Lecture 8 : Hypothesis Tests

D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2016

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:

$$\vec{x} = (x_1, ..., x_n)$$

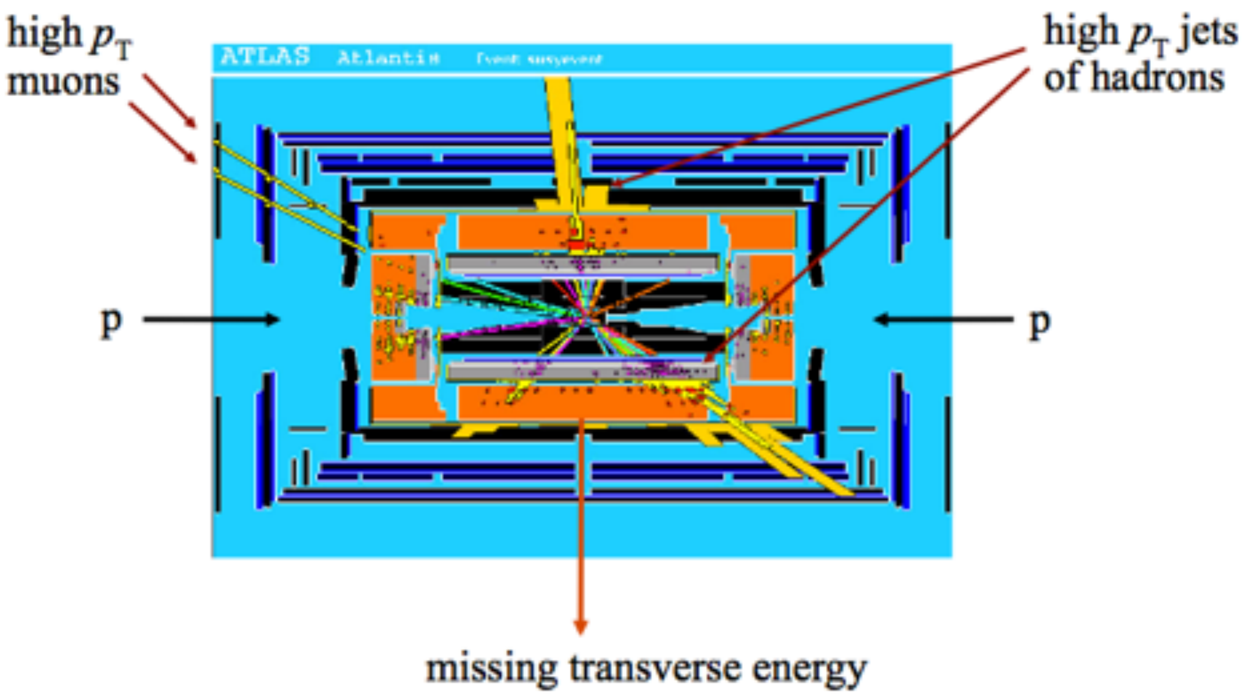
$$x_1 = \text{number of muons} \qquad x_2 = \text{number of jets} \; ...$$

  - 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:

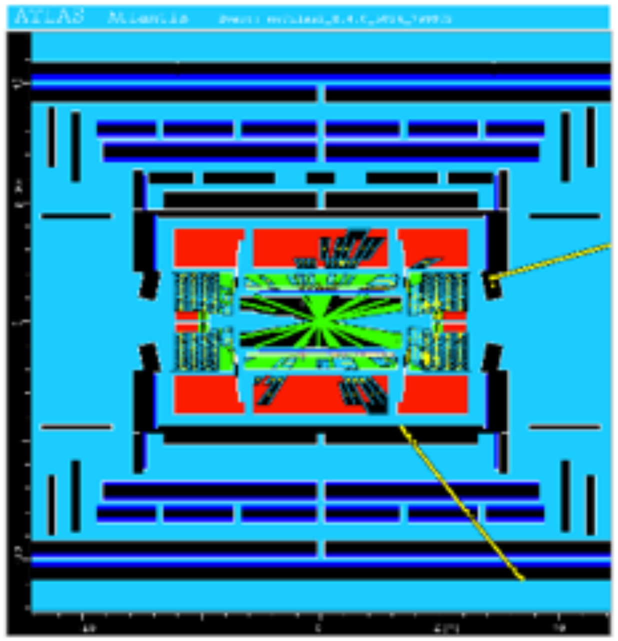
$$f(\vec{x}|H_0), f(\vec{x}|H_1), ...$$

  - 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



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 can not be proven true or false but you can determine the probability for obtaining the observed result if you assume the hypothesis is true.

