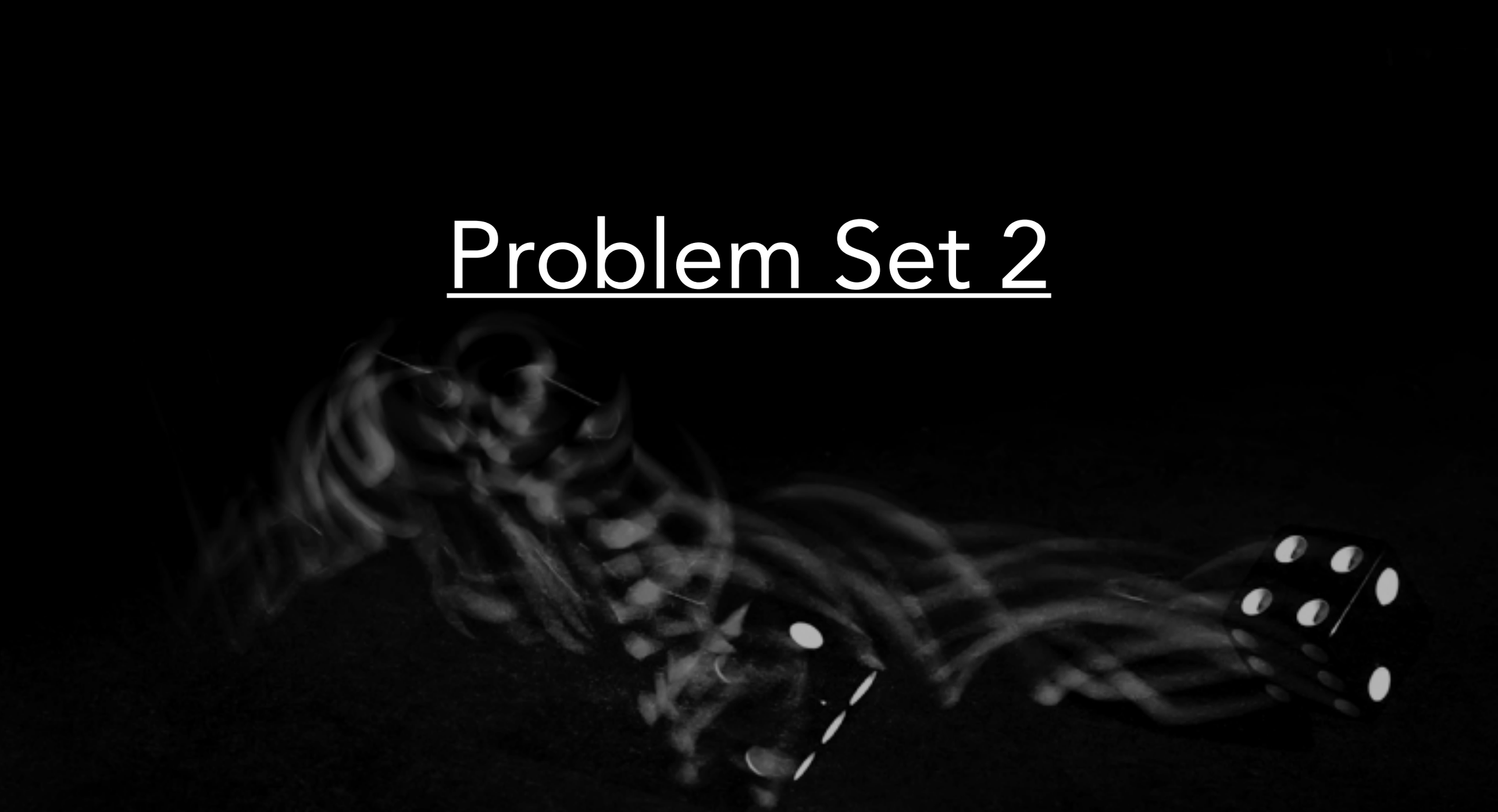
Problem Set 2



D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2019

Photo by Howard Jackman

Format

- The submission is:
  - A write-up as a PDF document, which includes any plots, diagrams, tables, pictures, and explanations
  - In a separate “file”, submit all code used to derive the results
    - Tarball, zipped directory, lots of individual files w/ self-explanatory titles, etc.
    - Do NOT include lines of code in your write-up. If results are dependent on coding choices then include those comments in the write-up.
  - Include any original data files or how the data was accessed
    - If you use a internet scraping tool, note the date when you retrieved the data
    - If you can save the data to a file, do so and submit the data file. There is no need to change the format, e.g. HTML, XML,

Starting points

- On the first page of your write-up include your full name, date, name of this course, UCPH ID, and the title of your problem set submission
- The file name for you submission should include your UCPH ID, e.g. “AMAS_ProbSet2_xdn365.pdf” or “xdn365.pdf”
- If you have any issues/problems email Jason or Jean-Loup
- Good luck!!!

