Extra Problems

D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2017

Photo by Howard Jackman

Info

- The following are two extra problems for those interested in more prep work for the course and/or exam

Lists of Distributions

- The data in Problem 1 comes from one of the distributions at right
- Note that these functions may be unnormalized

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

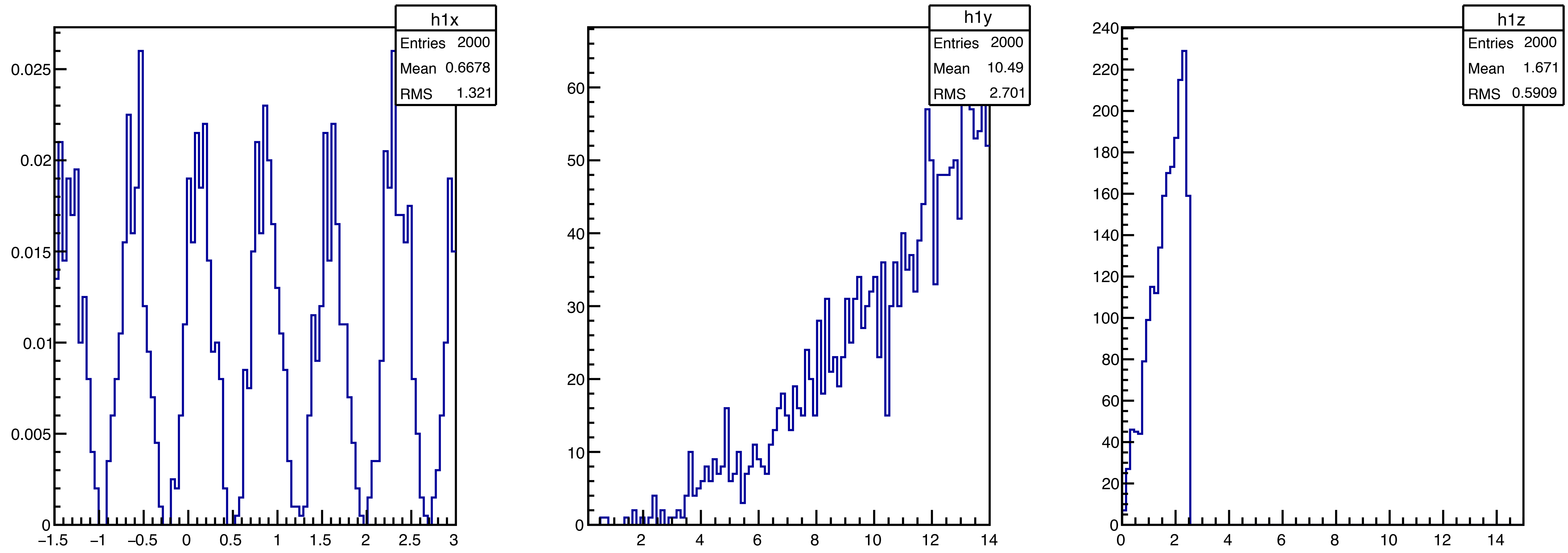
$$f(k) \propto \begin{array}{lr} \binom{n}{k} p^k (1-p)^{n-k} & \textit{binomial} \\ \dfrac{\lambda^k e^{-k}}{k!} & \textit{poisson} \\ \dfrac{-1}{\ln(1-p)}\dfrac{p^k}{k} & \textit{logarithmic} \end{array}$$

Extra Problem 1

- There is a file online which has multiple independent variables (columns) associated with each event (row).
- The independent variables ranges may be truncated.
- For the data in the first column, find the correct distribution type and fit any/all parameters.
- Note: there may be multiple distributions which are statistically compatible with the data. You need to only find one. E.g. cos(x) = sin(a+x) for certain values of 'a'.
- http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2016/data/Extra_Prob1_data.txt

