Lecture 7 : Hypothesis Tests

D. Jason Koskinen

koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics

Feb - Apr 2020

Photo by Howard Jackman

Statistical Tests - General Idea

- General idea - Particle Physics context
  - Given the measurement of an individual event, one has a collection of numbers: $\vec{x} = (x_1, ..., x_n)$

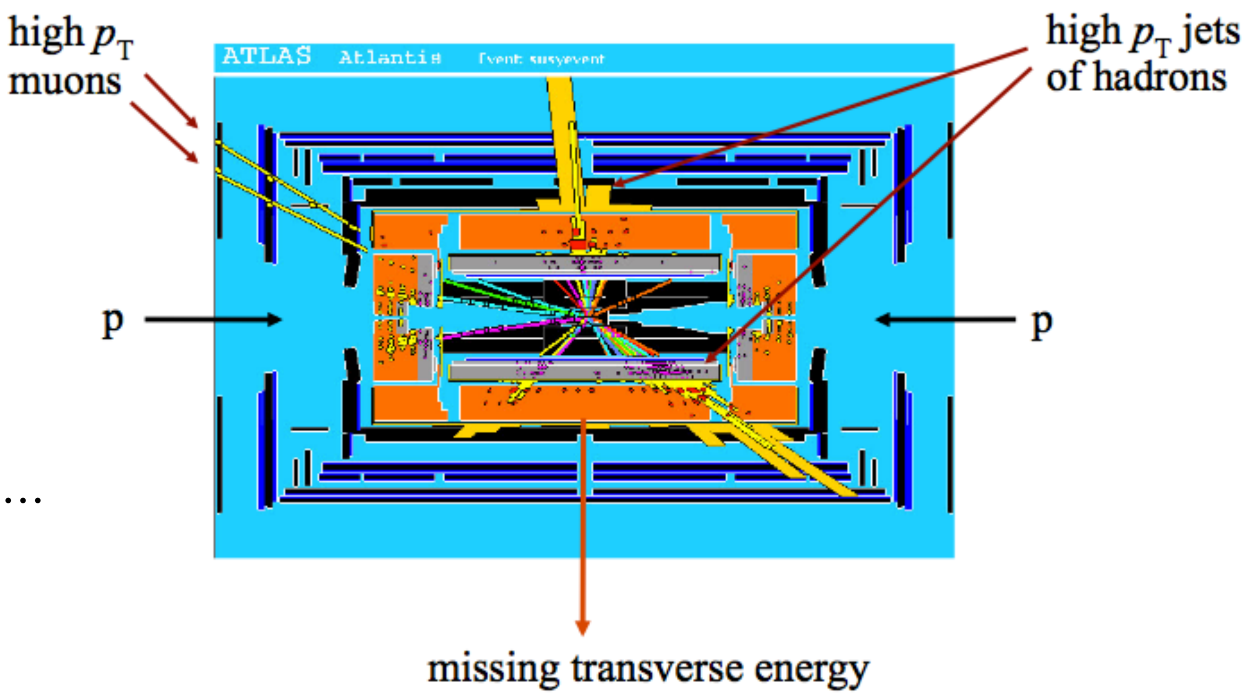
$$x_1 = \text{number of muons} \qquad x_2 = \text{number of jets} \;...$$

  - The set of measurements follow some n-dimensional PDF that depends on the type of event produced. For each reaction we can consider a hypothesis for the PDF. Example:

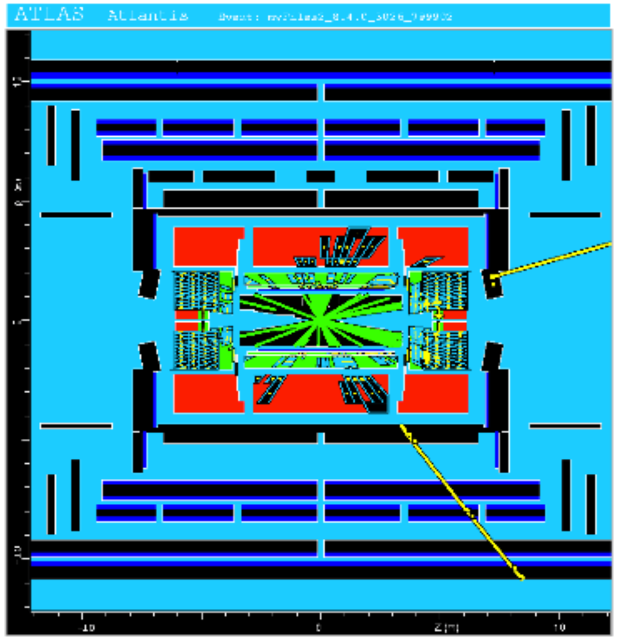
$$f(\vec{x}|H_0), f(\vec{x}|H_1), ...$$

  - We call H_0 the null (background) hypothesis (the event type we want to reject) and H_1 the alternate (signal) hypothesis

A simulated SUSY event


high p_T muons
ATLAS Atlantis Event: susyevent
high p_T jets of hadrons
p
p
missing transverse energy

Background events



This event from Standard Model ttbar production also has high p_T jets and muons, and some missing transverse energy.

→ can easily mimic a SUSY event.

Statistical Tests - General Idea

- Hence, rather than estimating an unknown parameter, the results of an experiment may be used to determine if a given **theoretical model** is acceptable given the observations. For example, suppose a model estimates the lifetime of a nucleus. Is a set of data compatible with the model(?):

$$H_0 : \tau = \tau_0$$

$$H_1 : \tau \neq \tau_0$$

- The above is an example of a parametric test. Typically a hypothesis cannot be proven true or false but you can determine the probability for obtaining the observed result if you assume the hypothesis is true.

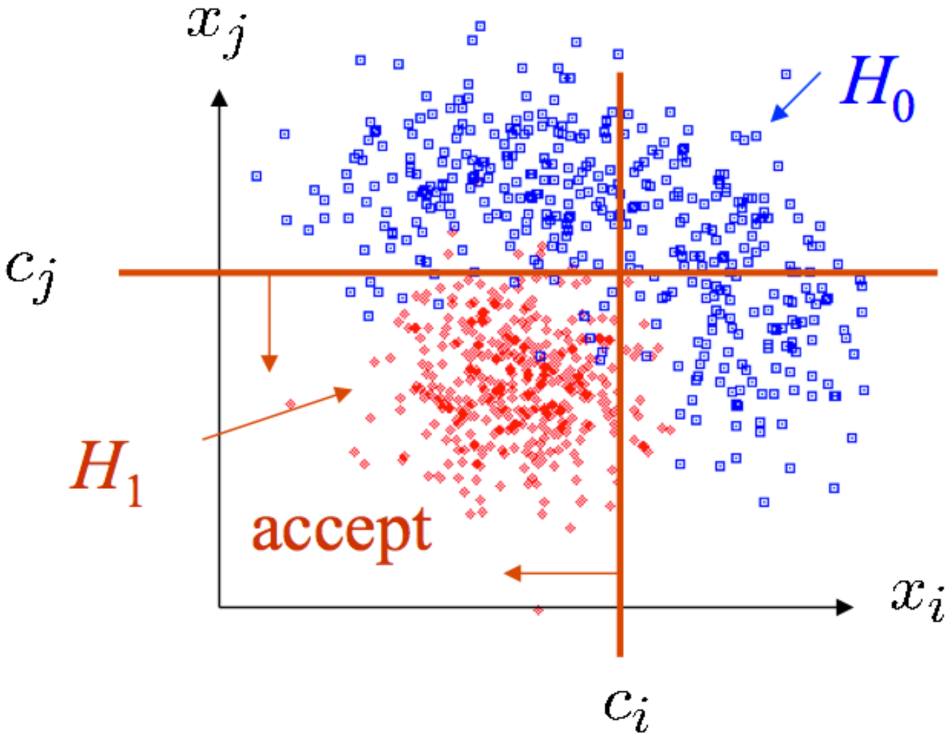
- Hypothesis testing is also a part of data analysis when, for example, you decide if a specific observed event is signal or background. Suppose you have a data sample with two kinds of events that correspond to the null and alternate hypotheses and you want to select those that are of the type corresponding to the alternate hypothesis. Then each event is a point in the space and we define a decision boundary of where to accept/reject events belonging to each of the event types.

Statistical Tests

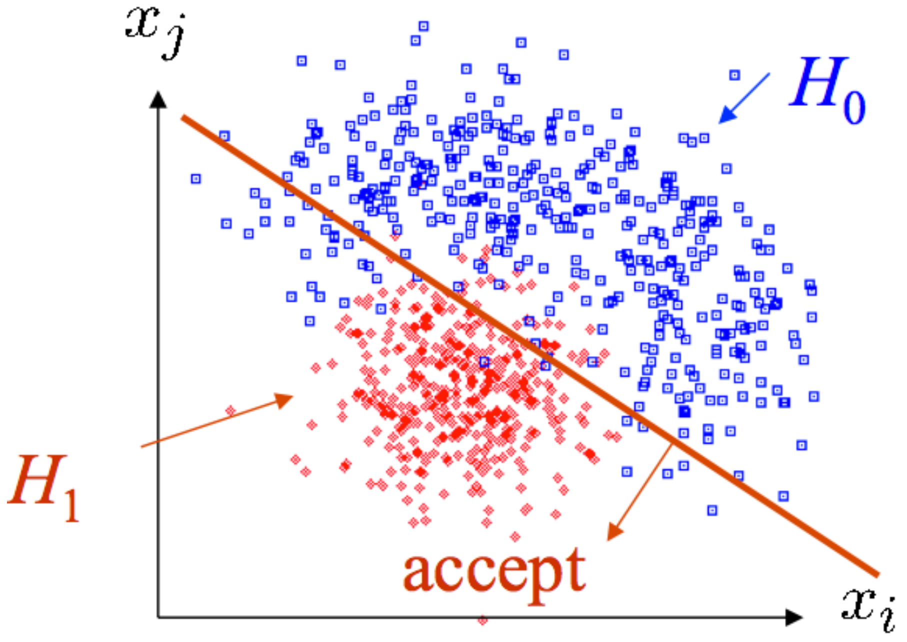
- Event Selection
  - selection cuts for events, e.g.

    $x_j < c_j \qquad x_i < c_i$

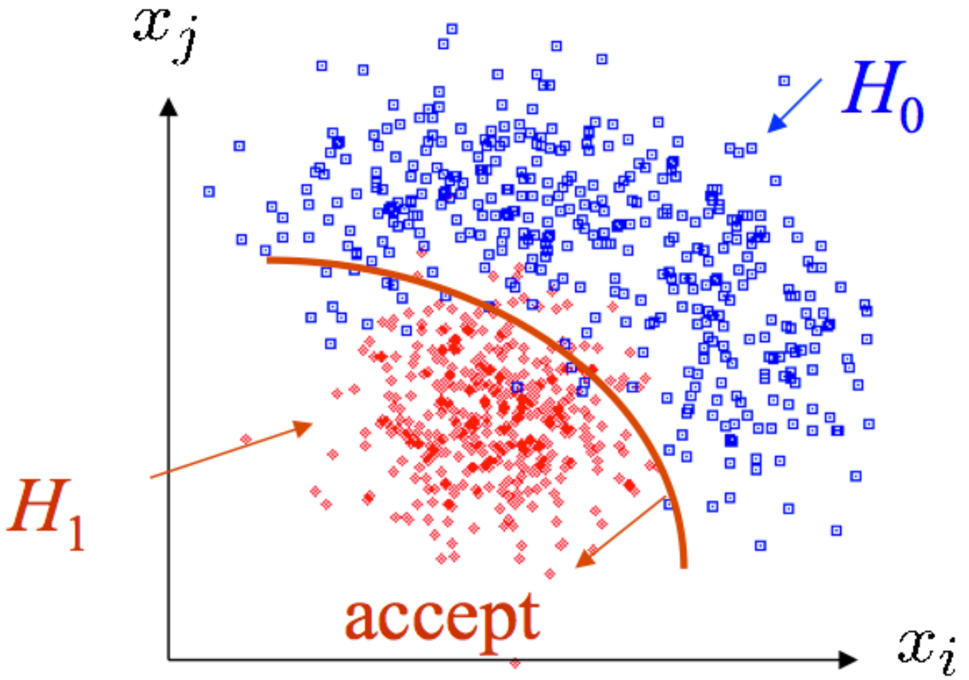
  - We would like to optimize this process…



linear



or nonlinear



*G. Cowan

Decision Boundary and Test Statistic

- A decision boundary can be defined using an equation or function that can be used to discriminate signal (H_1) from background (H_0):

$$T(\vec{x}) = t$$

"t" can be a multidimensional vector

- What we want is a single valued test statistic (t) which reduces lots of data or information to a single quantity
  - A likelihood value is an example of taking lots of discrete data points and reducing the ensemble to a single quantity
  - For discrete data points we can define a function which reduces the number of dimensions without losing the ability to separate 'signal' from 'background'. E.g. in the previous slide we could use radius from the origin where $r_{i,j} = \sqrt{x_i^2 + x_j^2}$ becomes the test statistic, i.e. t=r.

