# Exam


D. Jason Koskinen

koskinen@nbi.ku.dk


*Advanced Methods in Applied Statistics*

*Feb - Apr 2016*

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.
- Working in groups or any communication about the problems is prohibited. Using the internet as a resource is encouraged, but soliciting any help is also prohibited.
- Please do not zip the files for submission
- Some questions have multiple parts. For full credit, all parts must be done.

# Grading

- Each problem is worth 20% of the Final Exam grade
- At the discretion of the grader, small amounts of bonus credit, e.g. 1-2%, can be acquired for exceptionally insightful or thorough answers, even when the answer might not be fully correct
- Questions with 3 parts, each have -7% for each part, but cannot go past -20% if all 3 parts are skipped

# Problem 1

- There is a file posted online which has 5 columns, each representing a physical observable of interest generated from some underlying function. There are thousands of entries, i.e. rows.
  - http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2016/data/Exam_Prob1.txt
  - The first 3 variables/columns are independent distributions with no correlation to the other variables
  - The last two variables/columns are unused
  - 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 one of the functions shown at right
- Note that these functions may be unnormalized
  - Hint: Some will require a normalization to convert them to probability distribution functions

$$f(x) \propto \begin{array}{l} \frac{1}{x+5}\sin(ax) \\ \sin(ax)+1 \\ \sin(ax^2) \\ \sin(ax+1)^2 \\ x\tan(x) \\ 1+ax+bx^2 \\ a+bx \\ e^{-\frac{(x-\mu)^2}{2\sigma^2}} \end{array}$$

---

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

# Problem 1a

- Using the separate data from the first three columns, identify the function on slide 4 from which each was generated and find the best-fit values for that distribution using the maximum likelihood method
    - 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
    - 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 is necessary.
- Data in columns 1 and 2 have artificially truncated ranges
    - Column 1 is only sampled in the independent variable from 0 to 2
    - Column 2 is only sampled from -1 to 1
    - Column 3 is not truncated

# Problem 1a Solutions

- Distribution 1 is
  - F = ampl * numpy.sin( x_*freq) + ampl
  - ampl_true = 0.87 and **freq_true = 12.11±0.017**
  - I don't actually request the uncertainty, so it's optional for all answers to Problem 1. Maybe change for future years.
  - The amplitude variable doesn't actually matter too much here because it's only used for scaling, and therefore any value would produce the same data distribution

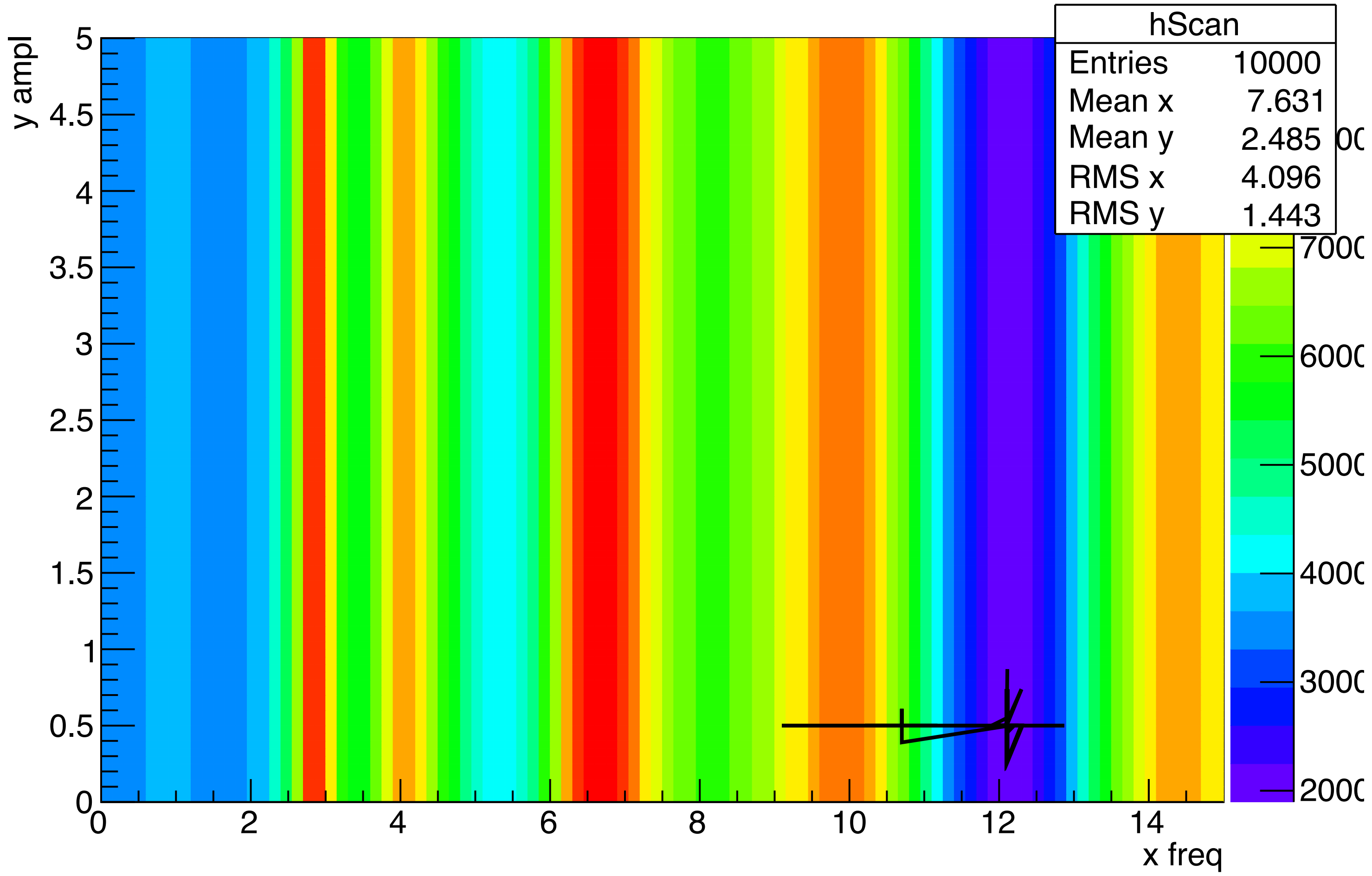
```
|      | Name  |  Value    | Para Err |    Err-    |    Err+    |  Limit-   |  Limit+   |          |
----------------------------------------------------------------------------------------------------
|    0 |  ampl =  0.5      |  0.7178  |            |            |  0.01     |  1        |          |
|    1 |  freq =  12.1     |  0.01723 |            |            |  0                  |  15       |          |
----------------------------------------------------------------------------------------------------

************************************************************************
{'freq': 12.101346311072076, 'ampl': 0.5000000000002252}
best-fit LLH:  1882.37351494
true LLH:  1882.528733
```

# Problem 1a Solutions

- It is also possible to use the $\sin(ax+1)^2$ distribution with the solution of $a \backsimeq 5.9$

# Problem 1a Solutions

- Distribution 1 is
  - The freq. dimension of the LLH landscape has a lot of minima, so if people start their minimizer near zero, they'll get a bad fit.
  - The likelihood is easy enough to scan over though, so they should be able to drop their minimizer into the right region or else use a MCMC



