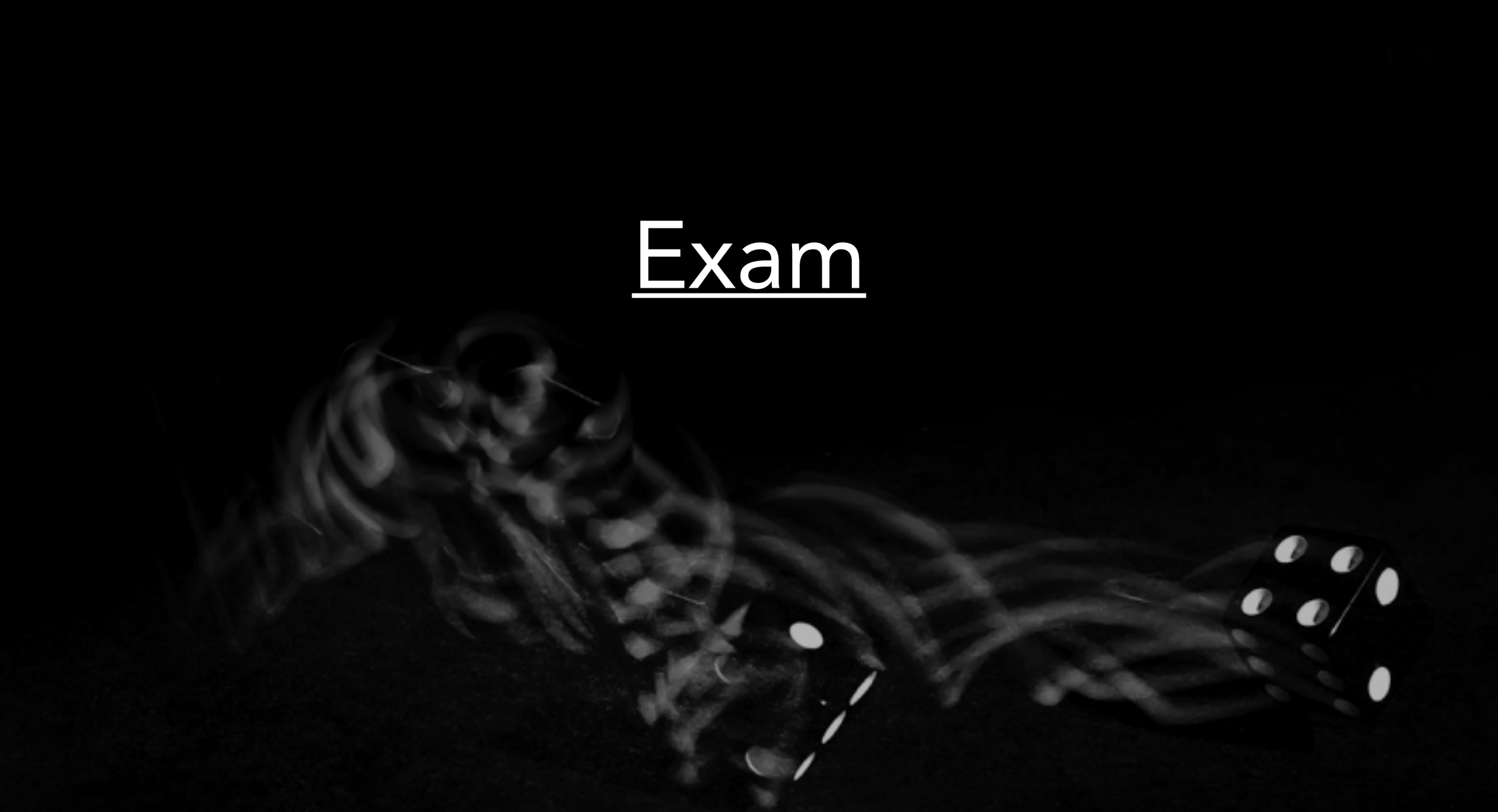
Exam



D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2020

Photo by Howard Jackman

Info

- In submitting the solutions there is no need to rephrase the problem. “Solution for 1a” is sufficient.
- The submission format for explanations and plots is a PDF file. Also, include any and all software scripts used to establish your answer(s) and/or produce plots in a **separate** file(s).
- Working in groups or any communication about the problems is **prohibited**. Using the internet as a resource is encouraged, but soliciting any help is also prohibited.
- Some questions have multiple parts. For full credit, all parts must be done.

