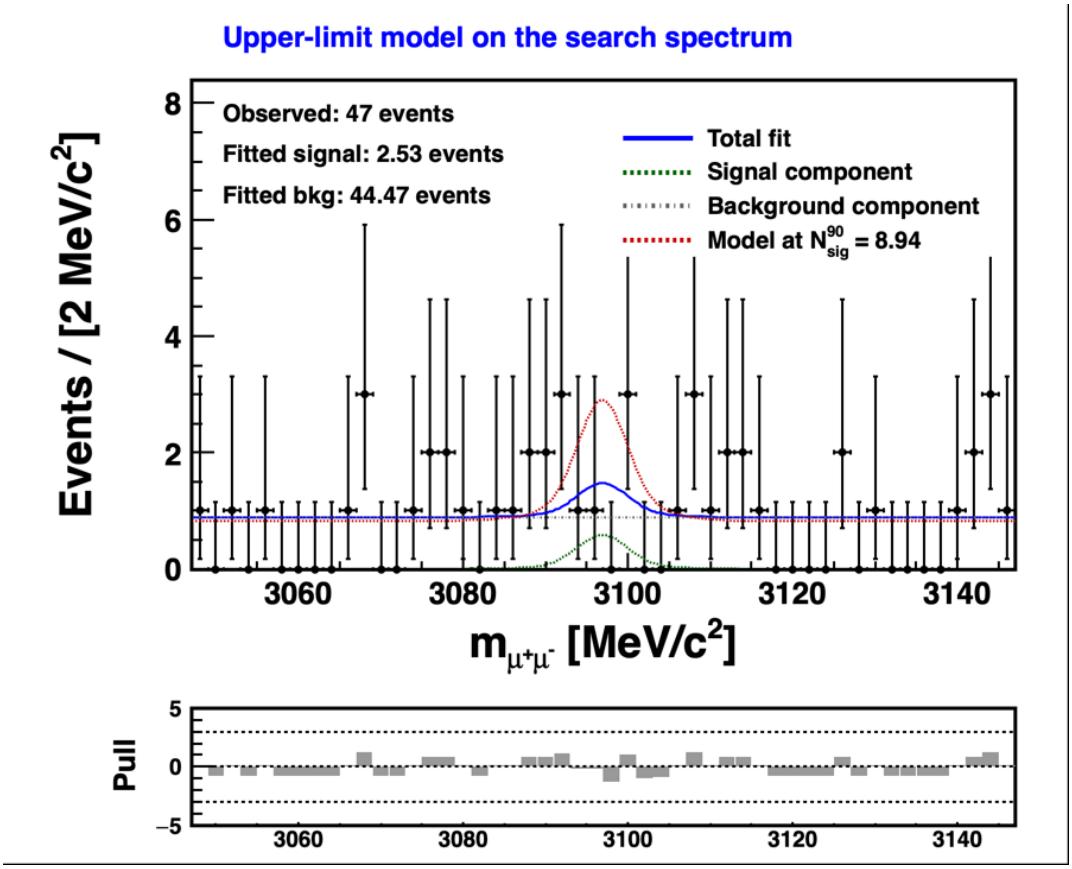


图 4 红线是把信号固定为轮廓似然积分上限后，重新拟合本底得到的整谱预测。

8.94 限制的是在当前模型与选择条件下的平均重建信号产额。可以把这个值放回拟合模型，直接看这样的信号峰在质量谱上有多大。实际分析常进一步把事例数上限换成分支比。以简单的单通道归一化为例，

$$\mathcal{B}^{\mathrm{UL}} = \frac{s_{\mathrm{UL}}}{N_{\mathrm{parent}}\,\varepsilon\,\prod_j \mathcal{B}_j}.$$

其中 N_{parent} 是对应样本中的母粒子数，ε 是信号重建和选择效率，$\mathcal{B}_j$ 是需要计入的子衰变分支比。具体归一化依照分析定义。

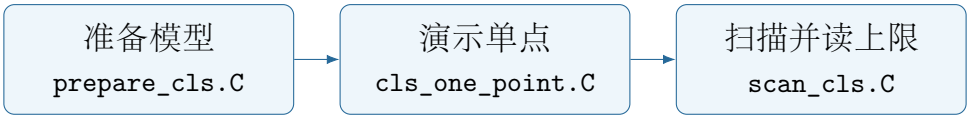
第三部分

CLs 实操

CL_s 实操流程与对应脚本

前面我们已经拟合了双缪子不变质量谱，并用似然积分给出了一个信号产额上限。现在我们沿用这份质量谱和同一个拟合模型，换一种方式提问。假如这里平均会重建出 5 个、8 个或者 12 个信号事例，眼前这份数据能排除哪些产额？

我们分三步来做。第一步只是准备材料，用 `prepare_cls.C` 把刚才的质量谱、信号和本底模型整理成 RooStats 能读的工作区。第二步用 `cls_one_point.C`，只挑 5 个事例这一点，把一个 CL_s 数字是怎么来的看清楚。第三步用 `scan_cls.C`，把一系列测试产额画成曲线，最后读出上限。后两个脚本都读第一步的输出，所以单点脚本主要是课堂演示，完整扫描可以从第一步直接开始。



同一张质量谱，怎样用 CL_s 给上限?

