概率统计：第3次编程计算题

Due：2024/6/7

要求：打包上传**完整代码**，以及**一个pdf 文件：**包含直方图截图和关键程序截图以及说明，如题目要求理论计算，则同样需要包含计算过程。编程语言不限，c/c++/ROOT,Python,Matlab等均可。

1. Higgs粒子的质量测量（100pt）

- 本题中我们将利用伪数据，和似然函数的方法测量Higgs粒子质量（包括误差）；
- 在参考文章中， https://arxiv.org/pdf/2002.06398.pdf， 总的数据被分成了7个独立的种类(category); 最终的结果基于联合似然函数（Combined Likelihood)

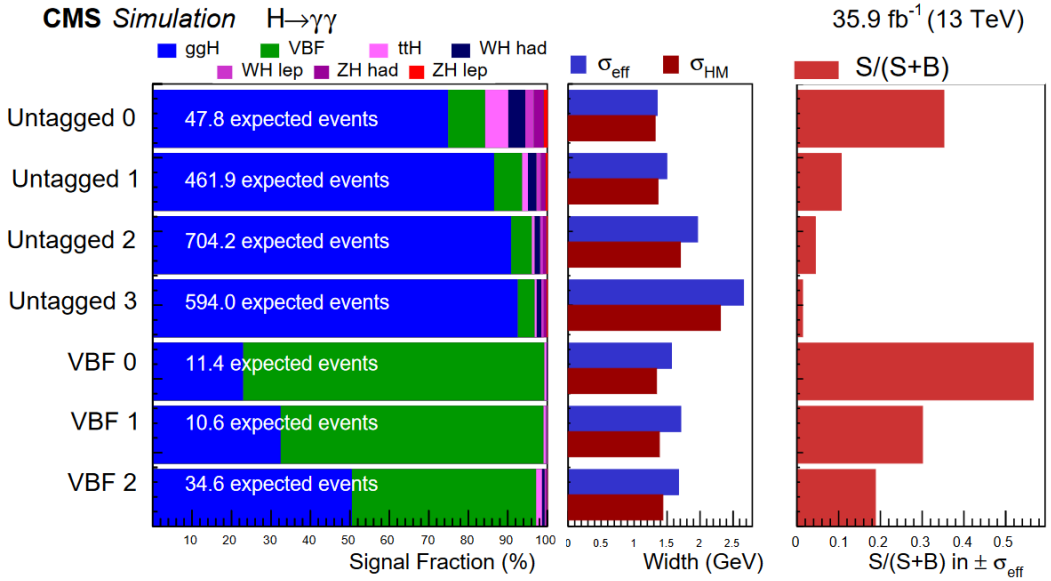

Figure 6: The expected number of signal events per category and the percentage breakdown per production mode. The σ_{eff} value (half the width of the narrowest interval containing 68.3% of the invariant mass distribution) is also shown as an estimate of the $m_{\gamma\gamma}$ resolution in that category and compared directly to the σ_{HM}. The ratio of the number of signal events (S) to the number of signal plus background events (S+B) is shown on the right-hand panel.

数据和模型

category	ns	σ (GeV)	nb
0	48	1.4	1520
1	462	1.5	66180
2	704	2.0	201599
3	594	2.7	257028
4 (VBF 0 in last page)	11	1.6	123
5	11	1.6	382
6	35	1.7	2238

- 每一个category的伪数据从oc上下载，包含ns个信号+nb个本底
- 本题中假设本底概率密度函数为（每一个category都一样）

$$f_b = c \times m^{-4.5}$$

c是一个你需要在代码中计算的常数（保证在[100, 180]满足概率归一化的要求，保留足够计算精度）

- 每一个category中，假设信号满足正态分布$\mathrm{N}(\mu, \sigma^2)$

似然函数

- 有了密度函数，每一个category中，似然函数就可以定义为

$$L_k = \prod_{i=1}^{n}(fs \times pdf_s + (1 - fs) \times pdf_b)$$

$k = 0, \ldots, 6$ 代表哪一个category

n 是这一个category的总的数据个数(上一页中ns+nb)

fs 即信号比例,ns/(ns+nb);

[说明：真正的分析中，这一个数实际上需要从数据中拟合得到，似然函数也需要稍微修改；本题中为简化起见，看成已知的]

pdf_s and pdf_b 指的是信号和本底的概率密度函数(特别注意pdf_b 归一化要求），这样构造得到的似然函数的每一项也自动保证是一个密度函数

- 7个category 联合的似然函数为

$$L_{comb} = \prod_{k=0}^{6}(L_k)$$

计算要求

- 编写代码计算似然函数的对数 $\ln L(\mu)$；
- 利用每一个category的数据，分别描绘一个 $2\,(\ln L(\hat{\mu}) - \ln L)\; vs.\; \mu$的曲线；纵轴覆盖0到4.5左右；$\hat{\mu}$ 为μ的极大似然估计值
- 68.3%CL 区间采用 $2\,(\ln L(\hat{\mu}) - \ln L) = 1$ 的方法
- 采用手动扫描的方式求解极值，绘图的时候扫描步长step不超过0.01；在扫描$\hat{\mu}$附近，以及68.3%CL 区间端点的时候 step不超过0.001；（最后的$\hat{\mu}$ 以及区间估计结果保留小数点后两位有效数值）
- 也可用求解极值的软件（本题无太大必要）；但须和手动扫描结果比较；
  （用ROOT的请参考 https://root.cern.ch/doc/v608/NumericalMinimization_8C.html）
- 报告每一个category得到的$\hat{\mu}$ 以及μ的68.3%CL 区间； 哪3个category的测量精度最高？
- 利用联合的似然函数，描绘$2\,(\ln L(\hat{\mu}) - \ln L)\; vs.\; \mu$的曲线;报告最终的 $\hat{\mu}$ 以及μ的68.3%CL区间； 和文章中的结果比较 125.78 ± 0.21，如下

$(\mathrm{ggH}, \mathrm{t\bar{t}H}) \to \gamma\gamma$ and $(\mathrm{VBF}, \mathrm{VH}) \to \gamma\gamma$ processes are free to vary. The best-fit mass of m_{H} is observed to be $m_{\mathrm{H}} = 125.78 \pm 0.18\,(\mathrm{stat}) \pm 0.18\,(\mathrm{syst})\,\mathrm{GeV}$, while it was expected to have a statistical uncertainty of $\pm 0.21\,\mathrm{GeV}$ and a systematic uncertainty of $\pm 0.18\,\mathrm{GeV}$. The signal

2. “计数”实验信号的上限(50pt)

- 设本底期望为b=3.2;
- 设一个信号期望s的值，在s+b的期望下，产生n个事例；（n满足泊松分布），计算s的90%CL的上限；模拟产生1000000次，计算覆盖概率p;
- 改变s的值（从0.1 到 20，步长0.1）;重新计算覆盖概率p；描绘p vs. s 的曲线;
- 基于课堂上（ppt附页）得到的经典频率论和贝叶斯论的s的上限公式，分别计算和绘制曲线；对这两个方法计算得到的s上限，是否有足够的覆盖概率（置信度）？
- 对曲线的主要特征做一个解释（假设你在给一个报告，你如何讲）;