# Problem 1a Solutions

- Distribution 2 is
  - F = (1 + alpha*x_ + beta*x_**2) where the normalization factor influences the best-fit (2 + 2*beta/3)
  - a_true = 0.20 and b_true = 0.40
  - **a = 0.195±0.042** and **b = 0.419±0.088**
  - We covered this exact likelihood in class so there's no excuse for anyone to get this wrong

```
---------------------------------------------------------------------------------------------
|      | Name  |  Value    | Para Err |   Err-    |   Err+    |  Limit-  |  Limit+  |          |
---------------------------------------------------------------------------------------------
|    0 | alpha =  0.1953   |  0.04239 |           |           |          |          |          |
|    1 |  beta =  0.4189   |  0.08844 |           |           |          |          |          |
---------------------------------------------------------------------------------------------
```

```
**********************************************************************
{'alpha': 0.1952721738139312, 'beta': 0.4188894802371008}
Beta best-fit LLH:  4519.26969334
Beta true LLH:  4517.66933547
```

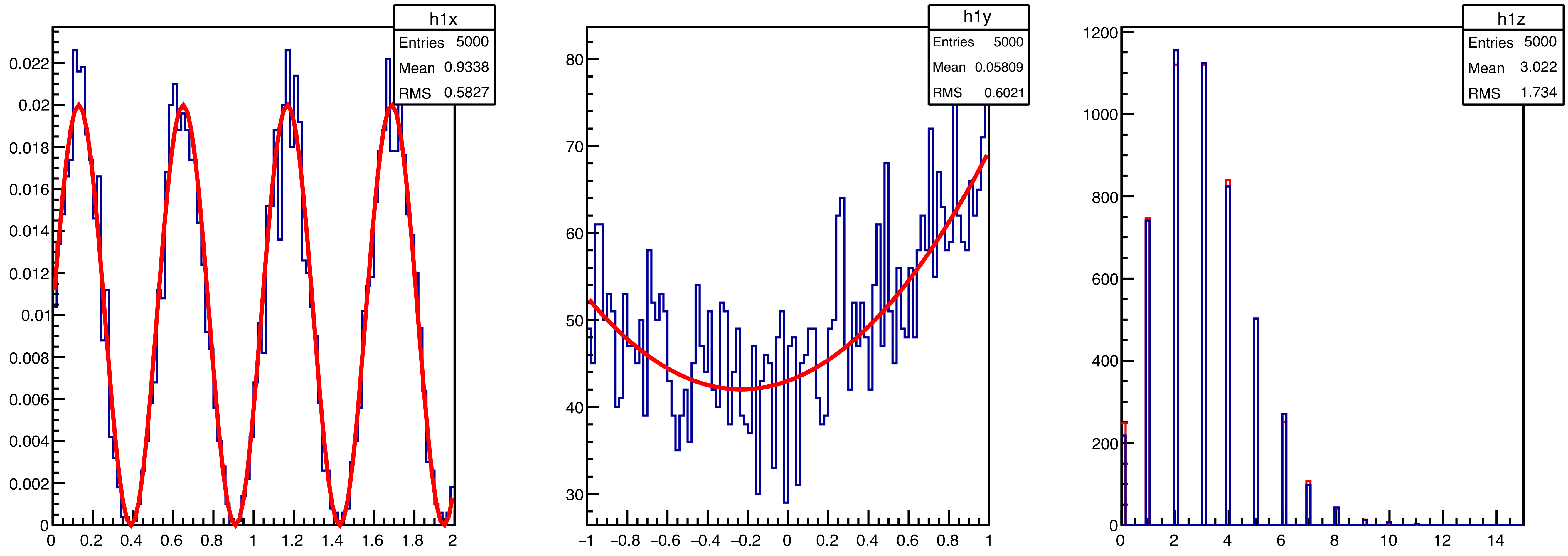
# Problem 1a Solutions

- Distribution 3 is a discrete poisson with an expectation of 3
    - Don't even need to fit this if you know what it looks like
    - But, it can also be a binomial for large values of N and low value of p, e.g. $\mathcal{O}(140)$ and $\mathcal{O}(0.02)$ respectively.

# Problem 1b

- Plot the data and the corresponding best-fit function on the same plots
  - 3 separate 1-dimension 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

# Solution 1b



h1x
Entries 5000
Mean 0.9338
RMS 0.5827
0.022
0.02
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8 2
h1y
Entries 5000
Mean 0.05809
RMS 0.6021
80
70
60
50
40
30
−1 −0.8 −0.6 −0.4 −0.2 0 0.2 0.4 0.6 0.8 1
h1z
Entries 5000
Mean 3.022
RMS 1.734
1200
1000
800
600
400
200
0
0 2 4 6 8 10 12 14

# Problem 2

- A cancer study in 1991 conducted in Wisconsin collected data from ~700 patients. There were 9 variables associated with the digitized image of a fine needle aspirate biopsy sample of a tissue mass. Each variable has discrete values from 1-10. There was also the patient identifier (code number) and whether the sample mass was ultimately benign or malignant.

| # Attribute | Domain/Range |
|---|---|
| 1. Sample code number | id number |
| 2. Clump Thickness | 1 - 10 |
| 3. Uniformity of Cell Size | 1 - 10 |
| 4. Uniformity of Cell Shape | 1 - 10 |
| 5. Marginal Adhesion | 1 - 10 |
| 6. Single Epithelial Cell Size | 1 - 10 |
| 7. Bare Nuclei | 1 - 10 |
| 8. Bland Chromatin | 1 - 10 |
| 9. Normal Nucleoli | 1 - 10 |
| 10. Mitoses | 1 - 10 |
| 11. Class: | (2 for benign, 4 for malignant) |

# Problem 2a

- There are two files online: training data and blind data
  - The following training data includes the aforementioned variables as well as whether the biopsy was benign or malignant (www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2016/data/breast-cancer-wisconsin_train-test.txt)
  - The following data includes the same variables, but the information of whether the biopsy was benign or malignant has been removed, i.e. a blind sample, (www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2016/data/breast-cancer-wisconsin_mod_real.txt)
- Using some method (straight cuts, support vector machine, boosted decision tree, etc.) and the training data, come up with a classification algorithm which uses the 9 variables to identify malignant and benign tissue samples
  - Note: the separate variables can be assumed to have the same features and shape between the training and blind sample

# Problem 2a cont.

- With a developed algorithm, run the classifier over the training sample and calculate the efficiency of identifying a malignant mass
    - It is possible to get 100% efficiency, provided the method is overtrained
    - But, you will have to use the same settings for classifying the blind sample in Problem 2b
- Calculate the overall classification efficiency for the whole training sample
    - (classified_true_malignant + classified_true_benign)/(total_trainingtest_sample)

# Problem 2a Solution

- Calculate the overall classification efficiency for the whole training sample
    - (classified_true_malignant + classified_true_benign)/(total_trainingtest_sample)
- Students should be at least getting better than 90%, and possibly ~99%, all without overtraining.

