Statistical Methods for Experimental Physics (Part-I)
Assessment 2

Nicholas Wardle

October 2024

Answer all parts of at least question 1.1. The additional problem (marked with a $*$) is *not for credit* but you can attempt them as practice in solving these types of problems.

Your answers must be submitted in the form of a pdf converted from a .ipynb notebook. You should run the notebook so that all outputs are produced (please be careful not to print information for every toy generated or for every likelihood evaluation!) and then you can convert this to a pdf directly inside Juypter. **You must name your submission** Name_CID_A2.pdf.

Note, if you cannot convert directly to a pdf, you can first download as a .html file and then convert this to a pdf by printing to pdf in a web browser. Make sure your code is visible in the pdf, as this can cut off in the conversion process (you can orient the page to landscape to help with this).

1 Assessed Problems

1.1 The Calorimeter Resolution

A calorimeter is a generic term for a device that measures the energy of some object. Traditionally this was done by measuring small temperature changes in some material due to the energy deposited by the object, but in particle physics experiments, since the energies are very small, the devices use sophisticated processes to measure the energies of incoming particles.

Designing a high performing calorimeter is crucial in the design of a good high energy particle physics detector. Having a good energy resolution is a key design challenge.

In this question, you will need to read in a dataset and estimate the energy resolution as a function of energy from the data.

In the file data_calorimeter.csv (which you can find on Blackboard) is a dataset of recorded energy measurements of high energy electrons from a calorimeter. This is calibration data for the experiment so the true energy e_true is known for every electron. The recorded energy e_measured is provided by the calorimeter. We want to measure the resolution of the calorimeter using e_measured - e_true as a function of the true energy.

You should use a Jupyter notebook to help you structure your answers. You should clearly label any plots and present results clearly - **marks will be lost for poor quality figures or unclear results!**. You can start by using something like the cell below to import the modules you'll need.

```
%matplotlib inline

import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
from scipy.stats import norm
from iminuit import minimize
from scipy.interpolate import interp1d

plt.rcParams.update({'font.size': 10})
```

First, read in the data. You could read this in using Pandas or convert the data into some other format if you prefer. The file contains rows with different beam energy (`e_true`) values. These should be the following energies $e_{true} = 10, 15, 20, 25, 30, 35, 40, 45, 50$ GeV.

i [**1 mark**] For each `e_true` value draw draw a histogram of $e_{diff} = e_{measured} - e_{true}$.

ii [**2 marks**] We can assume that due to the resolution of the calorimeter, e_{diff} will be distributed as a Gaussian distribution $\phi(\mu_{e_{true}}, \sigma_{e_{true}})$, with parameters that are different for different values of e_{true}. For each

value of e_{true} find the maximum likelihood estimators $\hat{\mu}_{e_{true}}$ and $\hat{\sigma}_{e_{true}}$. You can do this numerically using `scipy.optimize.minimize` or any other method. HINT: remember that to find the maximum likelihood estimator, you first need to write the likelihood function $L(\mu_{e_{true}}, \sigma_{e_{true}})$ for a particular dataset and then maximise it. Remember, it's nearly always easier to minimize $-\ln(L)$.

iii [**2 marks**] Plot the resulting Gaussian distribution $\phi(\hat{\mu}_{e_{true}}, \hat{\sigma}_{e_{true}})$ on top of the histograms. Remember, for this to look right, your histograms should be drawn as *densities* (e.g. you can specify `density=True` in the plotting of the histogram). Also, plot the ratio of maximum likelihood estimator to the value of e_{true} i.e plot $\frac{\hat{\sigma}_{e_{true}}}{e_{true}}$ as a function of e_{true} .