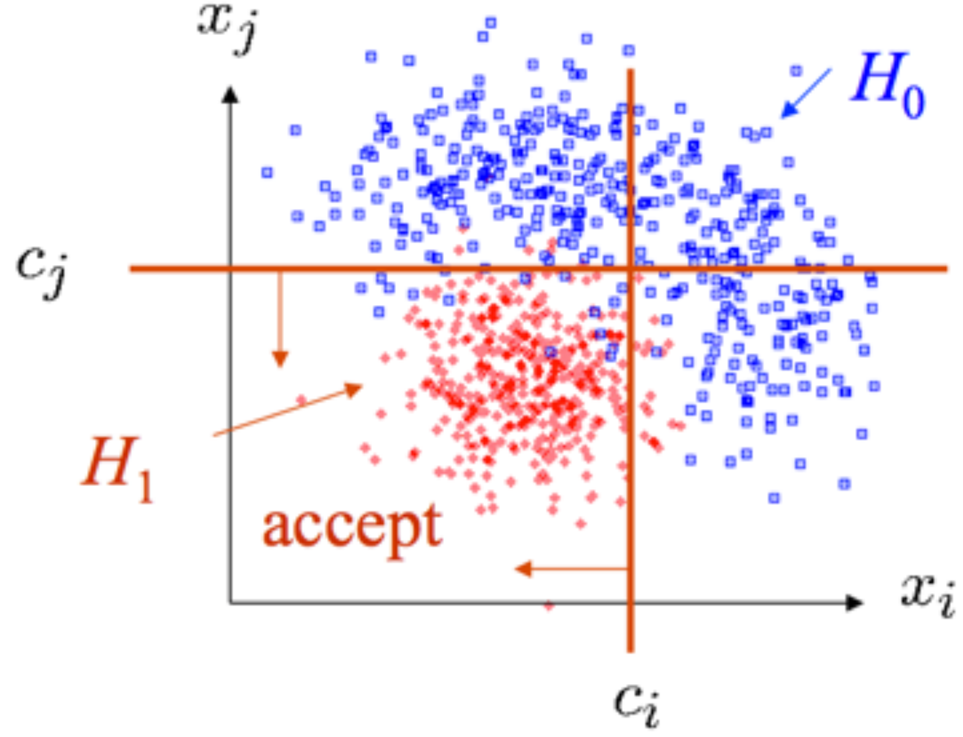
- Hypothesis testing is also a part of data analysis when, for example you decide if a specific observed event is signal or background. Suppose you have a data sample with two kinds of events that correspond to the null and alternate hypotheses and you want to select those that are of the type corresponding 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

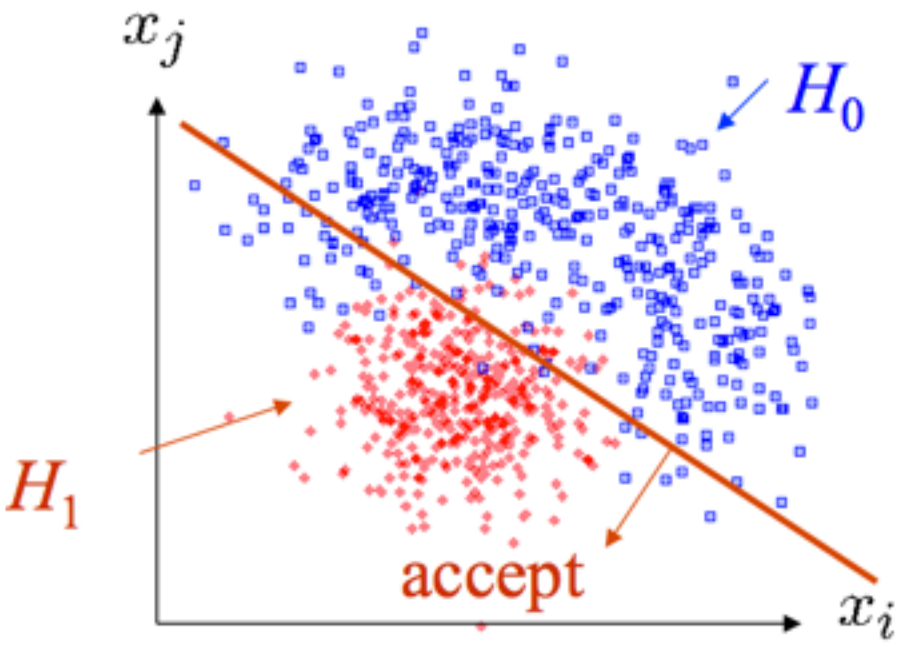
- Event Selection
    - selection cuts for events, e.g.

      $x_j < c_j \qquad x_i < c_i$

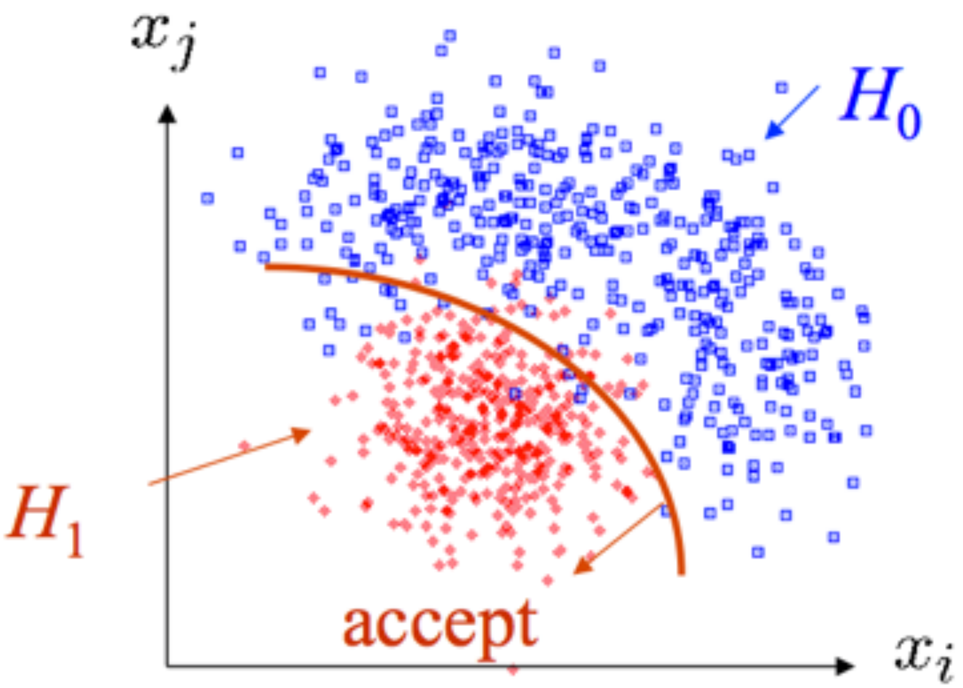
    - We would like to optimize this process…



