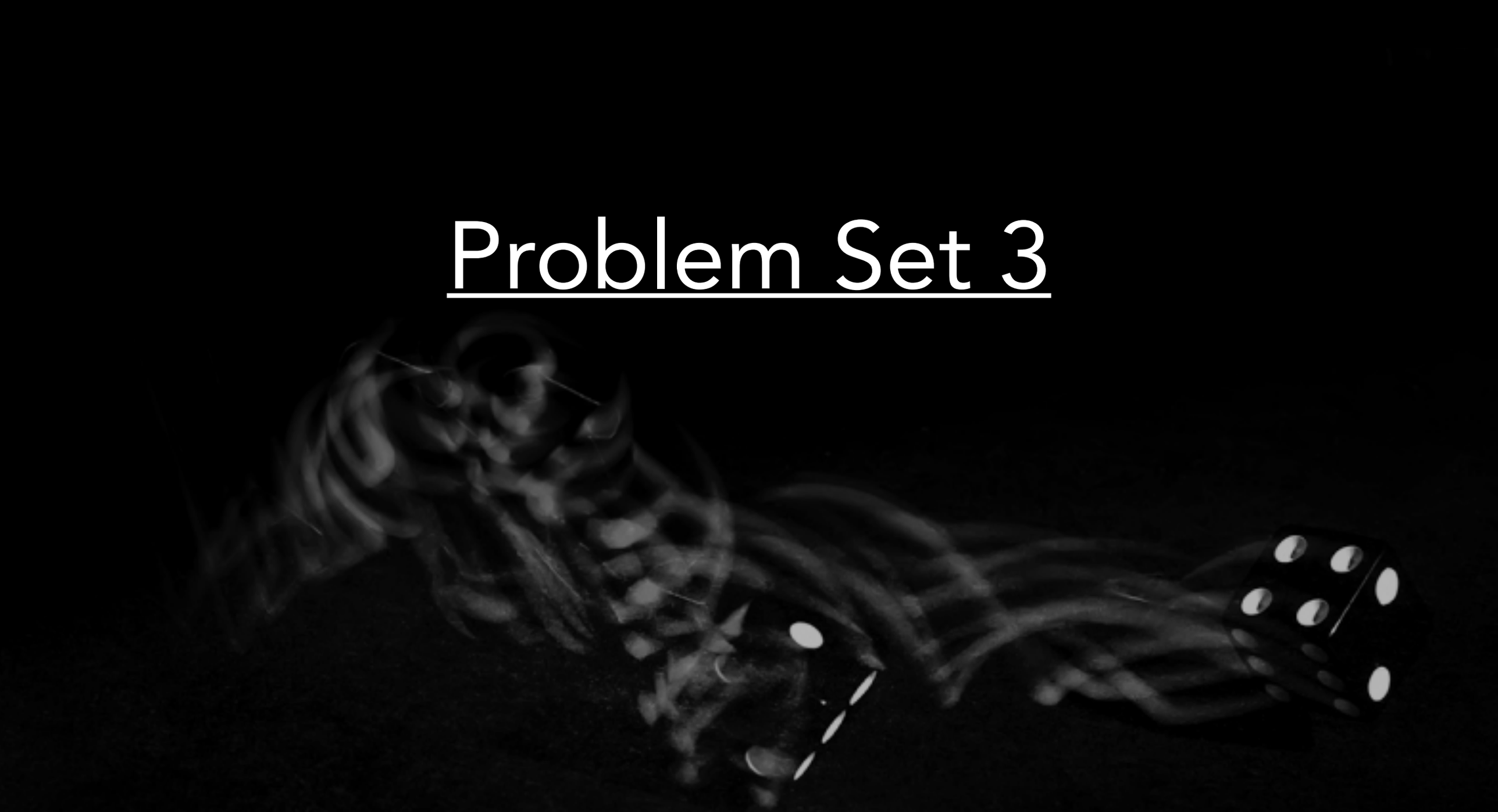
Problem Set 3



D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2023

Photo by Howard Jackman

Problem 1 (0.5 pts.)