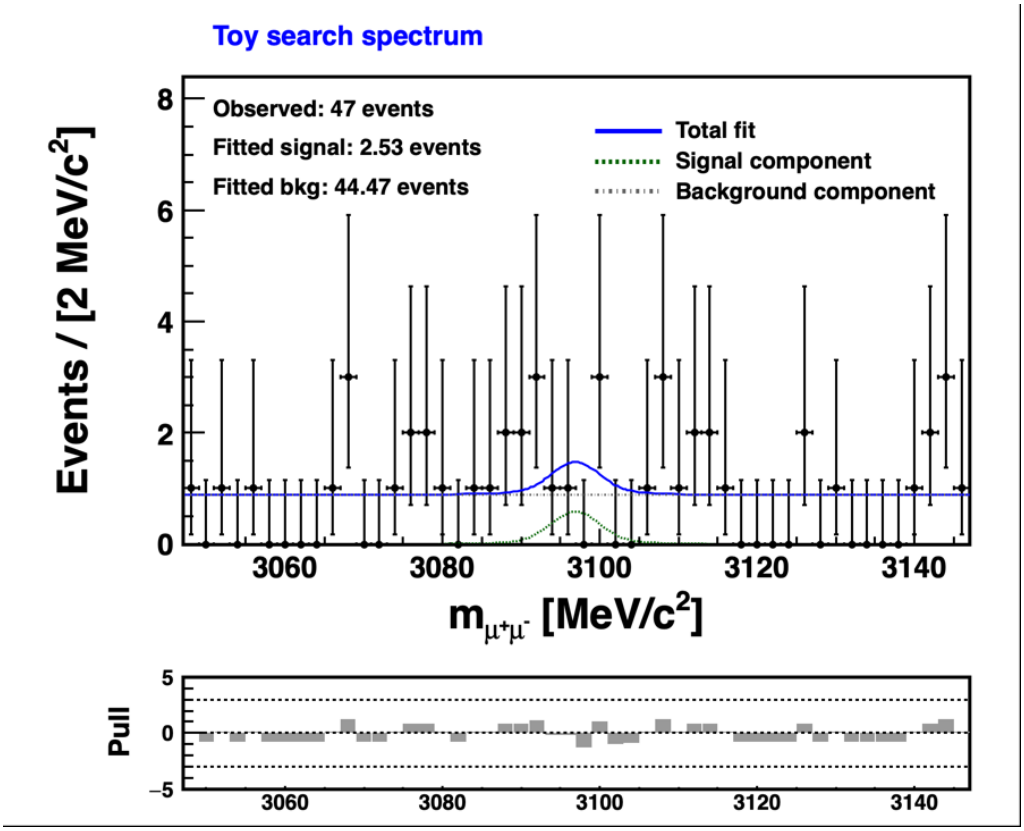


图 5 课堂搜索样本的双缪子不变质量谱。信号的双高斯峰形沿用 signal MC 拟合结果。

横轴是双缪子不变质量，点代表这份样本的分布。这里一共有 47 个模拟候选事例，课堂上就把它当作我们手里唯一一份观测数据。信号峰形已经用前面的 signal MC 确定了。现在测试不同信号产额时，峰形沿用原来的双高斯，本底产额可以重新拟合。

先拿平均重建 5 个信号事例举例。我们要准备两种假设，一种是「这个峰平均贡献 5 个事例，旁边还有本底」，另一种是「完全没有这束信号，只有本底」。两个假设用同一套质量谱模型来描述。第一条命令把这两种假设和搜索样本保存成 `cls_model.root`。

```
cd /gpfs/group/belle/users/wenpeng/winSchool_2026/UL/method2
root -l prepare_cls.C
```

已经有真实数据，为什么还要做 toy MC?

实际实验给我们一份数据。我们可以拟合它，得到一个数来描述这份数据对「平均 5 个信号」支持得怎么样。但是要判断这个数是否反常，还得知道同一种假设下，如果实验重做很多次，这个数通常会落在哪儿。

想一想，如果真有平均 5 个信号，每次都一定会恰好出现 5 个吗？有时会少几个，有时会多几个。本底数和每个候选事例的质量也会变。即使底层模型完全一样，重新采一次数据，质量谱和拟合结果也会起伏。

Toy MC 就是在电脑里按模型模拟这些重复实验。每份 toy 是一整份模拟搜索样本，和观测样本走同一套拟合、同一套检验，最后贡献一个统计量。手里的数据给我们一个观测值，大量 toy 让我们知道这个观测值在假设下有多极端。

为什么分别生成两组 toy MC?

第一组按「平均 5 个信号加本底」生成。我们想知道，如果真有这么大的信号，有多少次重做实验会像我们这份数据一样不支持它，甚至更不支持它。

第二组只按本底生成，但生成之后，还是用同样的规则去检验「平均 5 个信号」。这一组告诉我们，仅靠本底波动，有多常见会得到这样的检验结果。两组各生成 2,000 份，每一份都算一个统计量。

还可以和前面的 signal MC 对照一下。Signal MC 帮我们确定峰在哪里、峰有多宽。这里的 toy MC 拿着已经确定的峰形和本底模型去模拟整次搜索，看看检验统计量怎么波动。

生成假设与检验目标

两组 toy 的生成信号产额分别是 5 和 0。生成以后，对两组都计算用于检验 5 个信号事例的同一个统计量 q_s。这样才能把两个分布放到同一个横轴上比较。

固定 5 个信号，用同一规则检验

我们把「这份质量谱有多不支持 5 个信号」变成一个数字，记作 q。怎么做呢？先让信号和本底产额都浮动，找最佳拟合。再把信号固定在 5，让本底重新找它的最佳值。比较两次拟合，看固定在 5 以后拟合变差了多少。

这里做的是排除偏大的信号产额，所以只关心一个方向。如果自由拟合出来的信号产额反而比 5 大，数据往信号更强的方向走了，这种情况不给它排除 5 的分数，q 取 0。往另一个方向看，固定在 5 代价越大，q 就越大。