Info

- The exam will be graded out of 10 possible points
  - It will count for 40% of the final course grade
- Submit all code used!! The software you write to complete the problem is **part** of the solution.
- The exam **MUST BE** electronically submitted via the Digital Exam website.
  - For catastrophic submission failures you can email the exam submission to Jason
- For any concerns, questions, or comments email Jason (koskinen@nbi.ku.dk)

Starting points (0.5 pts.)

- On the first page of your write-up include your full name, date, name of this course, UCPH ID, and the title of your exam submission
- Also type out (please don't copy/paste) " I (your name here) expressly vow to uphold my scientific and academic integrity by working individually on this exam and soliciting no direct external help or assistance."
- Finding help/solutions online is fine. But, for example, posting to a forum and receiving assistance is not okay.
- Good luck!!!

Problem 1 (3.0 pts.)

- There is a file posted online which has 5 columns, each representing a physical observable of interest generated from some underlying function. There are 5000 entries, i.e. rows.
  - http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2020/data/Exam_2020/Exam_2020_Prob1.txt
- The variables/columns are independent distributions with **no** correlation to the data in the other columns
- Be mindful about accounting for truncated ranges, as well as likelihood functions that have periodic components which will create local minima/maxima
  - There is at least one column of data which is generated from a function with local minima/maxima

Lists of Distributions

- The data in each column is produced from functions ***similar to***, or potentially exactly the same as, f(x) or f(k) shown at right
- Note that the displayed functions may be unnormalized
  - Hint: Some will require a normalization to convert them to probability distribution functions
  - The functions f(x) have bounds on their parameters a, b, and c

$$-10 \le a \le 10$$
$$-10 \le b \le 10$$
$$4000 \le c \le 8000$$

$$f(x) \propto \begin{cases} \dfrac{1}{x+5}\sin(ax) \\ \sin(ax)+1 \\ \sin(ax^2) \\ \sin(ax+1)^2 \\ x\tan(x) \\ 1+ax+bx^2 \\ a+bx \\ \sin(ax)+ce^{bx}+1 \\ e^{-\frac{(x-a)^2}{2b^2}} \end{cases}$$

$$f(k) \propto \begin{cases} \binom{n}{k}p^k(1-p)^{n-k} & binomial \\ \dfrac{\lambda^k e^{-\lambda}}{k!} & poisson \\ \dfrac{-1}{\ln(1-p)}\dfrac{p^k}{k} & logarithmic \end{cases}$$

Problem 1a

- Use the separate data from columns 1, 2, and 3 to identify the function on the previous slide from which each was generated. Find the *best-fit values* and *uncertainties* on those values for the distribution using a *likelihood method* (either bayesian or maximum likelihood is fine)
  - E.g. if f(x)=sin(ax+b)*exp(-x+c)+x/k! were one of the functions, then find the best-fit values for a, b, c, and k and their uncertainties
  - Degeneracies exist, e.g. sin(x)=cos(a+x), which can produce functionally identical data distributions
  - Any function, with associated best-fit parameters which is **statistically compatible** with the data in the files will be accepted as a proper solution. Only one solution is necessary, but needs to be **justified** as statistically compatible.
- Data in column 1 and 2 have artificially truncated ranges
  - Column 1 is only sampled in the independent variable from 20 to 27
  - Column 2 is only sampled in the independent variable from -1 to 1

Problem 1b

- Plot the data and the corresponding best-fit function on the same plots
  - 3 separate 1-dimensional plots
  - Plot as a function of the independent variable
  - Histogram the data, and scale the best-fit function to be ‘reasonable’ so that the features of both the data and best-fit function can be visually compared

Problem 2 (2.0 pts.)

