Problem Set 2

D. Jason Koskinen

koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2025

Photo by Howard Jackman

Information

- The submission is:
  - A write-up as a PDF document, which includes any plots, diagrams, tables, pictures, and explanations
  - In a separate “file”, submit all code used to derive the results
    - Tarball, zipped directory, lots of individual files w/ self-explanatory titles, etc.
    - Do NOT include lines of code in your write-up. If results are dependent on coding choices then include those comments in the write-up.
  - Include any original data files or how the data was accessed
    - If you use a internet scraping tool, note the date when you retrieved the data
    - If you can save the data to a file, do so and submit the data file. There is no need to change the format, e.g. HTML, XML, txt, JSON…
- Usage of AI, e.g. ChatGPT, is permitted to assist with coding and explaining concepts, but is expressly forbidden from being used to directly answer any of questions in the assignment.

Problem 1 (4 points total)

- Generate 2 separate sets of Monte Carlo data with the following features:
  - Using a PDF which is proportional to $f(x|\alpha,\beta) = 1 + \alpha x + \beta x^2$ (remember about normalizations to ensure that this function is a PDF)
    - x has the range from x_{min}=-1.02 to x_{max}=+1.11
    - True values $\alpha = 0.9$ and $\beta = 0.55$
    - 807 generated data points
      - For those doing "accept/reject" this equates to 807 accepted points.
  - Using a PMF which is proportional to a Poisson distribution
    - True poisson variable $\lambda = 3.8$
    - x goes from x_{min}=0 to x_{max}=+∞
    - 513 generated data points
- Find the maximum likelihood estimator (MLE) 'best-fit' values for $\hat{\alpha}$ and $\hat{\beta}$ for the first data set, and $\hat{\lambda}$ for the second data set.
- Submit the two Monte Carlo generated data sets along with your MLE estimates of $\hat{\alpha}$ and $\hat{\beta}$, and $\hat{\lambda}$.
  - The data sets will be checked that 1) they are consistent with true PDFs (to within statistical fluctuations), and 2) the MLE best-fit values are consistent between the student write-up and a cross-check on the data by the grader.

Problem 1 (4 points total) cont.

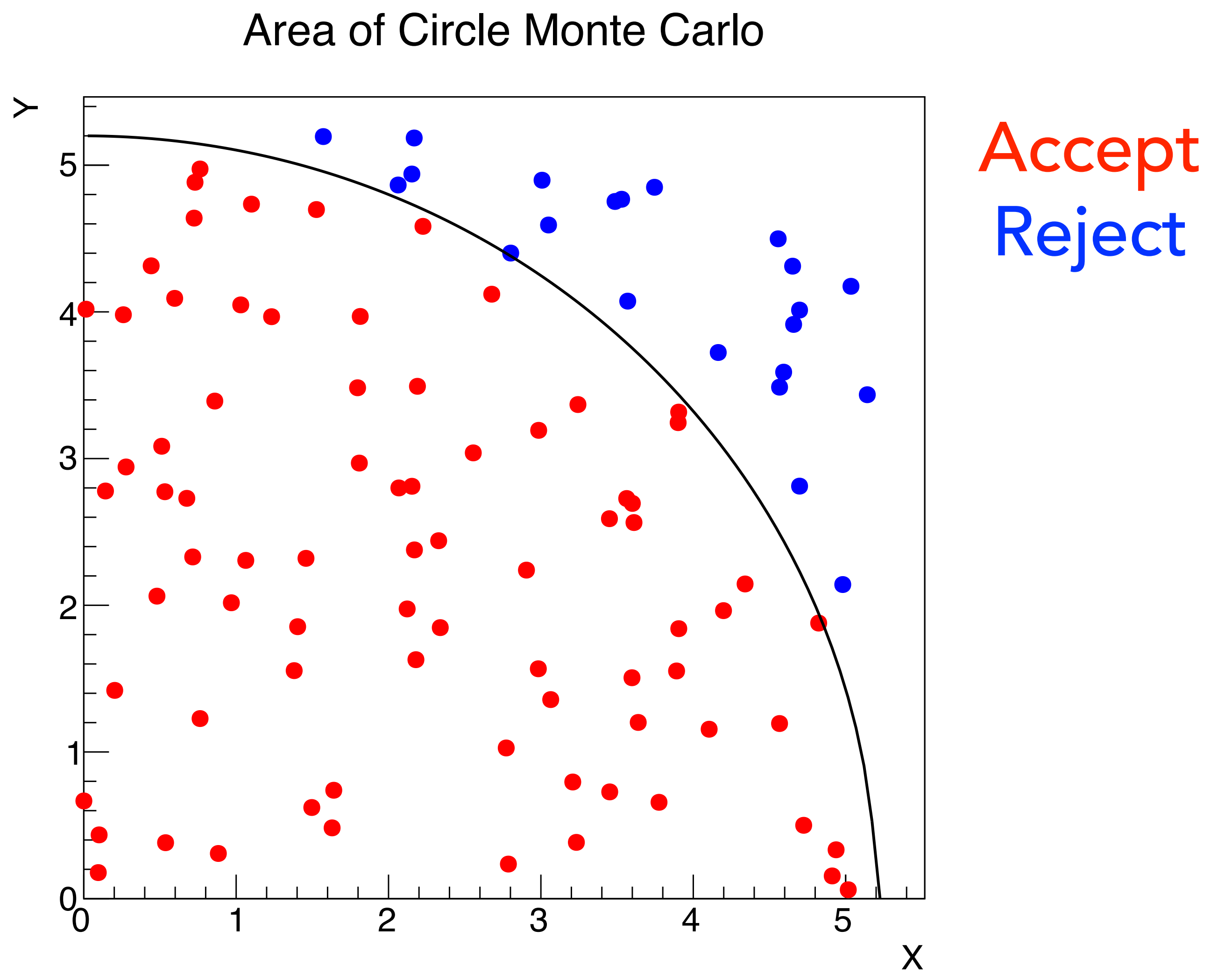
- Include plots of the histogrammed MC data as well as the best-fit function for each function. Thus, 1 set of plots for the $f(x \mid \alpha, \beta) = 1 + \alpha x + \beta x^2$ Monte Carlo and best-fit function and another set of plots for the Poisson.
  - The best-fit function plot should be appropriately scaled to be visually comparable to the histogrammed data.
- Explain how you ensure that the PDF is normalized when doing the fitting for $\hat{\alpha}$ and $\hat{\beta}$ for the first data set, and $\hat{\lambda}$ for the second data set.
- Save the Monte Carlo data sets as two separate .txt files with each data point on a separate line.
  - Files names should be “lastname_polynomial.txt” and “lastname_poisson.txt” all lower-case, e.g. “koskinen_polynomial.txt” and “koskinen_poisson.txt”.
- The file format should be very easy and straightforward to import the data in the files for the grader(s).