这里的 L 表示似然，s 是我们指定的测试信号产额，帽子表示拟合出来的值。分子里的 $\hat{\hat{b}}_s$ 是固定信号产额后重新拟合得到的本底产额，分母里的 $\hat{s},\hat{b}$ 是两个产额都放开后的最佳拟合值。

$$q_s = \begin{cases} -2\ln\dfrac{L(s,\hat{\hat{b}}_s)}{L(\hat{s},\hat{b})}, & \hat{s} \le s, \\ 0, & \hat{s} > s. \end{cases}$$

上式的第一种情况也可以读成「固定信号后的负对数似然减去自由拟合的负对数似然，再乘 2」。用 47 个候选事例算出来大约是 0.441。这就是文献里常用的单侧剖面似然比统计量 $\tilde{q}_\mu$ 的写法 [5]。随后每份 toy 都用这条公式算出自己的值，方便与观测值比较。

```
root -l cls_one_point.C
```

脚本开头已经设置了测试产额 5、每组 2,000 份 toy 和随机种子。这个命令跑完会给出观测统计量和两个尾部比例，下一页我们一起算一遍。

CL_s 是谁除以谁?

先用这 47 个候选事例算出一个观测统计量，记成 q_{obs}。我们把它当成一条分界线。两组 toy 都问同一个问题，有多少份的 q 大于或等于这条线? 越往右，就越偏向不支持 5 个信号的方向。

第一行是分子。我们在平均 5 个信号加本底的假设下生成了 2,000 份 toy，有 500 份落在右尾，所以分子是 500 除以 2,000，等于 0.2500。

第二行是分母。我们在仅本底假设下也生成了 2,000 份 toy，其中 1,681 份落在同一条线的右侧，所以分母是 1,681 除以 2,000，等于 0.8405。

比值中的位置	生成 toy 时的假设	满足 $q_s \geq q_{\text{obs}}$ 的比例
分子 CL_{s+b}	平均 5 个信号事例加本底	$500/2000 = 0.2500$
分母 CL_b	仅本底	$1681/2000 = 0.8405$

$$CL_s = \frac{CL_{s+b}}{CL_b} = \frac{0.2500}{0.8405} \approx 0.297.$$

拿第一个比例除以第二个比例，得到 CL_s 约等于 0.297。为什么要看分母? 因为我们还要知道这次的表现能不能帮我们区分两种假设。比如两个右尾比例都很小、却差不多，那么这种结果在两种假设下都少见，单看分子会显得证据比实际更强。把它们相除以后，比值仍会接近 1。

这个比值最早由 LEP 的物理学家提出[1]，并在高能物理中推广 [1]，其条件概率的论证可以追溯到 Zech 对计数实验的处理 [2]。它针对的是一个具体困难：标准单侧上限在观测事例数远低于预期时，会把物理上允许的信号产额一起排除掉，有界参数下的这个问题在统计文献里有专门讨论 [3]。除以仅本底的那条尾巴，相当于额外要求这份数据在「有信号」的假设下比在「纯本底」的假设下更不支持该产额；灵敏度低时两个比例会同步变小，比值自动接近 1。代价是得到的区间不再有严格的标称覆盖率：CL_s 区间的覆盖概率高于 $1-\alpha$[6]。

[1]文献里还常见 $\text{CL}_s = p_\mu/(1-p_b)$ 的写法，其中 p_b 是仅本底假设被拒绝的 p 值（PDG 综述式 (40.82)[6]，以及 Junk 的讨论 [4]）。本讲义和 RooStats 的 `CLb()` 用的是同一个统计量在仅本底下的右尾，即 $\text{CL}_s = p_{s+b}/p_b$。两种写法在渐近极限下数值接近，但分母的含义不同，读文献时注意区分。

怎么读 CL_s 这个数?

三件事：这个比值怎么读、拿本讲义的数字走一遍、以及它为什么比「$p \le 0.10$」更严格。

最省事的一句读法：$CL_s = 0.10$ 表示

$$\frac{\text{有信号时看到这么极端结果的概率}}{\text{纯本底时看到这么极端结果的概率}} = 10\%.$$

也就是说，如果真有这么强的信号，出现我们手上这种「不像信号」的结果，概率只有纯本底情况下的十分之一——于是可以按 90% 置信水平排除它。阈值取 0.10 而不是 0.90，是因为 CL_s 和 p 值一样「越小越排除」，而 $\alpha = 1 -$ 置信水平。

拿两个点走一遍。第 16 页那两个比例是 0.250 和 0.8405，比值 0.297，远大于 0.10，所以平均 5 个信号排除不了。沿曲线往右，到 8.06 个事例时比例变成约 0.087 和 0.867，比值 0.099，刚好跨过红线——这就是观测上限。

测试产额	有信号的右尾 CL_{s+b}	纯本底的右尾 CL_b	CL_s	结论
5.00	0.250	0.8405	0.297	不排除
8.06	0.087	0.867	0.099	排除

为什么这比「$p \le 0.10$」更严格？因为分母 $CL_b \le 1$，总有 $CL_s \le CL_{s+b}$：要排除一个产额，得同时过「分子足够小」和「相对本底足够小」两道关。代价是它不再是严格意义上有标称覆盖率的置信区间——实际覆盖概率高于 90%，术语叫超覆盖。好处是数据向下涨落得离谱时（比如零事例、本底却有 45 个），普通 p 值会把所有信号产额一起排除，而 CL_s 会把这种结论自动压回去，这正是上一页说的 Zech–Read 动机。

