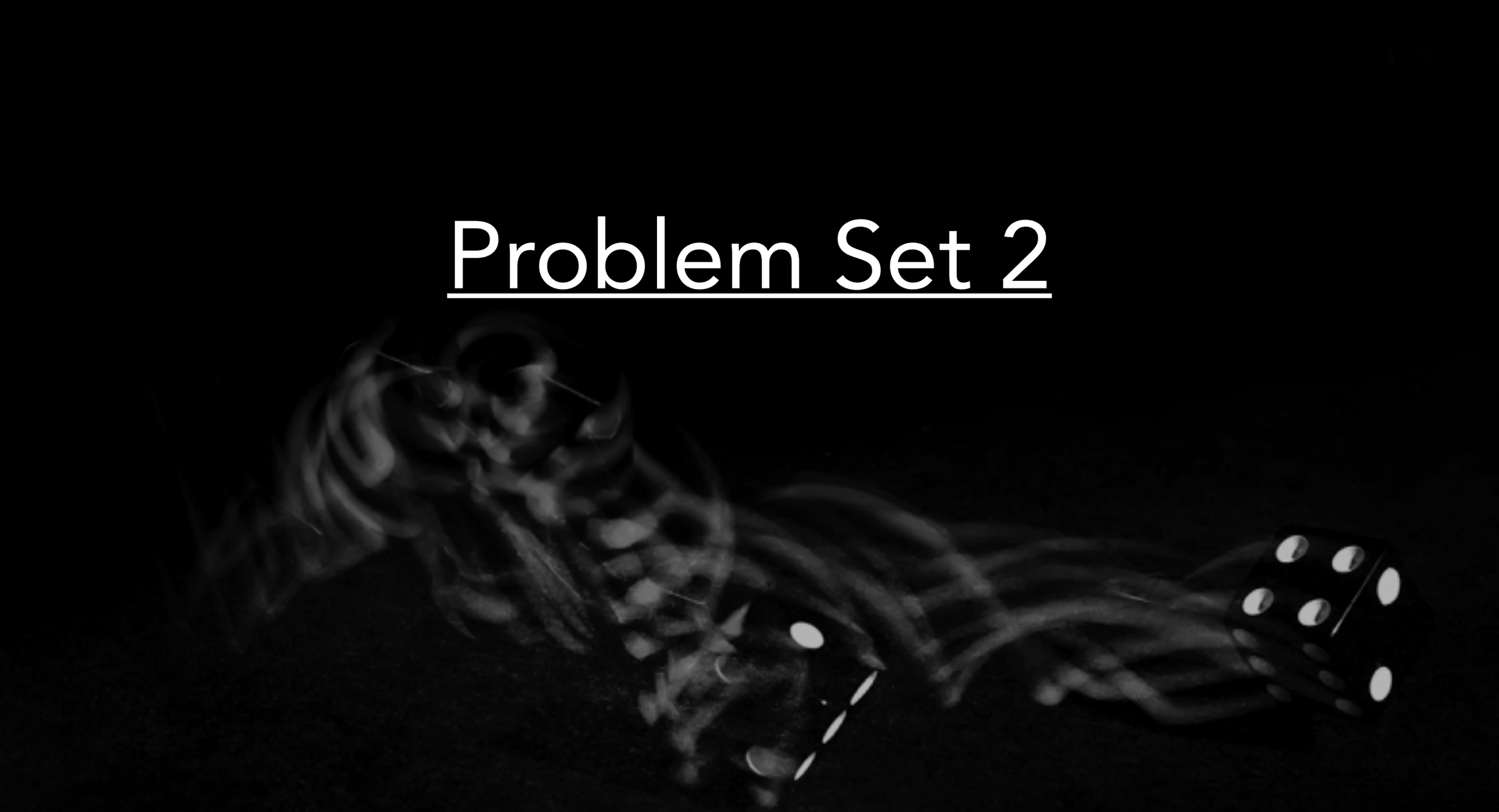
Problem Set 2



D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2018

Photo by Howard Jackman

Info

- The following problem set is due on, or before, Friday March 23 at 16:00 Copenhagen time.
  - The write-up should be in a PDF text document and should contain no code or software. There will be a deduction for extensive code in the write-up portion of the submission.
  - The write-up should be neat, clear, and contain any and all discussions, comments, reflections, results, tables, plots, captions, citations, html links, etc.