linear



or nonlinear



*G. Cowan

Statistical Tests - Decision Boundary

- The decision boundary can be defined using an equation of the form:

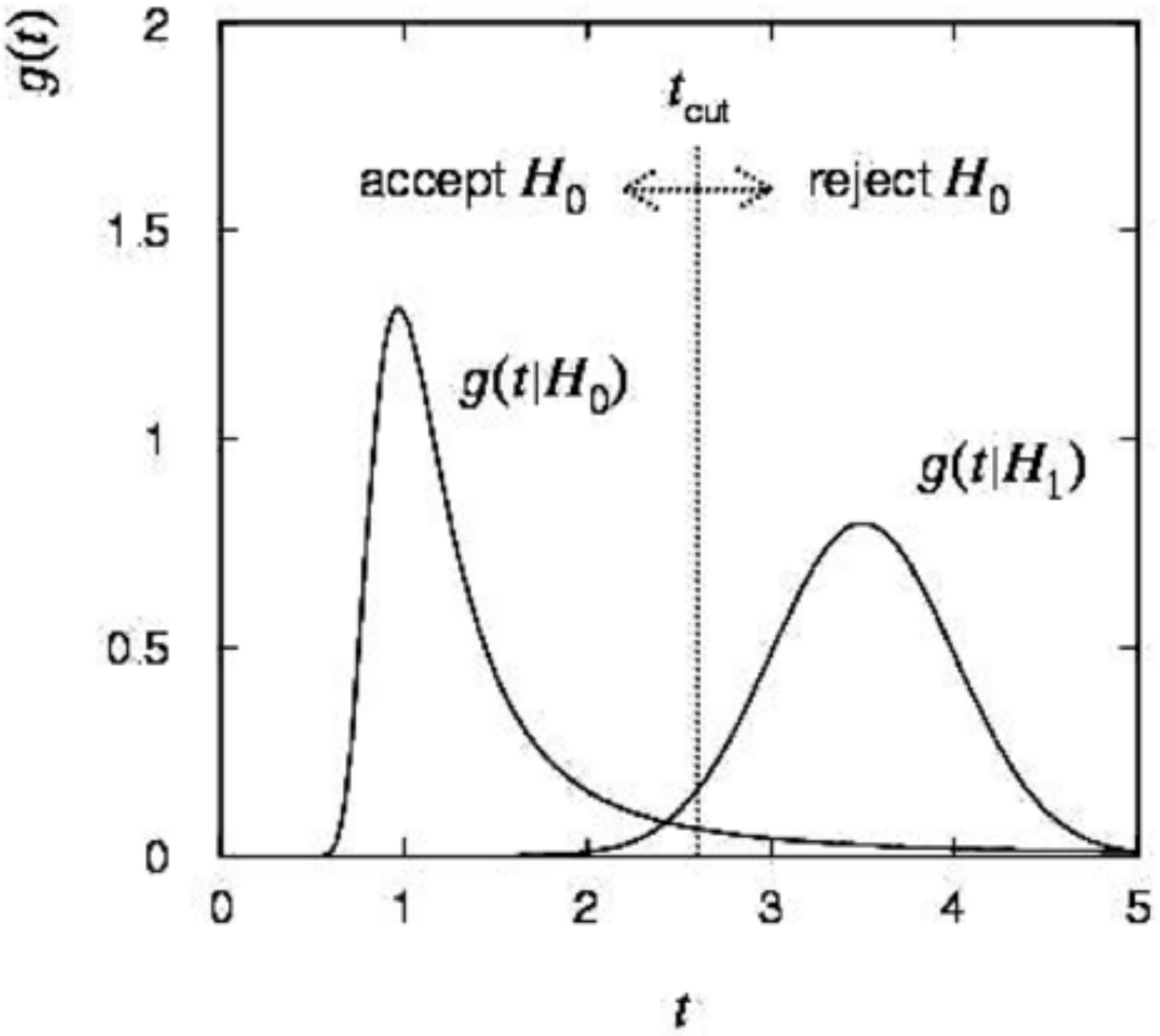
$$t(x_1, ..., x_n) = t_{cut}$$

- 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:

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



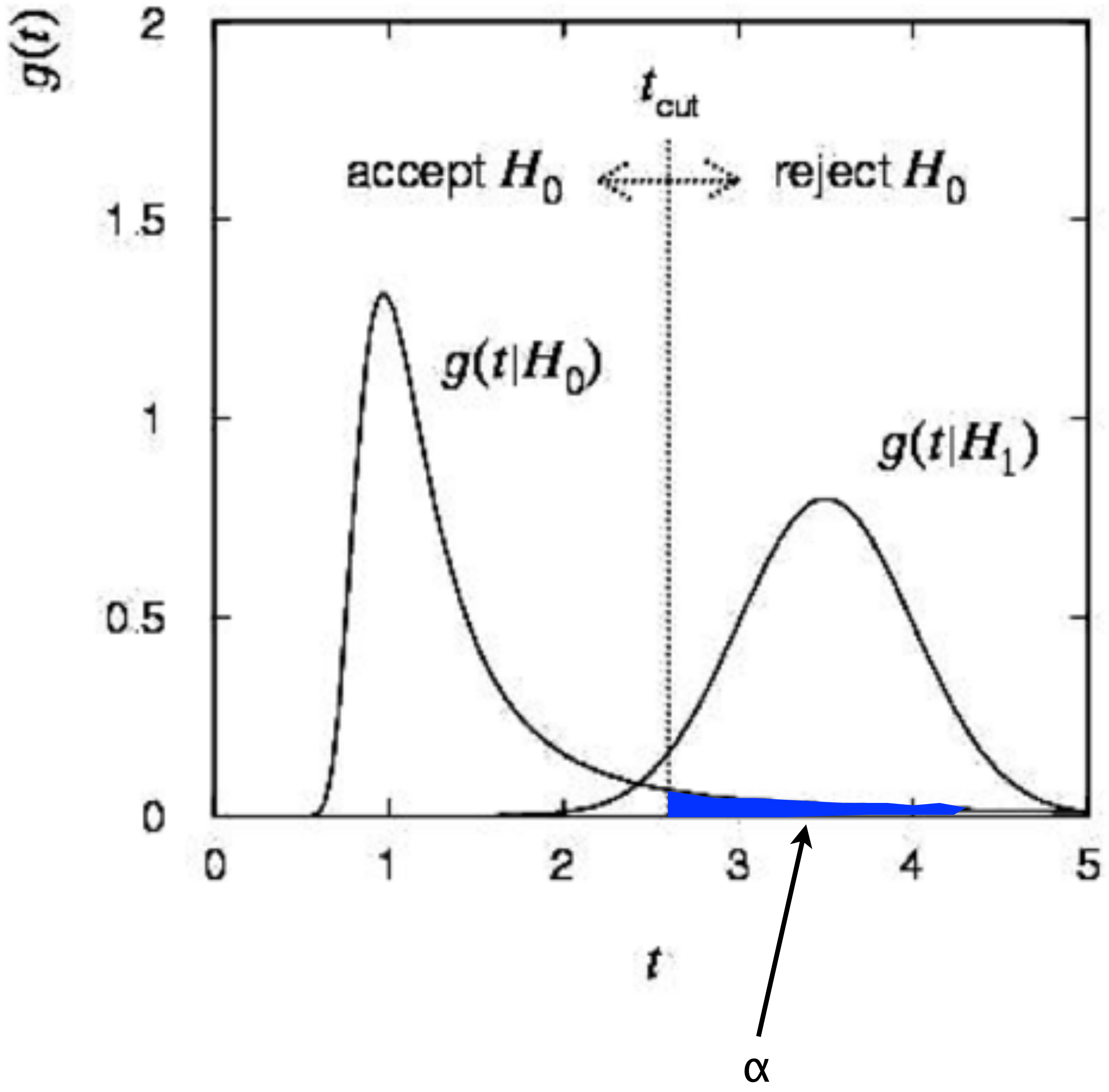
Statistical Tests - Decision Boundary

- The decision boundary defines a test. If the data falls into the critical region 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



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.
    - A 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 "$q_{\mu,\mathrm{obs}}$" is the TS value from the observed data

