Introduction to Data Processing

Radha Mastandrea
APS DSECOP Conference
06/22/2022

Berkeley
UNIVERSITY OF CALIFORNIA

The modules: Introduction to Data Processing

- The repository contains materials aimed to provide a high-level introduction to data processing as a tool to make developments in theory
- The material is aimed towards lower-year undergraduates with some exposure to coding
  - It might be a good accompaniment to a laboratory course, or a course in particle phenomenology
- The repository includes in-class assignments, quiz questions, and a longer homework assignment (all as jupyter notebooks)

In-class modules

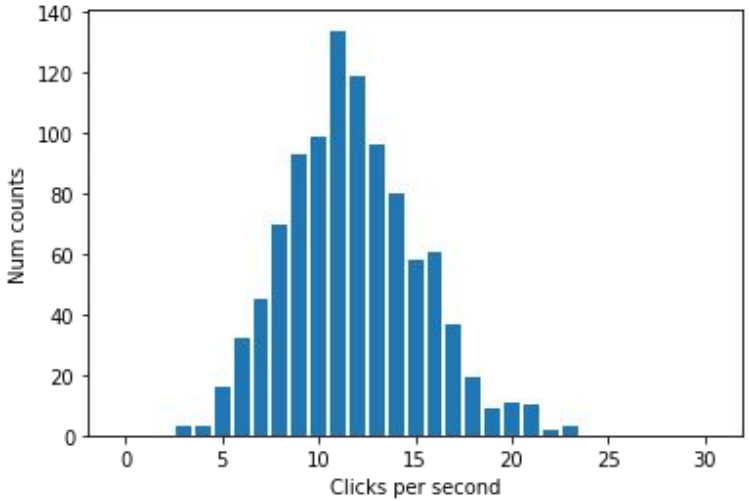
- Three in-class modules will guide students through the analysis of the radioactive decays of a toy “Uranium-241” sample
- Two datasets are provided for use in these modules
  - These are CSV .txt files containing a sample (~5000) of Poisson distributed values, meant to represent the number of clicks heard per minute from a Geiger counter placed near the sample

Module 1: Introduction to Histograms

Goal: Introduce the concept of a **histogram** as a tool to visualize large data sets

Students will practice creating histograms, both through writing their own code from scratch and from using built-in `numpy` / `pyplot` functions.

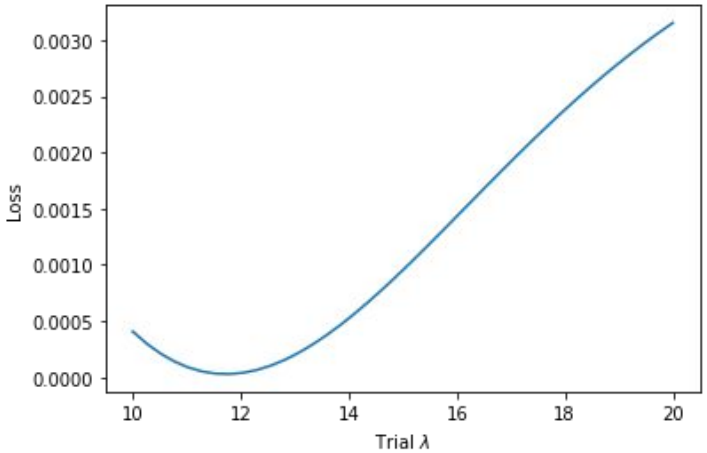
Ideally, students will write code to produce plots like this



