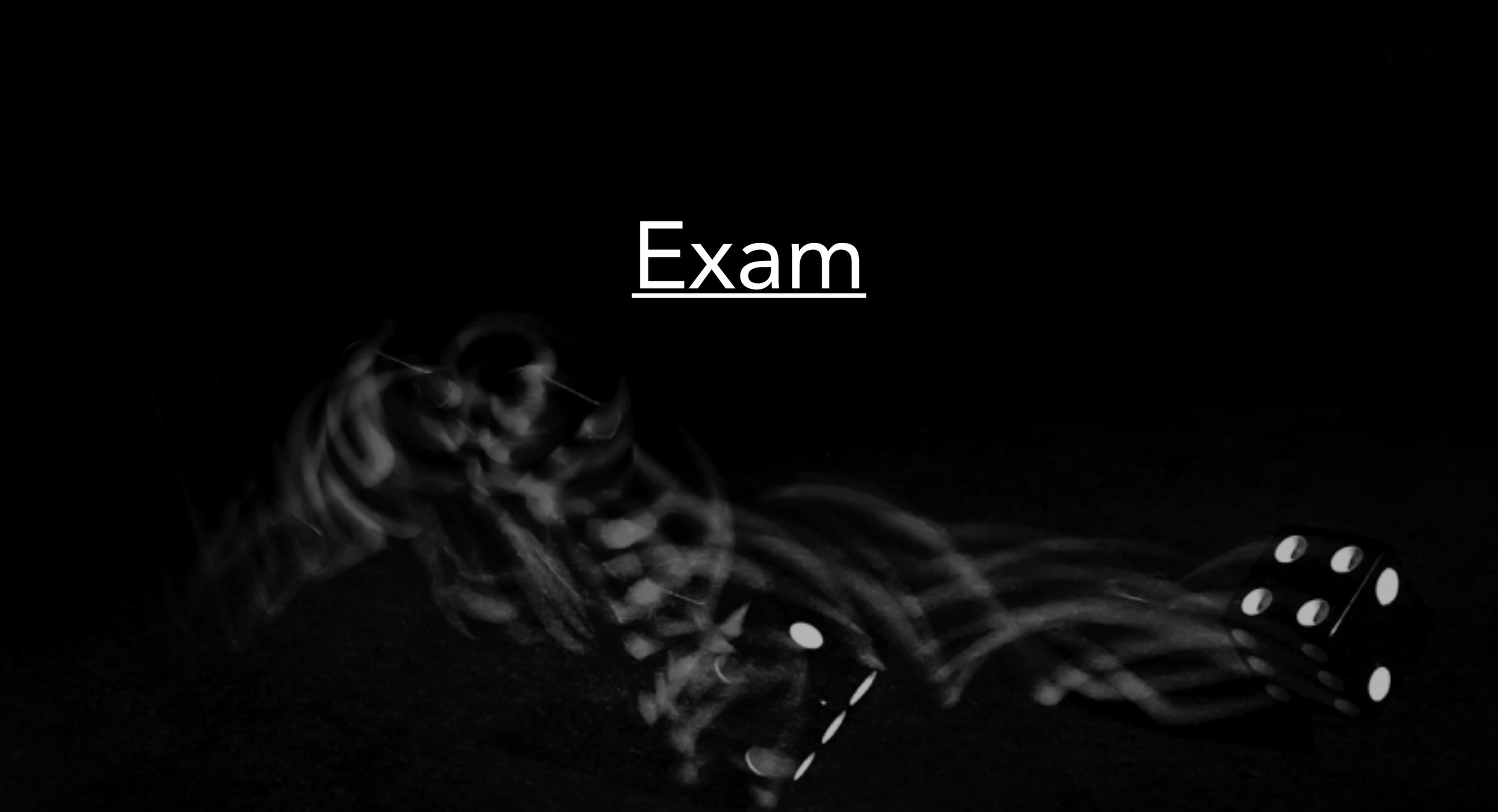
Exam



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 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, 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 a physical observable of interest generated from some underlying function. There are 5119 entries, i.e. rows.
  - http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2021/data/Exam_2021_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

- 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

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

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

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

Problem 1a

- Use the separate data from columns 1, 2, and 4 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 columns 1, 2, and 4 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
  - Column 4 is only sampled in the independent variable from 0 to 2.5

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/AdvancedMethodsInAppliedStatistics2021/data/Exam_2021_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

- 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.
  - Hint: isotropically distributed means ‘uniform’ **simultaneously** in azimuth and cos(zenith).