    - Solutions should have appropriate labels
    - Graphs have appropriate labels, titles, axes, legends, font size, line widths, marker sizes, etc.
    - Histograms have reasonable bin widths and ranges
  - Separate from the write-up, include the code
    - Zipped files, tarball, jupyter notebook, .C files, non-write-up PDF file, etc. are all fine
- The total points in this assignment will be appropriately scaled to account for 15% of the final course grade.

Problem 1 (5 pts.)

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

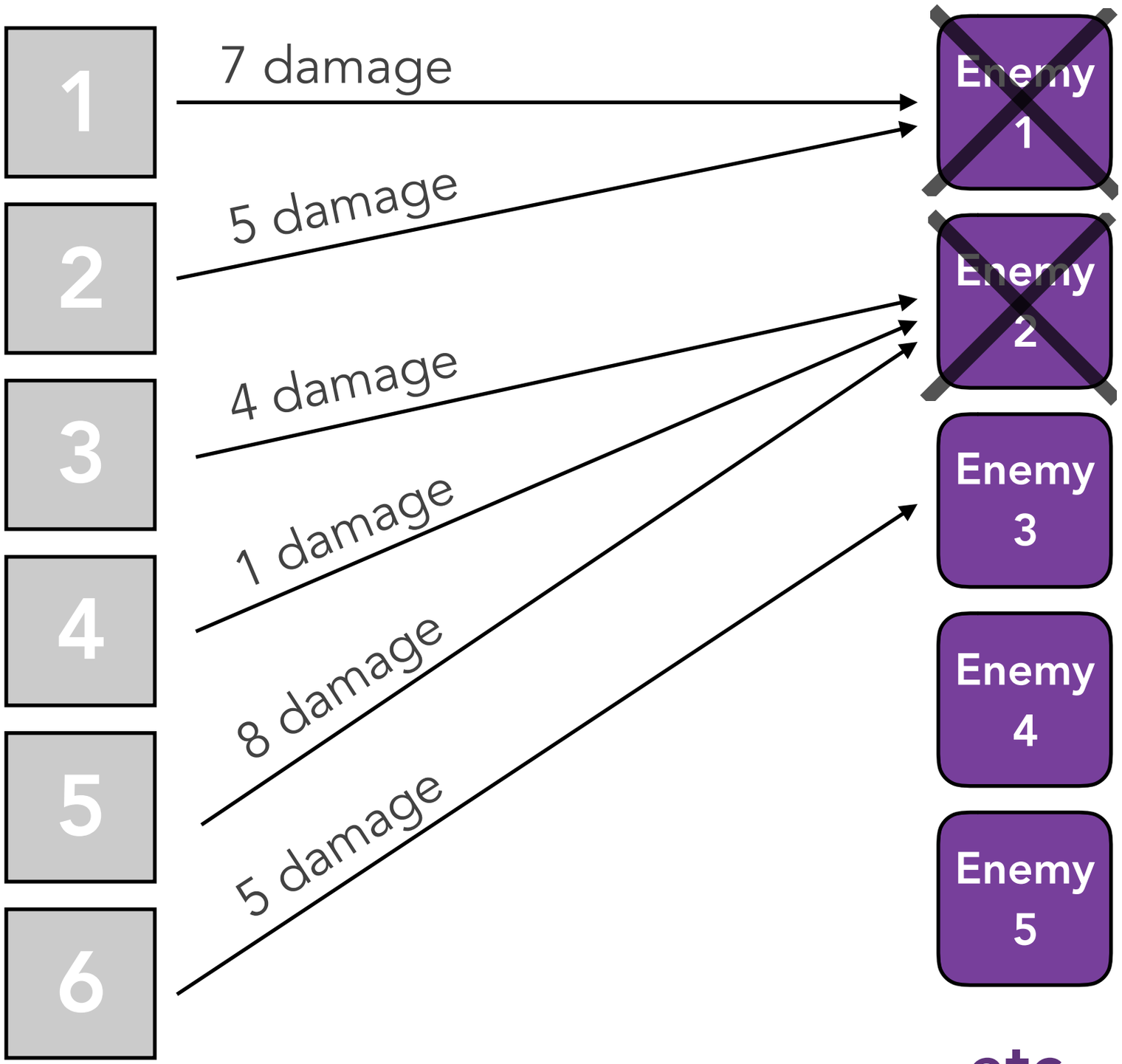
Problem 2 (20 pts.)