在 toy 分布图上找到这两个比例

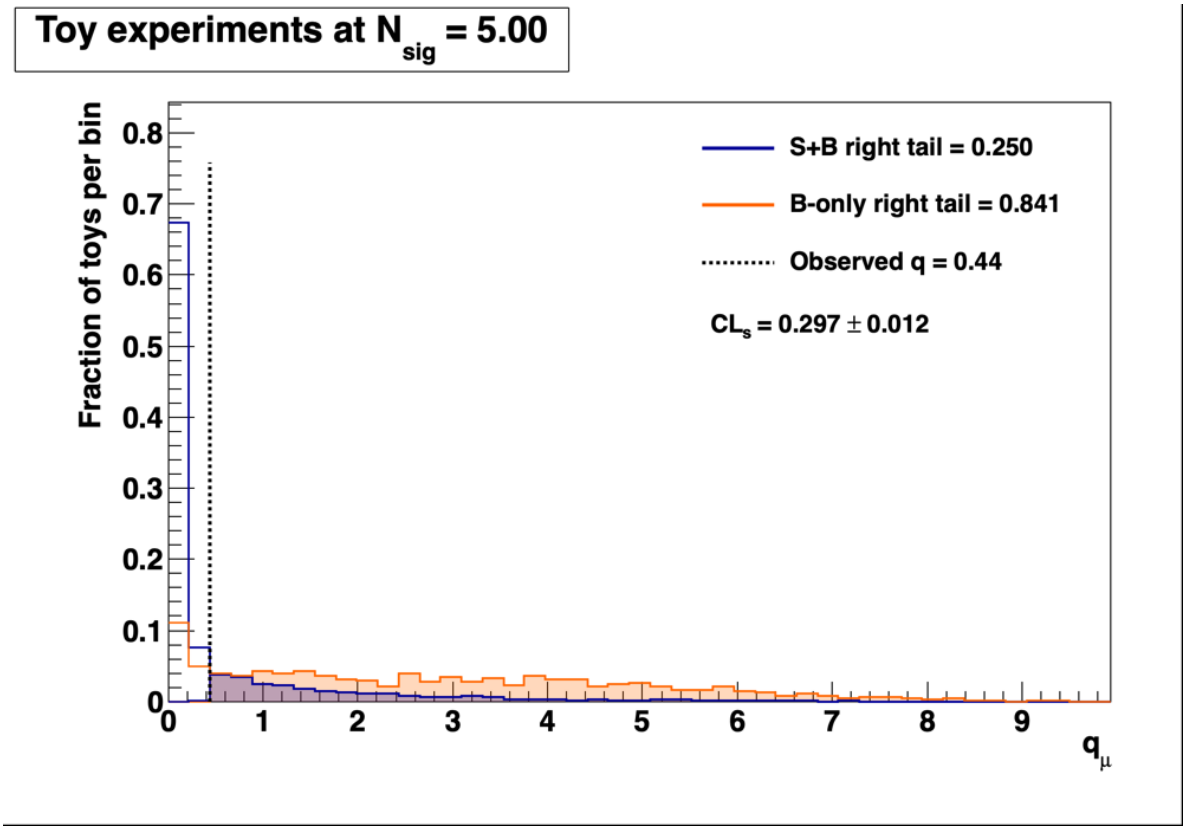


图 6 固定测试信号产额为 5 时，两组 toy 的统计量分布。黑色竖线标出观测值，蓝色和橙色阴影标出各自的右尾。

上一页是表格，这里把同一件事画出来。每份 toy 在横轴上贡献一个 q，纵轴表示每个小区间里 toy 占多大比例。蓝色是信号加本底，橙色是仅本底。黑色虚线在 q 约等于 0.441 的地方，是我们这份观测样本给出的值。

表格里的 500 和 1,681，就是分别数蓝色组和橙色组有多少份 toy 落在这条竖线右侧，等于竖线位置的也计入。两个比例相除是 0.297。图上还写了约 0.012 的误差，它反映用有限数量的 toy 来估计比例时的抽样波动。

我们事先选择了 90% 置信水平，对应的 CL_s 排除阈值是 0.10。0.297 比 0.10 高，所以这份数据还排除不了平均重建 5 个信号事例。那测 8 个会怎么样？测 12 个呢？接下来把测试值逐个换掉，就能找到上限。