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, 0.725π} and uniformly distributed in zenith over the range {0.30π, 1π}, and the remaining 85% 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

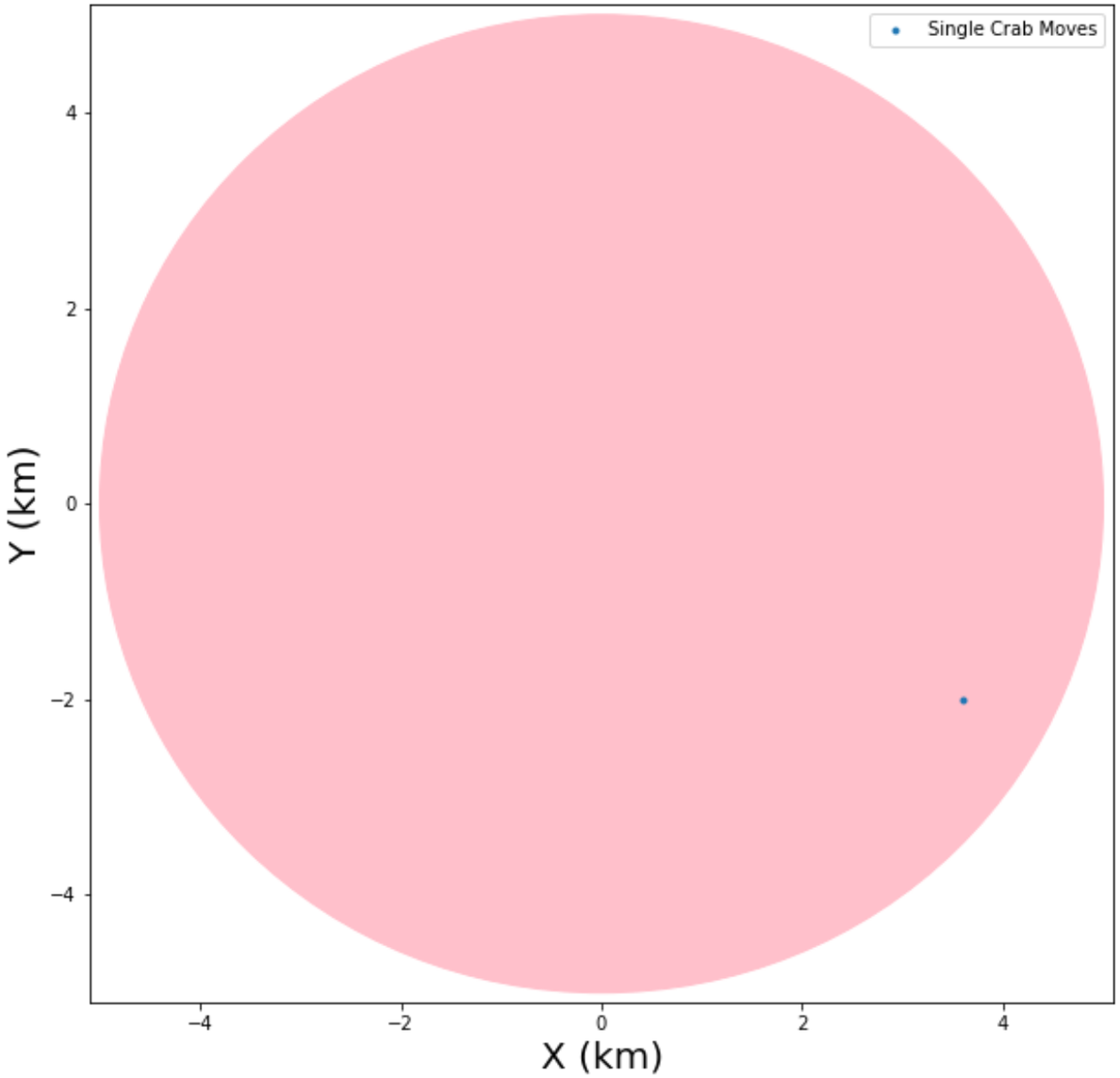
Problem 3 (2.0 points)

