

# Exam



D. Jason Koskinen  
[koskinen@nbi.ku.dk](mailto:koskinen@nbi.ku.dk)

*Advanced Methods in Applied Statistics*  
*Feb - Apr 2024*

# Info

- In submitting the solutions there is no need to rephrase the problem. For example, "Solution for 1a" is sufficient.
- The submission format for explanations and plots is a PDF file. Also, in a separate file (or files) include any and all software scripts used to establish your answers and/or produce plots.
- Working in groups or any communication about the problems is **prohibited**. Using the internet as a resource is encouraged, but soliciting any help is 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
- Look through all problems in the exam. Some problems are easier than others.
- For any concerns, questions, or comments email Jason ([koskinen@nbi.ku.dk](mailto:koskinen@nbi.ku.dk))

# Starting

- 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, academic, and moral 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 data of interest generated from some underlying function. There are 4121 entries, i.e. rows.
  - [http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2024/data/Exam\\_2024\\_Prob1.txt](http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2024/data/Exam_2024_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**

# Lists of Distributions

$$-10 \leq a \leq 10$$

$$-10 \leq b \leq 10$$

$$4000 \leq c \leq 8000$$

- 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  $c$

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

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

# Problem 1a (2.2 pts.)

- Use the separate data from **first**, **third**, and **fourth** columns 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 a **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.
- The first and second columns have artificially **truncated ranges**
  - First column is only sampled in the independent variable from 6 to 11

# Problem 1b (0.8 pts.)

- 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/AdvancedMethodsInAppliedStatistics2024/data/Exam\\_2024\\_Problem2.txt](http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2024/data/Exam_2024_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 139 paired data points in total
  - The values are in units of **radian**

# Problem 2a

- Correctly 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  $2\pi$ , and **uniform in  $\cos(\text{zenith angle})$  from -1 to 1**
    - Hint: you can use **Monte Carlo generated pseudo-experiments to produce a test-statistic distribution of a spherically isotropic distribution.**
    - Hint: isotropically distributed means '**uniform**' **simultaneously** in azimuth and  **$\cos(\text{zenith})$ .**
    - Hint: establishing that **a distribution is uniform in variable 1 and separately establishing that a distribution is uniform in variable 2 **does not conclusively establish** that the 2D distribution is uniform.**

# 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  $\{0.225\pi, 0.725\pi\}$  and uniformly distributed in  $\cos(\text{zenith})$  over the range  $\{0.30\pi, 1\pi\}$ , and the remaining 80% is fully isotropic
  - Hypothesis B: That 15% of the total sample is uniformly distributed in azimuth over the range  $\{0\pi, 1\pi\}$  and uniformly distributed in  $\cos(\text{zenith})$  over the range  $\{0.5\pi, 1\pi\}$ , and the remaining 85% is fully isotropic.
  - Report the two p-values:
    - $H_{\text{isotropic}}$  versus  $H_A$
    - $H_{\text{isotropic}}$  versus  $H_B$

# Problem 3 (1.5 pts.)

- With the function below as the defined test-statistic function we can use MultiNest, or some other nested sampling bayesian algorithm, to establish the best-fit test-statistic distribution and best-fit values of  $\theta_1$ ,  $\theta_2$ , and  $\theta_3$

$$\mathcal{L}(\theta_1, \theta_2, \theta_3) = 3 \left( \cos(\theta_1) \cos(\theta_2) + \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(\theta_3 - \mu)^2}{2\sigma^2}} \cos(\theta_1/2) + 3 \right)$$

- The range of allowed values is restricted for  $\theta_1$  and  $\theta_2$  to  $0-7\pi$  and for  $\theta_3$  from  $0-3$ . Also, set  $\mu=0.68$  as the true mean of the gaussian distribution and  $\sigma^2=0.04$
- The above function is unnormalized, and for this problem is not required to be normalized. As such, any distribution using values of  $\theta_1$ ,  $\theta_2$ , and  $\theta_3$  will not strictly be a 'posterior' or 'likelihood' distribution, but we can think of it as a 'posterior' for this problem.