# Problem 2b

- Using the same setting(s) as developed in Problem 2a, run the classifier over all the entries in the blind sample (breast-cancer-wisconsin_mod_real.txt)
    - Produce a text file which contains only the ID of the samples which your classifier classifies as malignant (last_name.malignant_ID.txt)
    - Produce a text file which contains only the ID of the samples which your classifier classifies as benign (last_name.benign_ID.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().
        - Example online at http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2016/data/koskinen.benign_ID.txt
    - Any and all duplicates, i.e. two samples with the same ID, should be kept and included in the text files and analysis

# Problem 2b

- I was able to get 90-91% training, so the breakdown will be
    - >87% Full Credit
    - 83-87%   -2%
    - 80-83%  -4%
    - < 79%  -6%

# Problem 3

- Data has the following probability distribution function G(x) with an unknown value of $f$:

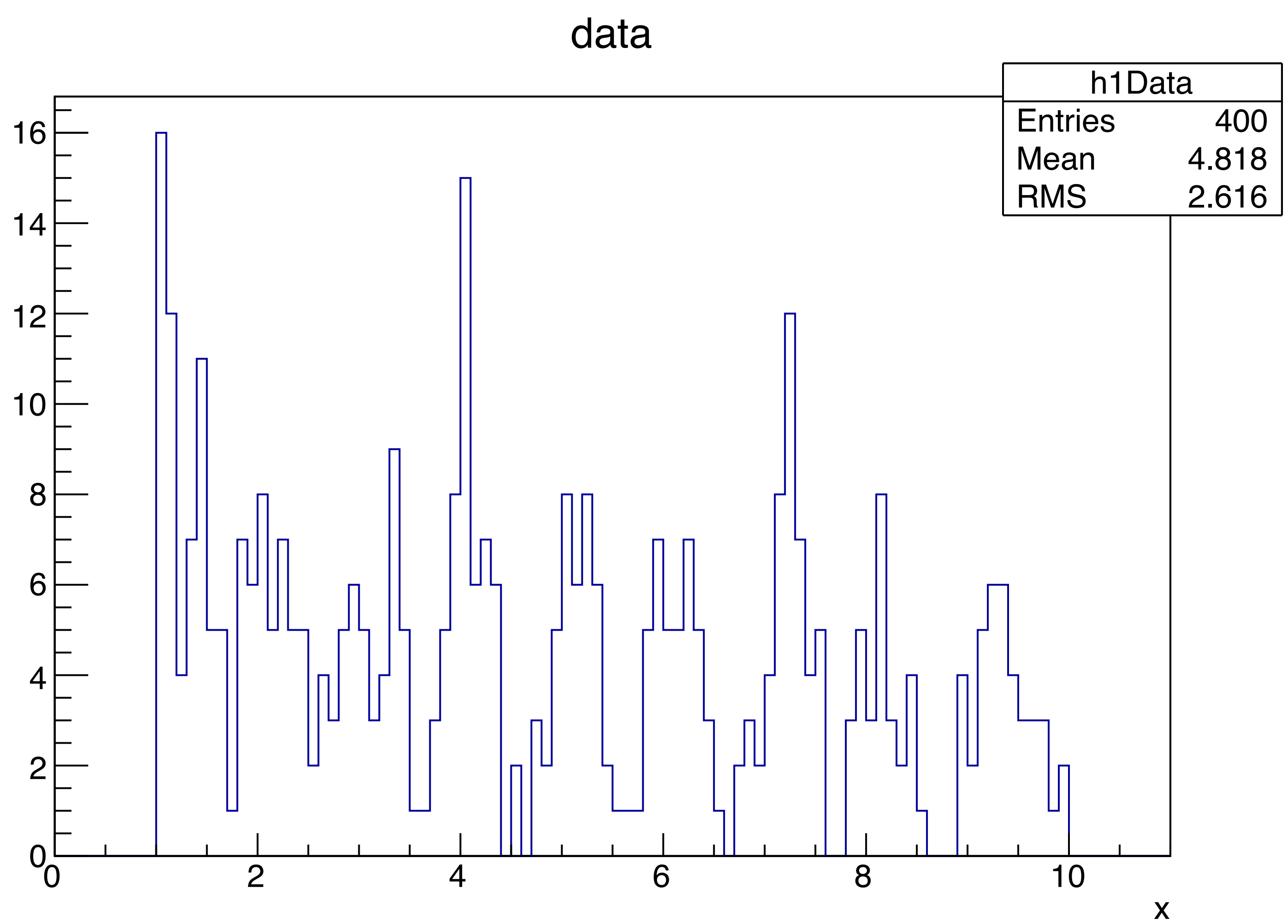
$$G(x;f) = \frac{1}{N}(\cos(f \cdot x) + \frac{3}{x+1} + 1) \qquad \text{for} \quad 1 \leq x \leq 10$$

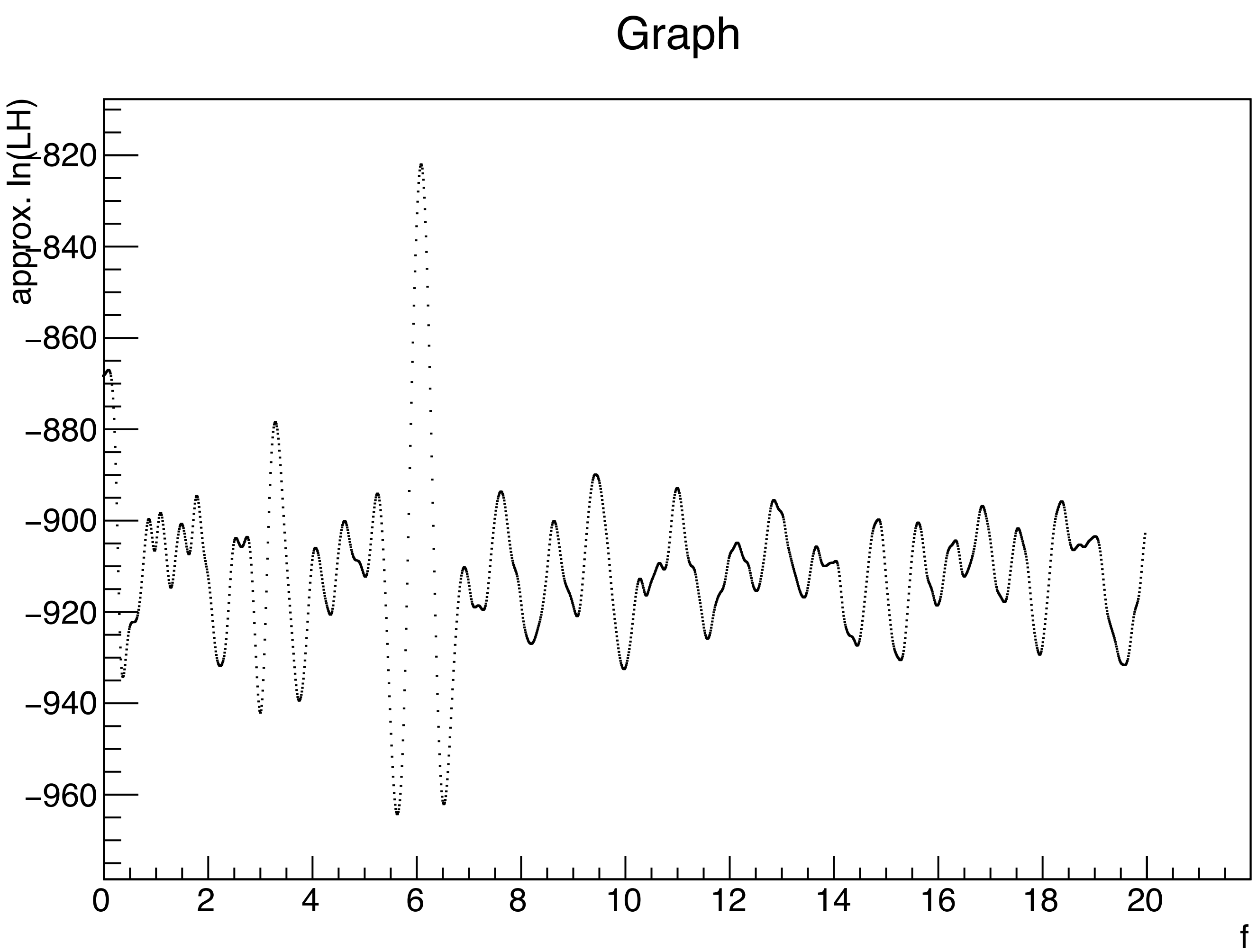
$$\text{normalization is} \qquad N = \frac{\sin(10f) - \sin(f)}{f} + 9 + \ln(\frac{1331}{8})$$