- Inspired by the Russian short story "Crabs on the Island" we will look at population evolution. Imagine an island with a 5 km radius inhabited by metal crabs, each with a mass of 1 kg. The metal crabs only consume other metal crabs.
- All crabs move in individual random directions (0-2π), once per day until:
  - they move 200 meters, or
  - they reach the edge of the island
- At the end of the day, any metal crabs within a certain distance will try and eat each other or defend from being eaten. Details are on later slides.

*outdoorsonparade.com

Problem 3a (0.5 points)

- For simplicity, we consider only a single crab on the island for this part of the problem. No other cannibal metal crabs… yet.
- Show a plot of a single crab starting at (3.6 km, -2.0 km) and moving for 200 days.
  - Put a single point for each stopping position of the crab
  - Example figure ONLY includes the starting point, i.e. day=0
  - There should be a total of 200 points, or 201 if you include the starting position.



Problem 3b (0.5 points)

- For simplicity, we continue considering only a single crab on the island.
- Run 500 pseudo-experiments with a single crab starting at the location (3.6 km, -2.0 km) and make a histogram of the 500 distances the crab travels before arriving at the edge of the island.

Problem 3 Crab Battle Details

- Crabs that end the day within 175 m of each other will battle. The larger crab will consume the smaller crab with odds of $\frac{M_{larger}^2}{1kg} : \frac{M_{smaller}^2}{1kg}$.
  - For example, a 5 kg crab eating a 2 kg crab has 25:4 odds, i.e. 25/(25+4)=86.2%. The smaller crab has a 4/(25+4)=13.8% of surviving, but might get eaten the next day (life is tough for small crabs).
  - If the smaller crab defends, then both crabs continue to exist.
    - Crabs of equal mass have 1:1 odds of one crab consuming the other. It is arbitrary which one is considered ‘larger’.
- If the larger crab eats the smaller crab its mass increases equal to the smaller crab it consumed. The position on the island of the ‘winning’ crab remains the same as where it stopped for the day, despite the battle. In the case of the smaller crab surviving, then both crabs stay in the same position as where they stopped for the day.
- Although rare, a single crab can do battle multiple times in a day if it is within 175 m of multiple other crabs. The smallest 2 crabs do battle first. If one of the crabs eats the other, its mass is updated for the potential second battle.

Problem 3 Crab Battle Tips

- Make sure that the sum of metal crab mass on the island is always 20 kg. If it's ever not equal to 20 kg, then there's a bug in your code.
- Eaten crab cannot battle. If a crab is consumed, somehow keep track that it cannot battle anymore. There are many, many, many ways to do the 'book keeping' for this. For example,
  - Change the mass of the consumed crab to 0 and the position to -infinity.
  - Remove the eaten crab from the array or list after the battle in which they are eaten.
  - Include a data element in the crab array, list, or class which notes whether it has 1) battled already that day, and/or 2) whether it has been eaten.

Problem 3c (0.5 points)

- There are 20 metal crab on the island with the starting positions provided in the file at https://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2021/data/Exam_2021_Prob3_CrabStartPositions.txt.
  - The first column is the x position and the second column is the y position.
  - Each row is the starting position of an individual crab.
- After 200 days:
  - What is the mostly likely number of individual crabs that remain alive?
  - What is the most likely mass of the largest crab?

Problem 3d (0.5 points)

- We will continue using the same 20 crab on the island with the starting positions provided in the file at https://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2021/data/Exam_2021_Prob3_CrabStartPositions.txt
- Show the distribution of days it takes until only 10 crabs remain alive
- What is the 1σ confidence interval on the number of days at which only 10 crabs remain?
  - If the distribution is asymmetric, then calculate the lower 1σ bound as the value where the lower tail contains (100%-68.27%)/2 of the distribution. Same thing for the upper 1σ bound calculated using the higher tail.
  - N.B. look at slide 14 from *Lecture 5: Parameter Estimation and Uncertainty* for a reminder about confidence intervals.