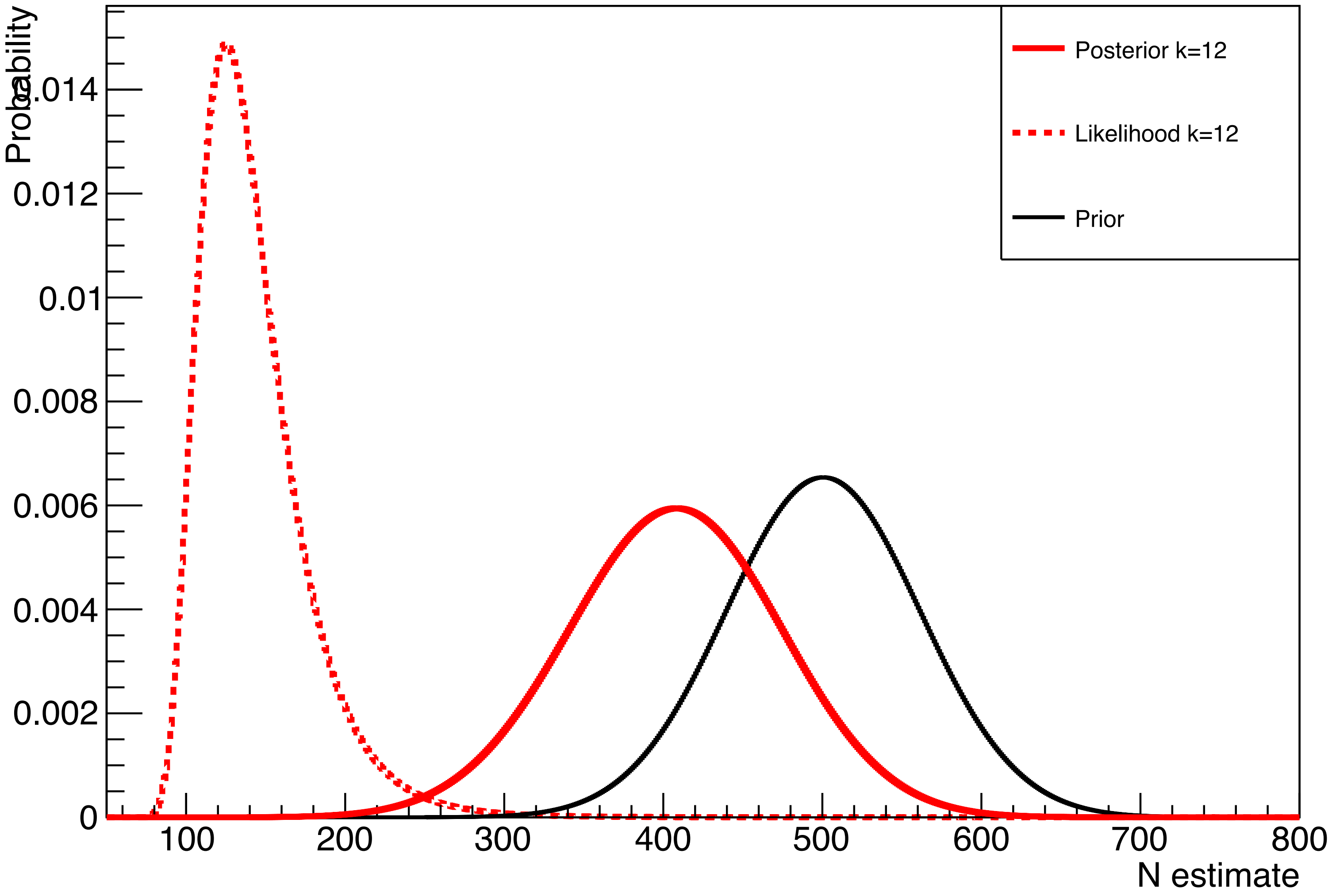
Even More Extreme

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)



Exercise #3 From Previous Lecture

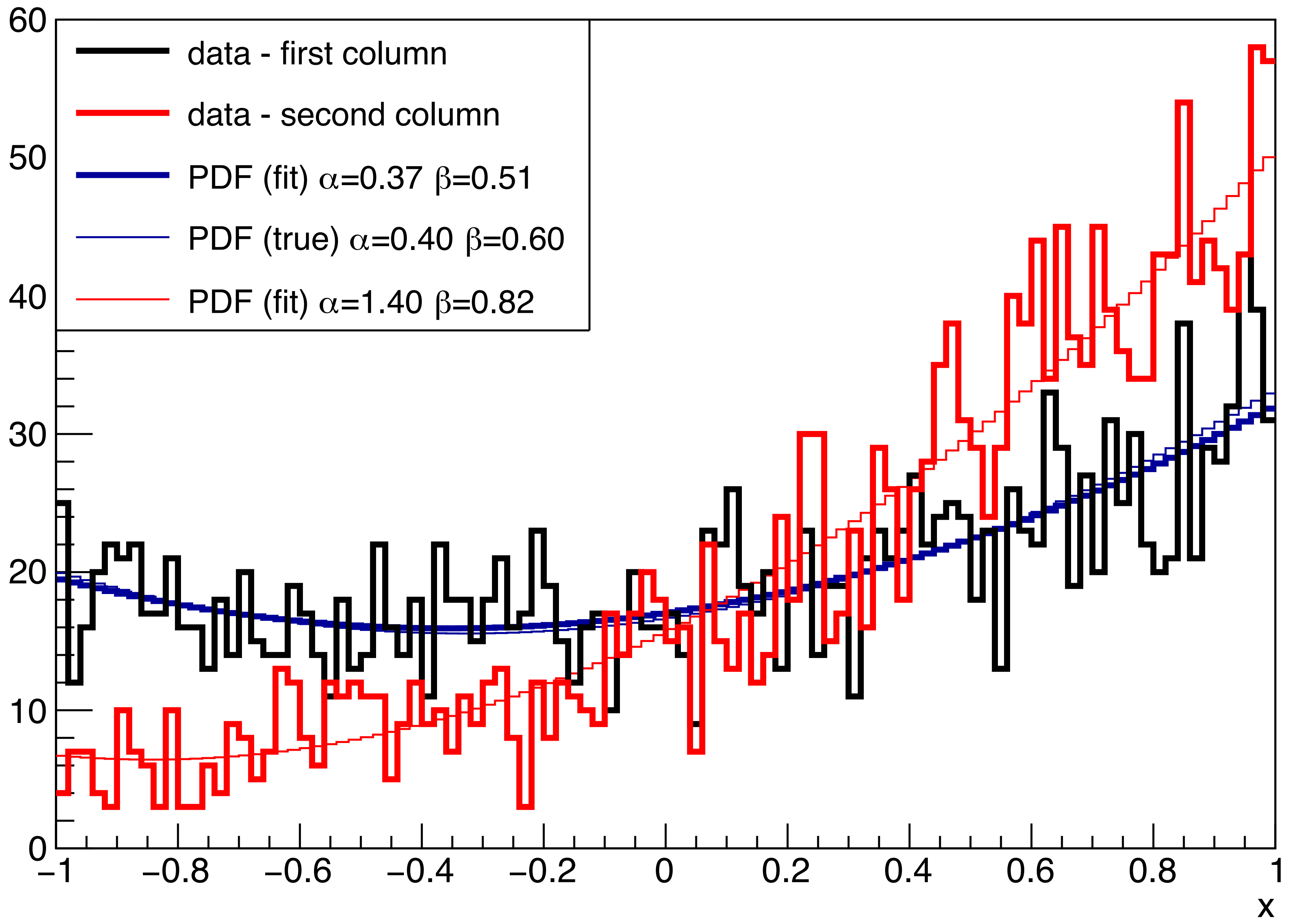
- There is a file posted on the class webpage for “Class 7” which has two columns of x numbers (not x and y, only x for 2 pseudo-experiments) corresponding to x over the range -1 ≤ x ≤ 1
- Using the function:

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

  - Find the best-fit for the unknown **α** and **β**
  - Calculate the reduced chi-square goodness of fit 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

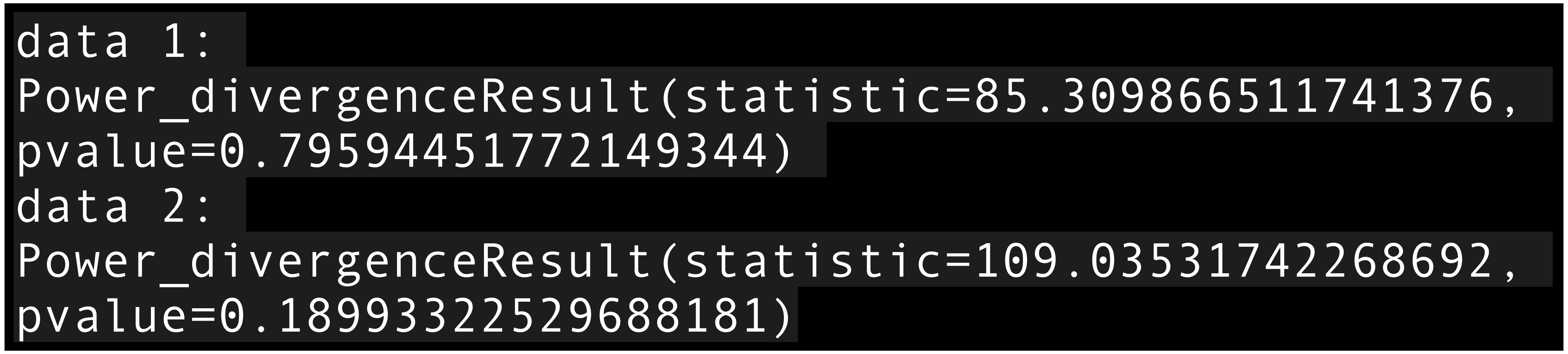
- Histograms: the x-values of the two pseudo-experiments, the expectation from PDF using the best-fit values and the true values (which only I knew because I generated the data)



Follow-up on Exercise

- In the last exercise I wrote to “Calculate the reduced chi-square goodness of fit by histogramming the data”, but it may be more informative to calculate the p-value
  - 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:

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

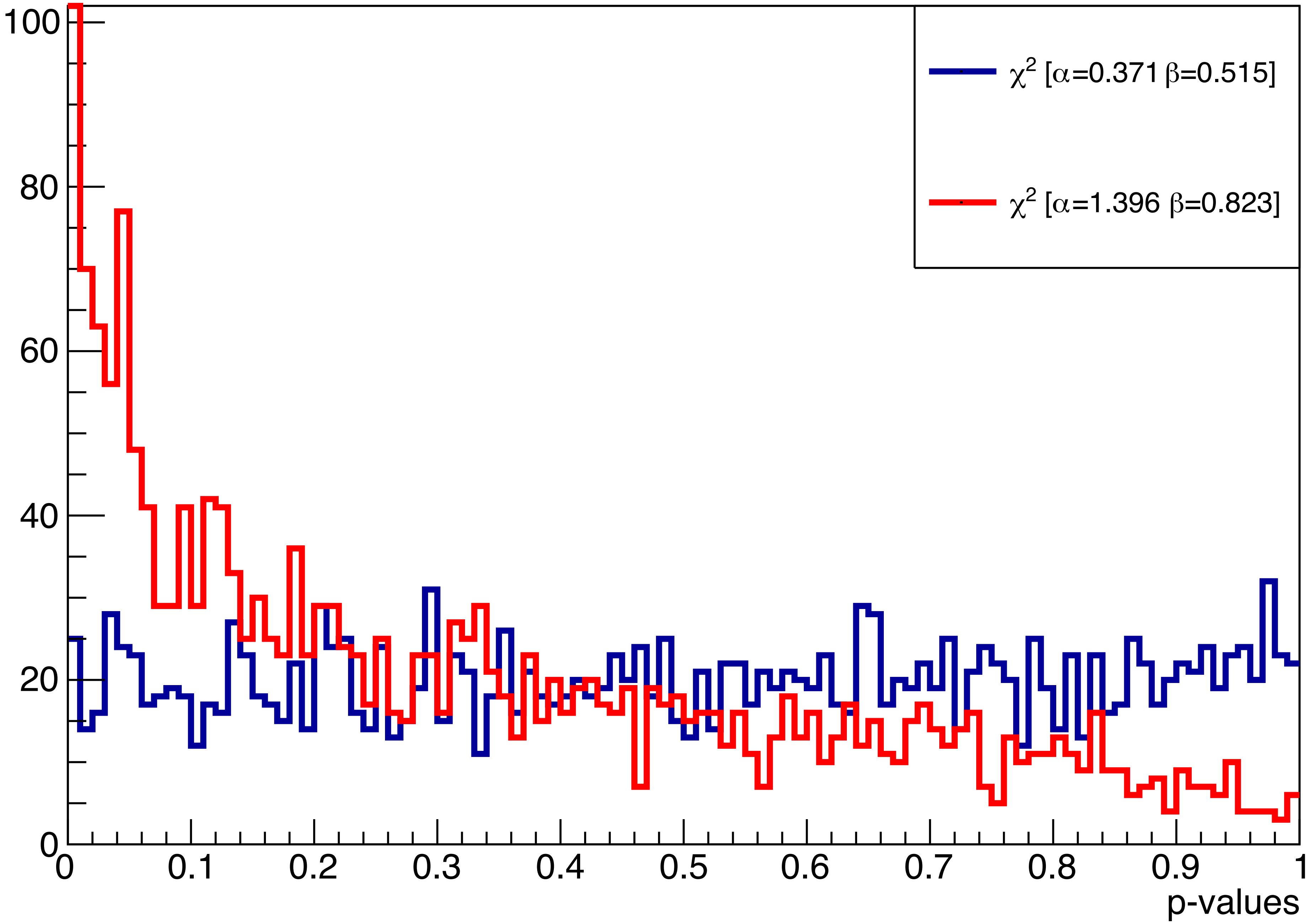


```
data 1:
Power_divergenceResult(statistic=85.309866511741376,
pvalue=0.79594451772149344)
data 2:
Power_divergenceResult(statistic=109.03531742268692,
pvalue=0.18993322529688181)
```

Funny Thing

- Someone 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 as before, but using the 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