- You are playing a strategic turn-based computer game and you want to better understand likely outcomes. You have 6 units which are fighting 6+ enemies. In a turn, your units act only once and inflict damage in successive iteration to the first enemy in the queue, until the enemy has 0 or negative health, whereby that enemy is defeated. Once an enemy has been defeated, any of your remaining units which have not acted now inflict damage to the next enemy in the queue, and on and on until all your units have acted
  - Your units only individually act once during a turn to inflict damage
  - Damage inflicted follows a poisson distribution

Problem 2a

- Find the mean number of enemies defeated in a single turn:
  - When the expected damage inflicted individually by each of your units is 5
  - Enemies each have 12 health and are defeated when their health ≤ 0
  - Your 6 units always individually inflict damage, i.e. any random samples of 0 should be rounded up to 1
  - An example illustration is on the next slide
- Plot the distribution of the number of ‘defeated enemies per turn’ for 1000 unique and independent trials/turns
  - Each trial is a fresh set of enemies, i.e. for each trial all of the enemies should start w/ 12 health

Single Turn Example



In this example, individual enemies can receive 12 damage before being defeated

In total 2 enemies were defeated for this turn

Problem 2b

- Using the same values from 2a now include that your 6 units vary in individual accuracy and have some probability to inflict damage, or miss thereby inflicting no damage
  - The probability per unit to inflict damage is [90%, 80%, 60%, 90%, 60%, 70%]
  - Follow the order in the above array for calculations/plots
- Out of 5000 trials, what percentage of the time will no enemies be defeated in a turn, and what is the uncertainty on that percentage?

Problem 2c

- Using the same setup and values from 2b, test the new reorderings below, of your units inflicting damage versus the ordering in 2b of [90%, 80%, 60%, 90%, 60%, 70%]
  - Sorted ascending [60%, 60%, 70%, 80%, 90%, 90%]
  - Sorted descending [90%, 90%, 80%, 70%, 60%, 60%]
- Are the ascending and descending statistically compatible with the original ordering for the number of enemies defeated per turn?
  - Quantitatively justify your conclusion

Problem 3 (15 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

- 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

- 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.