- The data is generated over a truncated range of 1 ≤ x ≤ 10 and can be found at
    - (www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2016/data/Exam_Prob3.txt)

Normalization from Wolfram|Alpha

# Problem 3 info



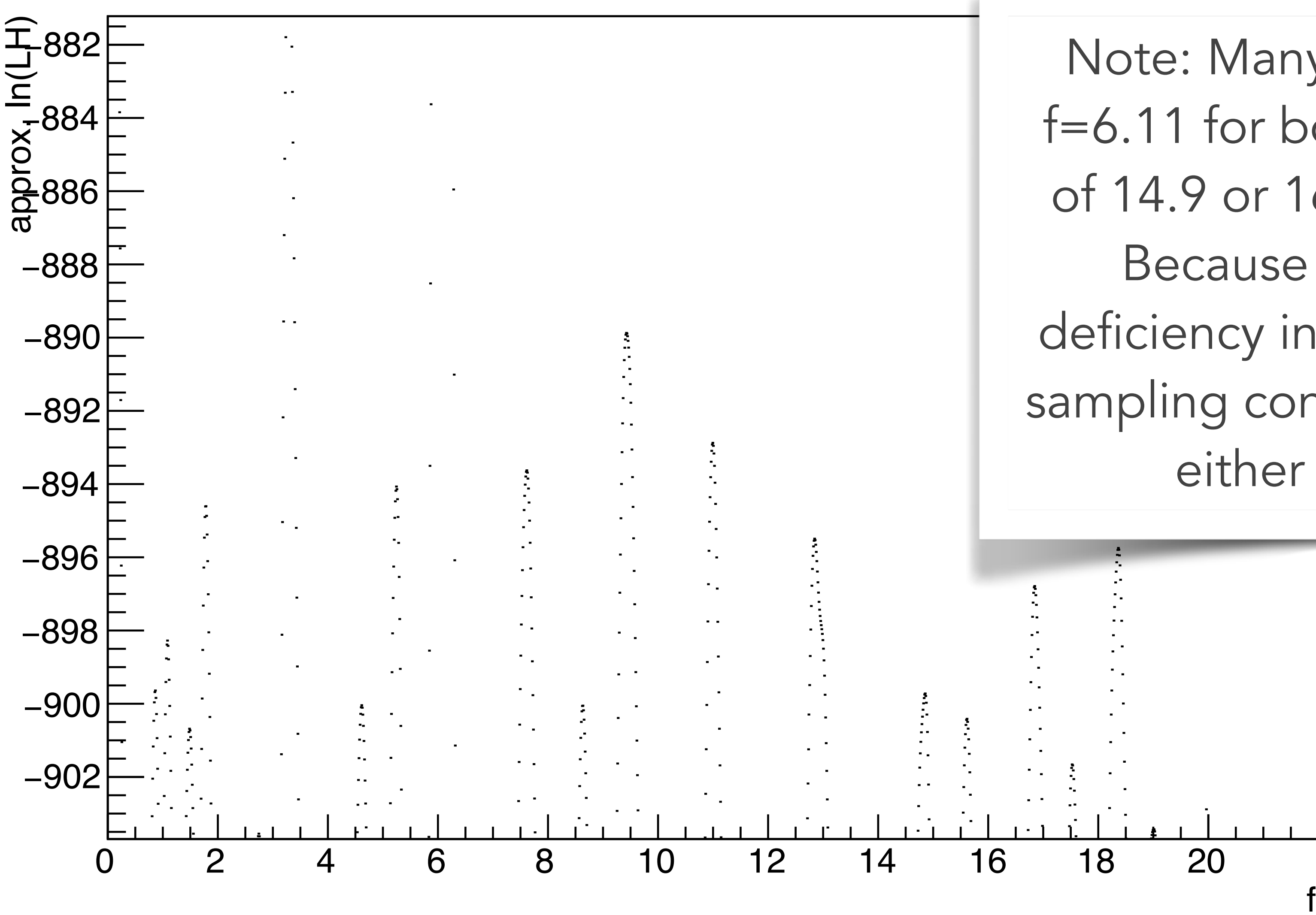



# Problem 3a

- We want to find f, but the likelihood as a function of f has many local minima/maxima. For this problem the search range will be constrained to $0 < f \leq 20$ for simplicity.
- From other sources, there is a bayesian prior on the 'true' value of f. The prior is a normalized gaussian with a gaussian width of 0.5, i.e. $\sigma_f = 0.5$, centered at f=15.
- Using the above prior and the function from the previous slide, find the maximum a posteriori (MAP) value of f, i.e. mode of the bayesian posterior
  - Use a Markov Chain Monte Carlo technique
  - Remember: from Lec. 4 $\mathcal{L}(\theta|x) = P(x|\theta)$ and $\mathcal{L}(\theta) = \prod_{i=0}^{N} f(x_i;\theta)$

# Solutions 3a

- Because the likelihood have 3-4 local maxima in the range of f from 12-18 (which is a 6 sigma range on the prior when f=15 and sigma=0.5) the MAP can be in somewhere near multiple values: 12.85, 14.9, 15.6, and 16.9