Problem 1 (0.5 points)

- Make a cover page which includes your name, UCPH ID, UCPH logo, date, appropriate title, and plot of the χ^2 probabilility distribution function w/ 1 DoF over the range of $0 \leq \chi^2 \leq 10$
  - These should be the ONLY things on the first page. All other solutions to later problems should start on a new page.

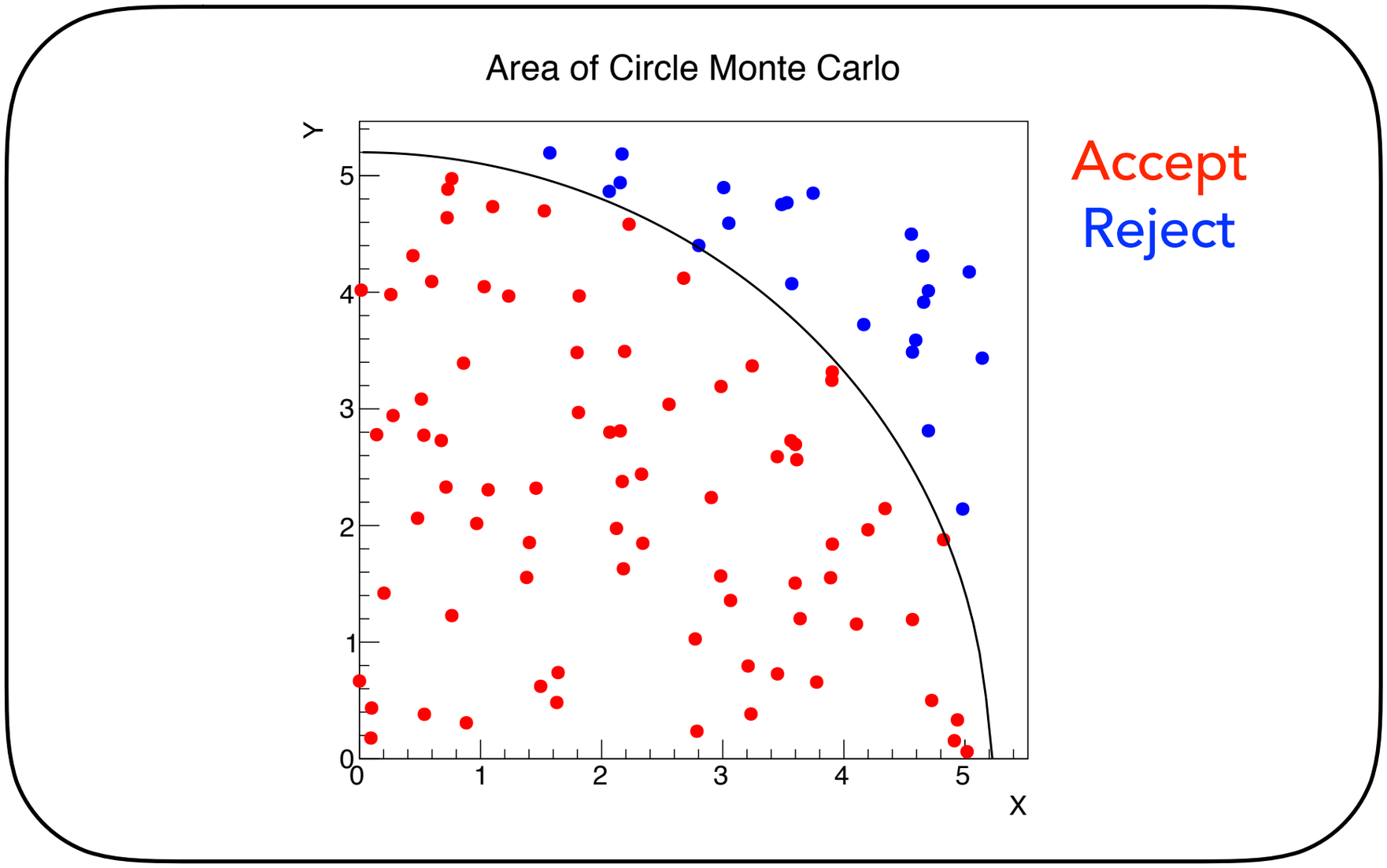
Problem 2 (1.5 points)

- There is a file posted online (http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2019/data/ProblemSet2_Problem2.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 2 (cont.)