# Problem 3a (0.5 pts.)

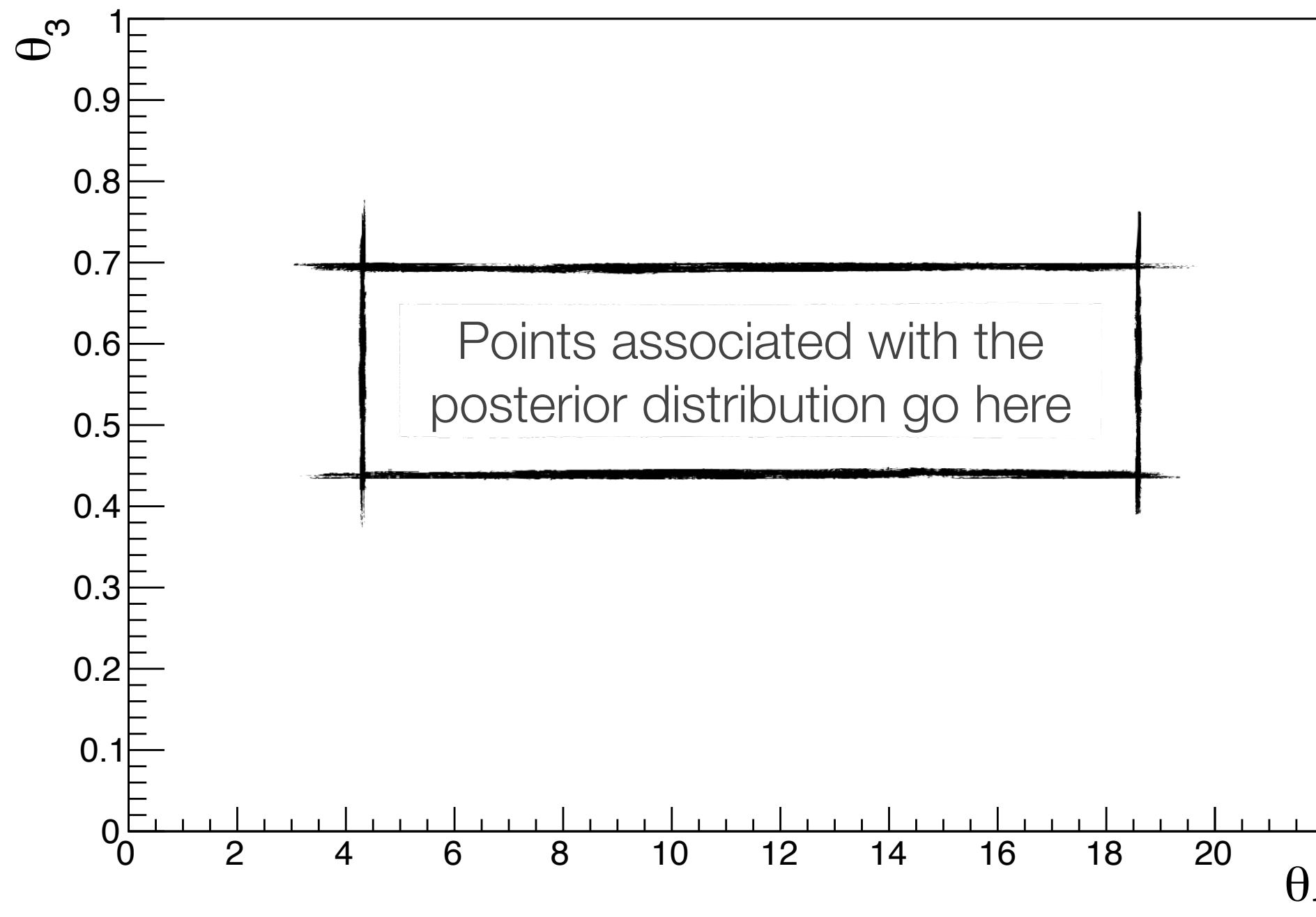
- What are the best-fit values  $\hat{\theta}_1$ ,  $\hat{\theta}_2$ , and  $\hat{\theta}_3$  that you find from maximizing the test-statistic function?