There are a number of different contributing factors to the resolution of a calorimeter but in calorimeters like the one at the CMS experiment, one large contribution is from the limited amount of scintillation light (number of photons) which is called *stochastic noise*. The resolution can be parameterised as

$$\sigma_{e_{true}} = S\sqrt{e_{true}}$$

where S is called the *stochastic term*. With the data, we can try to estimate S.

iv [**2 marks**] Write a new likelihood function $L(S, \vec{\mu}_{e_{true}})$ *combined* across the data from different e_{true} values. Remember that the total likelihood is the product over the individual likelihoods (or the sum if using $-\ln(L)$). The vector of values $\vec{\mu}_{e_{true}}$ are still free parameters but this time you should replace $\sigma_{e_{true}}$ in the Gaussian pdfs with $\sigma_{e_{true}} = S\sqrt{e_{true}}$. Calculate the maximum likelihood estimator $\hat{S}$. HINT: You will probably find it helpful to initialise the minimization with a value around $S \approx 0.02$. You might also want to impose boundaries on the minimizer with a line similar to that below.

```
boundaries = [[0.01,0.1]].extend([[-1,2] for m in mu_guesses])
res_free = minimize(,...,bounds=boundaries)
```

v [**2 marks**] What is the standard deviation on your estimator $\hat{S}$? To obtain this, you can use a bootstrapping method on the original dataset and calculate the sample standard deviation of the bootstrap samples - this may be quite slow so you should use no more than 100 bootstrap samples.

vi [**1 mark**] Finally make a plot of the function $\frac{\hat{S}\sqrt{e_{true}}}{e_{true}}$ overlaid on the individual fits you obtained in Part 1. How well do the points agree with your curve?

2 Additional Problems

2.1 Measure the Higgs boson mass*

For this question, you're going to be measuring the mass of the Higgs boson (m_H) using data taken at the CMS experiment at the LHC.

Due to its large mass, the Higgs boson decays almost as soon as it's produced meaning we can only infer its properties from its decay products. The decay of a Higgs boson into two photons is one of the most experimentally sensitive channels for studying Higgs boson properties.

To start you'll need to download the data which is in a simple .txt file. The file is on blackboard and is called diphoton_mass.txt. Each row represents the invariant mass of a pair of photons from a candidate Higgs boson decay.

i First read in the data and plot a histogram of the invariant mass distribution.

ii Next, you should try to fit a signal and background model to the events and extract the mass. You can assume that the signal component $p_S(m_{\gamma\gamma})$ is a Gaussian with a μ parameter m_H and a σ parameter fixed to 1.64.

$$p_S(m_{\gamma\gamma}) = \frac{1}{1.64\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(m_{\gamma\gamma} - m_H)^2}{1.64^2}}$$

The background component $p_B(m_{\gamma\gamma})$ should be a power law function.

$$p_B(m_{\gamma\gamma}) = K m_{\gamma\gamma}^{-\alpha}$$

You can write the total probability density function as,

$$p(m_{\gamma\gamma}) = f_S p_S + (1 - f_S) p_B$$

You'll need to figure out the normalisation K for the power law function (which will depend on the parameter α), over the range $100 < m_{\gamma\gamma} < 180$ GeV, and find the fitted values of m_H, α and f_S, by defining and (min)maximising a (log)likelihood function $L(m_H, f_S, \alpha)$.

iii Plot the probability density with parameters set to the values from the fit on top of data histogram - does it look reasonable to you - discuss the agreement of the data and the distribution you plotted.

iv What happens if you allow for the total rate of events to vary? To check, multiply your likelihood function by a Poisson term $\frac{\lambda^N e^{-\lambda}}{N!}$, where N is the total number of events and also profile λ. Discuss whether or not the fitted value of m_H changes with this modification. Hint, you can actually ignore the $N!$ since its a constant term in the likelihood.

v Make a plot of the value of $q(m_H) = -2 \times (\ln L(m_H) - \ln L(\hat{m}_H))$ as a function of m_H, where $\hat{m}_H$ is the fitted value of m_H from part 2. The other parameters, should be profiled (which means for each value of m_H you should find the values of α and f_S that minimise the log-likelihood function at m_H). Use Wilks' theorem to estimate a 68% CL interval for m_H. We'll cover this theorem in the last lecture but for now, just estimate the 68% CL interval as the range in m_H for which $q(m_H) < 1$