扫描信号产额，读出 CL_s 上限

```
root -l scan_cls.C
```

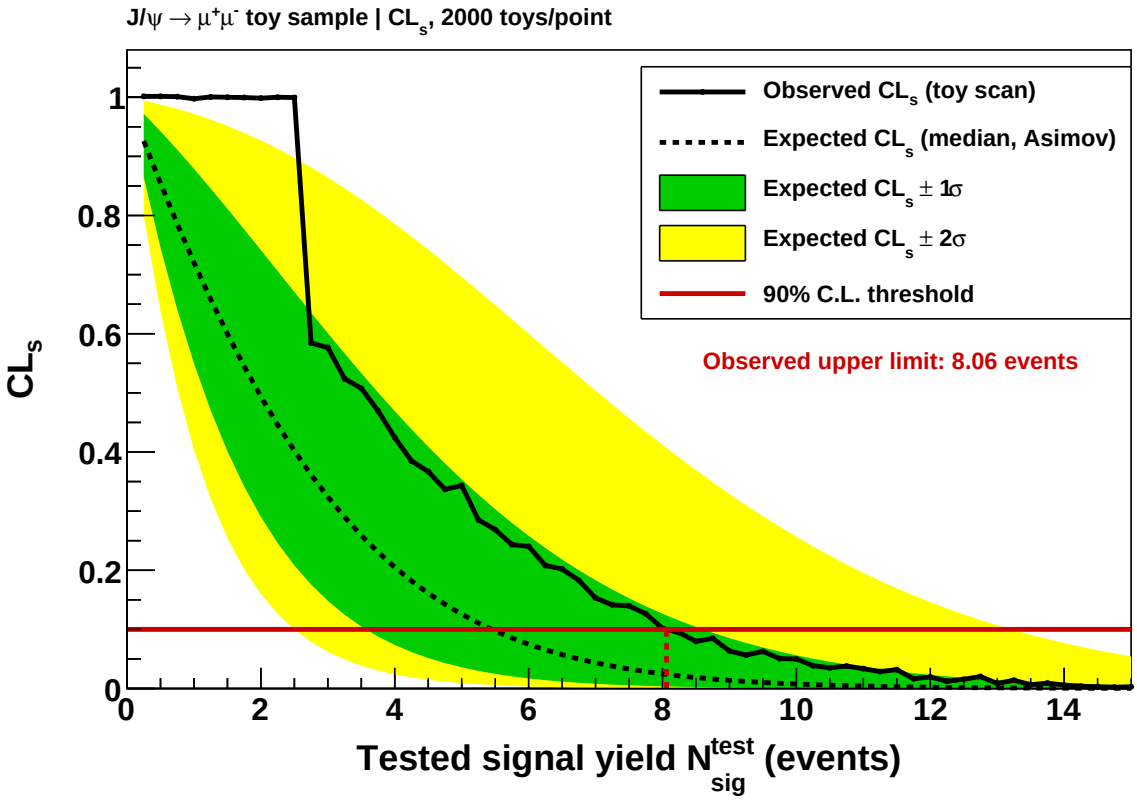


图 7 伪实验 CL_s 扫描，每点两组各 2000 个 toy。观测实线与 $CL_s = 0.10$ 的红色阈值线相交，给出观测上限 8.06 个事例。

横轴上的每一点都是一个待检验的平均信号产额，纵轴是在那个产额下得到的 CL_s。我们从 0.25 扫到 15 个事例，共扫 60 个点。刚才用 toy 演示了一个点的原理，这条曲线照同样的做法扫 60 个点，每点两组各 2000 个 toy，跑一遍约 50 秒。绿黄带是仅本底假设下的预期范围，由 Asimov 公式给出，这也是 LHC 实验的通行做法 [5]。

红线是我们在计算前就选定的阈值 0.10。沿着黑色观测曲线，看它什么时候穿过红线，再从交点往下读横坐标。这一次读到 **8.06 个事例**。每个点 2000 个 toy，交点本身还有约 0.2 个事例的统计误差，所以只读到两位有效数字。

两条黑线回答两个问题

黑色实线回答的是「这一次的数据能限制到什么程度」，所以读到 8.06。黑色虚线问的是另外一个问题，「假设只有本底，同样的分析通常能限制到什么程度」。虚线是预期中位线，它和红线相交的位置对应 5.45 个事例。