- 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

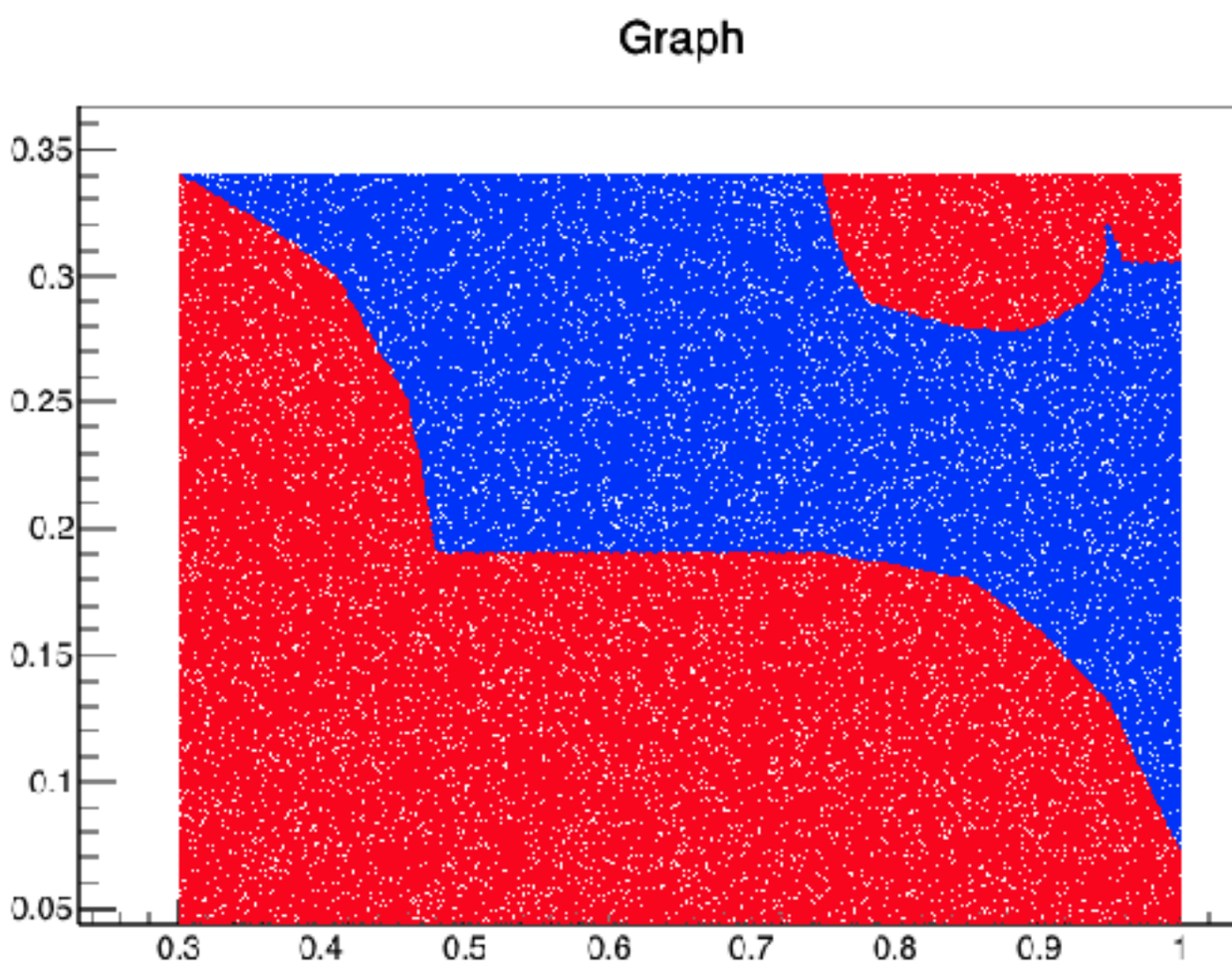


Problem 3a (Solution)

- Here I see that there is a reflection around x=1, and so I can calculate the area better using half of the points and then multiplying by 2.
- I made 2 linear splines and then did accept/reject
  - One spline was the lower outline of the symbol, and the other spline was the upper outline of the symbol
  - Accepted the (x, y) point if it was between the lower spline and upper spline
- The image is a crude outline of the batman symbol

Accept

Reject




```
The MC integrated area is: 0.164053
The actual area is: 0.163910
Between the MC integrated area and the true area, the
difference is: 0.000875
```

Problem 3 (3 points total)

- A pace-maker is generated by assembly facilities in 5 different countries with the following defective rates per factory and total worldwide production percentage

Facility	Total % produced	% Defective
A_1	35	2
A_2	15	4
A_3	5	10
A_4	20	3.5
A_5	25	3.1

Problem 3a (1 pt.)

- The world-wide distribution of the actual pace-makers is decoupled from the proximity to the production facilities
- Suppose a pace-maker is found to be defective (D), what is the probability it came from the A_2 facility?
- If a defective pace-maker is found, which facility is it most likely to be from?

Problem 3a setup

- $P(D) = P(D|A_1)p(A_1)+P(D|A_2)P(A_2) + \ldots$
- $P(A_2|D) = P(D|A_2)P(A_2)/P(D) = \ldots.$

```
W/ defective, the prob. of coming from A1: 0.213740
W/ defective, the prob. of coming from A2: 0.183206
W/ defective, the prob. of coming from A3: 0.152672
W/ defective, the prob. of coming from A4: 0.213740
W/ defective, the prob. of coming from A5: 0.236641
```

Problem 3b (1 pt.)