# Problem 3b (0.5 pts)

- Plot the 2D test-statistic distribution of the function for  $\theta_2$  versus  $\theta_1$ , and  $\theta_3$  versus  $\theta_1$ . [there is an empty example on the next slide].
  - Produce 2 separate 2D plots.
  - These can be weighted or unweighted points from a nested sampling algorithm or a weighted KDE-based contour-like 2D plot derived as part of the fitting routine, or even something else that produces a posterior-like distribution.

# Problem 3b [example]

Posterior (MultiNest)



# Problem 3c (0.5 pts.)

- The posterior distribution is proportional to the output of the test-statistic. Make two separate raster scan plots in 2-D of the output from the function over the same ranges as for the previous plot. Essentially, map the the landscape.
  - For the scan of  $\theta_2$  vs.  $\theta_1$ , fix  $\theta_3$  to the best-fit point found from your fit, i.e. an unchanging value  $\hat{\theta}_3$ . Similarly, for the scan of  $\theta_3$  vs.  $\theta_1$ , fix  $\theta_2$  to the best-fit point  $\hat{\theta}_2$ .
- Does the posterior distribution match the raster scan plots? Discuss why it should, or why it should not.



# Problem 4 ( 1.5 pts.)

- Data was taken to examine what variables (or combinations of variables) might be used to identify when a patient will miss their scheduled appointment, i.e. a 'No-show'.
- Create a classifier which separates patients that are likely to have a 'No-show' from those that are not likely to 'No-show'
  - Consider 'No-show=True' as the signal or real positive, and the 'No-show=False' as the background or real negative
- The data set has been divided:
  - Training/Testing data set is at:
    - [http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2024/data/Exam\\_2024\\_Prob4\\_TrainData.csv](http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2024/data/Exam_2024_Prob4_TrainData.csv)
    - [http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2024/data/Exam\\_2024\\_Prob4\\_TestData.csv](http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2024/data/Exam_2024_Prob4_TestData.csv)
  - The 'blind' analysis data set is at [http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2024/data/Exam\\_2024\\_Prob4\\_BlindData.csv](http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2024/data/Exam_2024_Prob4_BlindData.csv)
    - Only used in problem 4c
  - Include **ALL input files** when submitting your solution

# Problem 4 (cont.)

- There are many possible features (i.e. variables) to use, but we will restrict the **classification algorithm** to only use the following:

```
features_to_train = ['Gender',  
                    'ScheduledDay',  
                    'AppointmentDay',  
                    'Age',  
                    'TimeDifference',  
                    'Neighbourhood',  
                    'Diabetes',  
                    'Alcoholism',  
                    'Handcap',  
                    'SMS_received',  
                    'R1'  
                    ]
```