这次的观测上限比中位预期大，说明这份样本给的限制比仅本底下典型的一次宽松。要知道这种差距有没有大到少见，还要看仅本底时统计波动的大小。

曲线	对应的样本或假设	90% 置信水平上限
黑色实线	课堂固定的 47 个模拟候选事例	8.06 个事例
黑色虚线	仅本底假设下的中位预期	5.45 个事例

绿黄带画的是仅本底下的统计波动

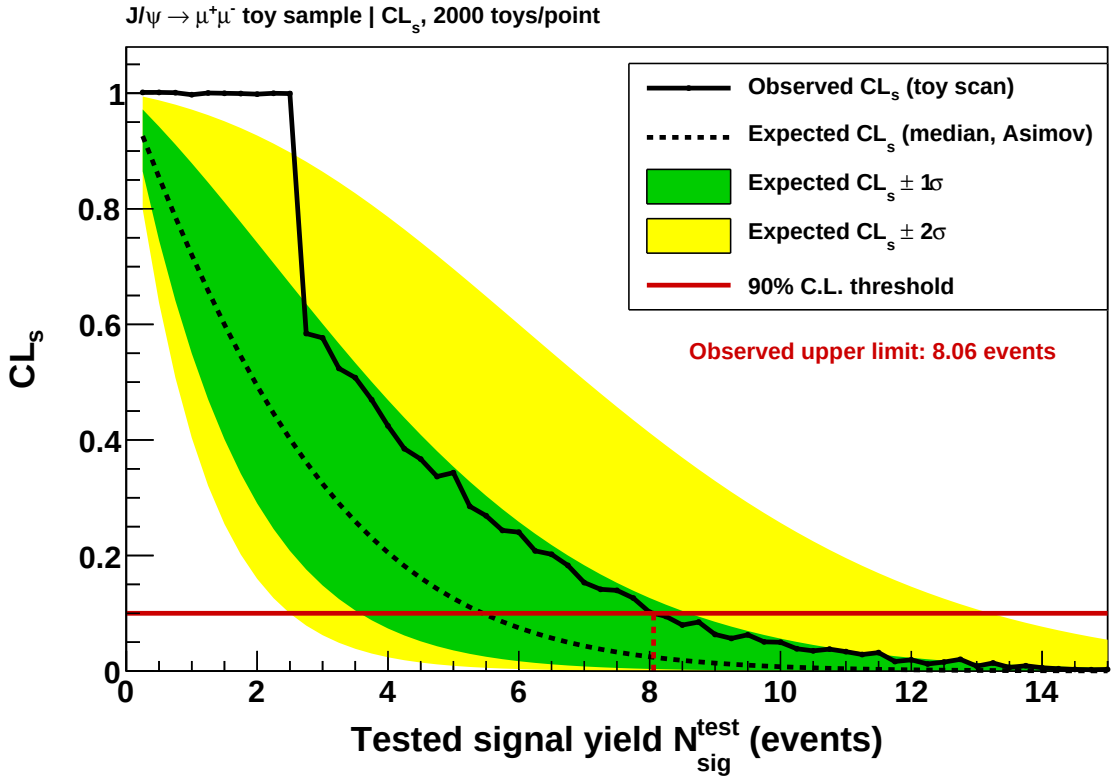


图 8 固定一个测试产额，竖着看绿带和黄带。它们给出仅本底假设下逐点的中央约 68% 和 95% 预期范围。

先固定一个测试产额，比如 8 个事例。想象只有本底时实验重做很多次，每次数据都稍有不同，算出的 CL_s 也会有大有小。黑色虚线给出这一列的中位数，绿带覆盖中央约 68% 的预期范围，黄带延伸到中央约 95%。

这里的读法是固定横坐标，竖着看。绿黄带描述的是仅本底假设下预期结果的波动，黑色实线代表我们手里这份样本。我们把两者放在同一张图上，是为了对照这次结果与本底波动的典型大小。单凭观测上限比预期大，还不能判断是否发现了信号。

伪实验扫描与渐近扫描差多少?

主线的扫描本身就用伪实验，所以可以直接问：它和渐近公式给的数差多少？同一份数据、同一个统计量、同样的 60 个扫描点：

做法	90% 上限（个事例）
伪实验扫描，每点 2000 个 toy（本讲义主线）	8.06
伪实验扫描，每点 8000 个 toy	8.15
渐近（Asimov 公式）	8.14

三种做法在 0.1 个事例内一致。真正会让人误判的是**单点**的 toy 结果：在 8.14 这个点上用 1000 个 toy 重复 8 次，算出的 CL_s 落在 0.084 到 0.137 之间，标称误差约 0.010。某一次单点读出 0.12，并不说明方法有偏差，只说明玩具数不够。

这正是下一页要讲的：toy 数决定交点能读几位。每点加到 8000 个 toy，扫描给出的 8.15 就与渐近的 8.14 重合；而 2000 个 toy 的扫描，交点本身有约 0.2 个事例的误差，读 8.06 时要留这个余量。

在 $b \approx 45$、$s \approx 8$ 这样的样本量下，Wald 渐近已经够用 [5, 6]；计数更少、或者模型更复杂时，伪实验才是必需的。

toy 数改变的是曲线的抖动，不是趋势

主线已经用伪实验扫描，这一页看 toy 数的影响。每个测试产额都要两组各 N 份 toy，也就是 $2N$ 次完整拟合。图 9 把同一份数据、同一个统计量在 $N = 200$、1000、5000 下的 toy 扫描与渐近曲线画在一起，右图是阈值附近的放大。

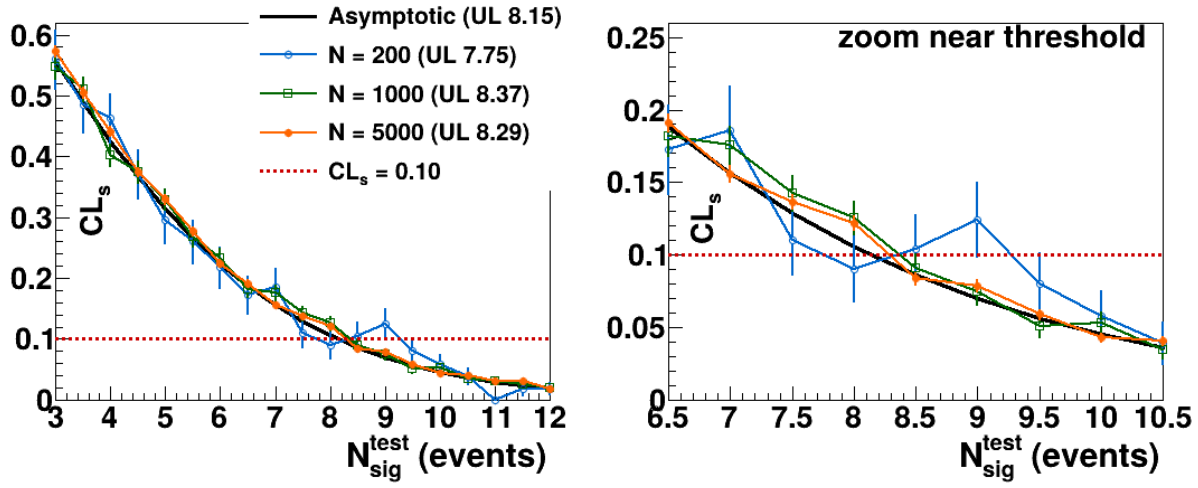


图 9 同一份搜索样本、同一个统计量，在不同 toy 数下的 CL_s 扫描。左图是完整范围，右图放大阈值附近。误差棒来自每个点的二项统计误差。

N 小的时候，每个点上只有几十到几百次计数，右尾比例本身带着明显的二项起伏。棒的长度还与 CL_s 本身大致成正比，所以越靠右的点棒越短：同样 2000 个 toy 时，5 个事例处约 0.012、阈值附近约 0.007、12 个事例处只有 0.004；图中 $N = 200$ 的棒约 0.02，$N = 5000$ 的约 0.005。$N = 200$ 的曲线在 7–9 之间来回穿过 0.10（7.5 附近 0.11、8.0 附近 0.09、8.5 又回到 0.11），交点被读成 7.75；$N = 1000$ 与 5000 分别给 8.37 和 8.29，与渐近的 8.15 只差百分之几；主线用的 2000 个 toy 给出 8.06，也落在这个范围里。