Statistical Tests - Decision Boundary

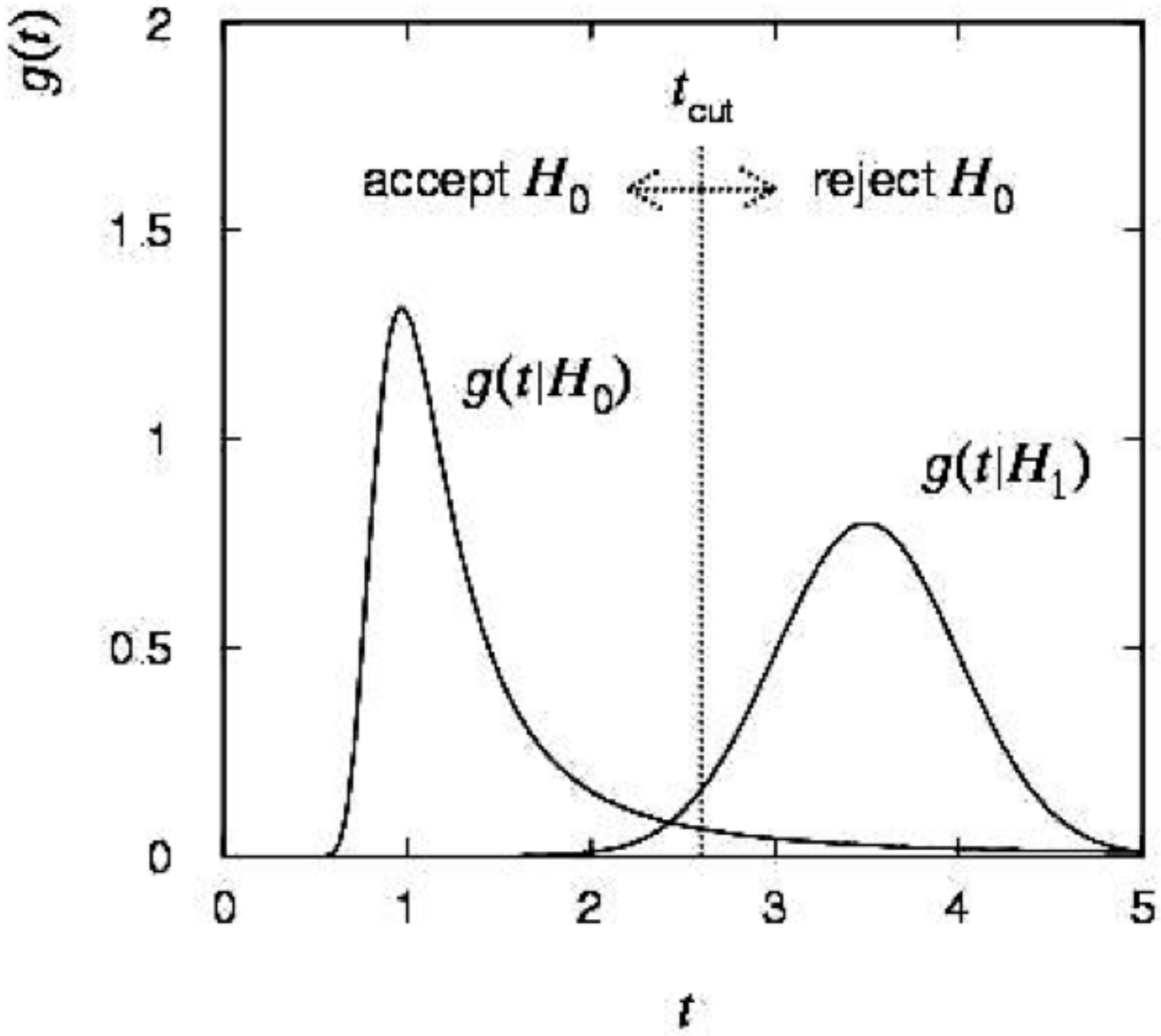
- The decision boundary can be defined using the test statistic to discriminate between hypotheses, e.g. signal or background

- Each hypothesis will imply a given PDF for the test statistic, t:

$$g(t;H_0) : \text{PDF for t under } H_0 \text{ true}$$
$$g(t;H_1) : \text{PDF for t under } H_1 \text{ true}$$

- Define:

$$t > t_{cut} \text{ Critical Region}$$
$$t < t_{cut} \text{ Acceptance Region}$$
$$t_{cut} \text{ Decision Boundary}$$



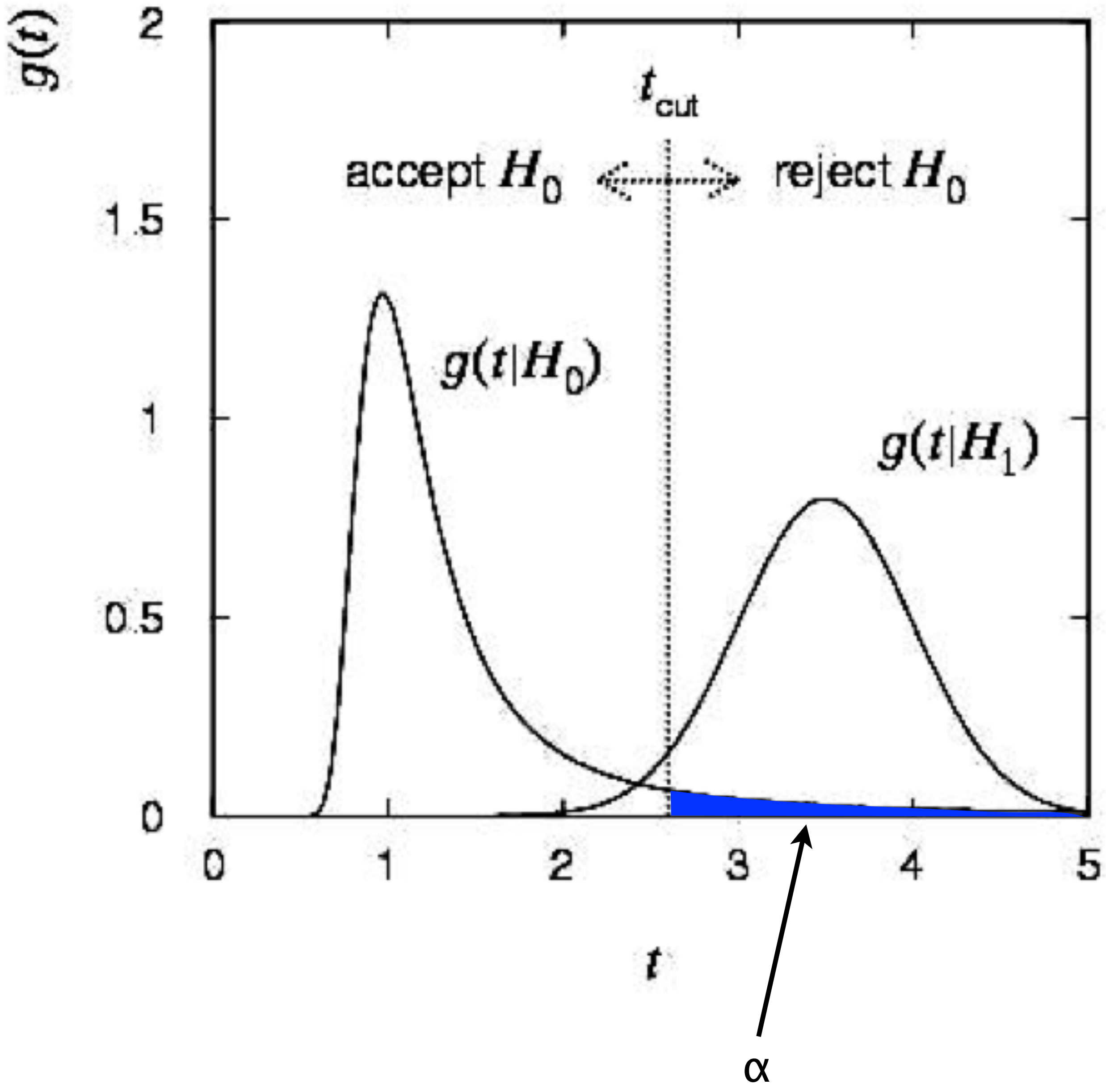
Statistical Tests - Decision Boundary

- The decision boundary defines a test. If the data falls into the critical region ($t>t_{cut}$) then we reject the null hypothesis.
    - But there is some probability that we **wrongly** reject H_0

- Define the error of the first kind (α) as a probability to reject the null hypothesis if the null hypothesis is true:

$$\alpha = \int_{t_{cut}}^{\infty} g(t;H_0)dt$$

- The statistical significance of rejection is given by the p-value



P-Value

- A p-value is the probability under the assumption of a specific model or hypothesis, generally H_0, of observing a test-statistic as compatible to, or less compatible with, the observed data
  - For example, consider we measure some value μ_{obs} and we want to see if it is statistically compatible with some other value of μ (H_0)
  - The test statistic (q_μ) reflects the level of agreement between the data and the hypothesized value of μ
  - The test statistic is generally constructed such that higher values represent increasing incompatibility of the model (H_0) with the data

$$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)dq_\mu$$

q_μ is the test statistic for a hypothesized value of μ, and "$q_{\mu,obs}$" is the TS value from the observed data

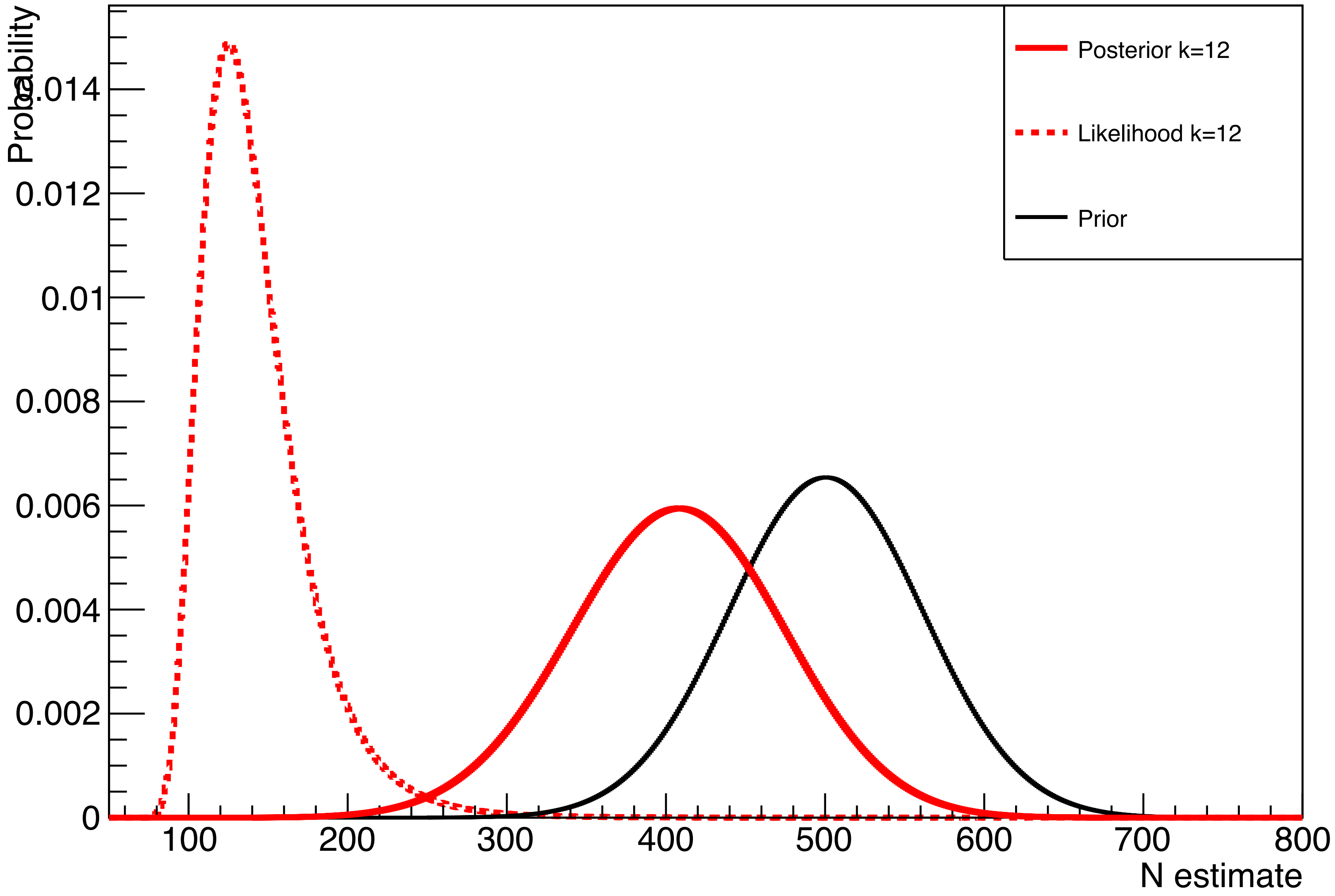
Even More Extreme

From Bayes Lecture

- For the instance where k=12, gaussian mean=500 and σ=61 we've got some some issues

- The bayesian posterior best estimate is ~409, but the best likelihood estimate is ~125.