- The CEO of Slightly Evil Inc. wants to ensure that all the facilities have the same probability of being identified with a failed pace-maker, but still wants the least total defects and with no changes to the per facility production. How must the percentage of defects change from each facility to accommodate this goal?
  - Make a list of the altered/updated defective rates for each facility
  - Defective product rates can **only** increase

```
W/ defective, the updated defective rate for A1: XXXX
W/ defective, the updated defective rate for A2: XXXX
W/ defective, the updated defective rate for A3: XXXX
W/ defective, the updated defective rate for A4: XXXX
W/ defective, the updated defective rate for A5: XXXX
```

Problem 3b solution

```
W/ defective, the updated defective rate for A1: 0.022143
W/ defective, the updated defective rate for A2: 0.051667
W/ defective, the updated defective rate for A3: 0.155000
W/ defective, the updated defective rate for A4: 0.038750
W/ defective, the updated defective rate for A5: 0.031000
```

Question 3c (1 pt.)

- Repeat question 3b using the new table below, which is in fractions and not percent

Facility	total produced	defective rate
A1	0.27	0.02
A2	0.1	0.04
A3	0.05	0.1
A4	0.08	0.035
A5	0.25	0.022
A6	0.033	0.092
A7	0.019	0.12
A8	0.085	0.07
A9	0.033	0.11
A10	0.02	0.02
A11	0.015	0.07
A12	0.022	0.06
A13	0.015	0.099
A14	0.008	0.082

Problem 3c solution

```
W/ defective, the updated defective rate for A1: 0.022037
W/ defective, the updated defective rate for A2: 0.059500
W/ defective, the updated defective rate for A3: 0.119000
W/ defective, the updated defective rate for A4: 0.074375
W/ defective, the updated defective rate for A5: 0.023800
W/ defective, the updated defective rate for A6: 0.180303
W/ defective, the updated defective rate for A7: 0.313158
W/ defective, the updated defective rate for A8: 0.070000
W/ defective, the updated defective rate for A9: 0.180303
W/ defective, the updated defective rate for A10: 0.297500
W/ defective, the updated defective rate for A11: 0.396667
W/ defective, the updated defective rate for A12: 0.270455
W/ defective, the updated defective rate for A13: 0.396667
W/ defective, the updated defective rate for A14: 0.743750
```

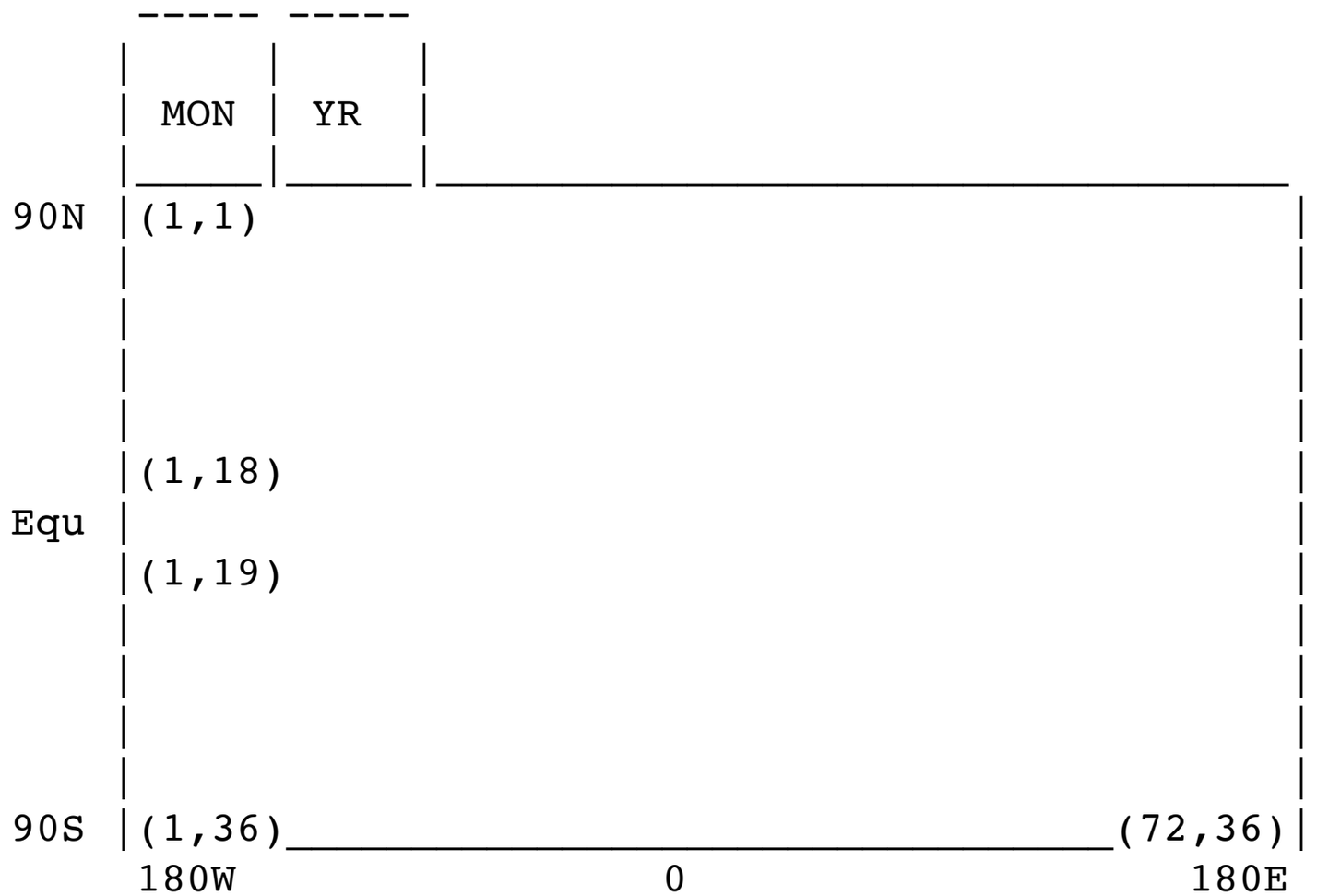
Problem 4 (2 points total)