Problem 4 (1.0 pts.)

- A fisherman named Alexa has 412 type-A fish at 08:00 in the morning in a bucket for sale at her pop-up shop. The shop is located next to the lake, so the fish have some probability of jumping back into the lake; a Poisson process with the mean rate of 23 fish per hour, which defines the *likelihood* $P(N_{\text{lost fish}} \mid t)$.
- At the same time she has a constant and unchanging inflow of customers (5 per hour), and every one of them buys 1 fish. Each type-A fish costs $6.
- Alexa has a deal with another fisherman, Bob, that when Alexa has about 240 type-A fish remaining, she will trade all of her type-A fish for 120 type-B fish, where a type-B fish sells for $10 each. After the trade, Alexa's customer inflow will decrease to a constant and unchanging 3 customers per hour, and the type-B fish will continue to jump to the lake with the same rate as the type-A fish.
  - For clarity, Alexa might trade 250 type-A fish for 120 type-B or maybe 236 type-A for 120 type-B fish. The amount of fish traded depends on the time of the trade between Alexa and Bob (further details on the next slide).
  - Alexa will ALWAYS trade all her type-A fish for 120 type-B fish.

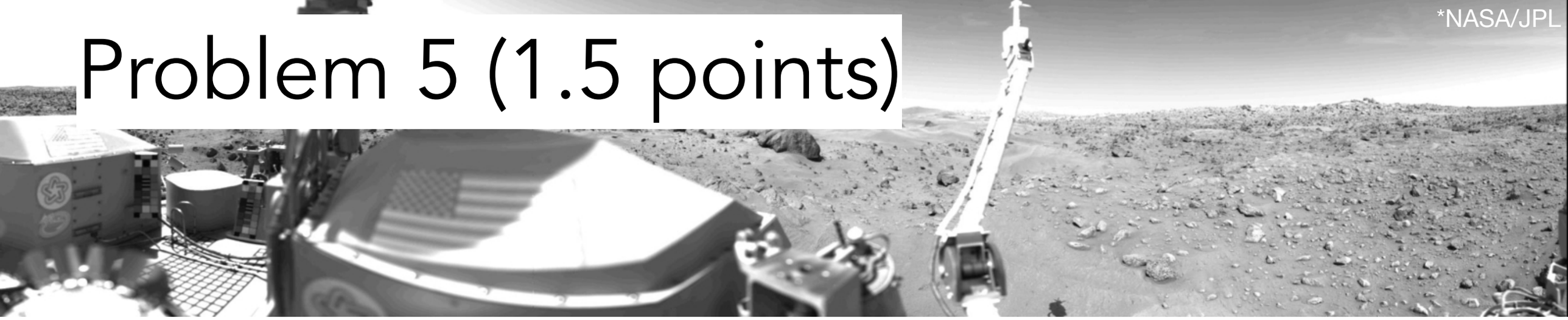
Problem 4 (1.0 pts.)

- Trading conditions:
  - Bob and Alexa are busy people and understand that sometimes the exchange may not be exactly 240 type-A fish during the trade: the *prior* probability $P(N_{\text{exchanged type-A fish}})$ of trading is a gaussian centered at $\mu=240$ type-A fish and has a standard deviation of $\sigma=10$ type-A fish.
  - Alexa starts her work day at 08:00 and ends it at 18:00, so the trade can happen anytime between 08:00 and 18:00. At 18:00:01, Bob comes to collect all the remaining type-B fish from Alexa and pays Alexa $4 for each remaining type-B fish.

Problem 4 (1.0 pts.)

- Questions (0.5 pts each):
  - At what time is the trade most likely to happen?
  - Plot the PDF for the total amount of money earned by Alexa from 08:00 to 18:00. Draw 50 samples from that PDF and submit as a text file, with one numeric entry per line and no extra information. (Hint: only integer numbers of type-A and type-B fish can be sold/lost, which should be reflected in the PDF and the drawn samples.)

Problem 5 (1.5 points)



- 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
  - 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, -114.3  ],
       [ 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 (0.5 points)

- Make a scatter plot 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 to 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?