- According to the likelihood PDF, how likely is it to have a value ≥ 409?
  - (hint integrate the tail of the likelihood distribution ≥ 409)

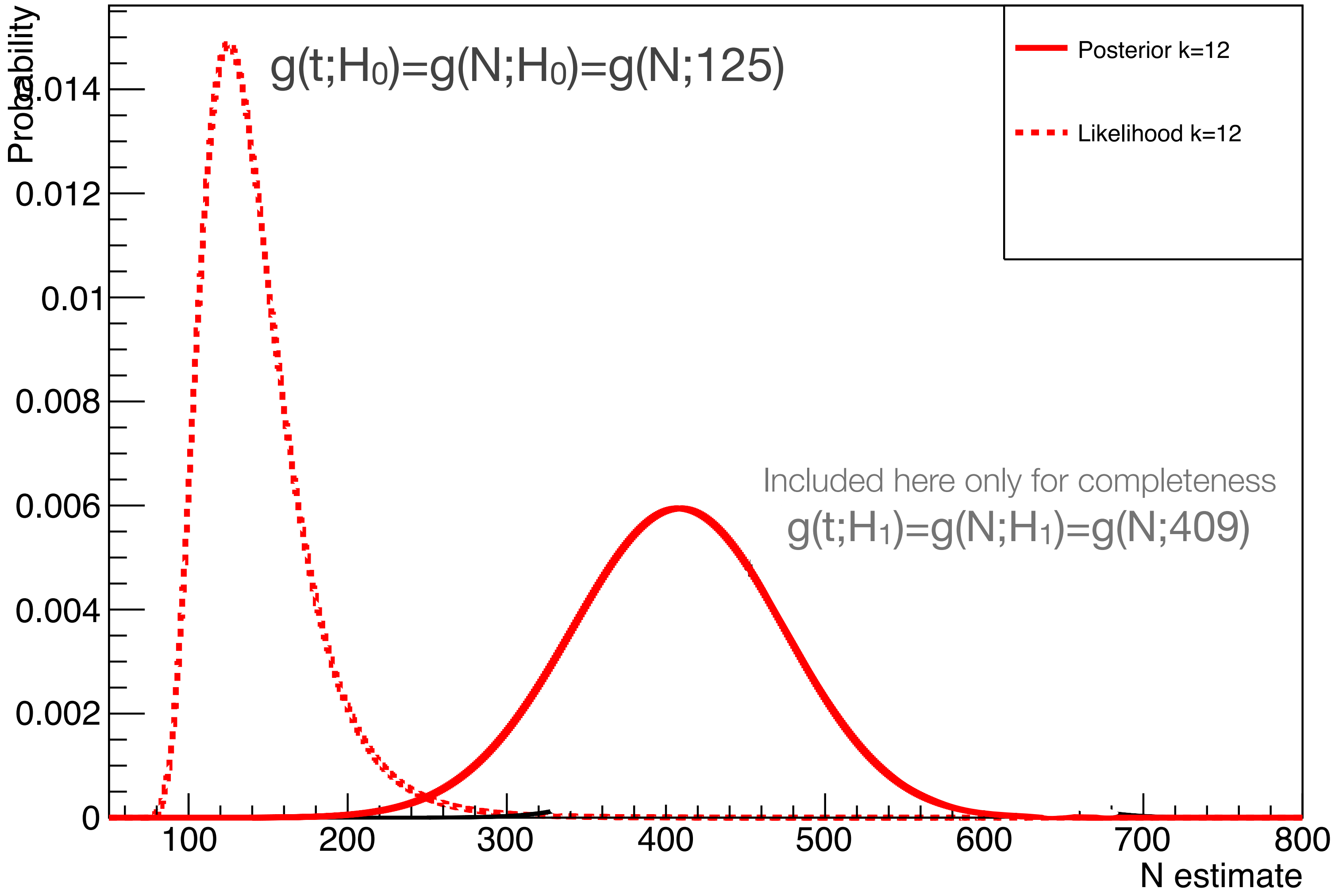


P-Value in Action

- For this example we consider N to be the test statistic (t=N), the *maximum a posteriori* value of 409 to be our alternate hypothesis (H_1), and value of 125 to be our null hypothesis (H_0).

- If we assume H_0 to be true, then $g(t;H_0)$ gives us the test statistic probability distribution function, and our p-value is:

$$\text{p-value} = \int_{409}^{\infty} g(N;125)dN =\sim 0.00017$$



Exercise #3 From a Previous Lecture

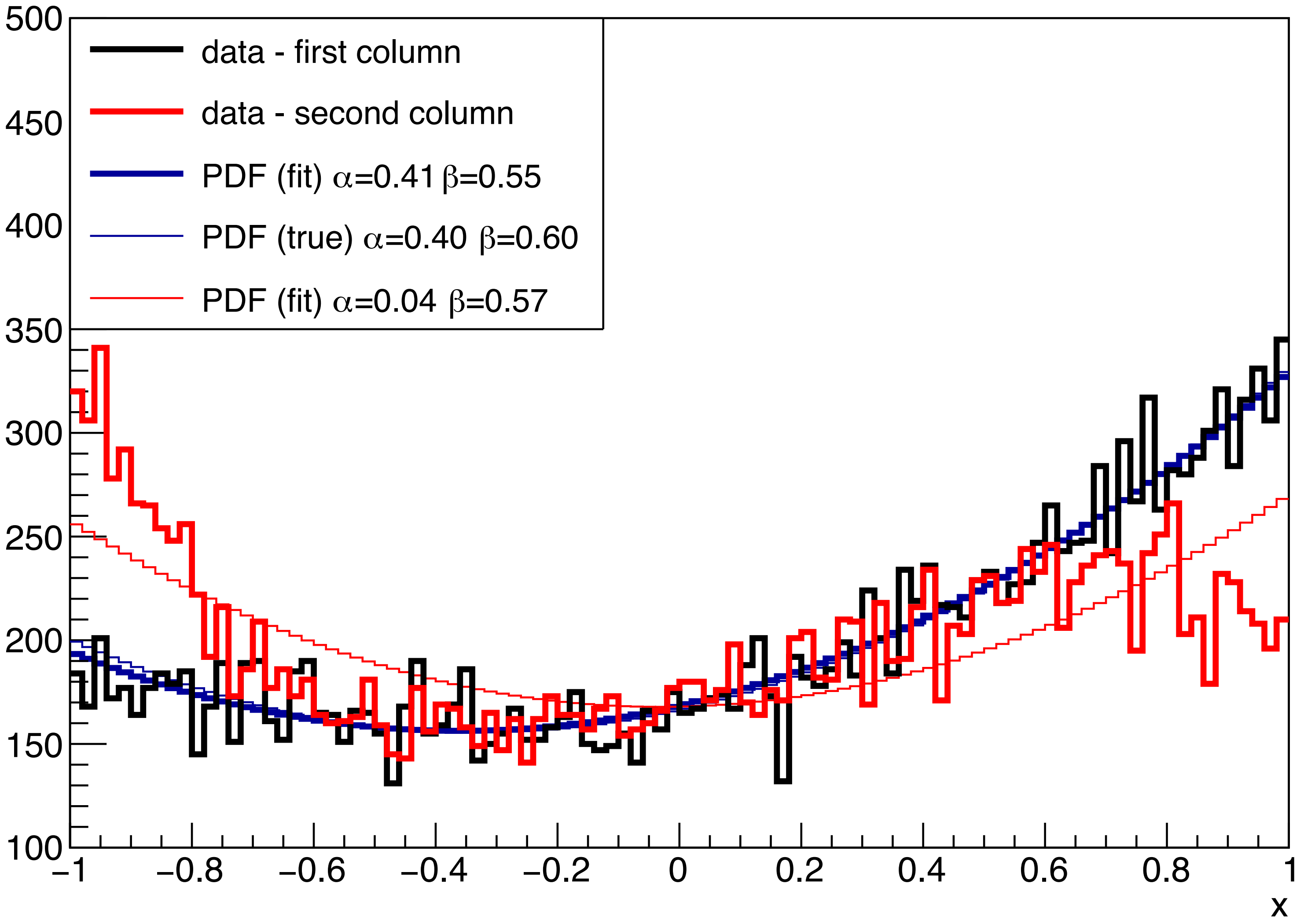
- There is a file posted on the class webpage from ”Parameter Estimation and Confidence Intervals” lecture which has two columns of x numbers (not x and y, only x for 2 pseudo-experiments) corresponding to x over the range -1 ≤ x ≤ 1
- Using the function:

$$f(x;\alpha,\beta) = 1 + \alpha x + \beta x^2$$

  - Find the best-fit for the unknown α and β
  - Calculate the reduced chi-square goodness of fit (p-value) by histogramming the data. The choice of bin width can be important.
    - Too narrow and there are not enough events in each bin for the statistical comparison.
    - Too wide and any difference between the ‘shape’ of the data and prediction histogram will be washed out, leaving the result uninformative and possibly misleading.

Previous Lecture Exercise

- For my own interest I generated an additional file, which is posted as "extra data file" for "Lecture on Parameter Estimation and Confidence Intervals"
- Histograms: the x-values of the two pseudo-experiments, the expectation from PDF using the best-fit values and the true values (which I knew because I generated the data)



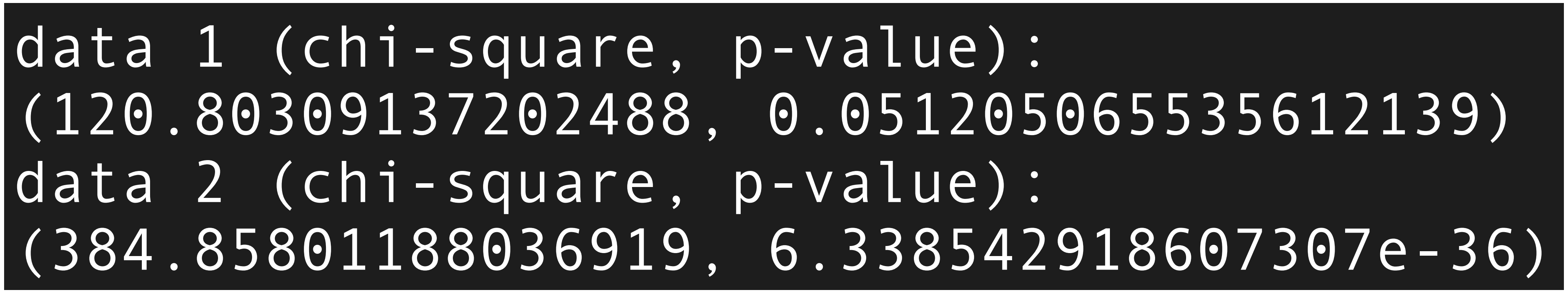
Follow-up on Exercise

- In exercise 3 from a previous class I asked to calculate the goodness-of-fit. The p-value from a chi-squared distribution is an appropriate choice.
  - Visually, the previous plot of the x data from the first and second column look to agree with the PDF using their best-fit values of α and β returned by the LLH minimization
  - The actual PDF for the data in the second column was:

$$f_2(x) \propto 1 + \alpha x + \beta x^2 - \gamma x^5$$
$$(\alpha = 0.4, \beta = 0.6, \gamma = 0.9)$$

  - But the fit was done for both data sets with the function

$$f(x; \alpha, \beta) = 1 + \alpha x + \beta x^2$$



```
data 1 (chi-square, p-value):
(120.80309137202488, 0.051205065535612139)
data 2 (chi-square, p-value):
(384.85801188036919, 6.338542918607307e-36)
```

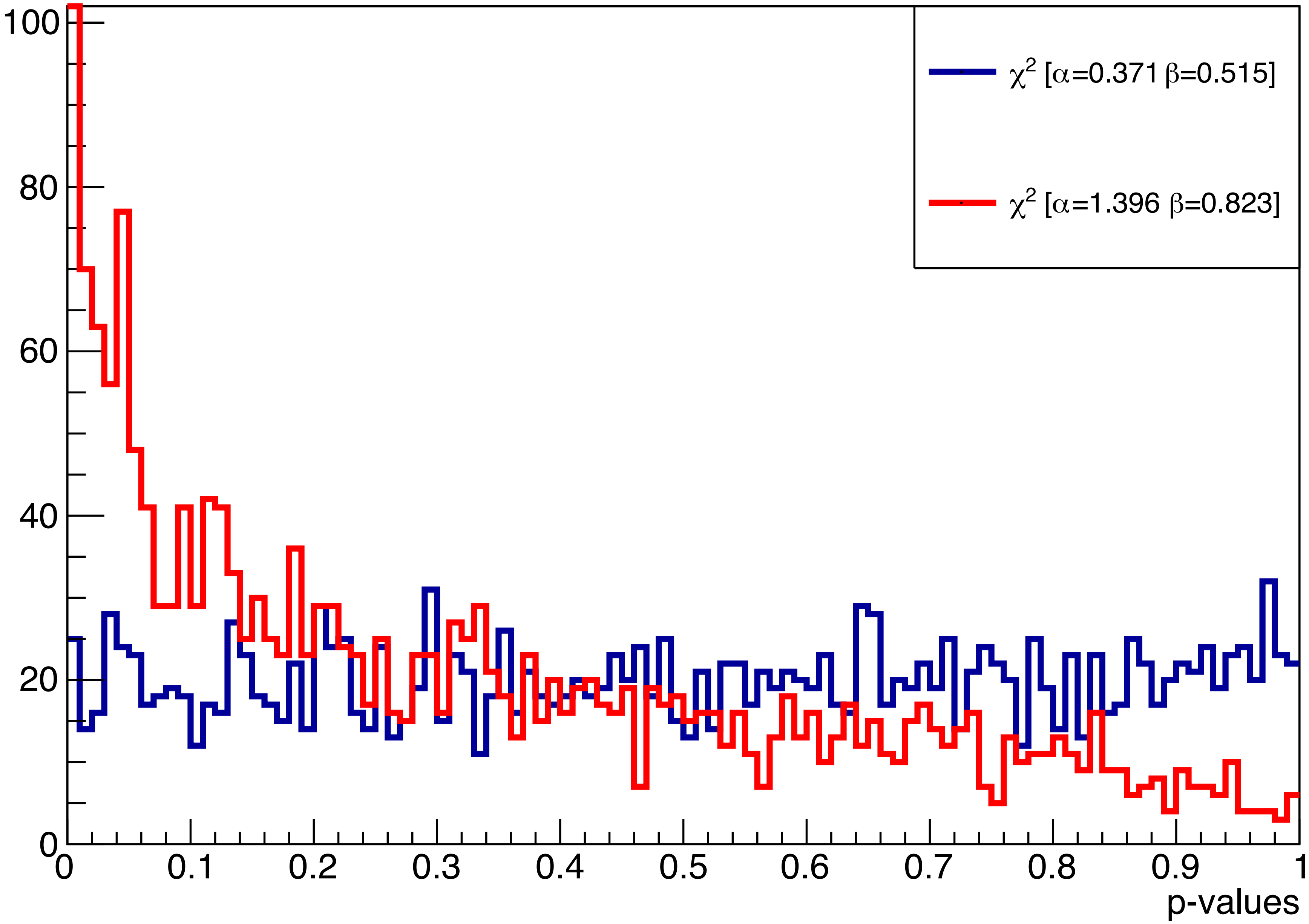
*from a binned histogram

Funny Thing

- In 2016 a student asked “For repetitions, what should a distribution of p-values look like?”, and I didn’t know
  - There are proofs that when the hypothesis is correct, the distribution of p-values is uniform from 0-1, i.e. flat
  - I wanted to check ‘uniformity’ using the same PDF, i.e. $(1+\alpha x+\beta x^2)/(2+\beta/3)$, as before but using different values of α and β
- Because we have Monte Carlo capability, we can randomly sample from the ‘correct’ PDF, and use the χ^2 as the test-statistic for the p-value calculations
  - By using Monte Carlo we are assured that the hypothesis we are comparing to the pseudo-experiments is correct