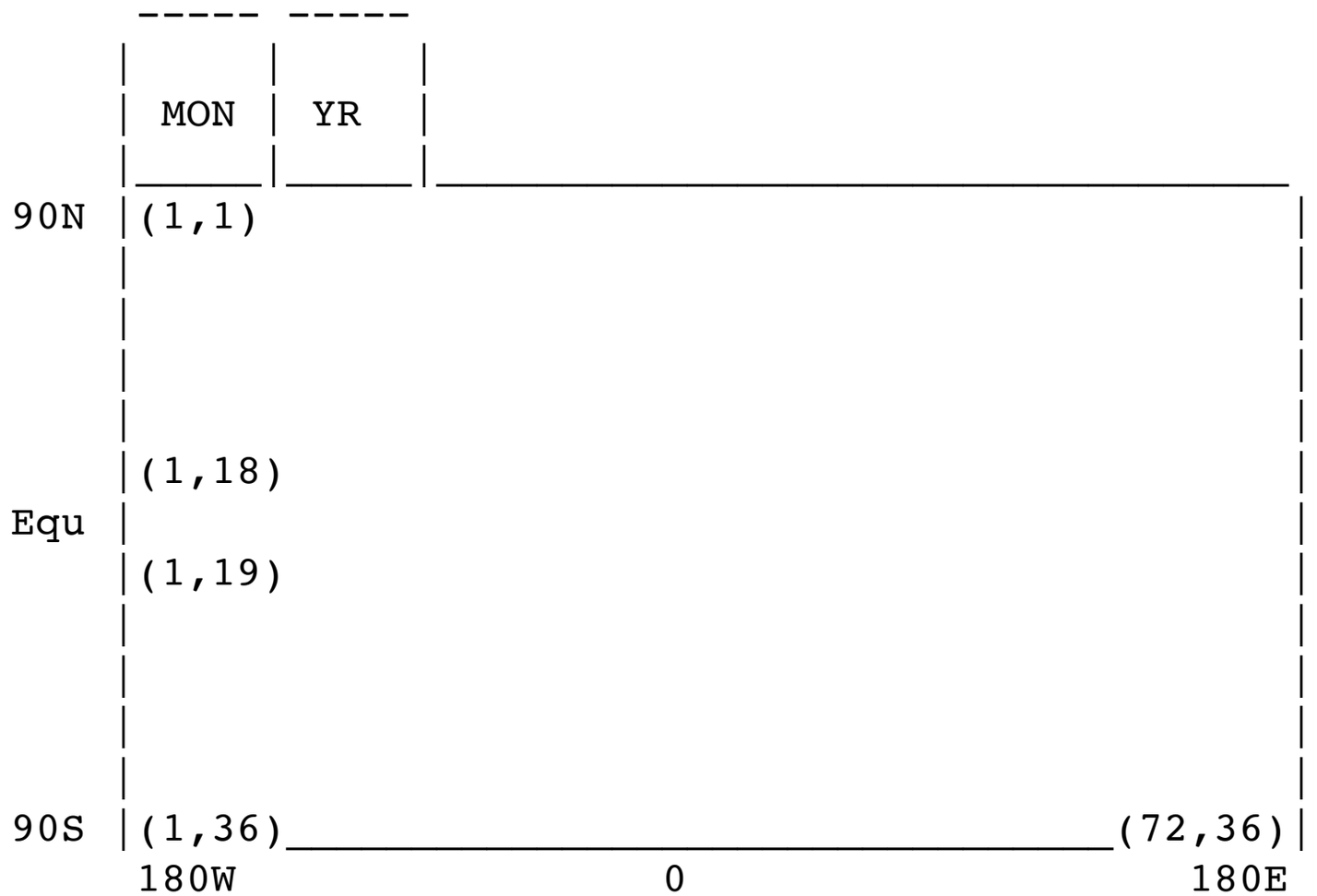
Problem 4 (25 pts.)

- There are two files which contain sea surface water temperatures from global monthly data from HadSST3
  - May 1997 at http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2018/data/GlobalTemp_1.txt
  - May 2017 at http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2018/data/GlobalTemp_2.txt
- Using data in the 8th row (including 1 line for the header info), construct a kernel density estimator using the Epanechnikov kernel with a bandwidth of 0.4
  - The 8th row is a band of constant latitude near Denmark
  - 1.07 C is the first entry in the 8th row for 2017, and 0.74 C for 1997
  - Do **not** include entries in constructing the KDE where there are **no** temperature measurements

http://hadobs.metoffice.com/hadsst3/

Problem 4 - Data Format

Data are stored in ASCII
Temperatures are stored as degrees C Land squares and missing data are set to -99.99 or, in the case of numbers of observations, 0
The month and year are stored at the start of each month.
Data Array (72x36) Item (1, 1) stores the value for the 5-deg-area centred at 177.5W and 87.5N Item (72, 36) stores the value for the 5-deg-area centred at 177.5E and 87.5S