Problem 2 (1.5 points)

- There is a file posted online which has the data points (x, y) that provide the outline of a contained area.
  - The file is at https://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2025/data/OutlineAreaSpline.txt
  - The outline is formed by linear interpolation between the data points.
  - The online data is in the correct and specific order to form the outline.

Problem 2 (cont.)

- Using Monte Carlo techniques, estimate the area that is contained within the outline.
- Include a visualization of the technique.



Included as an example visualization of Monte Carlo integration of a circle

Problem 3 (3.0 pts.)

- Suppose that there are genes which are individually ‘x' or ‘X’, and in combination determine some trait, e.g. hair color: xx is red, mixed genes (xX or Xx) are black, and XX is black. The population has a proportion of red-haired people equal to p^2 and mixed gene people equal to 2p(1-p), for $0<p<1$. Each parent gives a single gene to their offspring, with a 50:50 probability of x or X for mixed gene parents. We can assume a random mixture of parents within the population.

Problem 3a (1.5 pts.)

- Of children that are xX what is the proportion that come come from parents which both have black hair?
  - Hint this is a conditional probability.
  - The ordering of the gene pairs is irrelevant, e.g. xX=Xx.

Problem 3b (1.5 pts.)

- A person (parent A) that does have black hair and has parents w/ black hair produces N offspring w/ someone (parent B) that is known to have a xX gene combination. What is the posterior probability that parent A has a xX gene combination?
  - All N offspring have black hair.
  - The ordering of the gene pairs is irrelevant, e.g. xX=Xx.
  - Hint: “Chain Rule”

Problem 4 (1.5 points)

- In an earlier lecture we used a prior for the total number of fish in a lake that was based on the ratio (and related uncertainties) of the volume of a lake and the volume that fish singly occupied. The combined uncertainty, and the mean for the prior, do not follow canonical error propagation or simple estimates.

Problem 4a (1 point)

- Find the mean for the total fish population and the range of fish population which covers the interquartile range.
  - Lake volume estimate is gaussian with $5000 \pm 300\ m^3$
  - Fish volume estimate is gaussian with $10 \pm 1\ m^3$
  - The interquartile range covers the central 50% of the distribution, i.e. the range from 25% to 75%.
  - The distribution will be slightly non-gaussian, so do not assume that:
    - the uncertainties, and therefore the interquartile ranges, are symmetric
    - the distribution is gaussian

Problem 4b (0.5 pts.)

- The estimated mean from problem 4a very likely has an uncertainty if your solution used using Monte Carlo technique(s). Regardless of how you came to your answers for problem 4a, if a separate analysis suggested that the mean of the fish population was +4 larger than your estimated mean, i.e. $\bar{N}_{separate\ analysis} = \bar{N}_{your\ result} + 4$, would you consider this reasonable?
    - Explain and justify your conclusions