Problem Set #2

D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2016

Photo by Howard Jackman

Info

- The following problem set is due on, or before, Friday April 8 at 16:00 Copenhagen time via email to koskinen@nbi.ku.dk
  - The write-up should be in a PDF text document
  - Include the code with the email submission
    - Zipped files, iPython notebooks, .C files, etc. are all fine
  - I'll also accept submission up until 17:00 too ;)
- The work can be done in groups, but each student should submit their own write-up and supporting software code

Question 1

- Consider an experiment set up to measure the lifetime of an unstable nucleus, N, using the reaction: $A \to Ne\bar{\nu},\ N \to Xp$
- The creation and subsequent decay of N has a signature of an electron and proton. The lifetime of each N, that follows the PDF $f = \frac{1}{\tau}e^{-t/\tau}$ is measured from the time, observing the electron and proton with a gaussian resolution of σ_t
- The expected PDF is then the convolution of the exponential decay and the gaussian resolution is then:

$$f(t;\tau,\sigma_t) = \int_0^\infty \frac{e^{-\frac{(t-t')^2}{2\sigma_t^2}}}{\sqrt{2\pi}\sigma_t}\frac{e^{-t'/\tau}}{\tau}dt'$$

Question 1a

- Generate 200 events with τ=1 s and σ_t=0.5 s. Use the maximum likelihood method to find $\hat{\tau}$ and the uncertainty $\sigma_{\hat{\tau}}$. Plot the likelihood function in 1D as a function of τ. Separately, plot the resulting probability distribution function for the measured times compared to a histogram of the data, both on the same plot.
- Automate the maximum likelihood procedure to repeat this 100 times and plot only the distribution of $\frac{(\hat{\tau}-\tau)}{\sigma_{\hat{\tau}}}$ for the 100 experiments.
- Use the file of event times at (http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2016/data/ProblemSet2_NucData.txt) to:
  - Fit the values of $\hat{\tau}$ and $\hat{\sigma}_t$ using the same exponential and gaussian smeared PDF
  - Plot the 2D likelihood, or LLH, landscape using the best fit values from the file online over a reasonable range of $\hat{\tau}$ and $\hat{\sigma}_t$

Question 1b

- Using the best-fit values derived from the online file:
  - Plot the 50%, 68.3%, and 99.995% confidence intervals in 2-dimensions for τ and σ_t
  - What are the 1D numerical uncertainties for the same fits, i.e. $\sigma_t \pm x$ and $\tau \pm y$
- Note: after looking through some of the submissions, it became clear that there was some online file cross-contamination. As such, there are actually 'two' sets of solutions depending on which file was downloaded and at what times.

Question 2

- A pace-maker is generated by assembly facilities in 5 different countries with the following defective rates and total worldwide production percentage

Facility	Total % produced	% Defective
A_1	35	2
A_2	15	4
A_3	5	10
A_4	20	3.5
A_5	25	3.1

Question 2a

- The world-wide distribution of the actual pace-makers is decoupled from the proximity to the production facilities
- Suppose a pace-maker is found to be defective (D), what is the probability it came from the A_2 facility, i.e. $P(A_2|D)$?
- If a defective pace-maker is found, which facility is it most likely to be from?

Question 2b

- The CEO of Slightly Evil Inc. wants to ensure that all the facilities have the same probability of being identified with a failed pace-maker, but still wants the least total defects and with no changes to the per facility production. How must the percentage of defects change from each facility to accommodate this goal?
  - Make a list of the altered/updated defective rates for each facility
  - Defective product rates can only increase

```
W/ defective, the updated defective rate for A1: XXXX
W/ defective, the updated defective rate for A2: XXXX
W/ defective, the updated defective rate for A3: XXXX
W/ defective, the updated defective rate for A4: XXXX
W/ defective, the updated defective rate for A5: XXXX
```

Question 2c

- Repeat question 2b using the new table below, which is in fractions and not percent

Facility	total produced	defective rate
A1	0.27	0.02
A2	0.1	0.04
A3	0.05	0.1
A4	0.08	0.035
A5	0.25	0.022
A6	0.033	0.092
A7	0.019	0.12
A8	0.085	0.07
A9	0.033	0.11
A10	0.02	0.02
A11	0.015	0.07
A12	0.022	0.06
A13	0.015	0.099
A14	0.008	0.082

Question 3a

- Using the ice core data do a linear fit to find the slope and intercept (hint, regression):
  - http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2016/data/optical_log.up1.dat
  - Using the entries from 100000 to 140000, e.g. approx. depths from 1540.45 meters to 1602.21 meters
    - No outlier cleaning!
- Along with showing the equation of the linear fit, plot the data including the linear fit line
- Calculate the Pearson's chi-squared using the data and best-fit line for the entries from 100500 and 139500

Question 3b

- Using the same data and entries from 100000 to 140000, create a linear spline using every 100th entry
  - 400 total points
  - Entries 100100, 100200, 100300, 100400, etc.
- Calculate the Pearson's chi-squared using the data and linear spline for the entries from 100500 and 139500
  - All 39000 data points, including the 400 points used for the spline knots (which by construction should contribute zero to the `chi-squared')
  - For the uncertainty use sqrt(N), where N is the dust logger data which is in intensity
  - Is the chi-squared quantitatively meaningful for the spline 'fit' compared to the data? Explain your conclusion.

Question 4

- Using the same likelihood function and data from Question 1, use a Markov Chain Monte Carlo to estimate the values of $\hat{\tau}$ and $\hat{\sigma}_t$
  - Use the online file at (www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2016/data/ProblemSet2_NucData.txt)
  - Use a flat prior for both variables