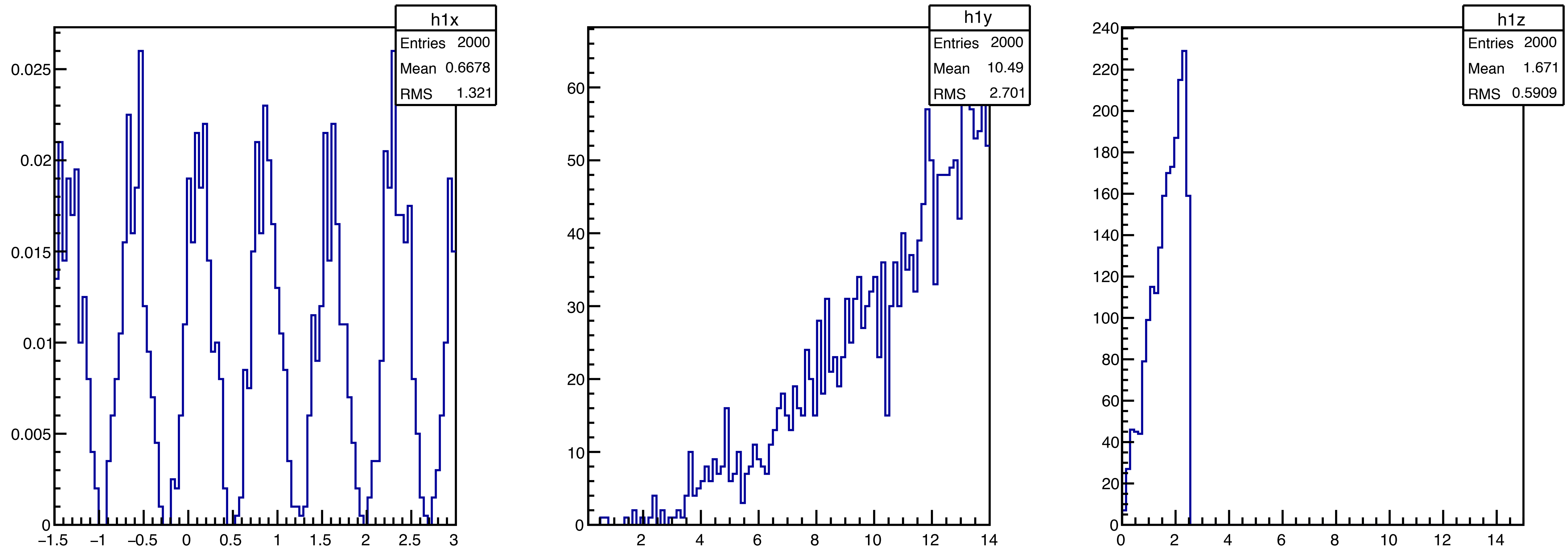
Solution to Extra Problem 1

- Here I plotted the data from the first 3 columns, even though I'm only fitting the left-most distribution
  - The first column is normalized, hence the different y-axis numbers versus the other two distributions
- It's a trig-function, and the data goes from -1.5 to 3.



Solution to Extra Problem 1

- The function doesn't seem to change amplitude, and at zero the distribution isn't producing zero events, which leaves only two of the sin() functions as reasonable possibilities



Problem 1 solution

- Integrating to get a PDF, via wolfram online, I end up with the pythonic

```
PDF or P(x) = numpy.sin( a*x+1)**2/
( (2.25*a+0.25*numpy.sin(2-3*a)-0.25*numpy.sin(6*a+2))/a  )
```

- Even so, I know that sin() functions can produce many local minima/maxima. Important, because I want to find out the value of ‘a’.
- So I’ve got two prominent options: 1) use a Markov Chain Monte Carlo, or 2) do a coarse likelihood scan across values of ‘a’ and start my minimizer near where I think the global minima/maxima is.
- I go with option 2 and find that a value of a=5.5 looks to be close to the global minima