- 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/AdvancedMethodsInAppliedStatistics2019/data/GlobalTemp_1.txt
  - May 2017 at http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2019/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 (1 pt.)

- 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)

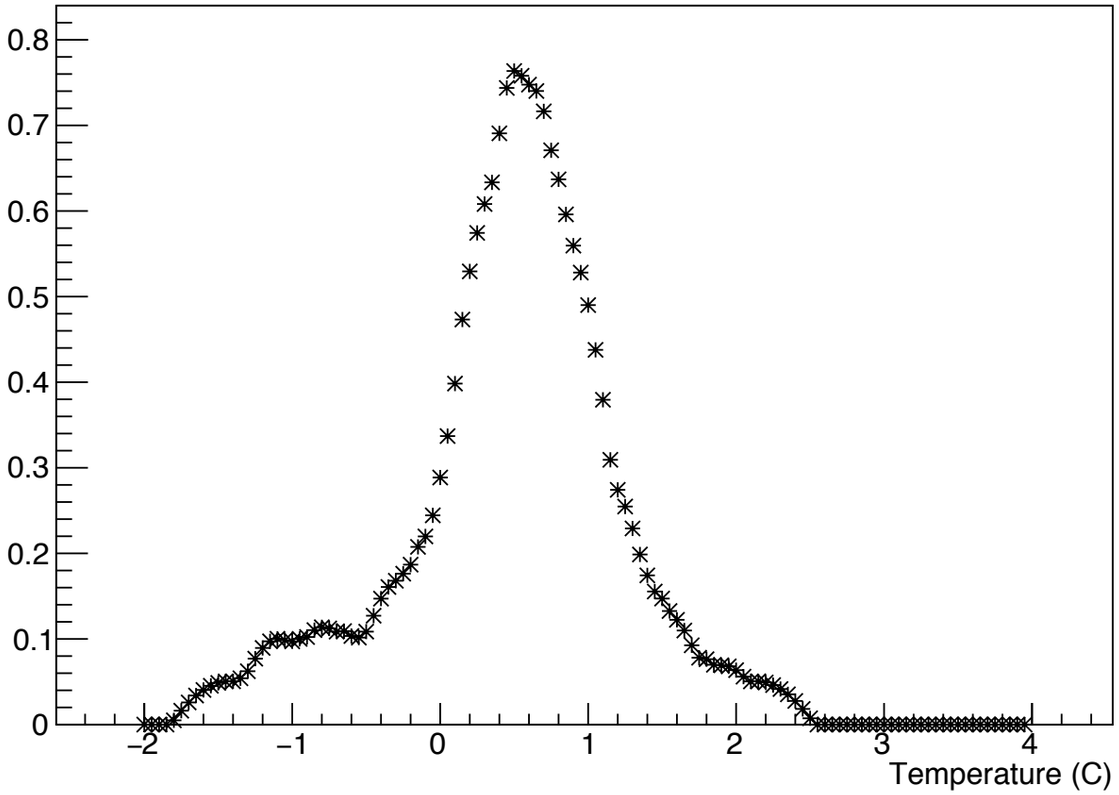
- Calculate the integral of P_{KDE}(temp) for 1997 and 2017:
  - over the range -2 C to +4 C
  - over the range of -2 C to 0 C