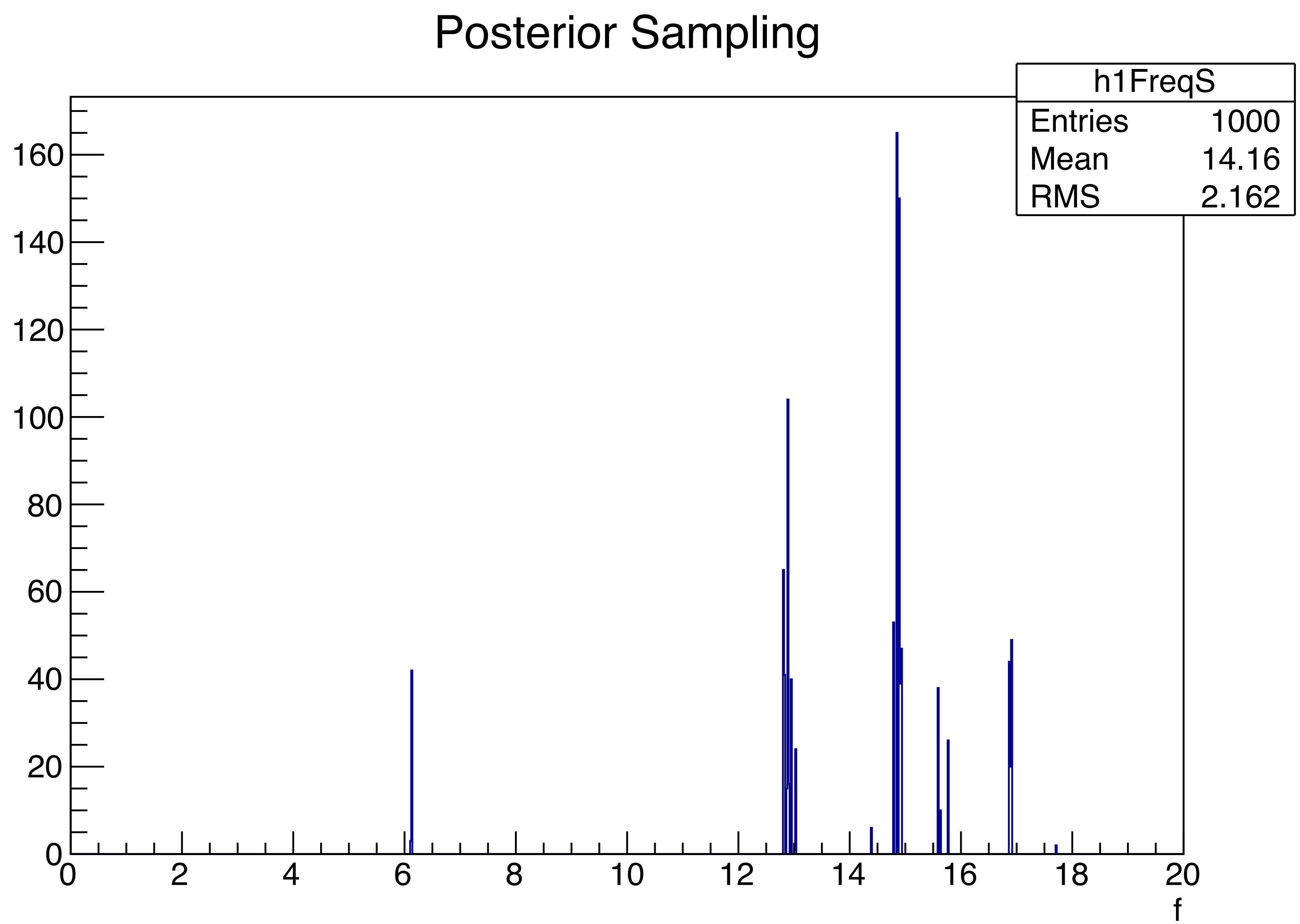

Note: Many students either got f=6.11 for both 3a and 3c, or one of 14.9 or 16.8 for both 3a or 3c. Because this shows a clear deficiency in the MCMC posterior sampling construction I considered either error to be -7%

# Problem 3b

- Using the same likelihood and prior from the previous slides for Problem 3, plot a histogram of the stable posterior distribution sampled points as a function of $f$
  - Whether to include or omit burn-in points/trials/steps is up to you
  - The histogram should include at least 500 sample points/trials/steps, but preferably less than 10001
  - Remember, the posterior distribution may have multiple local maxima/minima so you may end up with multiple bumps related to those minima/maxima

# Solution 3b

- Every so often a chain will find the likelihood best-fit value of 6.11, but it isn't a stable posterior because it's so very, very far away from the prior

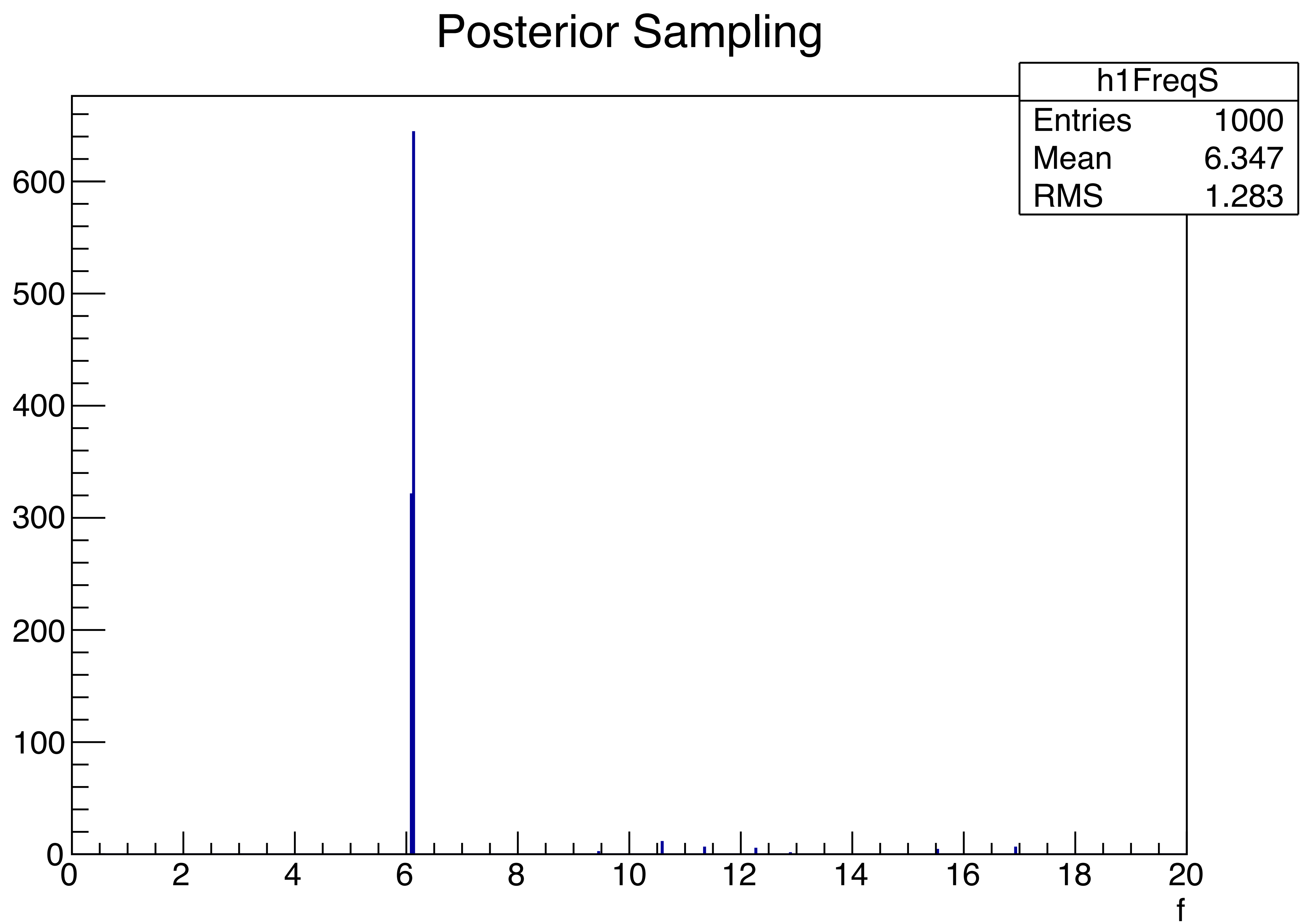


# Problem 3c

- Plot the posterior distribution using the same likelihood and the normalized gaussian prior center at $f$=15, but now the prior has a gaussian width of $\sigma_f = 2.5$
  - Does the posterior distribution change from when the gaussian width was 0.5?
  - If not describe why, and if so describe the changes.
  - What is the maximum a posterior value after including the new prior with a gaussian width of 2.5?

# Solution 3c

- With the sigma on the gaussian prior widened, the MCMC should find the true maxima of 6.11



