Statistical methods for data analysis

PHY683

October 16, 2024

- There are two exercises; please complete and submit both. Name your program file with your name for easy identification.
- Submit the Python code by 5 PM on November 5th, 2024.
- Do not use ChatGPT, as I may ask follow-up questions about the program.
- Read up on iMinuit.
- For questions that require explanation (e.g., 1(b)), you may submit a separate sheet with your written answer.

Exercise 1

This exercise concerns maximum-likelihood fitting with the minimization program MINUIT using either its python implementation `iminuit`. The exercise is carried out by modifying and running the program `mlFit.py`.

`https://github.com/navaneethphysics/PHY683/blob/main/mlFit.py`

To use python on your own computer, you will need to install the package `iminuit` (should just work with "`pip install iminuit`"). See:

Minuit is a minimisation algorithm, read about it

`https://pypi.org/project/iminuit/`

The program provided generates a data sample of 200 values from a pdf that is a mixture of an exponential and a Gaussian:

$$f(x;\theta,\xi) = \theta\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} + (1-\theta)\frac{1}{\xi}e^{-x/\xi},$$

The pdf is modified so as to be truncated on the interval $0 \leq x \leq x_{\max}$. The program `Minuit` is used to find the MLEs for the parameters θ and ξ, with the other parameters treated here as fixed. You can think of θ as representing the fraction of signal events in the sample (the Gaussian component), and the parameter ξ characterizes the shape of the background (exponential) component.

1(a)

By default the program `mlFit.py` fixes the parameters μ and σ, and treats only θ and ξ as free. By running the program, obtain the following plots:

- the fitted pdf with the data;
- a "scan" plot of $-\ln L$ versus θ;
- a contour of $\ln L = \ln L_{\max} - 1/2$ in the (θ, ξ) plane;
- confidence regions in the (θ, ξ) plane with confidence levels 68.3% and 95%.

From the graph of $-\ln L$ versus θ, show that the standard deviation of $\hat{\theta}$ is the same as the value printed out by the program.

From the graph of $\ln L = \ln L_{\max} - 1/2$, show that the distances from the MLEs to the tangent lines to the contour give the same standard deviations $\sigma_{\hat{\theta}}$ and $\sigma_{\hat{\xi}}$ as printed out by the program.

1(b)

Recall that the inverse of the covariance matrix variance of the maximum-likelihood estimators $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$ can be approximated in the large sample limit by

$$V_{ij}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right] = -\int \frac{\partial^2 \ln P(x|\theta)}{\partial\theta_i \partial\theta_j} P(x|\theta)\, dx,$$

where here θ represents the vector of all of the parameters. Show that V_{ij}^{-1} is proportional to the sample size n and thus show that the standard deviations of the MLEs of all of the parameters decrease as $1/\sqrt{n}$. (Hint: You can think of Fisher information and connection with Variance, and also remember log-likelihood function is a sum over the data points, the second derivative will also will be summed over the data point.) [you don't have to do anything in the program, just comment and you will understand it when doing 1c]

1(c)

By modifying the line

```
numVal = 200
```

rerun the program for a sample size of $n = 100$, 400, and 800 events, and find in each case the standard deviation of $\hat{\theta}$. Plot (or sketch) $\sigma_{\hat{\theta}}$ versus n for $n = 100$, 200, 400, 800 and comment on how this stands in relation to what you expect.

1(d)

By modifying the line

`parfix = [False, True, True, False]# change these to fix/free parameters`

find $\hat{\theta}$ and its standard deviation $\sigma_{\hat{\theta}}$ in the following four cases:

- θ free, μ, σ, ξ fixed;
- θ and ξ free, μ, σ fixed;
- θ, ξ, and σ free, μ fixed;
- θ, ξ, μ, and σ all free.

Comment on how the standard deviation $\sigma_{\hat{\theta}}$ depends on the number of adjustable parameters in the fit.

Exercise 2

1. Generate 100 data points from an exponential distribution with a probability density function (PDF):

$$f(t|\tau) = \frac{1}{\tau} e^{-t/\tau},$$

where τ is the true parameter of the distribution. For this assignment, τ is set to 1.

do the following

(a) **Generate Data:**

- Use Python to generate 100 random data points from the exponential distribution with $\tau = 1$.

(b) **Plot the Distribution:**

- Plot a histogram of the generated data points and overlay the theoretical probability density function (PDF) for $\tau = 1$.

(c) **Fit the Data:**

- Perform a maximum likelihood estimation (MLE) to fit the generated data and determine the best-fit value of τ.
- Compare the estimated τ to the true value of $\tau = 1$.