Results - Odd

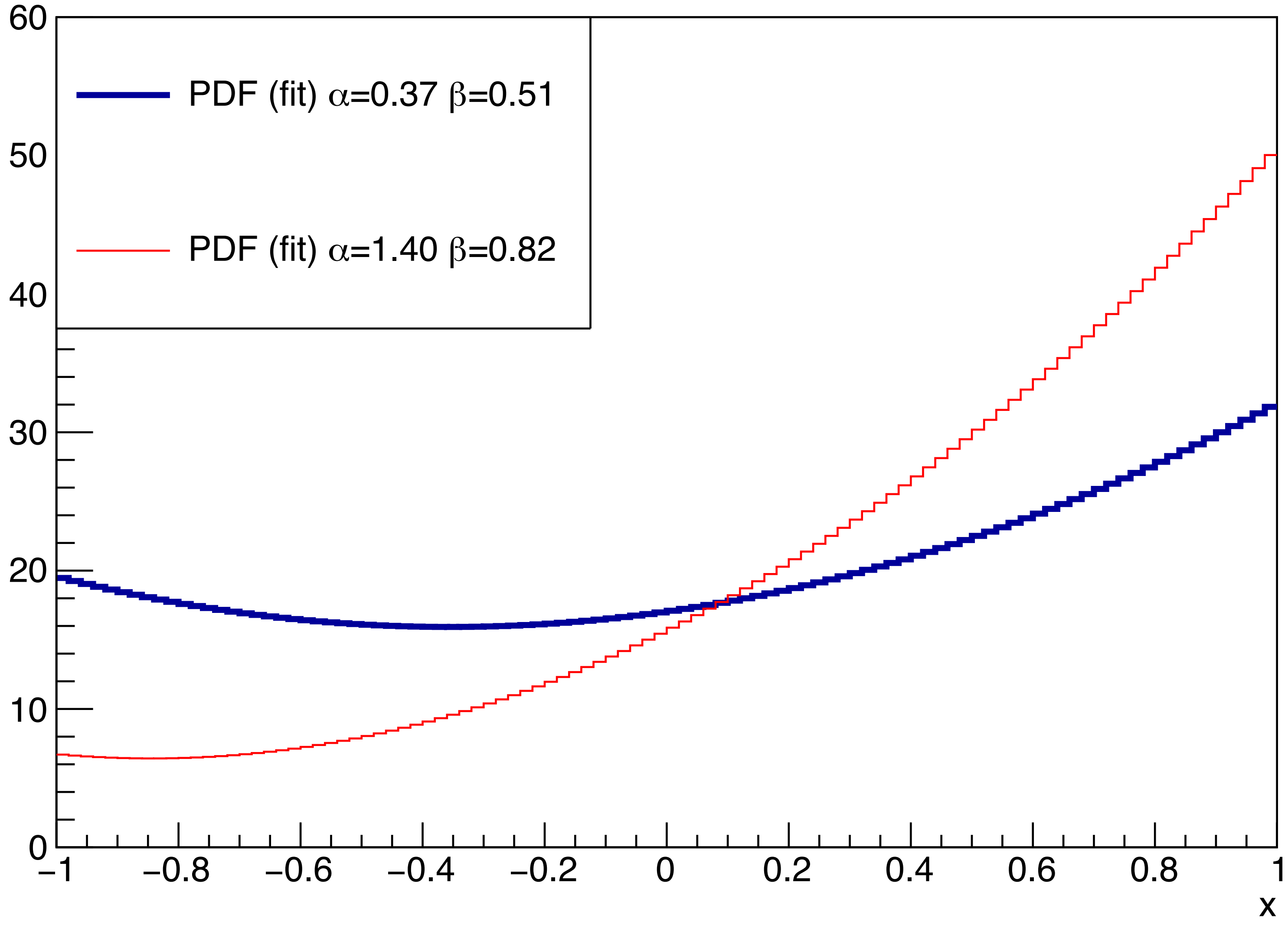
- For 800 pseudo-experiments (w/o any fitting), each having 2000 points, one set of α and β values produce uniform p-values while the other set does not, both using the same original PDF of $(1+\alpha x+\beta x^2)/(2+\beta/3)$



*Different file than what is posted online

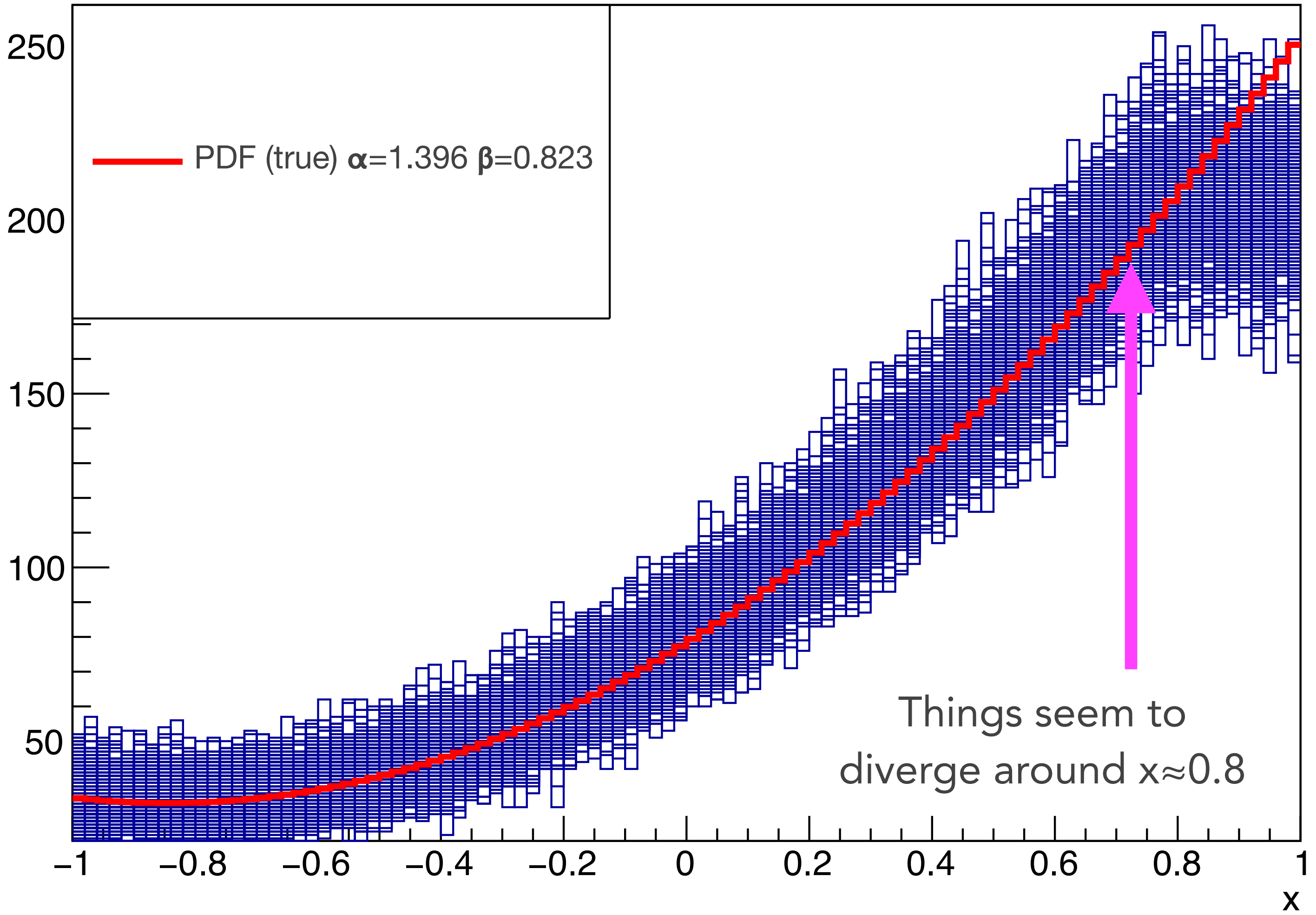
Debugging

- My first thoughts were to look at the underlying PDFs
  - The χ^2 test-statistic can be inaccurate in regions of low event rates
  - I increased the number of samples in each pseudo-experiment by a factor of 4 to 5… but there was no change



Clue

- I stopped trying to be clever and just brute force plotted things
    - I histogrammed the x values for 800 pseudo-experiments, each w/ 10k points and also plotted the underlying PDF
    - For **α**=1.396 and **β**=0.823 they didn't match at x values of 0.8-1.0



Things seem to diverge around x≈0.8

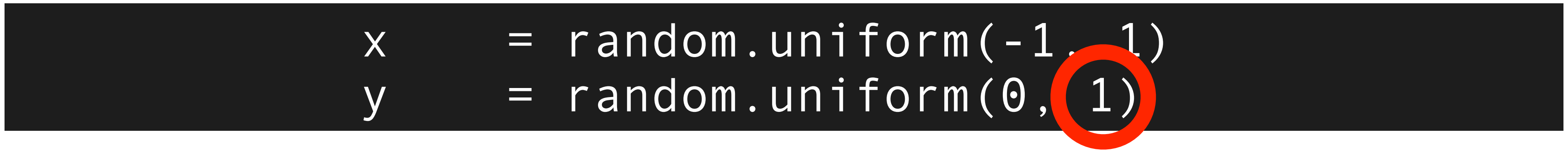
Only 1 of 800 pseudo-experiments had an upward fluctuation in the number of events for the bin $0.98 \leq x < 1.0$. But, I expect ~1/2 of the pseudo-experiments to have an upward fluctuation in any single bin

Solution

- So I went back to my PDF calculation and using **α**=1.396 and **β**=0.823 for:

$$f(x;\alpha,\beta) = \frac{1+\alpha x+\beta x^2}{2+2\beta/3}$$

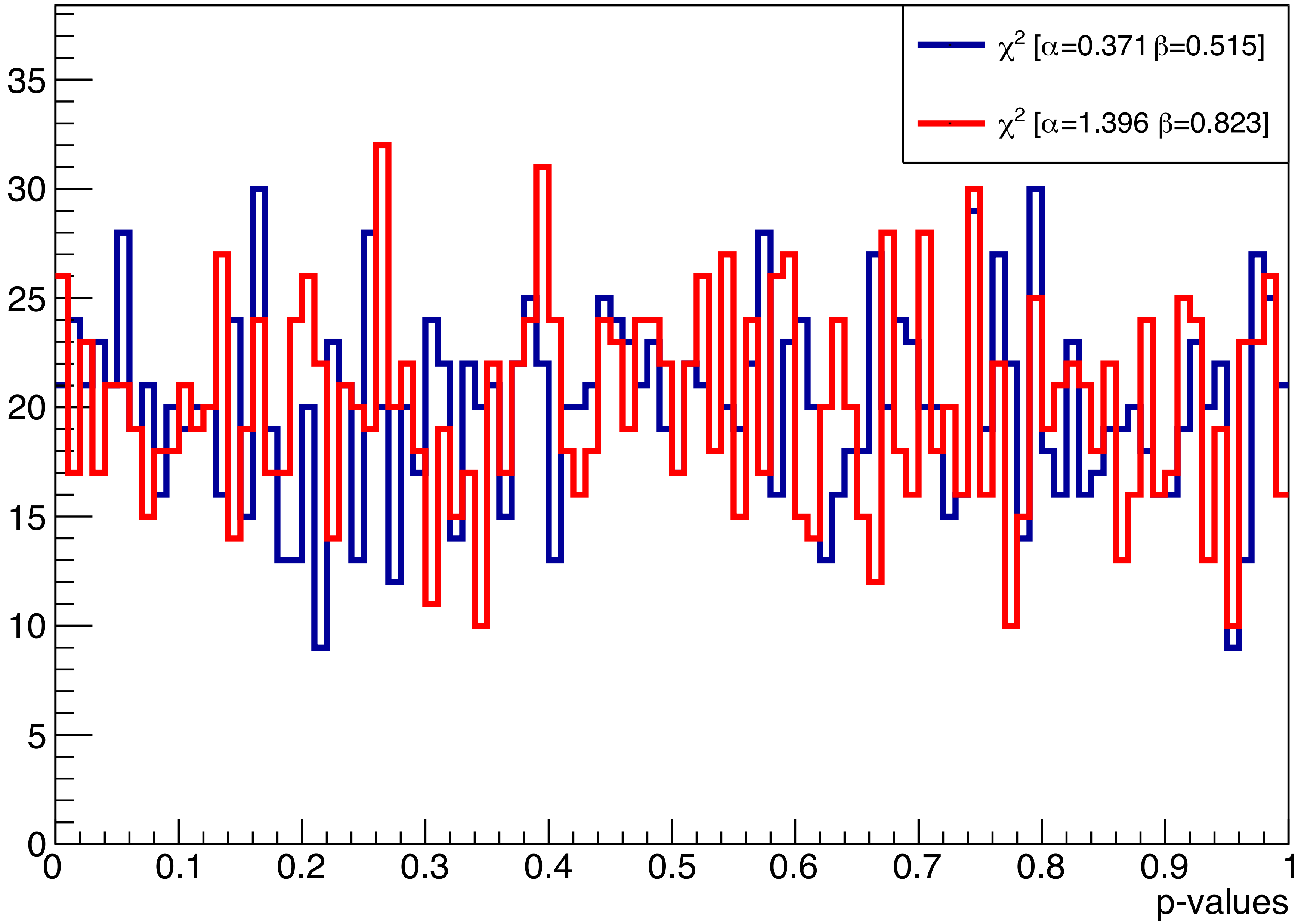
- What's so special about x≈0.8?
  - Well, f(x=0.8; **α**=1.396, **β**=0.823)=1.039
  - The distribution is normalized to 1, but the instantaneous probability density goes above 1 in the range of ~0.8-1
  - My accept/reject method of Monte Carlo sampling the PDF went from -1 to 1 in x, but only 0 to **1** in y



```
x     = random.uniform(-1, 1)
y     = random.uniform(0, 1)
```

Fixed

- Changing the bounds on my accept/reject sampling fixed the problem
- This was a silent failure mode, which can be incredibly difficult to debug. Be thankful when your code crashes, because then it's obvious.

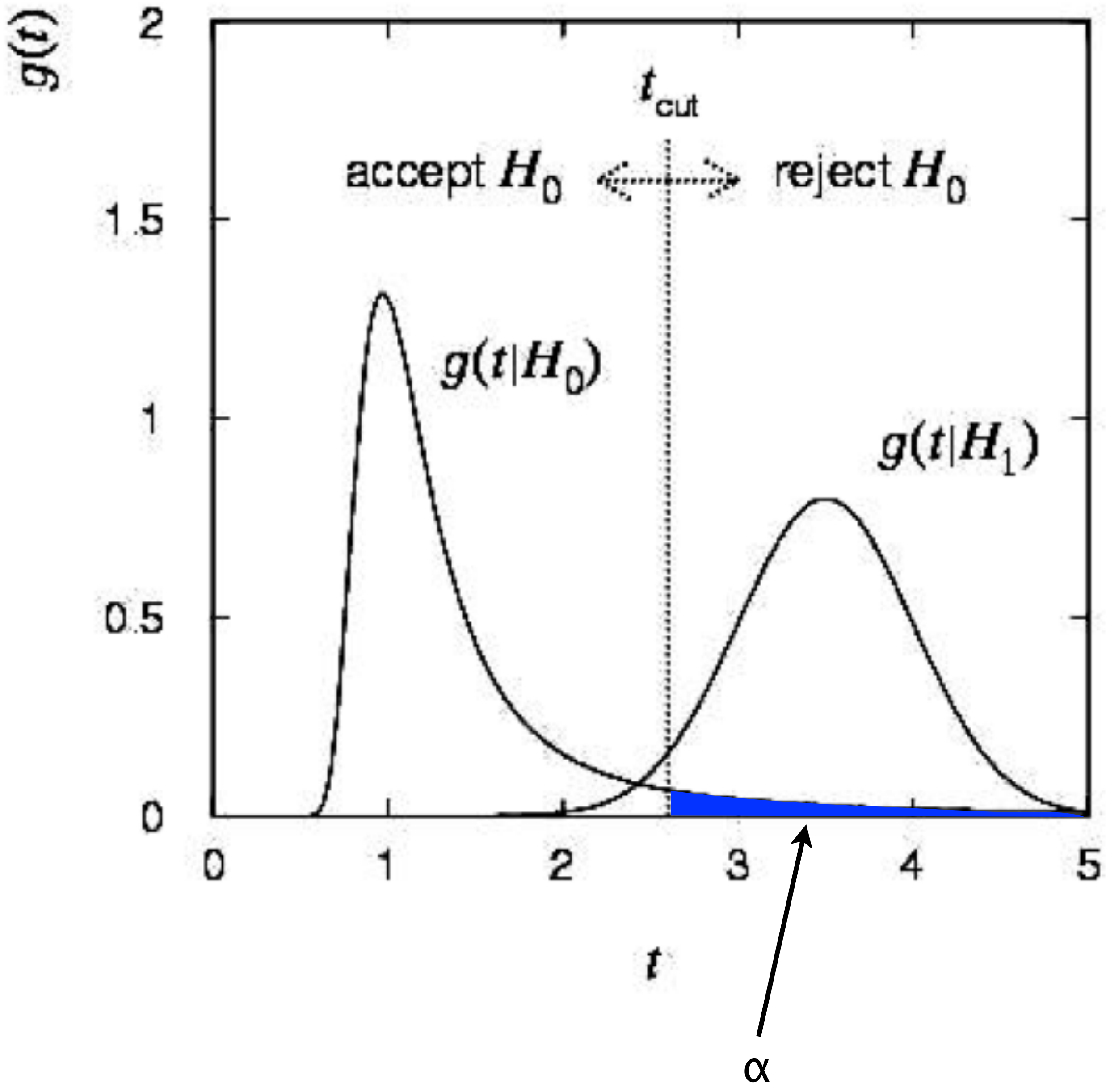


Statistical Tests - Decision Boundary

- The decision boundary defines a test. If the data falls into the critical region then we reject the null hypothesis.
- Define the error of the first kind as α as a probability to reject the null hypothesis if the null hypothesis is true:

$$\alpha = \int_{t_{cut}}^{\infty} g(t; H_0) dt$$

- The statistical significance of rejection is given by the p-value



Statistical Tests - Decision Boundary

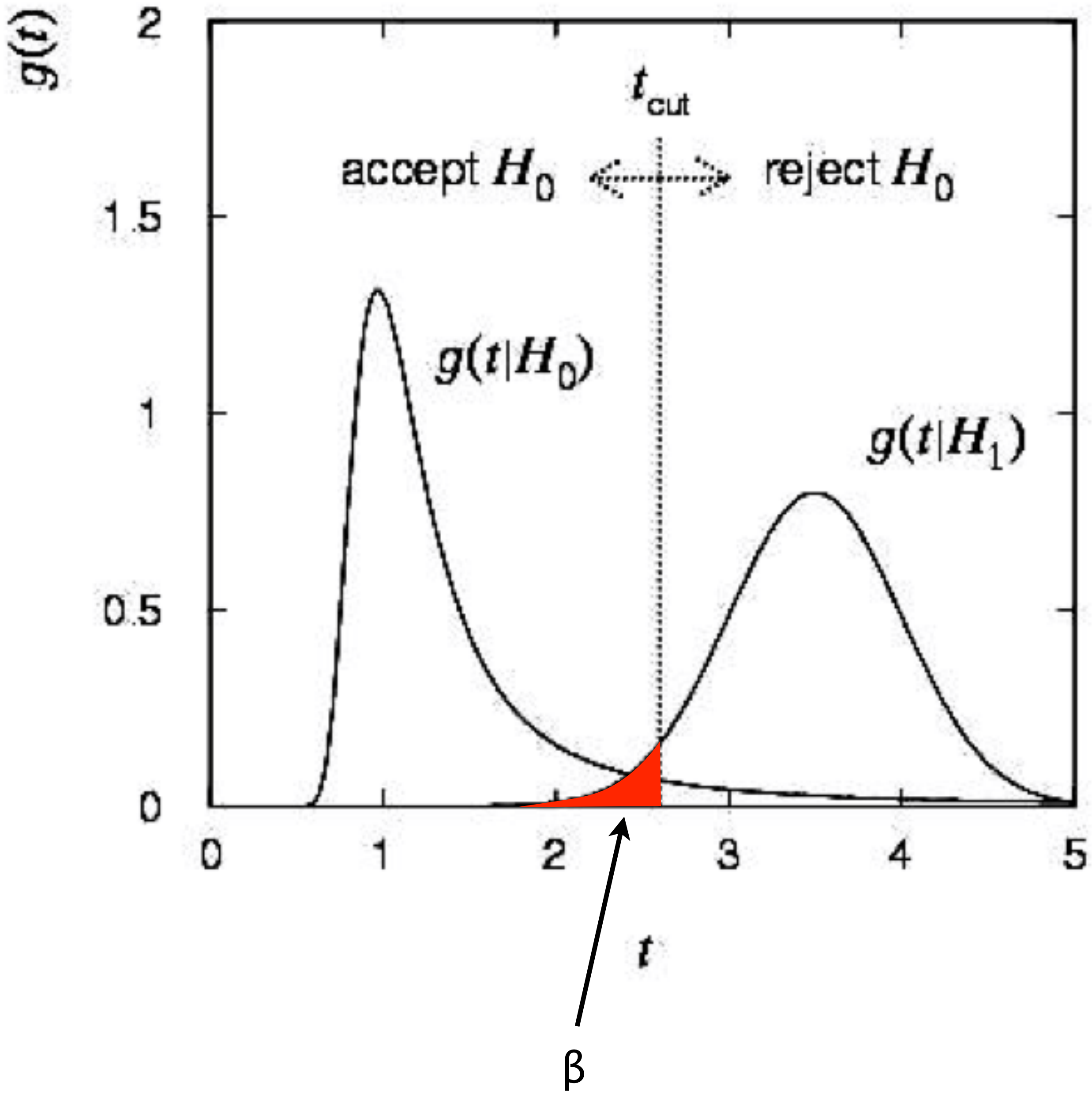
- Consider now the alternate hypothesis.

- Define the error of the second kind as β as a probability to accept the null hypothesis but the true hypothesis was the alternate hypothesis

$$\beta = \int_{-\infty}^{t_{cut}} g(t; H_1)dt$$

- The **power** of the test, probability of rejecting the null hypothesis when it is false, is (1-β).

- A more powerful test leads to: (1-β) = maximized. Aim for α and β small as possible.



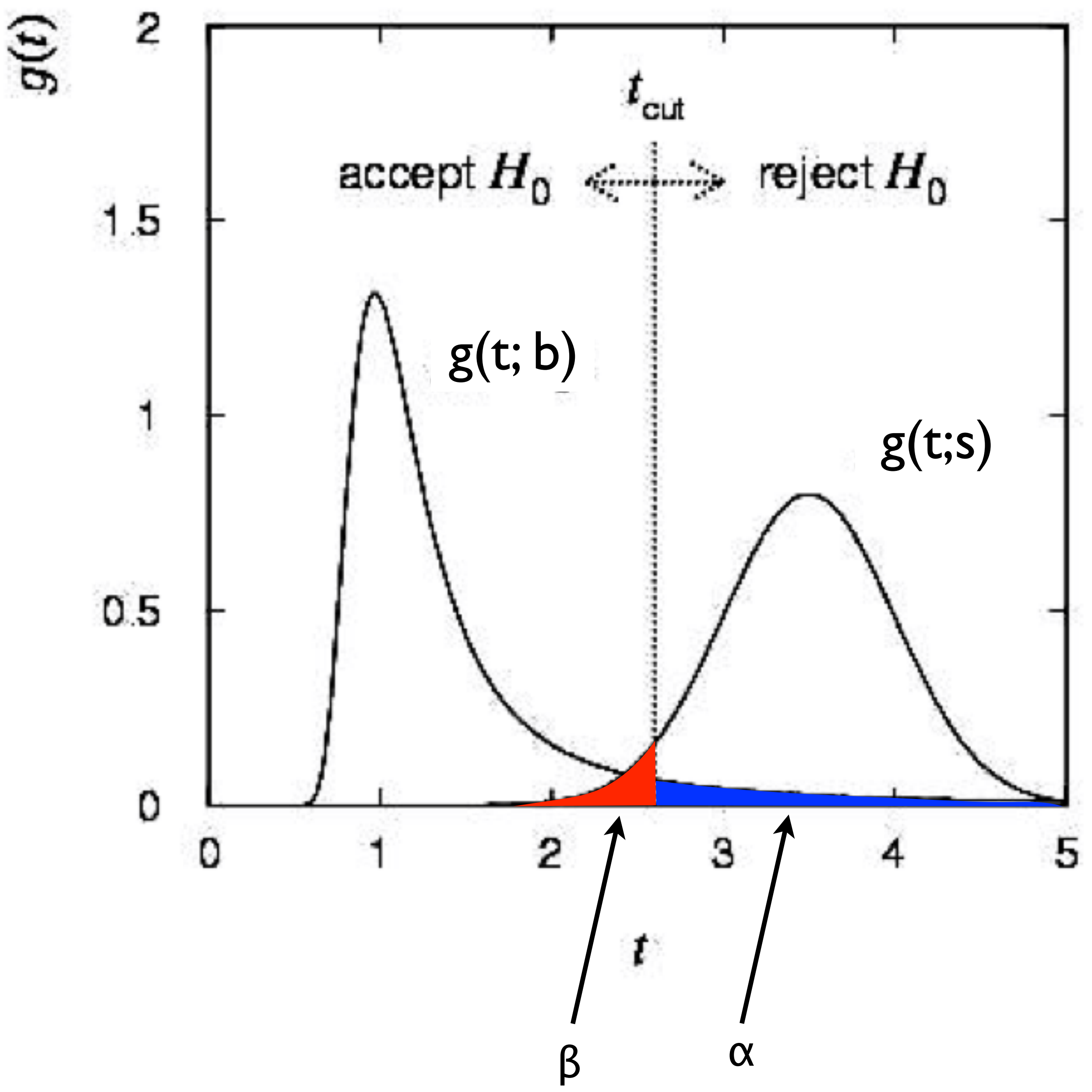
Statistical Tests - Signal & Background

- The probability to reject a background hypothesis for background events is:

$$\epsilon_b = \int_{t_{cut}}^{\infty} g(t;b)dt = \alpha$$

- The probability to accept a signal event as signal is the signal efficiency:

$$\epsilon_s = \int_{t_{cut}}^{\infty} g(t;s)dt = 1 - \beta$$



Statistical Tests - Test Statistic

- Constructing a test statistic
  - Keep in mind the goal is to choose a test's critical region in an optimal way
  - The Neyman-Pearson lemma states:

> To obtain the highest power for a given significance level in a test of the null/background hypothesis versus the alternate/signal hypothesis, choose the critical region such that:

$$\frac{f(x|\theta_1)}{f(x|\theta_0)} > k$$ **inside the region**

- - We can demonstrate this method by choosing a critical value for x and both the null and alternate hypotheses are simple (only two possible values):

$$\alpha = \int_R f(x|\theta_0)dx \qquad 1 - \beta = \int_R f(x|\theta_1)dx = \int_R \frac{f(x|\theta_1)}{f(x|\theta_0)} f(x|\theta_0)dx$$

- - To maximize the power, take the region of 1-β, and define the set of points according to the above condition. Note that k is determined from α.

Maximum Likelihood Ratio

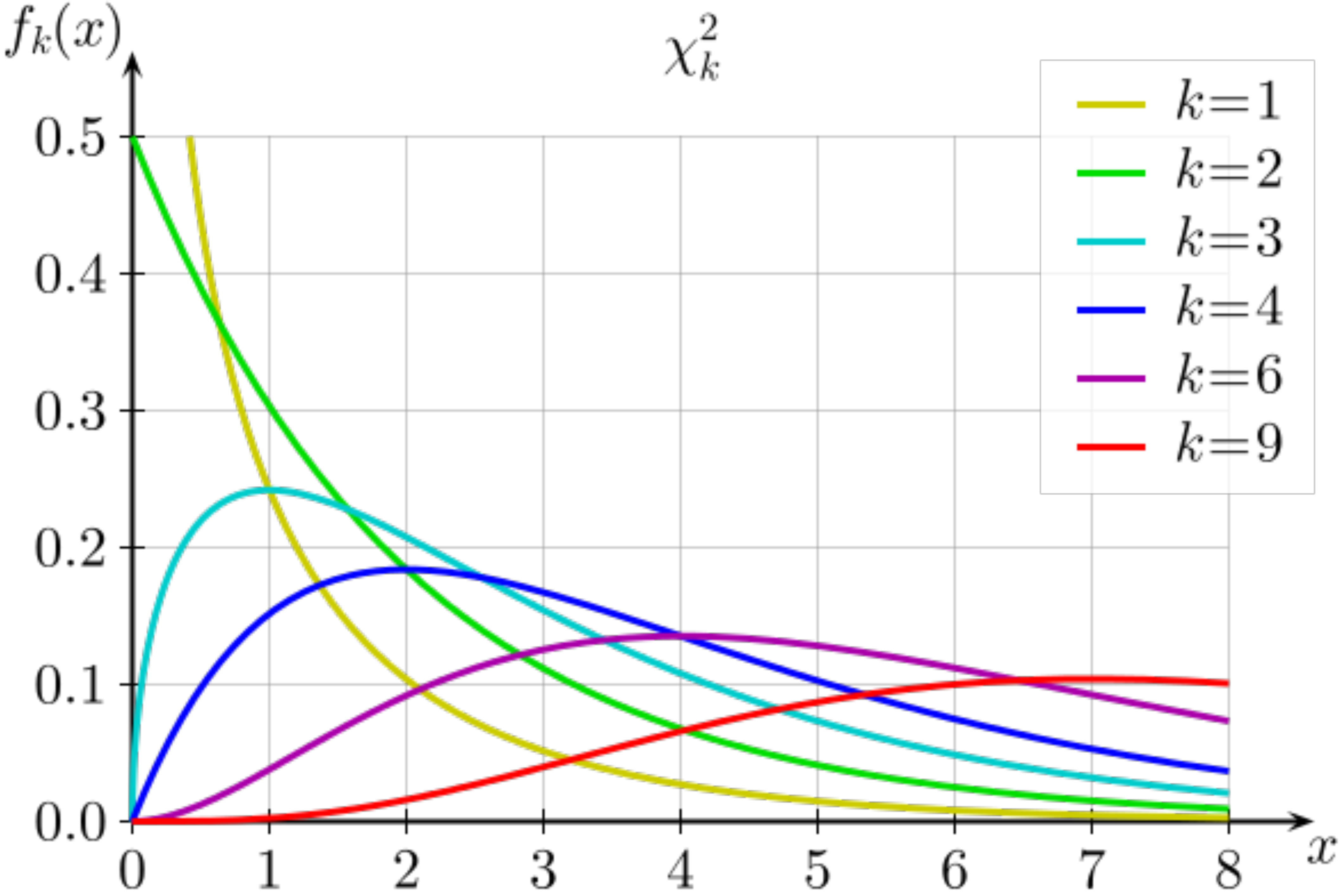
- An very common test-statistic for the likelihood ratio is:

$$\Lambda(\theta, x_{obs}) = -2\ln\frac{\mathcal{L}(\theta_0|x_{obs})}{\mathcal{L}(\hat{\theta}|x_{obs})}$$

  - Where the difference between the null hypothesis in the numerator and the alternative hypothesis in the denominator is that the null hypothesis has a fixed value of a single (or more) of the θ parameter(s) whereas the alternative hypothesis fits/maximizes the parameter.
  - The null hypothesis is named as such because it often has a parameter set to zero

- For a normal distributed, i.e. gaussian, variable the likelihood ratio follows a χ^2 distribution,
  - N_{DOF} = difference in dimensionality between the models
  - Also requires that Wilk's Theorem is satisfied (more later)

χ^2 Distributions

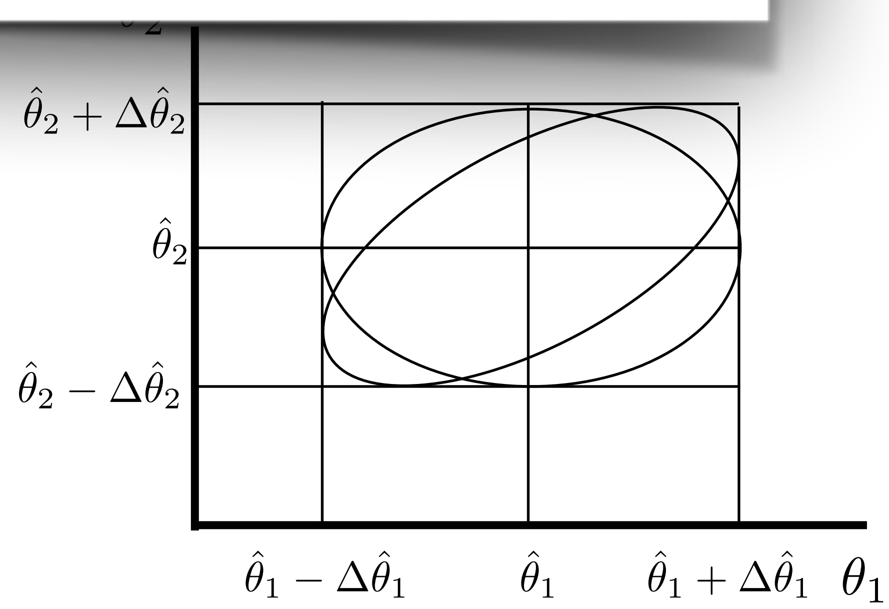


*wikipedia

Variance of Estimators - Graphical Me

From Parameter Uncertainty Lecture

- When the correct, tangential, method is used then the uncertainties are not dependent on the correlation of the variables.
- The probability the ellipses of constant $\ln L = \ln L_{max} - a$ contains the true point, θ_1 and θ_2 , is:


$\hat{\theta}_2 + \Delta\hat{\theta}_2$
$\hat{\theta}_2$
$\hat{\theta}_2 - \Delta\hat{\theta}_2$
$\hat{\theta}_1 - \Delta\hat{\theta}_1$
$\hat{\theta}_1$
$\hat{\theta}_1 + \Delta\hat{\theta}_1$
θ_1

correct

a (1 dof)	a (2 dof)	σ
0.5	1.15	1
2.0	3.09	2
4.5	5.92	3

Significance Values for Uncertainty Limits from Likelihood Values

The probability the ellipses of constant $\ln L = \ln L_{max} - a$ contains the true point, θ_1 and θ_2 , is:

a (1 dof)	a (2 dof)	σ
0.5	1.15	1
2.0	3.09	2
4.5	5.92	3

→

Multiply ‘a’ by 2 to get

The probability the ellipses of constant $2\ln L = 2\ln L_{max} - a$ contains the true point, θ_1 and θ_2, is:

a (1 dof)	a (2 dof)	σ
1	2.30	1
4	6.18	2
9	11.83	3

So, where do the values of ‘a’ come from?

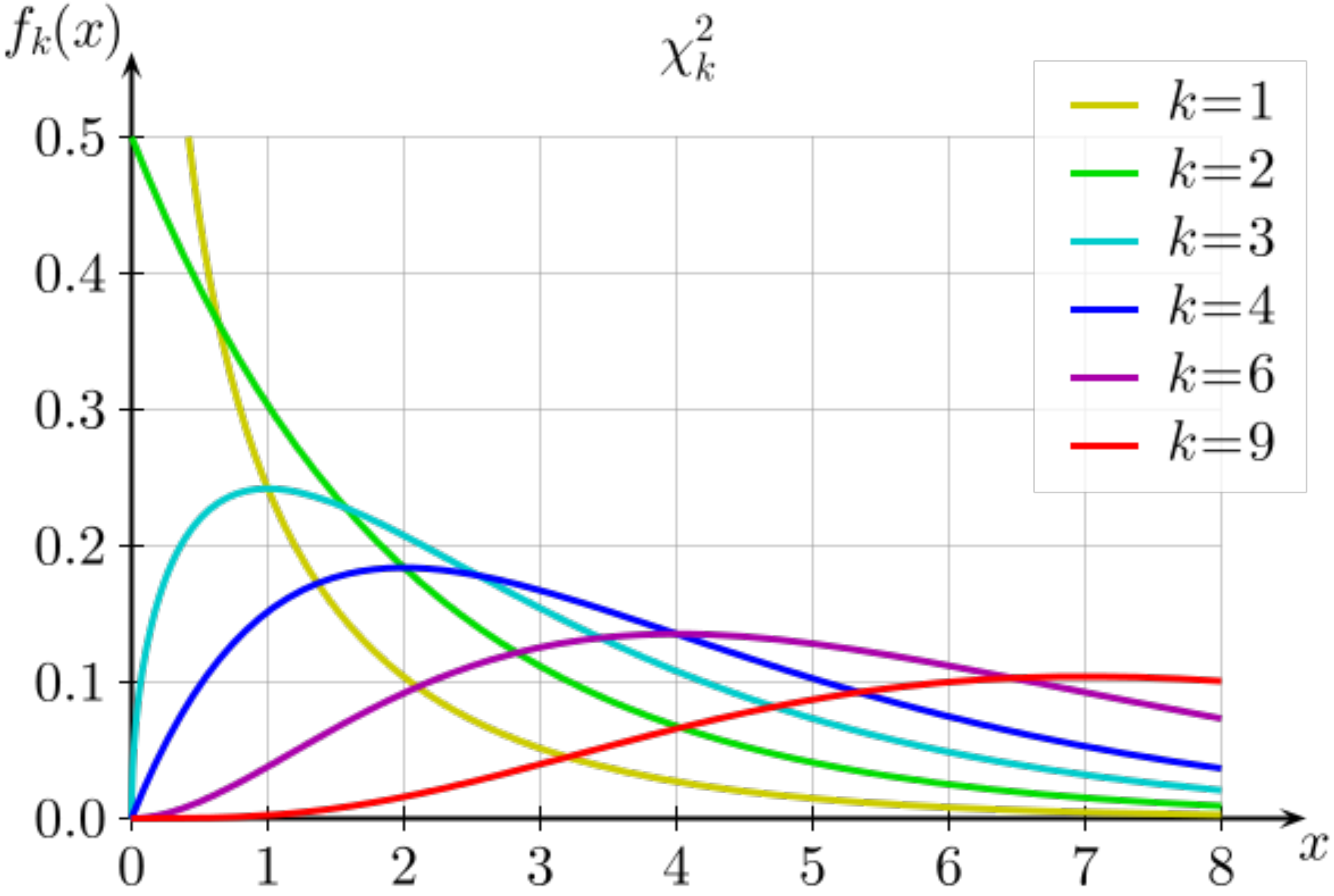
Significance Values for Uncertainty Limits from Likelihood Values