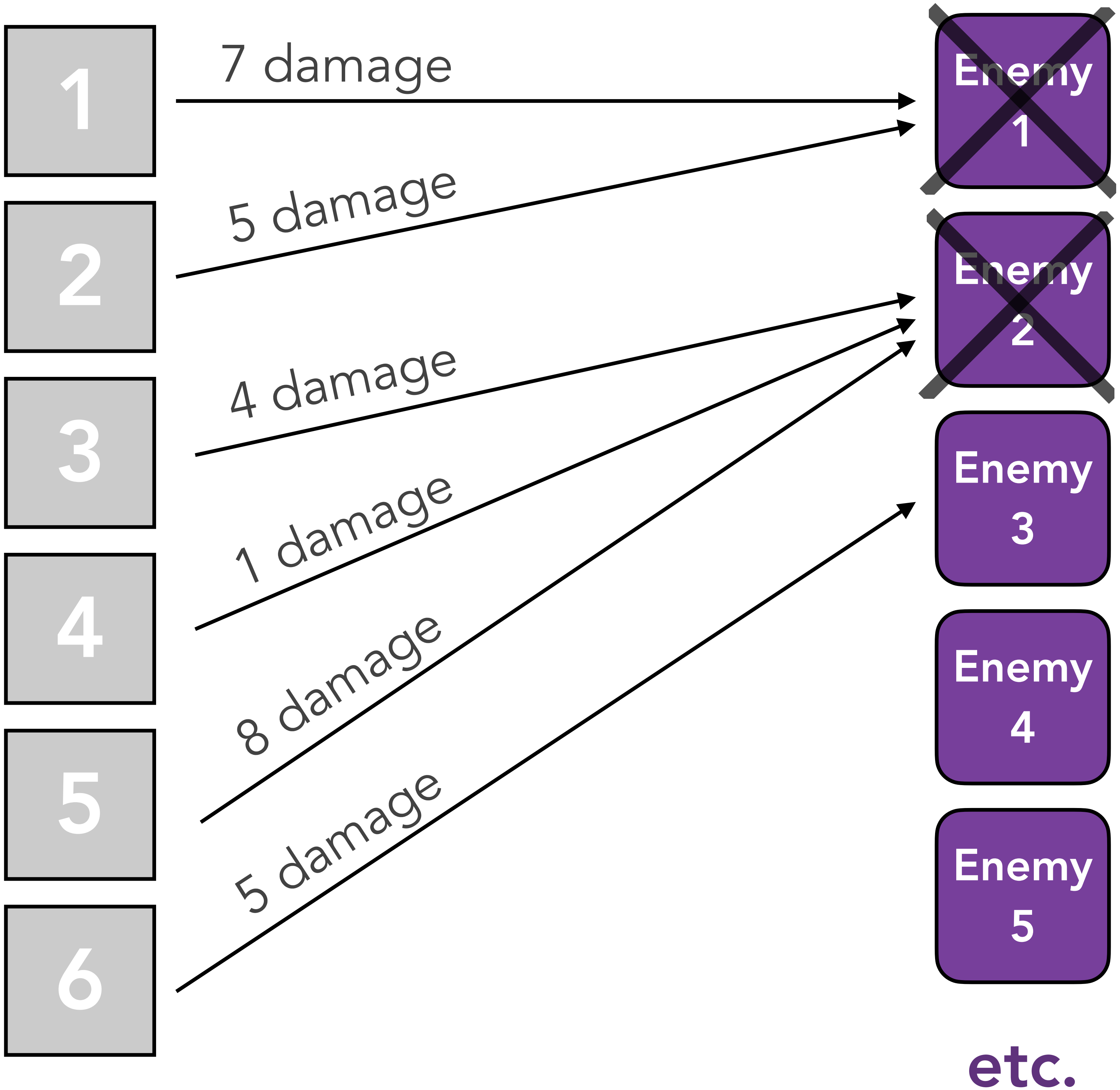
# Problem 4

- 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 likelihood

# Problem 4a

- 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 can individually take 12 damage before being defeated
  - 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
- Make a histogram 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