Problem 4a Solutions

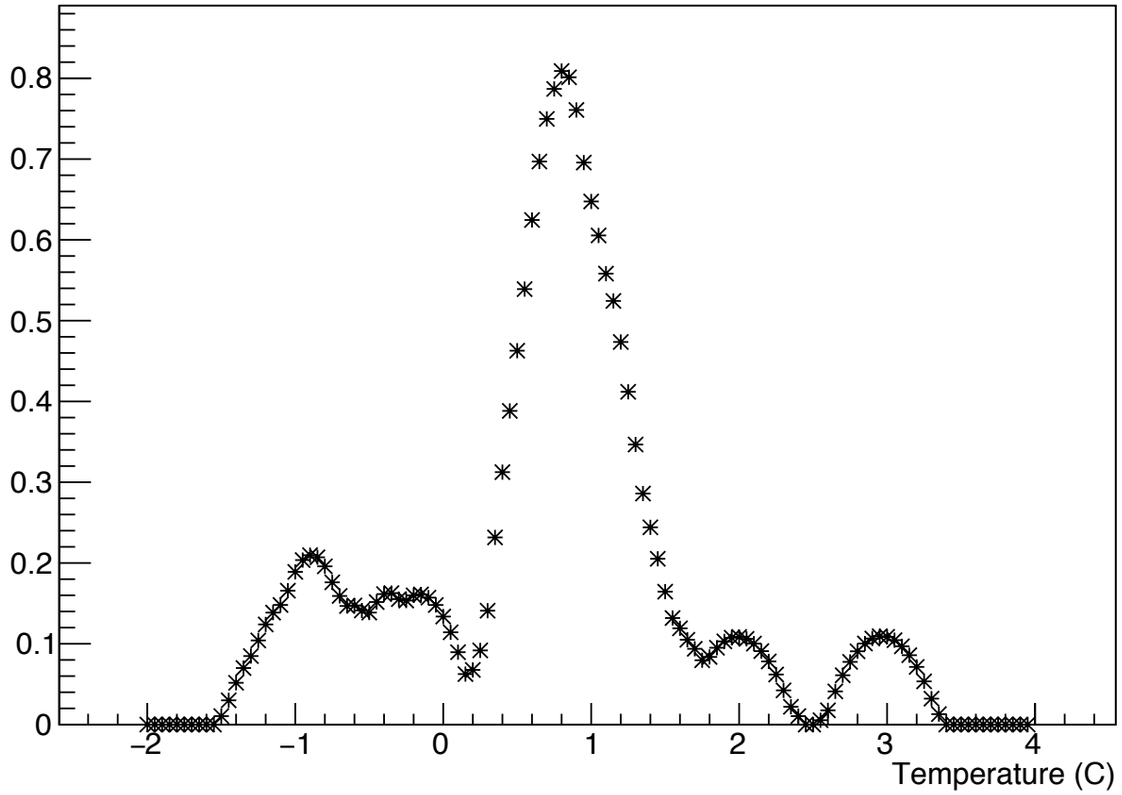
- For the integrals I did a fine-grained scan
  - Between -2 C to +4 C I get 0.9999984375 for both
  - Between -2 C and 0 I get the following

```
1997 KDE: 0.191753917335
2017 KDE: 0.215906055262
```

Joint KDE for 1997 w/ Epanechnikov kernel



Joint KDE for 2017 w/ Epanechnikov kernel



python code can be found online at
http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2019/ProblemSet2_Problem4_KDE.pdf

Problem 4b (1 pt.)

- 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

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

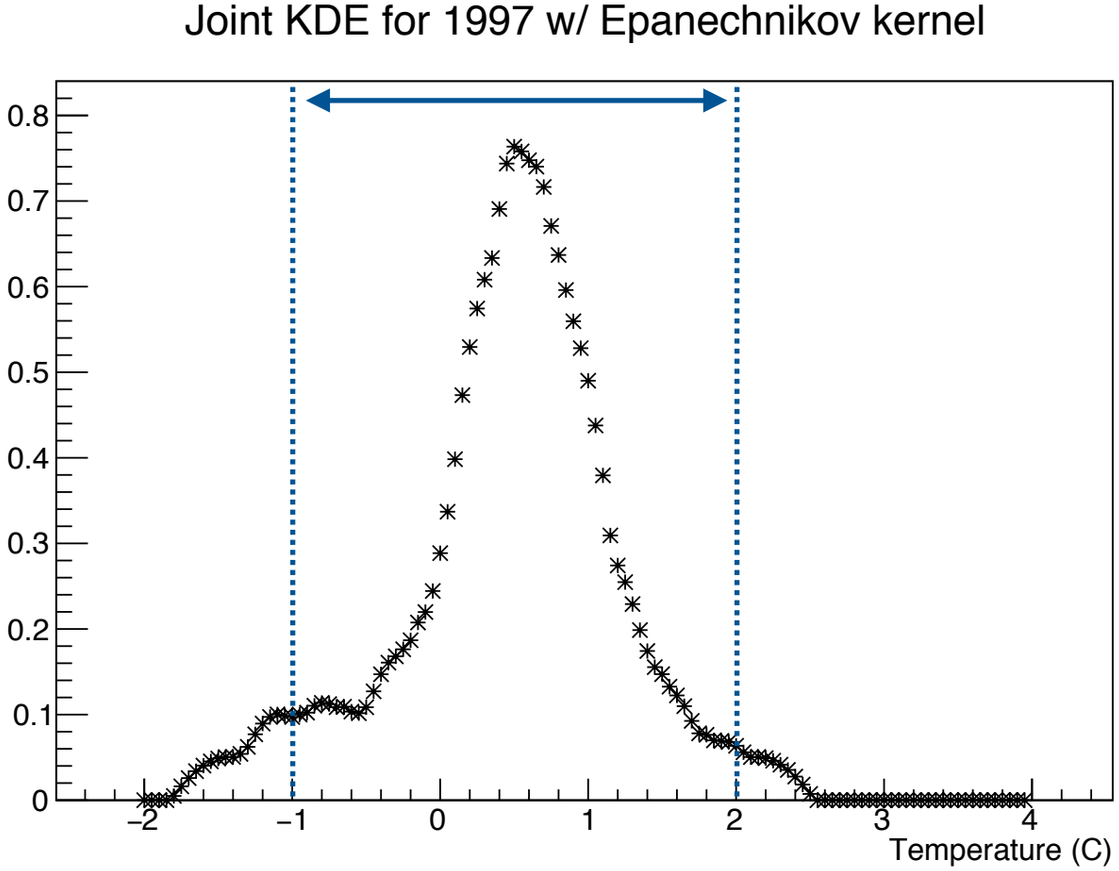
- Submit your 1000 samples as an ASCII txt file:
  - each entry on a separate line for 1000 total lines in the file
  - File name should be your last name and "_KDE_1000_samples.txt", e.g. "koskinen_KDE_1000_samples.txt"