- The probability the ellipses of constant $2\ln L = 2\ln L_{max} - a$ contains the true point, θ_1 and θ_2, is:

a (1 dof)	a (2 dof)	σ or %
1	2.30	1σ or 68.27%
4	6.18	2σ or 95.45%
9	11.83	3σ or 99.73%

- Because 2*ΔLLH is χ^2 distributed, the values of ‘a’ in the table above correspond to

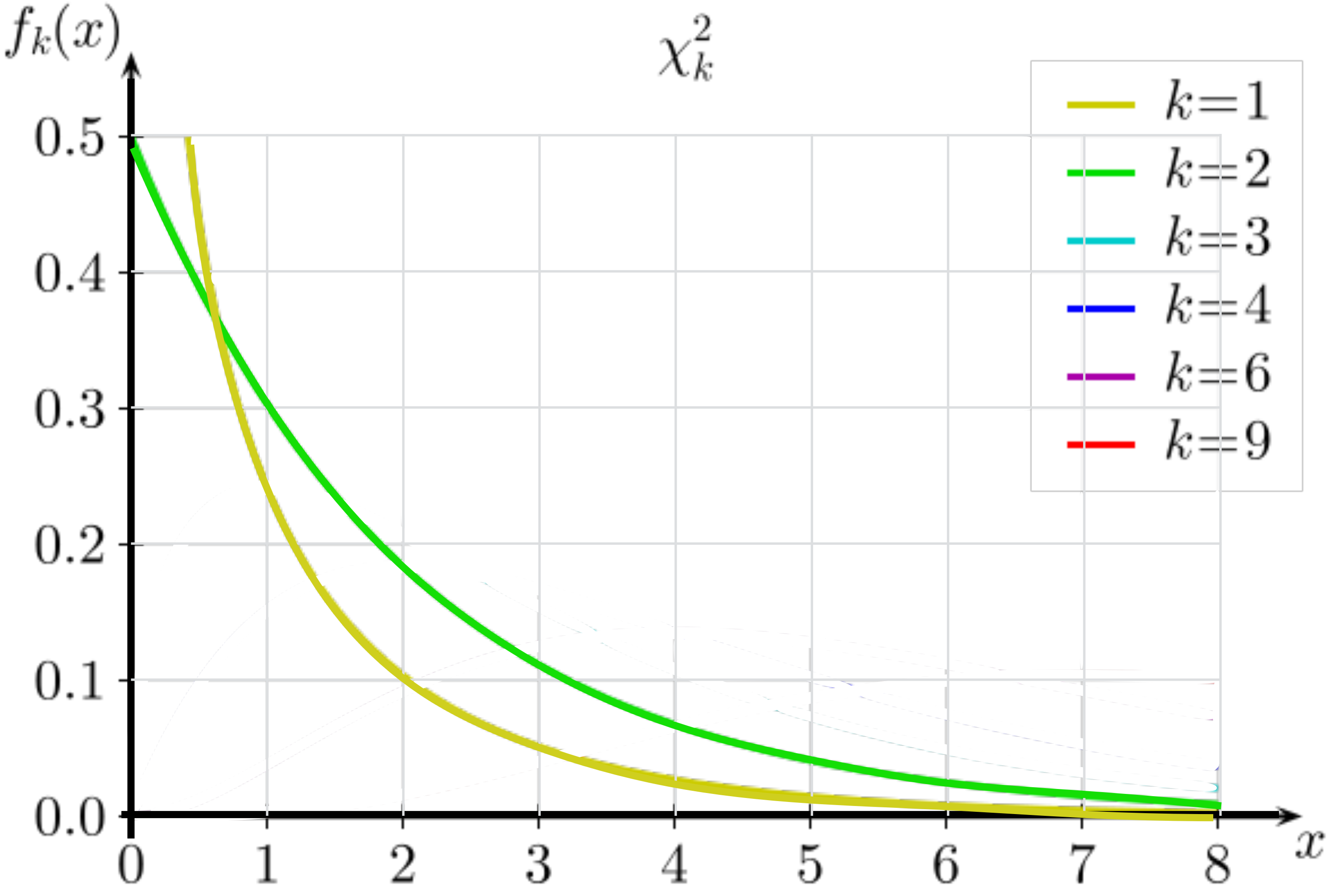
$$N\sigma = \int_0^a f_k(x)dx$$



Significance Values for Uncertainty Limits from Likelihood Values

- The probability the ellipses of constant $2\ln L = 2\ln L_{max} - a$ contains the true point, θ_1 and θ_2, is:

a (1 dof)	a (2 dof)	σ or %
1	2.30	1σ or 68.27%
4	6.18	2σ or 95.45%
9	11.83	3σ or 99.73%



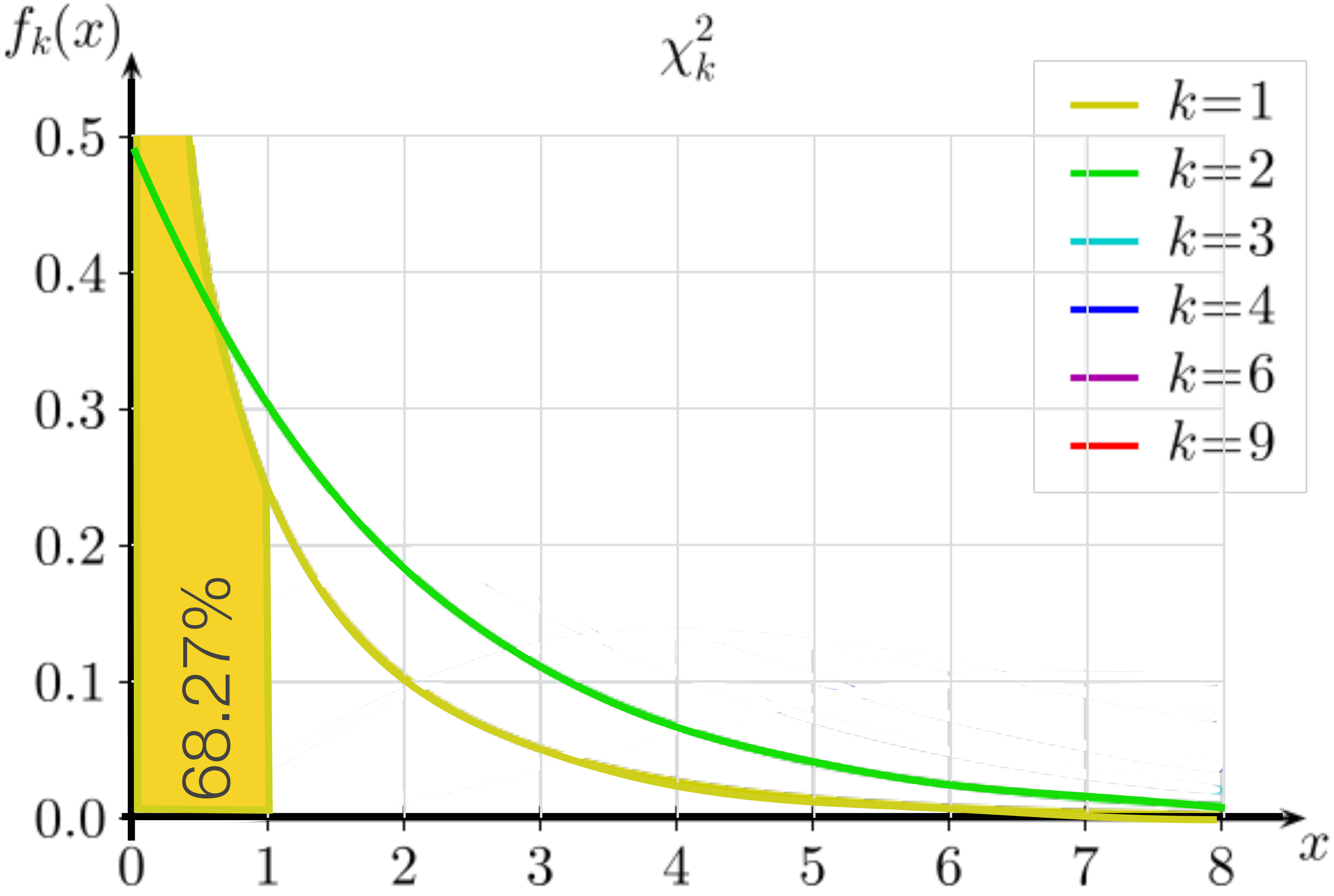
- Because 2*ΔLLH is χ^2 distributed, the values of 'a' in the table above correspond to

$$N\sigma = \int_0^a f_k(x)dx$$

Significance Values for Uncertainty Limits from Likelihood Values

- The probability the ellipses of constant $2\ln L = 2\ln L_{max} - a$ contains the true point, θ_1 and θ_2, is:

a (1 dof)	a (2 dof)	σ or %
1	2.30	1σ or 68.27%
4	6.18	2σ or 95.45%
9	11.83	3σ or 99.73%



- Because 2*ΔLLH is χ^2 distributed, the values of 'a' in the table above correspond to

$$N\sigma = \int_0^a f_k(x)dx$$

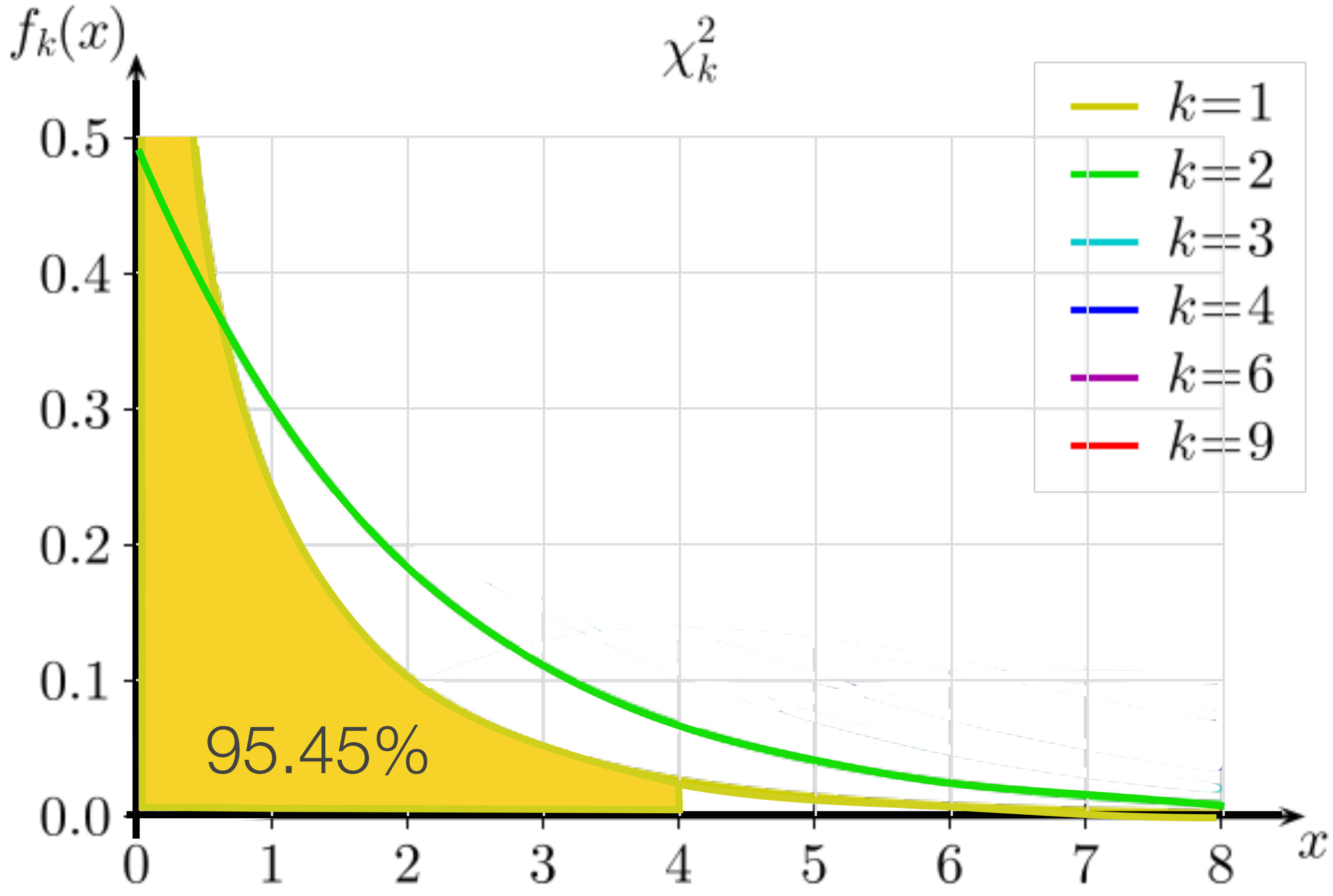
Significance Values for Uncertainty Limits from Likelihood Values