Problem 1 solution

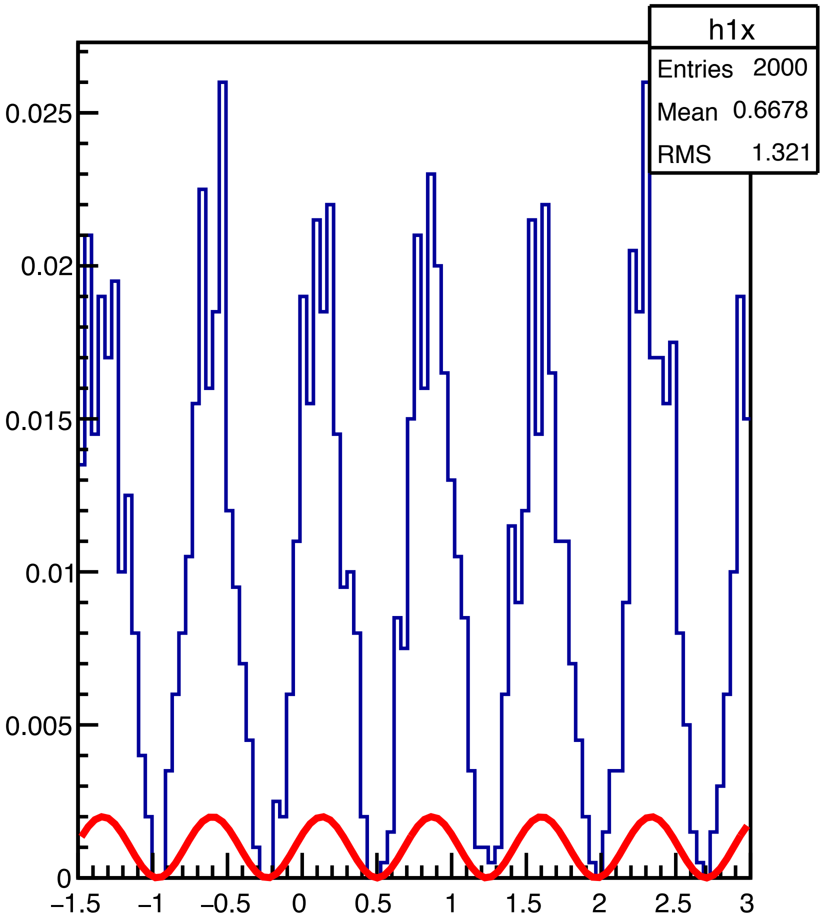
*Note: in my personal code I use ‘f’ instead of ‘a’ which is why the screen shot is ‘f’

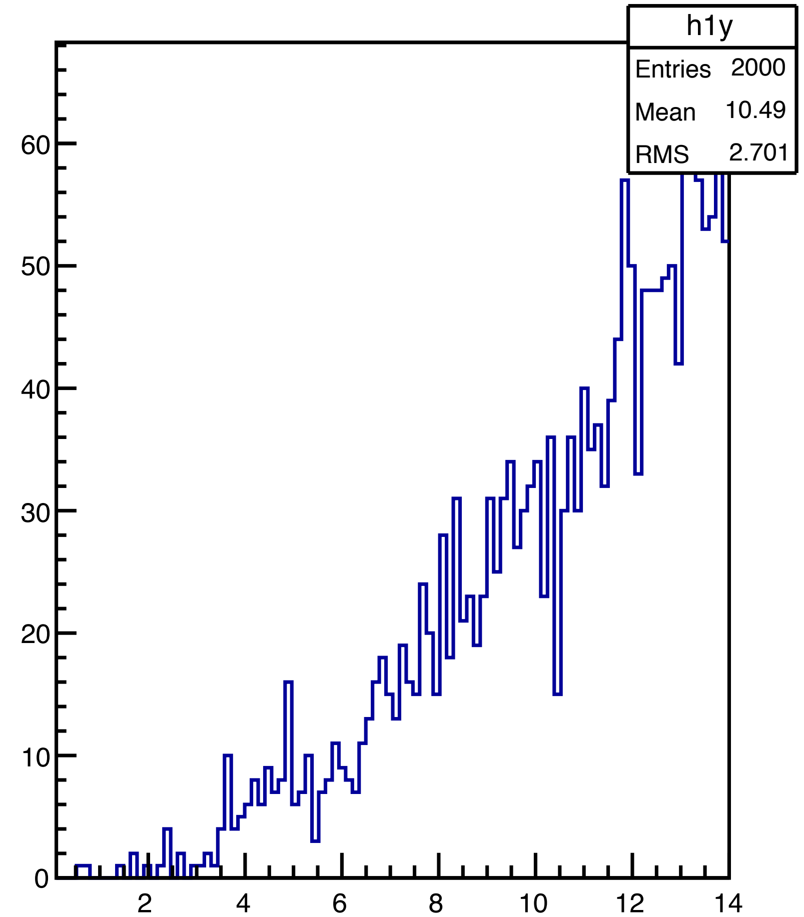
- I turn the distribution function into a ln-likelihood (LLH), i.e. sum the ln(P(x)) for all data, and then have MINUIT minimize the -LLH
- I use MINUIT and migrad as the minimizer and I start the minimizer at a=5.5
  - I know that a≠0, and it doesn’t look to rapidly oscillate, so I put a range of ‘a’ to be [0, 15]. But, if my fit returns 15, then I will expand the boundary, because that would be a sign that range is too narrow (or that I have a bug).
- I let the minimizer go and it returns a=4.279 ±0.0109