所以 toy 数的意义不在改变曲线该有的形状，而在决定每个点上「比例」这个估计有多准，从而决定交点能读几位、会不会被一次涨落带偏。

	$N = 200$	$N = 1000$	$N = 5000$	渐近
读出的 90% 上限（个事例）	7.75	8.37	8.29	8.15

CL_s 的误差按 $1/\sqrt{N}$ 缩小

在 8.14 这个点上用不同的随机种子重复 8 次独立实验，把 8 个 CL_s 的离散程度画出来，就是图 10 的蓝点（误差棒是 8 次重复本身的不确定度，约 25%）。它和解析式

$$\sigma_{CL_s} \simeq CL_s \sqrt{\frac{1-p_{s+b}}{N\,p_{s+b}} + \frac{1-p_b}{N\,p_b}}$$

一致：N 从 50 加到 10000，误差从 0.05 降到 0.003，正好是 $1/\sqrt{N}$。

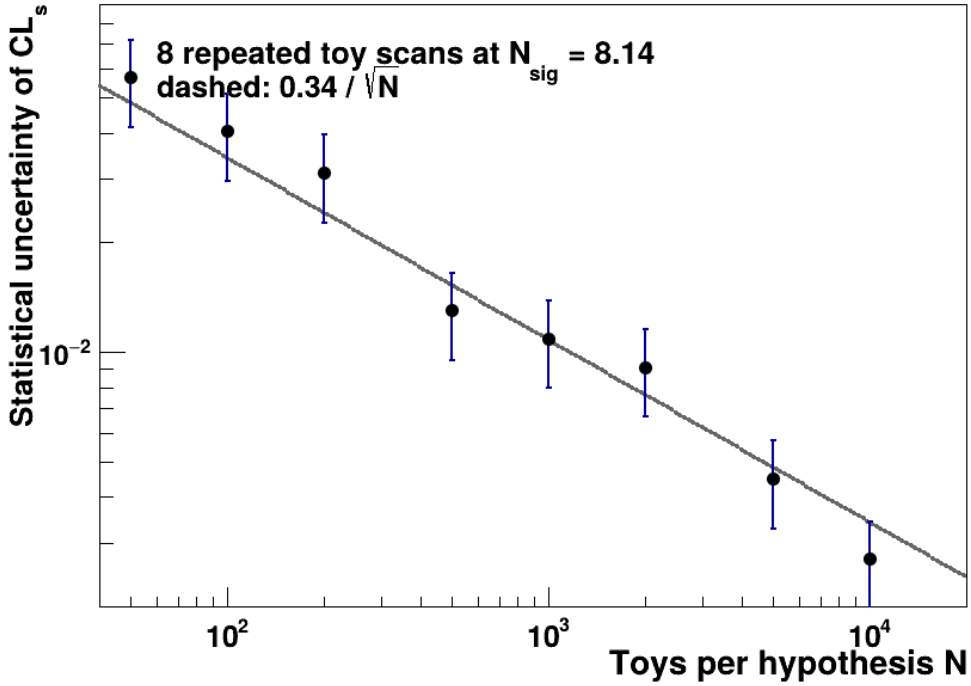


图 10 固定测试产额 8.14，CL_s 的统计误差随 toy 数的变化。虚线是 $1/\sqrt{N}$ 参考。

每组的 toy 数	50	200	1000	2000	10000
σ_{CL_s}（8 次重复实测）	0.057	0.031	0.011	0.009	0.003

课堂脚本用 2000：$\sigma_{CL_s} \approx 0.008$，足以分辨 0.10 与 0.12 的差别——第 16 页那个 0.297 ± 0.012 的误差就来自这里。继续加大收益按 $1/\sqrt{N}$ 递减：从 2000 到 10000 误差只压到 1/2.2，计算量却翻 5 倍。

因此报告用 toy 扫描得到的上限时，应该同时说明每个点用了多少 toy；否则交点的小数位没有意义。

上限对物理结论有什么用?

8.06 个事例怎么理解? 质量谱里的信号产额是模型预测的平均重建数，可以有小数。它描述我们在这套选择条件下能约束多大的信号峰。比如某个物理模型在同样的质量、峰形和选择条件下预测平均重建 12 个信号事例，它落在本例观测曲线的排除一侧，可以按这个准则排除该预测。

实际分析通常还要把它换成物理量。例如已知产生了多少母粒子、信号有多大比例能被重建和选中，以及所需的子衰变分支比，就能把「平均能重建多少个信号」换算成「这个衰变的分支比允许有多大」。实际换算时，用的是对应分析的归一化和系统误差处理。