- Make a cover page which includes your name, UCPH logo, date of submission, appropriate title, and plot of the χ^2 probability distribution function w/ 1 DoF over the range of $0 \leq \chi^2 \leq 10$

Problem 2 (1 pt.)

- For data and a probability distribution function with 4 fit parameters, e.g. $f(x\,|\,g,u,q,b) = (gx + ux^2 - qx^{3.7})/(13 - bx^{1.5})$, which satisfies Wilks theorem, what is the value of the 2*ln-likelihood difference ($2\Delta LLH$) that should be used to construct the 77.9% confidence interval from the best-fit point?

Problem 3 (4.5 pts.)

- A study of White sharks in the Northwest of the Pacific Ocean has data collected from 1951-2012**. The length of some of those sharks is included as a data file at https://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2023//data/WhiteSharkLength.txt

*PHOTOGRAPH BY KAREN SCHOFIELD, NATIONAL GEOGRAPHIC YOUR SHOT

**The Last Frontier: Catch Records of White Sharks (*Carcharodon carcharias*) in the Northwest Pacific Ocean

Problem 3a (2 pts.)

- Plot the shark length data as a histogram. As a separate plot, show the density estimate from a Kernel Density Estimation method where:
  - The KDE should have a gaussian kernel
  - The bandwidth should be 25 cm

- Using the constructed KDE what is the p-value of recording a White shark that is longer than 653 cm in length?

Problem 3b (1.5 pts.)

- Let's assume a new study is conducted where for White sharks:
  - The female:male ratio of White sharks is 50:50
  - Weights:
    - Mature female length probability is $\frac{1}{55 * \sqrt{2\pi}} e^{-(L-0.434*W)^2/(2*55^2)}$
    - Mature male length probability is $\frac{1}{62 * \sqrt{2\pi}} e^{-(L-0.293*W)^2/(2*62^2)}$
      - L is the shark length in cm and W is the shark weight in kg
- Including the previous KDE from Problem 3a and the new length probability equations—and assuming that the Pacific Northwest ocean White shark population is 50:50 for female:male—what is the probability that an observed White shark of weight 763 kg in the Pacific Northwest ocean will have a length greater than 337 cm?
- Plot the fully normalized probability distribution for White shark length over the range of 100 cm to 750 cm.

Problem 3c (1 pt.)

- Repeat 3b, but now consider only mature sharks, where mature sharks are those with length > 201 cm.
  - Any and all relevant functions (priors, posteriors, likelihoods) should be truncated such that any values or estimates less than 201 cm have a zero probability.
    - E.g. the mature length probabilities from 3b will be truncated gaussian distributions.
- The ratio between male:female is 50:50 for mature sharks with length > 201 cm.

Problem 4 (4.0 pts.)