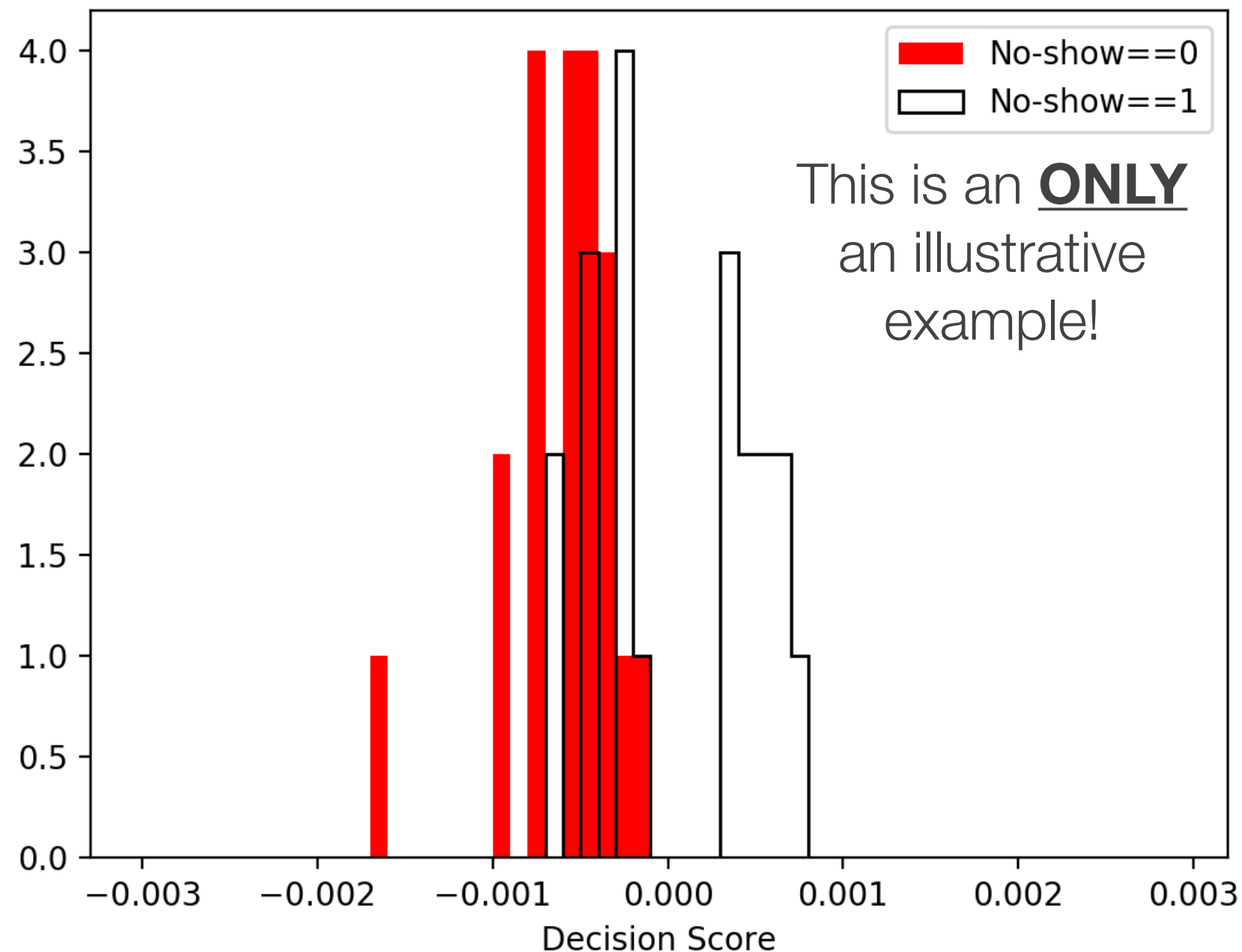
- The above features should be the only variables (besides "ID" and "No-show" for the training and testing) which are in the file

# Problem 4a (0.5 pts)

- Make a single plot with overlaid histograms using **all events** from the test file versus the test statistic; separated into 'No-show==1' and 'No-show==0'
- Separate the two populations and plot the No-show==1 patients in **black** and No-show==0 patients in **red**
- The x-axis should be the test statistic that is the output of the classification algorithm

# Problem 4a (example)

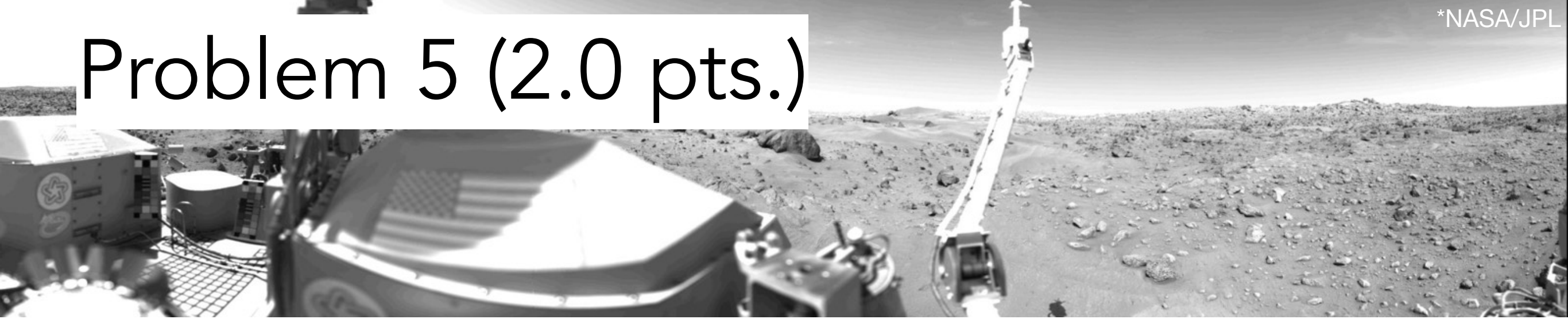
- Example here is an illustration for only 20 No-show==0 entries and 20 No-show==1 entries, your plot may look **very** different