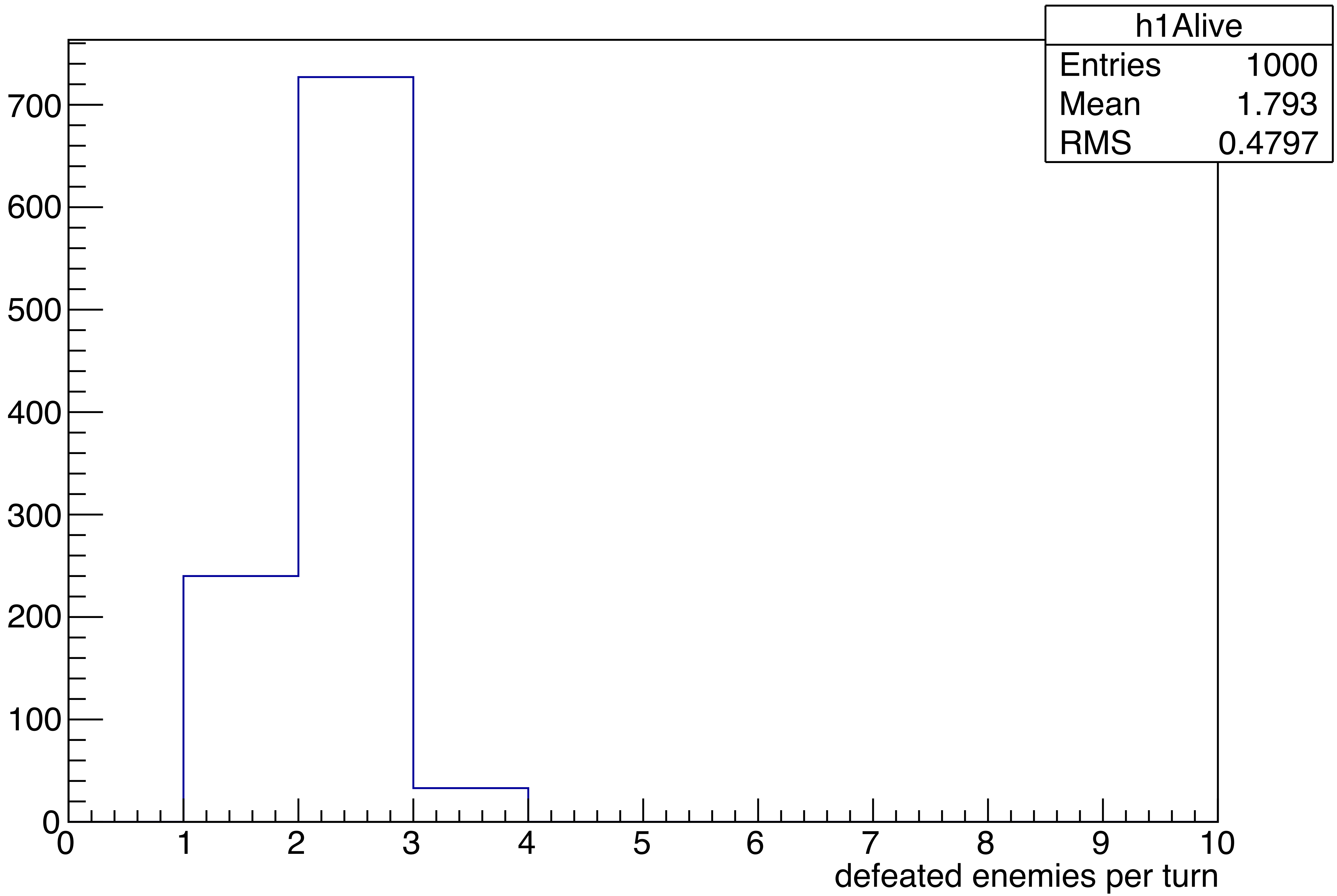
# Single Turn Example




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

In total 2 enemies were defeated for this turn

# Solution 4a

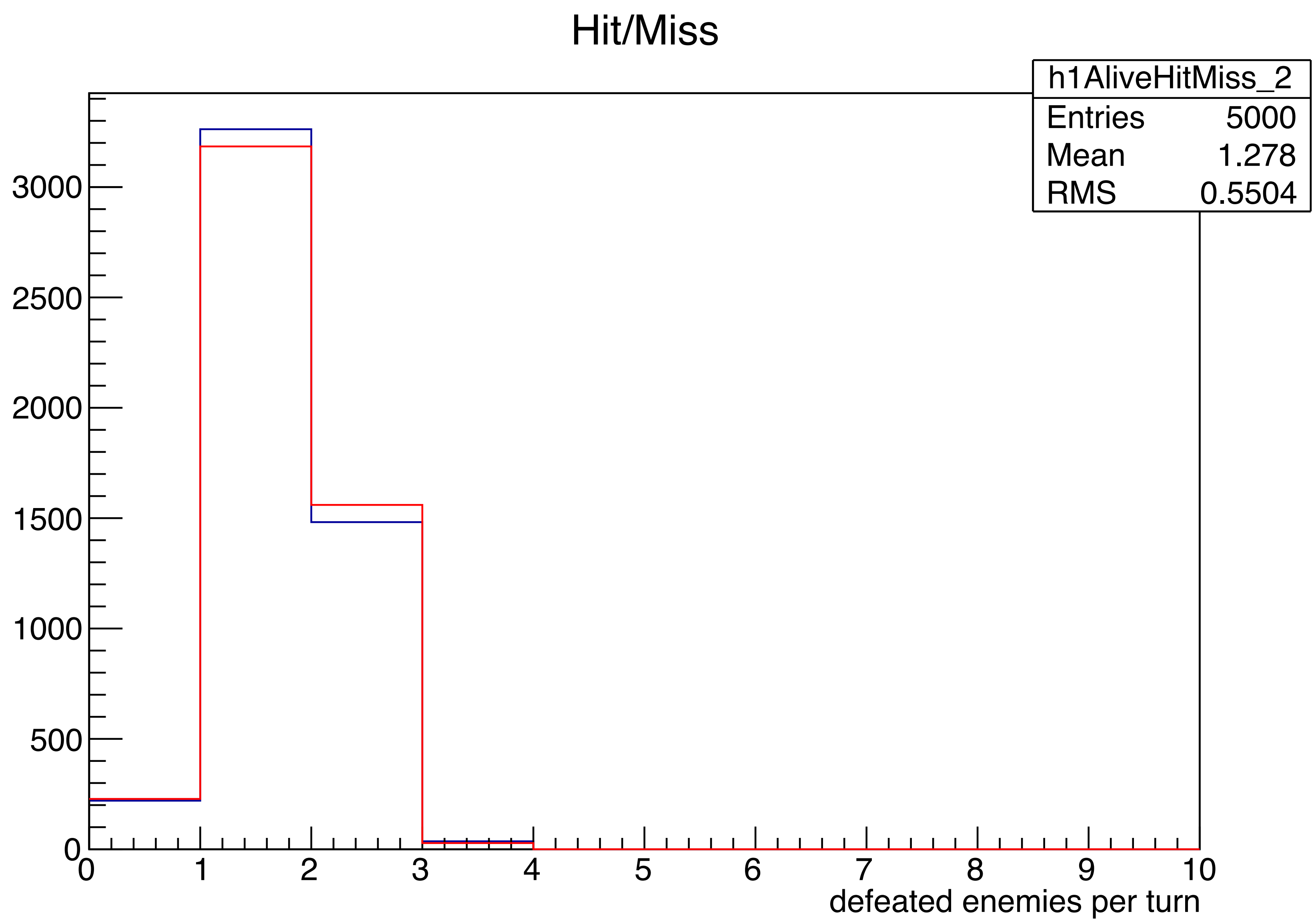

h1Alive
Entries 1000
Mean 1.793
RMS 0.4797
0
100
200
300
400
500
600
700
0
1
2
3
4
5
6
7
8
9
10
defeated enemies per turn

# Problem 4b

- Using the same values from 4a 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?

# Solution 4b

- There are 251/5000 pseudo-trials that defeat 0 enemies. So the percentage is 251/5000=0.0502 and sqrt(251)/5000=0.003168

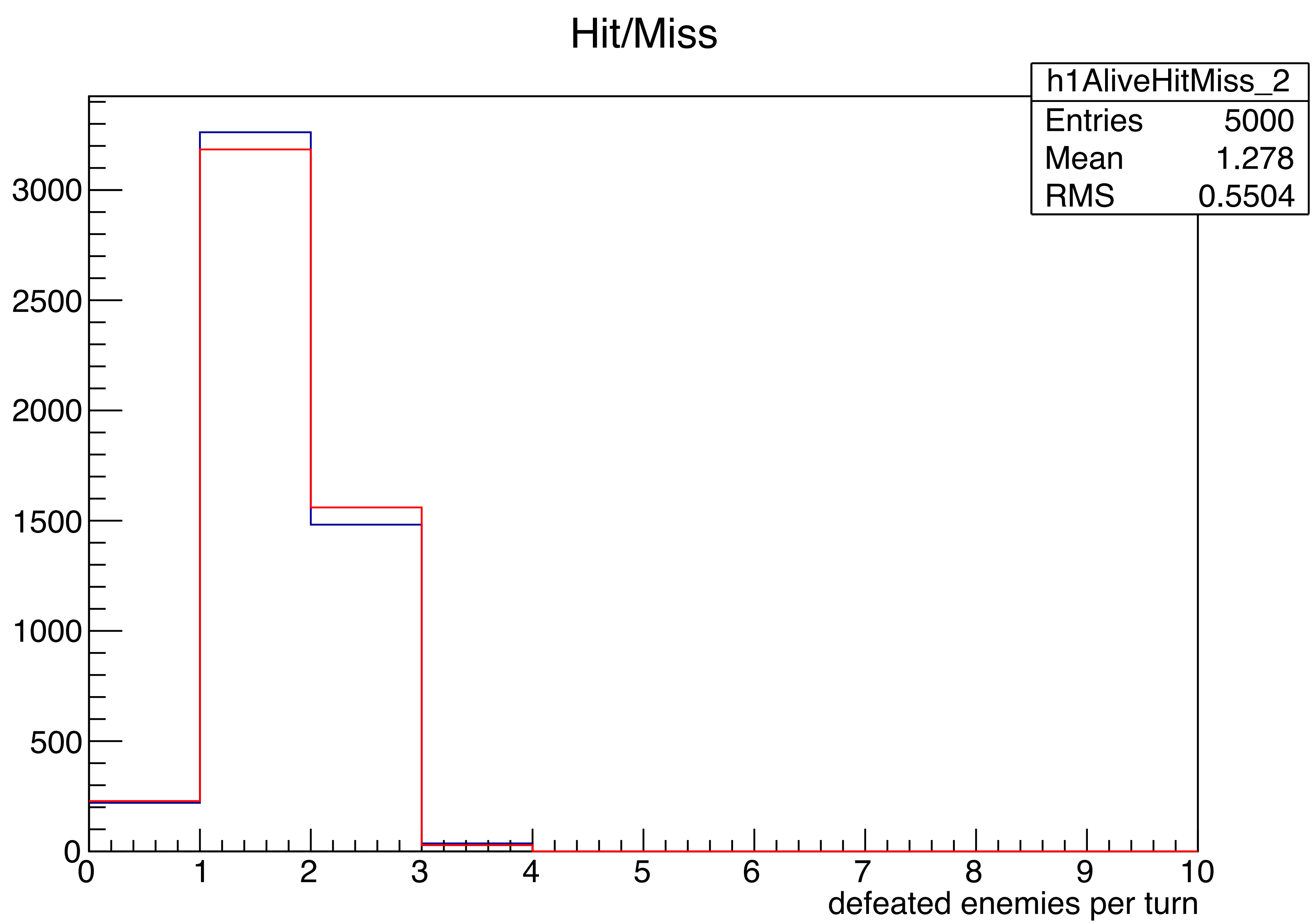


# Problem 4c

- Using the same setup and values from 4b, test the new reorderings below, of your units inflicting damage versus the ordering in 4b 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?
  - Show and/or briefly explain your results

# Solution 4c

- They are statistically compatible because they are well within the statistical uncertainty of the other distribution
  - Here red is ascending and blue is normal



Select only 1 of the following problems for submission. Do all the parts.