同一张质量谱，两种上限

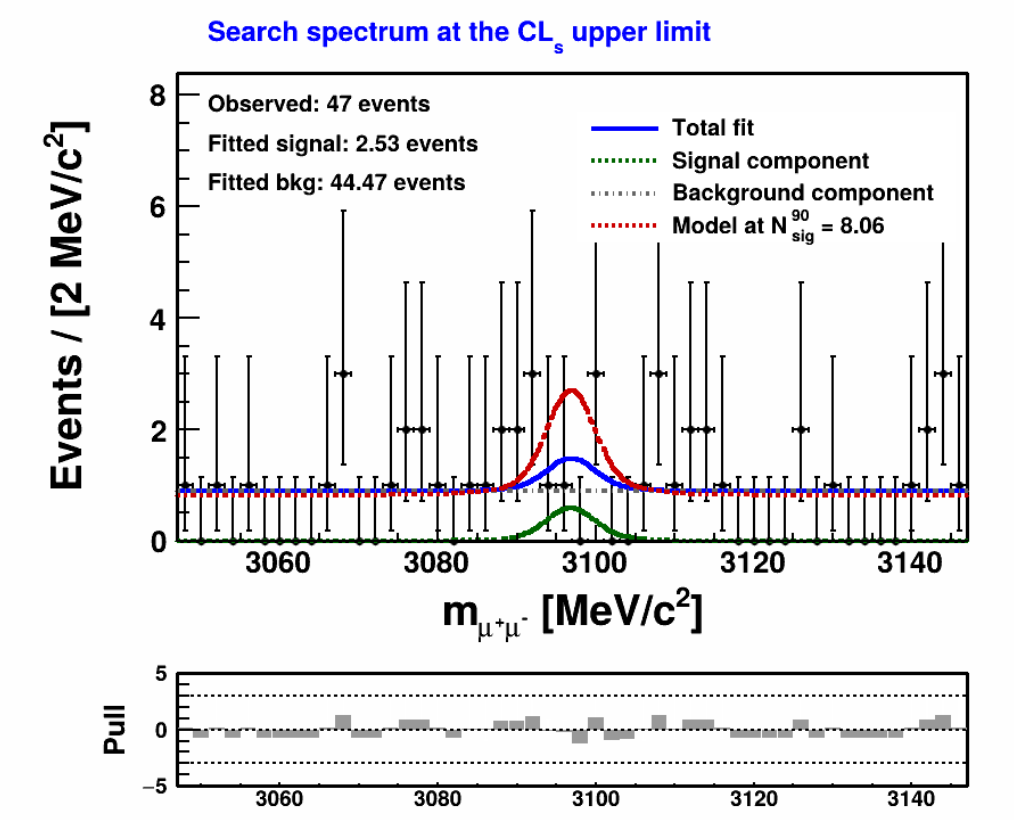


图 11 把 CL_s 观测上限叠回质量谱。红线使用固定在上限的信号产额，并重新拟合本底。

这份课堂模拟数据用轮廓似然积分得到 8.94 个事例，用伪实验 CL_s 得到 8.06 个事例。积分方法找的是，从信号产额为零开始累计，哪里达到总面积的 90%。CL_s 方法是在每个测试产额上做排除检验，找观测曲线降到 0.10 的位置。两种方法的计算规则不同，所以给出的数字也可能不同。

图上的红线把 8.06 放回了质量谱。我们把信号产额固定在这个值，再让本底重新拟合，画出这时的信号加本底模型。这样大家可以回到最初那张质量谱，从曲线的交点理解上限，也能直观看到这个上限对应多大的信号峰。

参考文献

[1] A. L. Read, *Presentation of search results: the CL_s technique*, J. Phys. G: Nucl. Part. Phys. **28** (2002) 2693–2704. doi:10.1088/0954-3899/28/10/313.

[2] G. Zech, *Upper limits in experiments with background or measurement errors*, Nucl. Instrum. Methods Phys. Res. A **277** (1989) 608–610. doi:10.1016/0168-9002(89)90795-X.

[3] M. Mandelkern, *Setting confidence intervals for bounded parameters*, Statist. Sci. **17** (2002) 149–159. doi:10.1214/ss/1030550859.

[4] T. Junk, *Confidence level computation for combining searches with small statistics*, Nucl. Instrum. Methods Phys. Res. A **434** (1999) 435–443. arXiv:hep-ex/9902006.

[5] G. Cowan, K. Cranmer, E. Gross, O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. C **71** (2011) 1554. arXiv:1007.1727.

[6] G. Cowan and L. Lista, *Statistics*, in *Review of Particle Physics* 第 40 章（2026 版，2025 年 6 月修订），pdg.lbl.gov/2026/reviews/rpp2026-rev-statistics.pdf；CL_s 的定义与讨论见 §40.4.2.5，式 (40.82)。

延伸阅读

- A. Birnbaum, *On the foundations of statistical inference*, J. Amer. Statist. Assoc. **57** (1962) 269–326. 关于条件性原理、似然原理与「证据强度」$\alpha/(1-\beta)$ 的经典论证；本讲义「除以本底那条尾巴」的直觉与这里的讨论一脉相承。
- R. N. Giere, *Allan Birnbaum's conception of statistical evidence*, Synthese **36** (1977) 5–13. 对上述证据观念的一篇简明评述，适合想弄清「CL_s 为什么不是一次假设检验」的读者。
- D. A. S. Fraser, N. Reid, A. C. M. Wong, *Inference for bounded parameters*, Phys. Rev. D **69** (2004) 033002; arXiv:physics/0303111. 讨论有界参数下频率学区间为何会「跑偏」，与本例 $s \geq 0$ 带来的困难直接相关。
- R. D. Cousins, *Negatively biased relevant subsets induced by the most-powerful one-sided upper confidence limits for a bounded physical parameter*, arXiv:1109.2023. 说明为什么单侧上限在低计数时会给出「把所有产额一起排除」的结果。
- RooStats 类参考手册：`https://root.cern.ch/doc/master/namespaceRooStats.html`。第 18、23–24 页的 toy 演示与 toy 数对比，用的是其中的 `FrequentistCalculator` 类。
- `AsymptoticCalculator` 与 `HypoTestInverter` 两个类，对应第 19–22 页的扫描、预期带与上限读取。