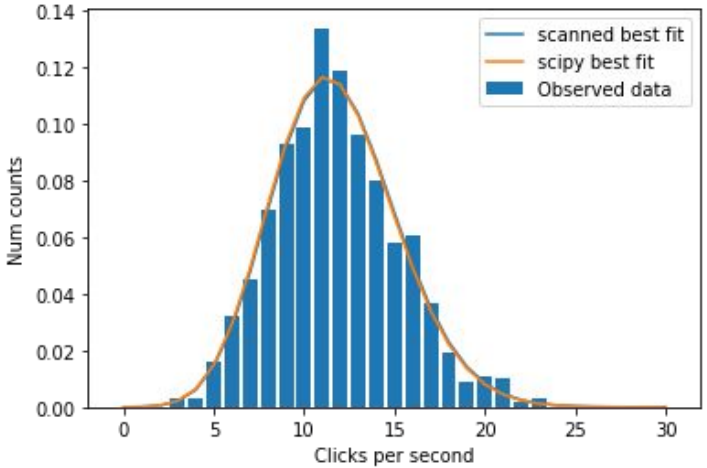
Module 2: Introduction to Curve Fitting

Goal: Introduce the concept of **curve fitting** (including the idea of a **loss function**) to extract model parameters from a histogram

Students will practice curve fitting to a histogram, both through writing their own code from scratch and from using built-in `numpy` / `pyplot` functions



Ideally, students will write code to produce plots like these



Module 3: Evaluating goodness-of-fit

Goal: Explore how a given **fit to data can be evaluated**, specifically though the **χ^2 test statistic** and the **Kolmogorov-Smirnov test**. Students will also learn about the language of **hypothesis testing**.

Students will learn about the theory behind the χ^2 test statistic, calculate it for a specific dataset, and use it to accept or reject a certain hypothesis. Students will also apply the Kolmogorov-Smirnov test to the same dataset.

```
$\chi^2$ for gaussian fit = 84.49167839824504 , with dof = 29
$\chi^2$ for poisson fit = 26.188485487957443 , with dof = 30
```

Quiz: Conceptual questions

Goal: Expose students to problems that **challenge implicit assumptions** they might have developed while completing the modules

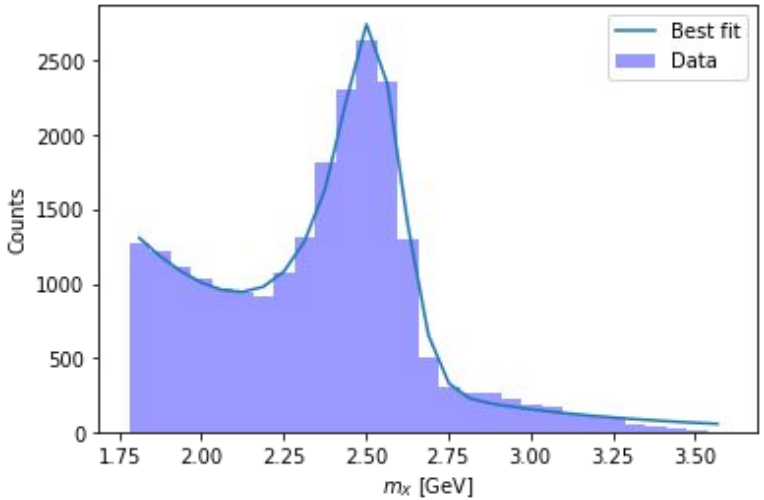
Questions explore…

- how many bins is optimal for a given histogram
- the limitations for using the χ^2 test statistic
- the nuances of the language of hypothesis testing

Homework: *Searching for particle resonances*

Goal: Provide a **realistic data analysis** task to students that draws upon the material introduced in the in-class modules.

Students will be given a dataset of $ee \rightarrow abcd$ collisions. They will be told that the decay particles are produced through a mystery particle X. They will be guided through the task of calculating and histogramming m_X candidates, fitting the distribution, and comparing their recovered value of m_X with the value that their “theorist friend” has predicted.



Areas for improvement

- Module 3: while the χ^2 test statistic is taught in virtually every intro stats class, I have found that it's not always a useful metric (especially in particle physics). I do introduce the Kolmogorov goodness-of-fit test, but that test isn't very common either…what tests are most common in your fields?
- Quiz questions: I would like to add a few more conceptual questions.
- Currently, the notebooks are self-contained. Would separate slides be useful?

Thank you to Prof. Simone Pagan Griso for writing the code skeleton to generate the homework module dataset!

Module feedback: bit.ly/DSECOP-feedback