- There is a file posted online (http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2020/data/Exam_2020/Exam_2020_Problem2.txt) with data.
  - The first column is the azimuth angle of the data point
  - The second column is the zenith angle of the data point
  - There are 100 paired data points in total
  - The values are in units of radian

Problem 2a

- Quantify whether the data is spherically isotropically distributed
  - Include any supporting plots, discussion, and numbers
  - A spherically isotropic distribution is uniform in the azimuth angle from 0 to 2π, and uniform in cos(zenith angle) from -1 to 1
  - Hint: you can use Monte Carlo generated pseudo-experiments to produce a test-statistic distribution of a spherically isotropic distribution.

Problem 2b

- Test whether the data fits the two following alternative hypotheses better than the isotropic hypothesis:
  - Hypothesis A: That 20% of the total sample is uniformly distributed in azimuth over the range {0.225π, 0.55π} and uniformly distributed in zenith over the range {0.30π, 1π) and the remaining 80% is fully isotropic
  - Hypothesis B: That 15% of the total sample is uniformly distributed in azimuth over the range {0π, 1π} and uniformly distributed in zenith over the range {0.5π, 1π) and the remaining 85% is fully isotropic.
  - Report the two p-values: $H_{isotropic}$ versus H_A as well as $H_{isotropic}$ versus H_B

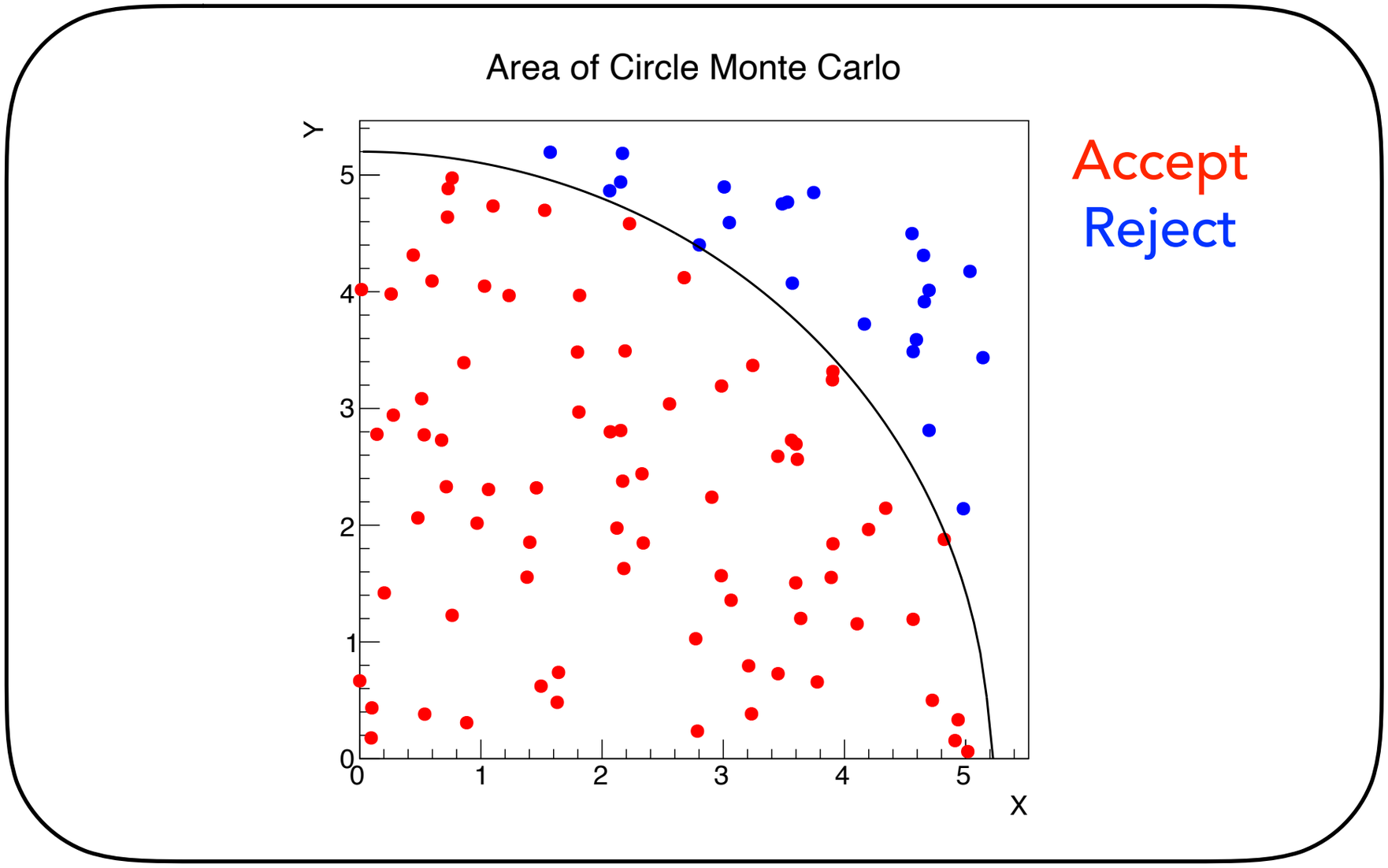
Problem 3 (1.5 points)