Problem 4b solutions

- For 1000 samples, the likelihood will probably go beyond computer precision, and the ensuing ratios will go to zero which would be incorrect. But, doing the comparison as the difference in the natural log and then going back to the likelihood ratio via e^(diff) should not go beyond computer precision
- Some students will produce the ln-likelihood ratio, i.e. $\ln(L_0)/\ln(L_1)$ which is **not** the same as $\ln(L_0/L_1)$
- The likelihood ratios should be renormalized over the temperature range of -1 C to +2 C for this portion

Problem 4b solutions

- The integral of the two P should be 1.0 from the ranges of -1 C to +2 C. This is just a scale factor, and doesn't necessarily change the Monte Carlo generation, but will change the likelihood ratios.



Problem 5 (3 pts. total)

- Consider an experiment set up to measure the lifetime of an unstable nucleus, N, using the reaction: $A \to Ne\bar{\nu},\ N \to Xp$
- The creation and subsequent decay of N has a signature of an electron and proton. The lifetime of each N, which follows the PDF $f = \frac{1}{b}e^{-t/b}$, is measured from the time, observing the electron and proton with a gaussian resolution of σ_t
  - Normally the lifetime would be represented by 'τ' instead of 'b', but this becomes a disaster when dealing with t, t', and τ
- The expected PDF is then the convolution of the exponential decay and the gaussian resolution:

$$f(t; b, \sigma_t) = \int_0^\infty \frac{e^{-\frac{(t-t')^2}{2\sigma_t^2}}}{\sqrt{2\pi}\sigma_t} \frac{e^{-t'/b}}{b} dt'$$

Problem 5 (cont.)

- Neither b nor σ_t are explicitly known, and we want to test whether b=1 *second* can be rejected. We can do so via a hypothesis test, where the two hypotheses H_0 and H_1 are given as:

$$b_0 = 1.0\ s$$

$$H_0 : b = b_0$$

$$H_1 : b \neq b_0$$

- Use the likelihood ratio test:

$$\lambda = \frac{\mathcal{L}(\hat{\omega})}{\mathcal{L}(\hat{\Omega})} \qquad \begin{aligned} &\omega \text{ given by } b = b_0,\ 0 < \sigma_t < \infty \\ &\Omega \text{ given by } 0 < b < \infty,\ 0 < \sigma_t < \infty \end{aligned}$$

- Where $\mathcal{L}(\hat{\omega})$ is the value of the null hypothesis likelihood calculated using the maximum likelihood estimator(s) $\hat{\omega}$

Problem 5a (1 pt.)

- There are 20000 events in the online file below, which corresponds to 100 simulated pseudo-experiments where each pseudo-experiment has 200 events
  - The data is at http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2019/data/ProblemSet2_Prob5_NucData.txt
    - The first 200 data points are the first pseudo-experiment, the second 200 data points are the second pseudo-experiment, the third 200 data points are the third pseudo-experiment, etc.
    - The data entries are given in units of seconds (s)
- For each of the 100 pseudo-experiments find the values of the ln-likelihoods that are maximized for the two hypotheses
- As a histogram, plot the values of $-2\ln(\lambda)$