# Problem 5

- Apply the likelihood ratio test to the experiment from question 1 of Problem Set 2, with the PDF given below.
  - I have changed the previously poor notation to avoid some confusion, and 'τ' is now replaced by 'b'

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

- Neither b nor $\sigma_t$ 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:

$$b_0 = 1.0\ s$$

$$H_0 : b = b_0$$

$$H_1 : b \neq b_0$$

# Problem 5 (cont.)

- Use the likelihood ratio test:

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

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

$$-2\ln\lambda = -2[\ln(L(\hat{\omega})) - \ln(L(\hat{\Omega}))]$$

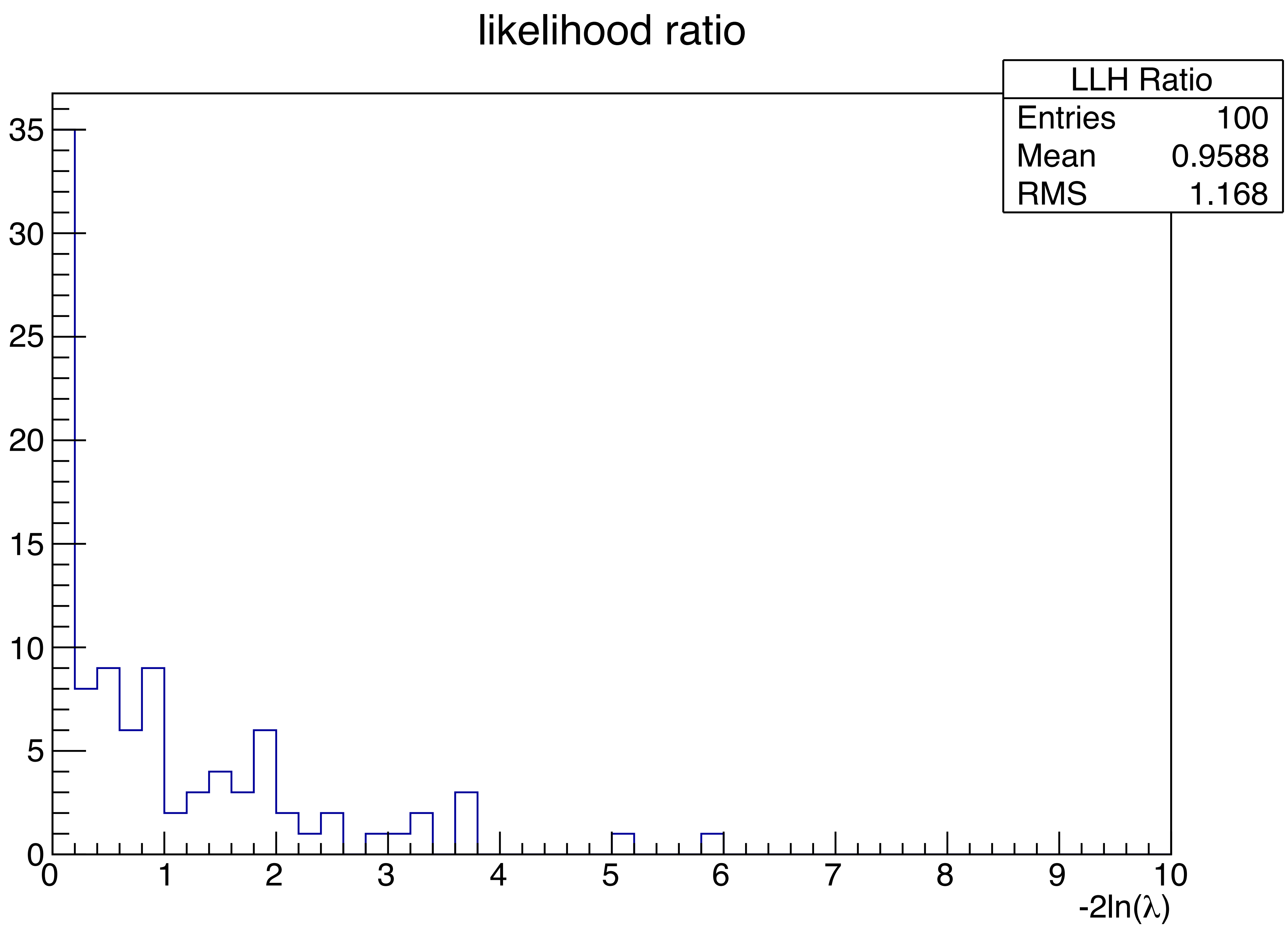
$$\text{if } \lambda \approx 1, \text{ then the null hypothesis cannot be excluded}$$

$$\text{if } \lambda \approx 0, \text{ then the null hypothesis is unlikely true}$$

# Problem 5a

- There are 20000 events in the online file below, which corresponds to 100 simulated pseudo-experiments where each pseudo-experiment has 200 events.
- For each of the 100 pseudo-experiments find the values of the $\ln$-likelihoods that are maximized for the two hypotheses, i.e. $\ln(L(\hat{\omega}))$ and $\ln(L(\hat{\Omega}))$ and calculate -2ln($\lambda$)
- As a histogram, plot the values of -2ln($\lambda$)
- The data is at http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2016/data/Exam_Prob5_NucData.txt

# 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

- Assuming that $-2\ln(\lambda)$ is chi-squared distributed, how many pseudo-experiments of the 100 are expected to have $-2\ln(\lambda) > 2.706$?
- How many pseudo-experiments actually have $-2\ln(\lambda) > 2.706$?
- Bonus question: Why did I choose 2.706?

# 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

# Problem 5c

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

# Solution 5c

- The full fit gives -2ln($\lambda$) = 2.275, which is not even at the 1σ level, and well below the 3σ level. Ergo, the null hypothesis is not rejected at 3σ.

Note: Some students reinterpreted the -2ln($\lambda$) to a p-value and then sometimes (errantly) back to a sigma-like quantity. While the methodology was sometimes wrong, the student received much of the benefit of the doubt if the answer was correct. I didn't ask for justification, but I will in future years.

# Problem 6

- The artist Jackson Pollock was famous for creating paintings that look fractal, or scale invariant. On the next slide is an image of his piece “One: Number 31.”
- A script has been used to convert the entire image to black and white and then write a single horizontal row of pixels into a text file at (http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2016/data/One_row.txt)
  - Values are in 8-bit grayscale
  - 0 is black
  - 255 is white

# Problem 6 - Photo

- Lower resolution image than the image posted on the class webpage (http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2016/data/One_Number31.png)

# Problem 6 (cont.)

- Wavelet coefficients are labelled by two indices: a scale (or level), and a positional index. The RMS of those coefficients that all have the same scale index provides a measure of the activity in a signal at that given scale.
- Enlarging a small section of a scale-invariant signal will result in a new signal that has the same characteristics as the original signal from which the small piece was taken. This means that the size of fluctuations in a scale-invariant signal must be proportional to the scale of the signal.

# Problem 6a

- Take the single row of Jackson Pollock data in the file below and plot the RMS of the wavelet coefficients vs. the scale of the wavelet coefficients.
    - (http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2016/data/One_row.txt)
- What would you expect this plot to look like if the signal were purely scale-invariant?
- Does Pollock's artwork deviate from scale invariance, and if so, how?

# Problem 6b

- White noise is not scale invariant because the size of fluctuations is constant vs. scale, meaning the power spectrum is flat in the frequency domain. However, other types of noise are scale invariant, an example being the sound of the sea.
- Use the Haar and D4 wavelet bases to generate two scale-invariant noise samples, each 256 bins long. What are the differences between these two samples?
- Bonus Question: A frequent cigarette smoker while working, many Pollock paintings have cigarette stubs embedded in the paint. Would it be worthwhile to develop a “Pollock” wavelet bases to account for non-paint debris?