- There is a file posted online (http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2020/data/Exam_2020/Exam_2020_Problem3.txt) which has the data points (x, y) that provide the outline of a contained area.
  - 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 3a

- 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 4 (3.0 pts.)

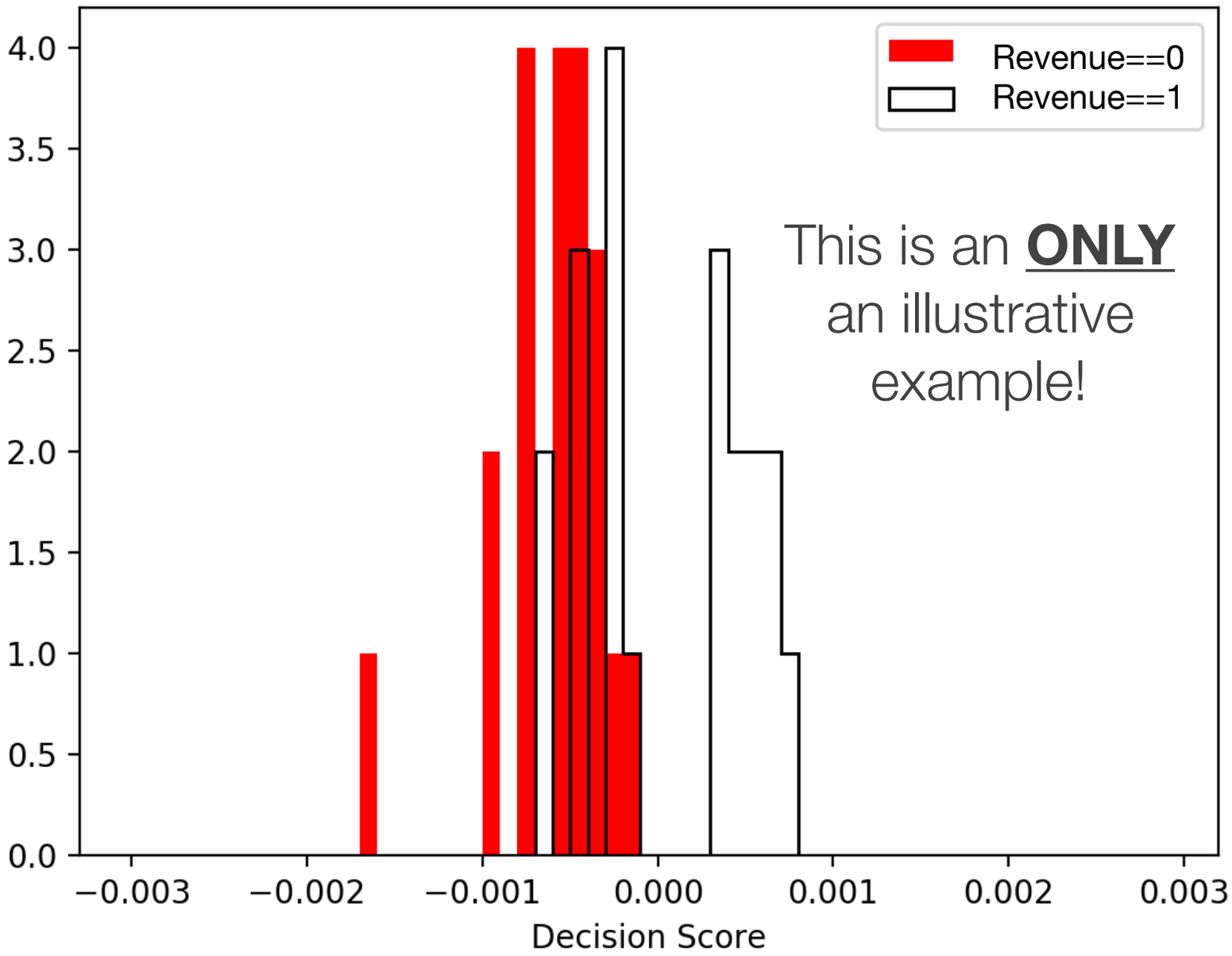
- Anonymous data was collected regarding whether online activity results in revenue, e.g. purchases at a website
- Create a classifier trained on the training data files which separates those online user sessions which do create revenue from the online user sessions which do not create revenue
- The data set has been divided:
  - Training/Testing data set is at:
    - http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2020/data/Exam_2020/Exam_2020_Prob4_TrainData.csv
    - http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2020/data/Exam_2020/Exam_2020_Prob4_TestData.csv
  - The 'blind' analysis data set is at http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2020/data/Exam_2020/Exam_2020_Prob4_BlindData.csv
    - Only used in problem 4c