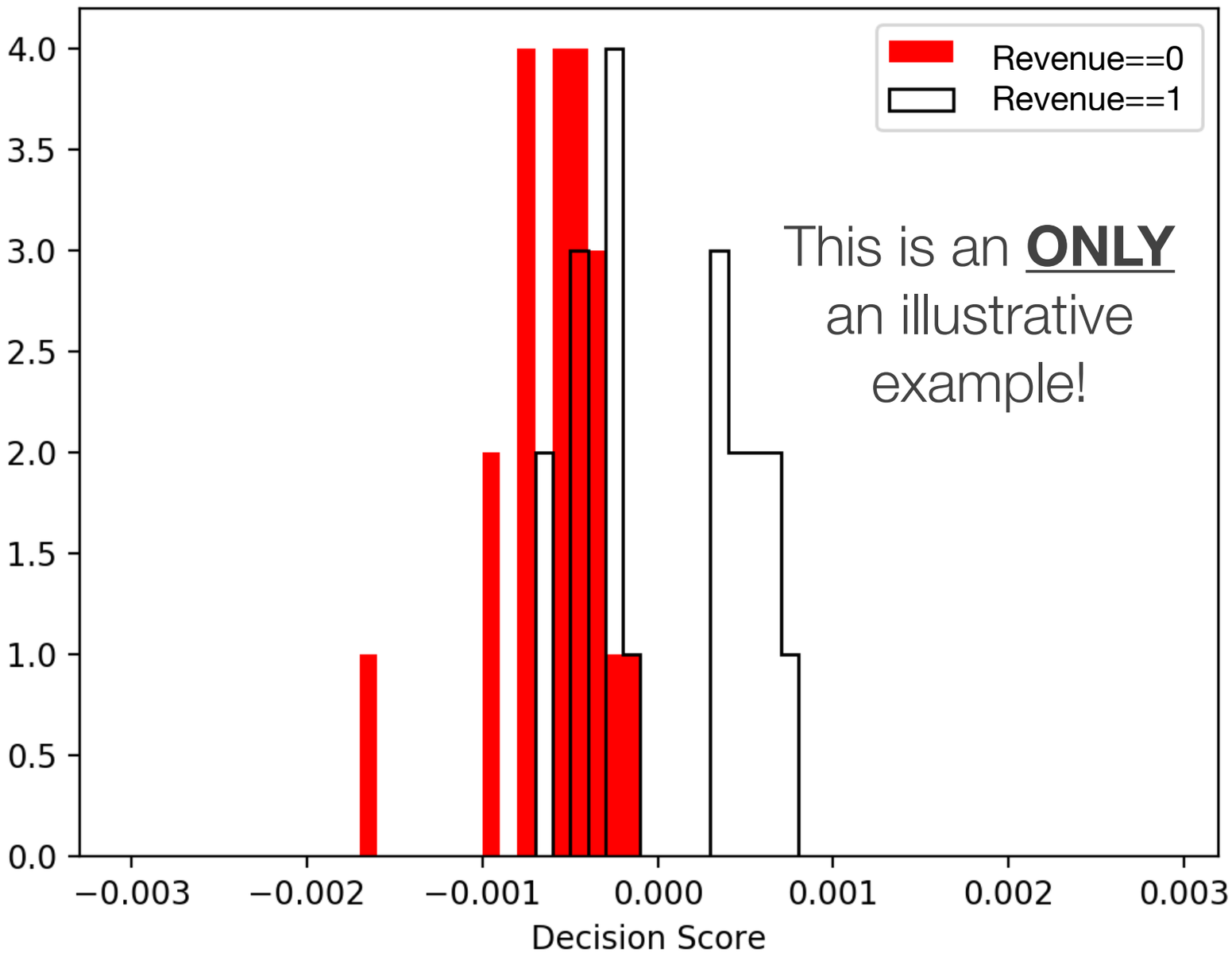
- Anonymous data was collected regarding whether online activity results in revenue, e.g. purchases at a website
- Create a classifier trained on the training data files which separates those online user sessions which do create revenue from the online user sessions which do not create revenue
- The data set has been divided:
  - Training/Testing data set is at:
    - https://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2023/data/Set3_Prob4_TestData.csv
    - https://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2023/data/Set3_Prob4_TrainData.csv
  - The 'blind' analysis data set is at https://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2023/data/Set3_Prob4_BlindData.csv
    - Only used in problem 4c

Problem 4a (1.5 pts.)

- 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 (0.5 pts.)

- Discuss how to identify and avoid overtraining in supervised machine learning algorithms, and what checks you made to ensure that your trained algorithm does not exhibit signs of overtraining

Problem 4c (2 pts.)

- Using the same classifier developed in Problem 4a, run the classifier over all the entries on the blind sample
  - https://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2023/data/Set3_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.Problem4.RevenueTrue.txt)
    - Produce a text file which contains **only** the IDs which your classifier classifies as **Revenue==0** (last_name.Problem4.RevenueFalse.txt)
    - The file names **MUST BE EXACT**. For two submissions from Jason Koskinen these would be "koskinen.Problem4.RevenueFalse.txt" and "koskinen.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.