# Problem 4c (1.0 pt.)

- 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/AdvancedMethodsInAppliedStatistics2024/data/Exam\\_2024\\_Prob4\\_BlindData.csv](http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2024/data/Exam_2024_Prob4_BlindData.csv)
  - Results will be graded on the **classification accuracy**
  - The new data file has a unique ID number for every patient
    - Produce a text file which contains **only** the IDs which your classifier classifies as **No-show==1** (last\_name.AMAS\_Exam.Problem4.NoShowTrue.txt)
    - Produce a text file which contains **only** the IDs which your classifier classifies as **No-show==0** (last\_name.AMAS\_Exam.Problem4.NoShowFalse.txt)
    - The file names **MUST BE EXACT**. For two submissions from Jason Koskinen these would be "koskinen.AMAS\_Exam.Problem4.NoShowFalse.txt" and "koskinen.AMAS\_Exam.Problem4.NoShowTrue.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.

# Problem 5 (2.0 pts.)



- The success of planetary exploration on Mars is highly dependent on the ability of any exploratory vehicle, e.g. Mars Perseverance rover and Ingenuity helicopter, to survive the low temperatures as well as the temperature changes.
- Viking 1 was the first successful Mars landing and took temperature data that was used to inform future Mars missions.
- We will use temperatures recorded at different time intervals from Viking 1
  - The first entry is the sol (the Mars analog of an Earth day) and the second is the temperature in Celsius. We will assume the temperature uncertainty is  $\pm 0.001\text{C}$ .
  - For example [203.41, -89.37] is data taken on sol 203.41 and the temperature is -89.37 C.

```
array([[ 203.41 , -89.37 ],
       [ 203.435, -94.88 ],
       [ 203.46 , -101.25 ],
       [ 203.484, -106.52 ],
       [ 203.509, -108.66 ],
       [ 203.534, -114.25 ],
       [ 203.558, -114.30 ],
       [ 203.583, -117.66 ],
       [ 203.608, -122.45 ]])
```

# Problem 5a (0.5 points)

- Use the data provided on the previous page to create 2 splines:
  - A linear spline
  - A cubic spline or a Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) spline. Create one of these non-linear splines, but not both.
- What are the estimates of the temperature on sol 203.570 from the created linear-spline, as well as the created cubic/PCHIP-spline?

# Problem 5b (1 point)

- Make a scatter plot of temperature versus time of the interpolated temperatures from the linear-spline as well as the cubic/PCHIP-spline covering the time range from 203.410 sol to 203.608 sol.
  - To ensure we can see any interesting features of the linear and cubic/PCHIP splines, make sure there are at least 200 points for each interpolation.
- During sol 203.410 to 203.608, if we know that the temperature should be continuously dropping, are there any regions of time that we should look at more closely for the cubic/PCHIP spline to make sure that the interpolated temperature is monotonically decreasing?



# Problem 5c (0.5 points)

- Imagine that proposed electronics components for a Mars rover, or other Mars planetary-surface exploration vehicle, are unable to sustain temperature changes of more than 0.09 C within 0.0004 sol.
- Would new electronics be needed with more robustness to temperature fluctuations according to your interpolation(s) to the Viking 1 data?