Problem 4a

- Make a single plot with overlaid histograms using **all events** for the test file. The x-axis should be the classifier algorithm test-statistic and the plotted data should be separated into 'Revenue==1' and 'Revenue==0'
  - Separate the two populations and plot the **Revenue==1** entries in ***black*** and **Revenue==0** in ***red***

Problem 4a (example)

- Example here is an illustration for a very small sample with Revenue==0 entries and Revenue==1 entries. Your plot may look **very** different



Problem 4b

- Rank the variables starting with most important to least important
  - Discuss any variables that have similar discrimination power
  - Provide the ranked list
- Discuss how to identify and avoid overtraining in supervised machine learning algorithms

Problem 4c

- Using the same classifier developed in Problem 4a, run the classifier over all the entries on the blind sample
  - http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2020/data/Exam_2020/Exam_2020_Prob4_BlindData.csv
  - Results will be graded on the **classification accuracy**
  - The new data file has a unique ID number for every entry
    - Produce a text file which contains **only** the IDs which your classifier classifies as **Revenue==1** (last_name.AMAS_Exam_2020.Problem4.RevenueTrue.txt)
    - Produce a text file which contains **only** the IDs which your classifier classifies as **Revenue==0** (last_name.AMAS_Exam_2020.Problem4.RevenueFalse.txt)
    - The file names **MUST BE EXACT**. For two submissions from Jason Koskinen these would be "koskinen.AMAS_Exam_2020.Problem4.RevenueFalse.txt" and "koskinen.AMAS_Exam_2020.Problem4.RevenueTrue.txt"
  - Basic text files. No Microsoft Word documents, Adobe PDF, or any other extraneous text editor formats. Only a single ID number per line in the text file that can be easily read by numpy.loadtxt().
    - One entry per line and no commas, brackets, parenthesis, etc.