- The probability the ellipses of constant $2\ln L = 2\ln L_{max} - a$ contains the true point, θ_1 and θ_2, is:

a (1 dof)	a (2 dof)	σ or %
1	2.30	1σ or 68.27%
4	6.18	2σ or 95.45%
9	11.83	3σ or 99.73%

- Because 2*ΔLLH is χ^2 distributed, the values of ‘a’ in the table above correspond to

$$N\sigma = \int_0^a f_k(x)dx$$



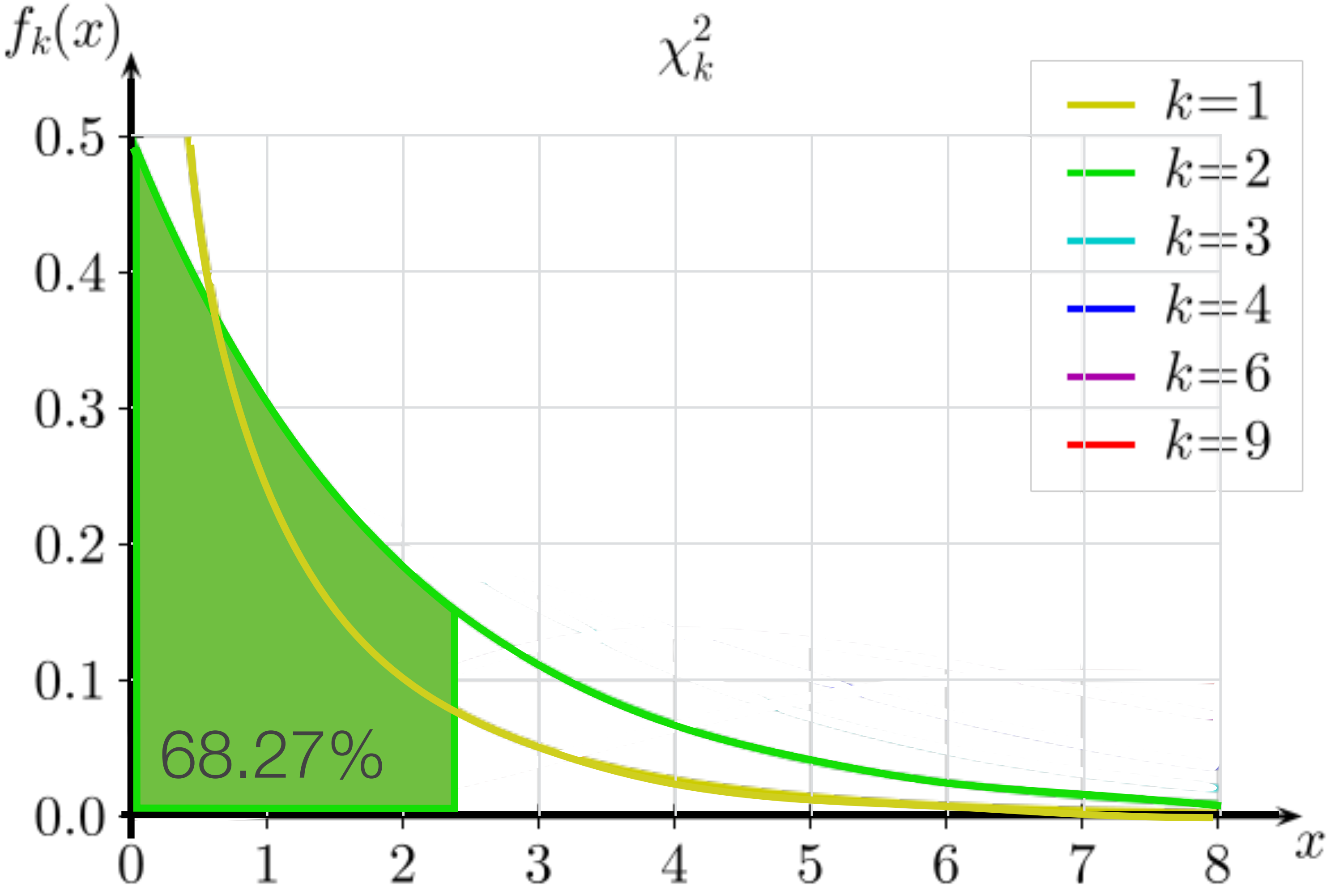
Significance Values for Uncertainty Limits from Likelihood Values

- The probability the ellipses of constant $2\ln L = 2\ln L_{max} - a$ contains the true point, θ_1 and θ_2, is:

a (1 dof)	a (2 dof)	σ or %
1	2.30	1σ or 68.27%
4	6.18	2σ or 95.45%
9	11.83	3σ or 99.73%

- Because 2*ΔLLH is χ^2 distributed, the values of 'a' in the table above correspond to

$$N\sigma = \int_0^a f_k(x)dx$$



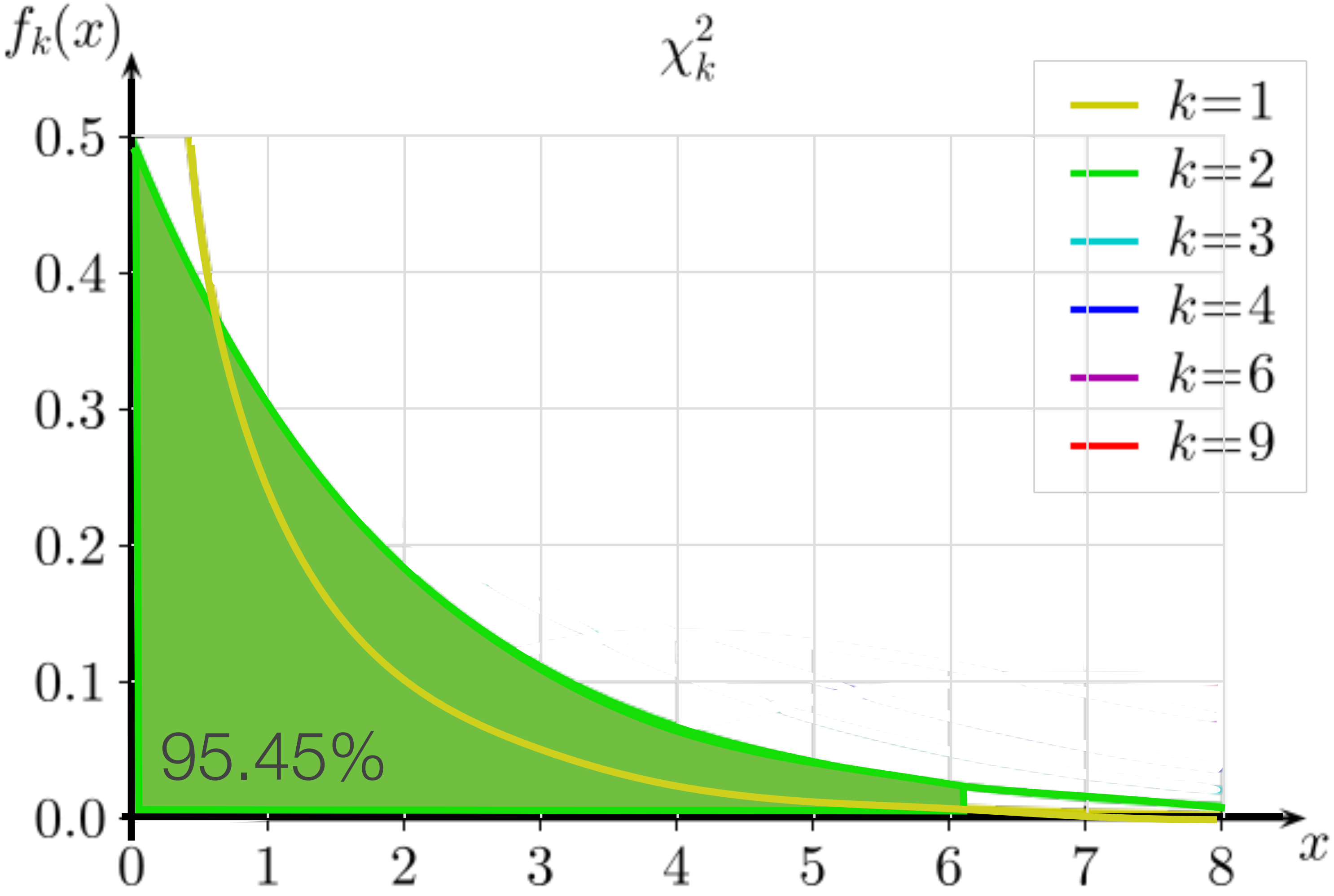
Significance Values for Uncertainty Limits from Likelihood Values

- The probability the ellipses of constant $2\ln L = 2\ln L_{max} - a$ contains the true point, θ_1 and θ_2, is:

a (1 dof)	a (2 dof)	σ or %
1	2.30	1σ or 68.27%
4	6.18	2σ or 95.45%
9	11.83	3σ or 99.73%

- Because 2*ΔLLH is χ^2 distributed, the values of ‘a’ in the table above correspond to

$$N\sigma = \int_0^a f_k(x)dx$$

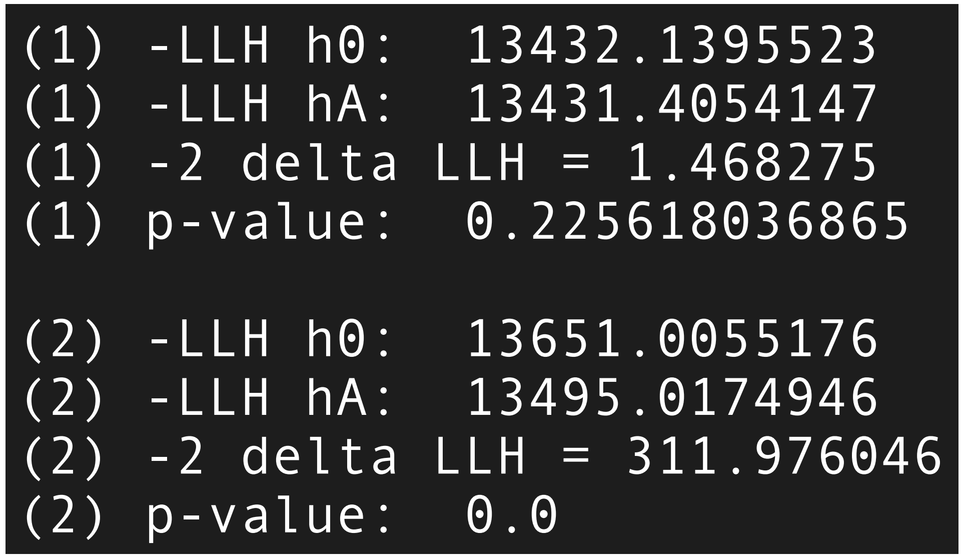


Quick Note

- For any arbitrary percent threshold and degrees of freedom, the critical chi-squared value can be calculated from the inverse survival function
  - scipy.stats.chi2.isf(1-C.L. as percent/100, DoF)
  - For a 68.27% interval w/ 2 DoF
    scipy.stats.chi2.isf(1-0.6827,2)=2.2958

Exercise #1

- From the files posted on the class webpage for this lecture, use the $\ln$-likelihood ratio and calculate the p-value of each data set for -1 ≤ x ≤ 1:
  - The null hypothesis is the PDF from $f(x;\alpha,\beta) = 1 + \alpha x + \beta x^2$
  - The alternative hypothesis is $f(x;\alpha,\beta,\gamma) = 1 + \alpha x + \beta x^2 - \gamma x^5$



```
(1) -LLH h0:  13432.1395523
(1) -LLH hA:  13431.4054147
(1) -2 delta LLH = 1.468275
(1) p-value:  0.225618036865

(2) -LLH h0:  13651.0055176
(2) -LLH hA:  13495.0174946
(2) -2 delta LLH = 311.976046
(2) p-value:  0.0
```

Wilk's Theorem… Kinda

- As the number of data points approaches infinity, the ln likelihood ratio converges to a χ^2 distribution if H_0 is true

$$\Lambda(\theta, x_{obs}) = -2 \ln \frac{\mathcal{L}(\theta_0|x_{obs})}{\mathcal{L}(\hat{\theta}|x_{obs})}$$

- But there are regions where the gaussian, and therefore Wilk's and our use of χ^2, breaks down
  - Low number of events where the probability switches from gaussian to poisson
  - Bounds on the model parameters, e.g. as n→infinity the parameter does not smoothly vary, but has some truncation or discrete behavior
  - Parameters that have a near-infinite variance

Real World Application

- The tests so far have been within the realm of Monte Carlo perfection and do not include any systematic uncertainties that are found in real experiments. In practice, i.e. when including systematics, χ^2 and p-values and other tests tend to give better agreement between data and hypothesis/simulation/fits than what is expected.
  - Systematic uncertainties are almost always conservative, i.e. too big
  - Fitting procedures try to make the model/simulation/etc. look like the data as best as possible (maximum likelihood)
  - Fitting procedures will use systematic parameters to ‘damp’ statistical under- and over-fluctuations

Conclusion

- Hypothesis testing is good
- Take time to go back through previous class exercises if you have not already
- Nice link about quickly interpreting distributions of p-values
  - http://varianceexplained.org/statistics/interpreting-pvalue-histogram/
- Nice material about the Neyman-Person lemma and the power of the likelihood ratio
  - https://onlinecourses.science.psu.edu/stat414/node/307
  - Original paper is at https://doi.org/10.1098/rsta.1933.0009