*from the README file

Problem 4a

- Plot the P_{KDE}(temp) as a function of temperature for both 1997 and 2017 over the range of -2 C to +4 C
  - P_{KDE}(temp) is the data driven kernel density estimated probability distribution function (PDF)

- Calculate the integral of P_{KDE}(temp) for 1997 and 2017:
  - over the range -2 C to +4 C
  - over the range of -2 C to 0 C

Problem 4b

- Produce 1000 Monte Carlo draws/samples/events from the 1997 P_{KDE} over a temperature range from -1 C to +2 C
- Calculate the likelihood ratio for the 1000 Monte Carlo samples where H_0 uses the KDE from 1997 and H_1 uses the KDE from 2017

$$\frac{\mathcal{L}(H_0|x)}{\mathcal{L}(H_1|x)}$$

- Submit your 1000 samples as an ASCII txt file:
  - each entry on a separate line for 1000 total lines in the file
  - File name should be your last name and "_KDE_1000_samples.txt", e.g. "koskinen_KDE_1000_samples.txt"

Problem 5 (35 pts.)

- There is a distribution which has 3 types of events. Events of type A and B do **not** have analytic PDFs, but they do have some data points from which the PDFs can be estimated. Events of type C are gaussian distributed.
  - The gaussian PDF for events of type C have an expectation from 0-13 (often noted as the 'mean' for gaussian distributions) and a width from 0.05-2.
  - All PDFs should be smooth and continuous in the range of 0-13
    - 'Smooth' here means the first derivative is continuous in the range of 0-13
- A and B have true PDFs which are continuous and smooth, so the estimated PDFs should also be continuous and smooth
  - There is a file at (http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2018/data/ProblemSet2_Prob5_InputDataPDF.txt) which details the data points from which to estimate the PDFs for A and B.
  - The data is paired, i.e. the first column is the independent variable and the second column is the dependent variable (x, f(x))
  - The true PDF is known only by Jean-Loup and Jason ;-)

Problem 5a

- Assume that the relative fraction of events (f_A:f_B:f_C) is (0.4 : 0.5 : 0.1) and the gaussian PDF for C is centered at 8.2 and has a width of 0.95
  - Show the individual normalized PDFs on a single plot. Make it readable and pretty. The PDFs for A and B must be estimated from the data.
  - Show the joint PDF for the entire distribution; all three event types combined. Use the appropriate fractions.

Problem 5b

- There is a file online at (http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2018/data/ProblemSet2_Prob5b_data.txt) which is a sample of data that contains all three event types. The relative fraction of event types is unknown. Also the parameters for the gaussian PDF are unknown. Using the joint PDF, what are the maximum likelihood estimator ,i.e. best-fit, values for the relative fraction of each event type, as well as the location and the width of the gaussian for event type C.
  - The PDFs for events A and B should be the same as what was shown in 5a

Problem 5c

- Using a relative fraction of (0.58 : 0.31 : 0.11) for (f_A:f_B:f_C) and setting the position of the gaussian at 8.4 and width equal to 0.67:
  - Create 100 pseudo-experiments, each w/ 3488 Monte Carlo data samples from the joint PDF from $0 \leq x \leq 13$
  - Using 'fixed' values for the gaussian parameters, find the MLE values for the relative fractions and report the 1σ uncertainty for each of f_A, f_B, and f_C
    - The position and width of the gaussian should **not** be fit. Set them, respectively, at 8.4 and 0.67.
    - This could take a while, so make sure everything works for a few pseudo-experiments before doing the full 100
    - My first minimization software setup took ~30 seconds per pseudo-experiment fit, but after some optimization I got it down to a more manageable 2 seconds
- Show a histogram of the fit values of f_A, f_B, and f_C for the 100 pseudo-experiments
  - Note: depending on your likelihood construction, the returned value by the minimizer may not sum to one, but the relative ratios might still be correct. Make the histogram knowing that $f_A+f_B+f_C=1$.