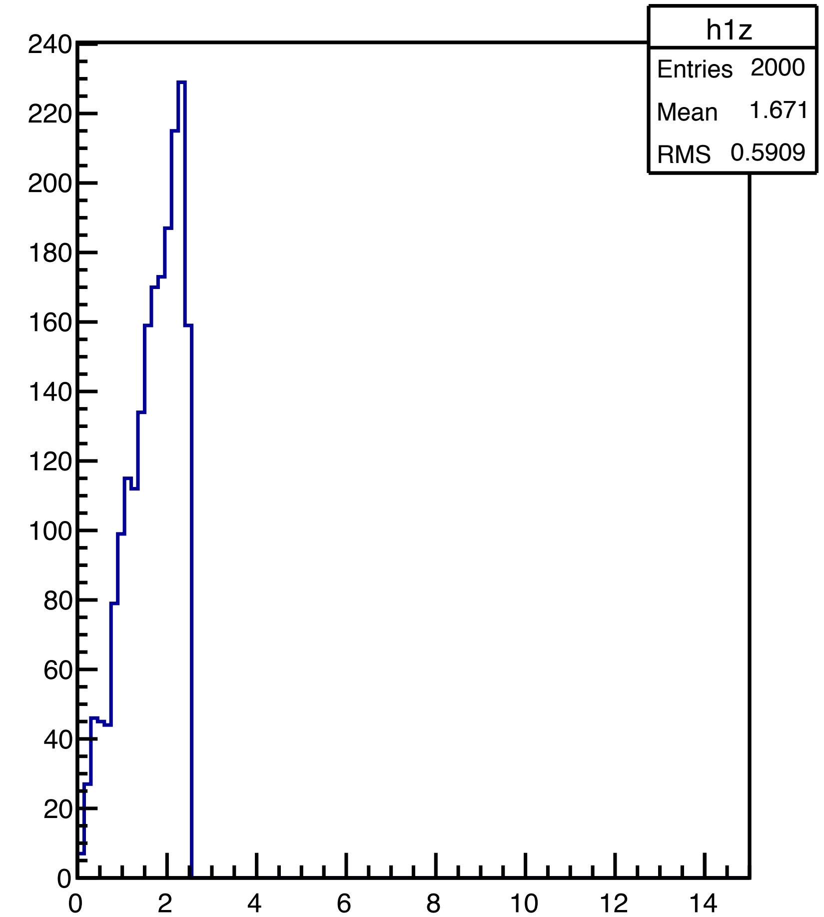
```
--------------------------------------------------------------------------------------------
|      | Name   |   Value    | Para Err |    Err-    |    Err+    |  Limit-   |  Limit+   |
--------------------------------------------------------------------------------------------
|    0 |      f =   4.279    |  0.0109  |            |            |  0        |  15       |
--------------------------------------------------------------------------------------------
```

Problem 1 solution

- Using my best-fit value of a=4.279, I plot $f(x)=\sin(a*x+1)^2$ to see if the distribution passes the ‘does it look okay?’ test.
- Yup, looks good, but I still need a quantifiable metric.







Problem 1 wrap-up

- Quantifiable metric:
  - KS-test
  - Chi-squared test using the analytic function with a=4.279
  - Create many Monte Carlo pseudo-experiments using $\sin(4.279x+1)^2$ and doing a data-data KS-test
  - etc.
- Any test-statistics or metric should return that $f(x)=\sin(4.279x+1)^2$ is statistically compatible w/ the data, especially because the true value is a=4.28. If the test metric had not been good, e.g. a p-value of 0.004, I would have started the minimizer in many different places, switched to an MCMC, or move on to $f(x)=\sin(ax)+1$

Extra Problem 2

- There is a machine learning algorithm database hosted by UC-Irvine and included is a German credit data set for testing. The file format is a little cleaner in the link below
  - Info: https://onlinecourses.science.psu.edu/stat857/node/215
  - File: https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/german_credit.csv
- The first column signifies the persons credit risk; 1=good and 0=bad
- Use a classifier algorithm, e.g. decision tree, to identify credit-worthy customers.
  - Train using the only ~30% of the data set, and use the other ~70% to test the training

Extra Problem 2

- Solutions are everywhere online, due to this being a classic problem
- Using a BDT, I was able to get around 67% accuracy on a training sample for correctly identifying credit worthiness.