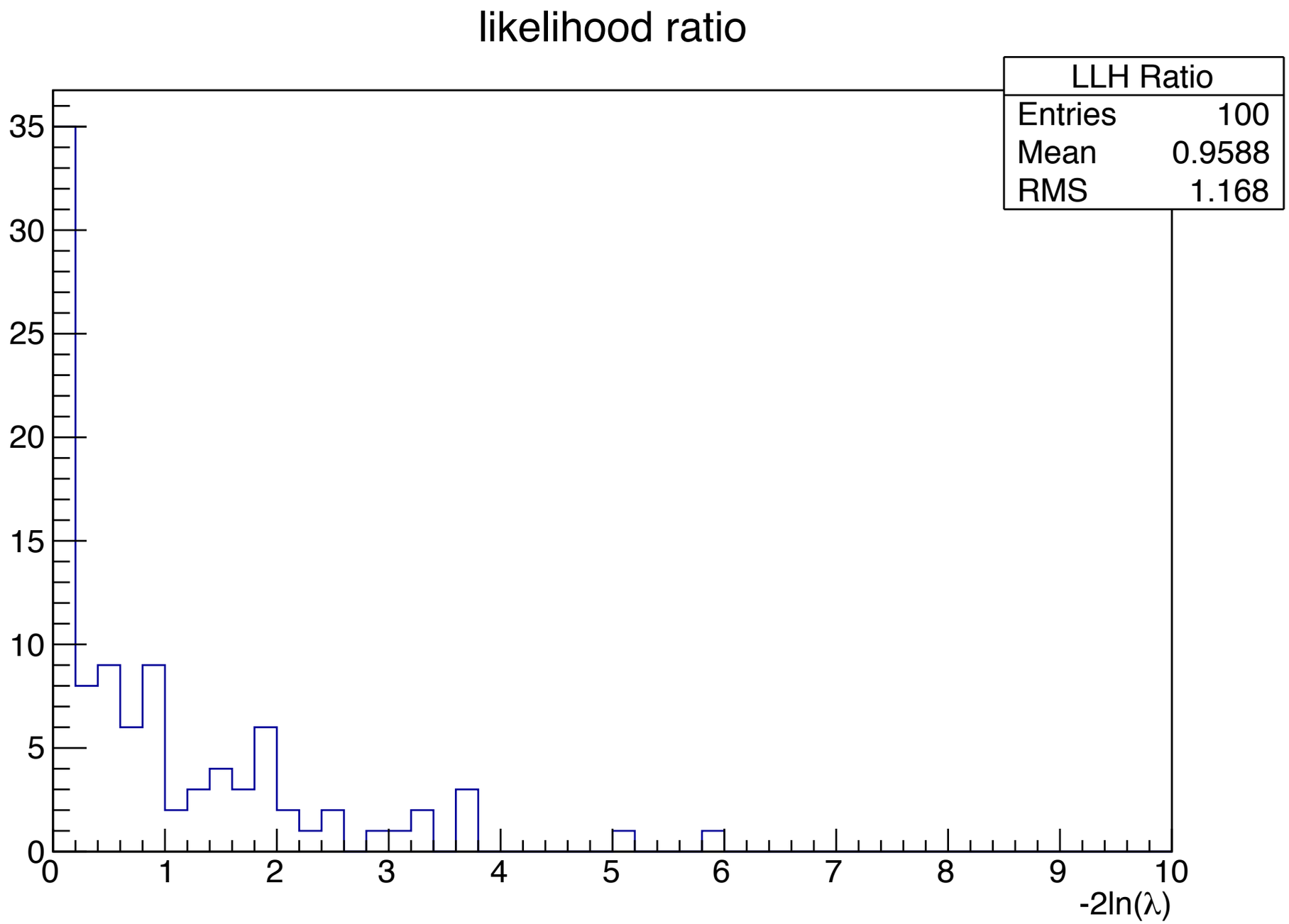
Problem 5a setup

- Simulation to measure the lifetime of an unstable nucleus. The probability distribution function:

$$= \frac{1}{2b} \exp\left(\frac{\sigma_t^2}{2b^2} - \frac{t}{b}\right) \mathrm{erfc}\left(\frac{\sigma_t}{\sqrt{2}b} - \frac{t}{\sqrt{2}\sigma_t}\right)$$

  - Which is kinda a pain to solve and figure out
- So the above ends up being the probability distribution function from which to sample
- Alternatively, you can use numerical integration to get the normalization, or just calculate it for each b and σ_t step taken by the minimizer

Solution 5a


likelihood ratio
LLH Ratio
Entries 100
Mean 0.9588
RMS 1.168
35
30
25
20
15
10
5
0
0 1 2 3 4 5 6 7 8 9 10
-2ln(λ)

Problem 5b (1 pt.)

- Is the distribution of -2ln(λ) chi-squared distributed?
  - Be sure to use the correct number of degrees of freedom
  - Justify and explain your answer
- How many pseudo-experiments have -2ln(λ) > 2.706?
- Is the number of pseudo-experiments with -2ln(λ) > 2.706 consistent with the expectation from a chi-squared distributed test-statistic and 100 data 'points'?

Solution 5b

- The distribution is chi-squared distributed and with 1 degree of freedom 2.706 is 90%, so I expect 10 pseudo-trials
- In actuality I get 9 pseudo-experiments above 2.706
- Yes. 10 and 9 are pretty similar.
    - Jean-Loup got 11 trials. This is odd, because it's all the same data and the fitting shouldn't be too different.

Problem 5c (1 pt.)

- Using all 20000 events as a single pseudo-experiment, can the null hypothesis (H_0) be rejected at 3σ confidence?

Solution 5c

- Using all 20000 events as a single data set I calculate the likelihood ratio test-statistic ($-2\ln(\lambda) \cong 0.68$). Because this test-statistic is chi-squared distributed and the difference in the number of parameters between H_0 and H_1 is 1, the value of $-2\ln(\lambda)$ must be >9 in order to reject H_0 at $> 3\sigma$.
- The value of 0.68 is absolutely less than 9, so we **cannot** reject H_0 at $> 3\sigma$.