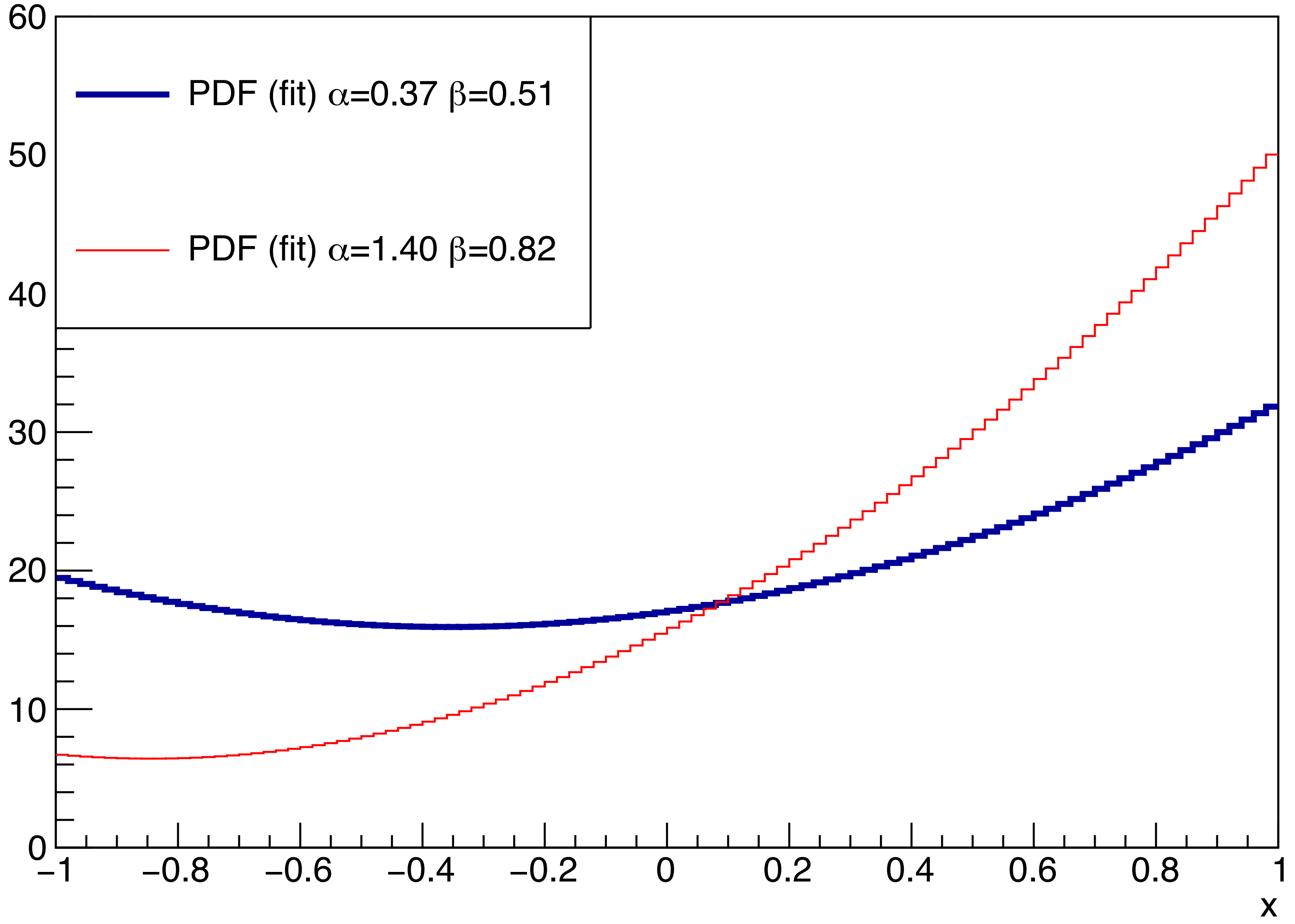
Results - Odd



- 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.

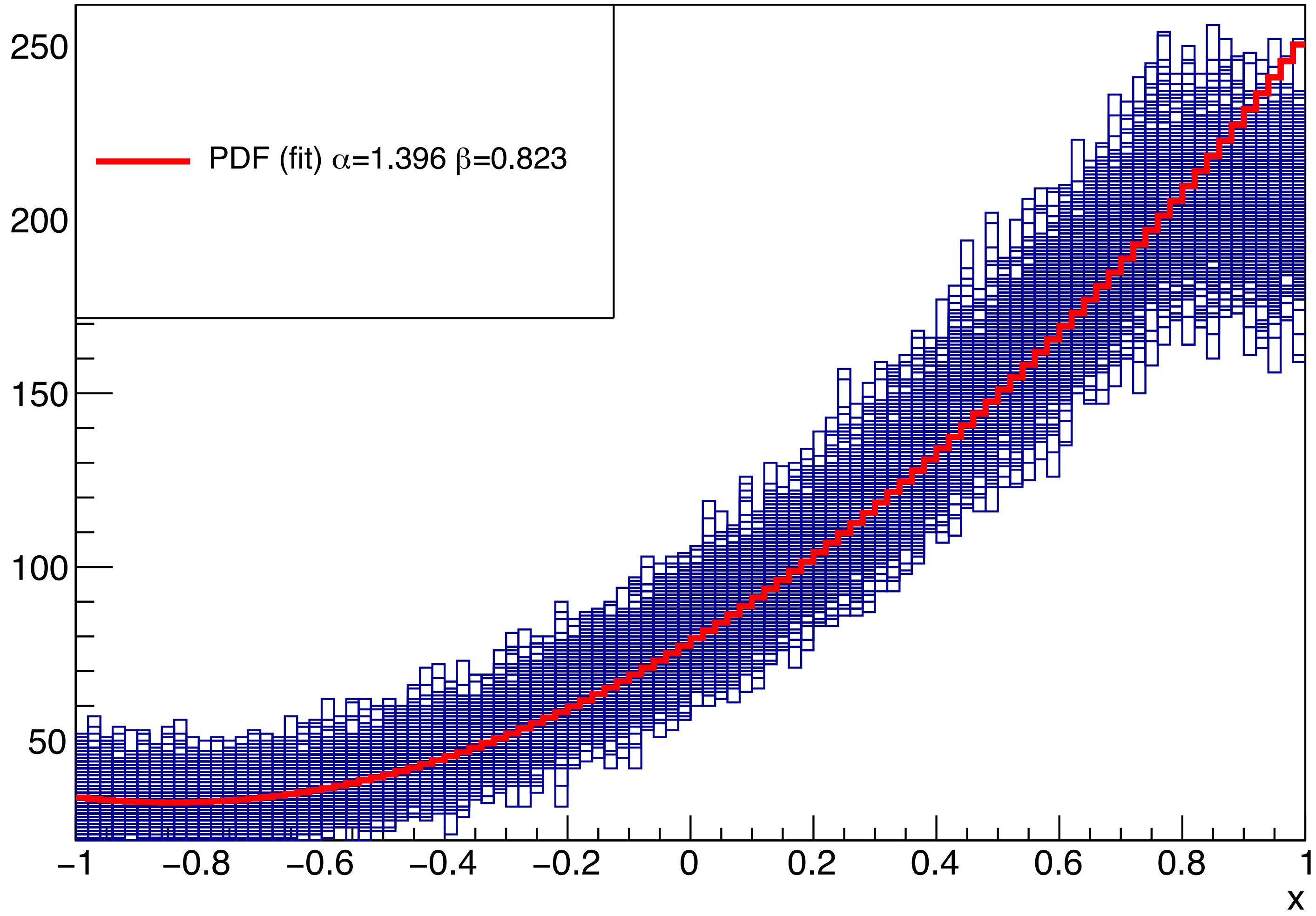
Debugging

- 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 points 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 plotted the x values for 800 pseudo-experiments, each w/ 10k points and also the underlying PDF
  - For **α**=1.396 and **β**=0.823 they didn't match at x values of 0.8-1.0



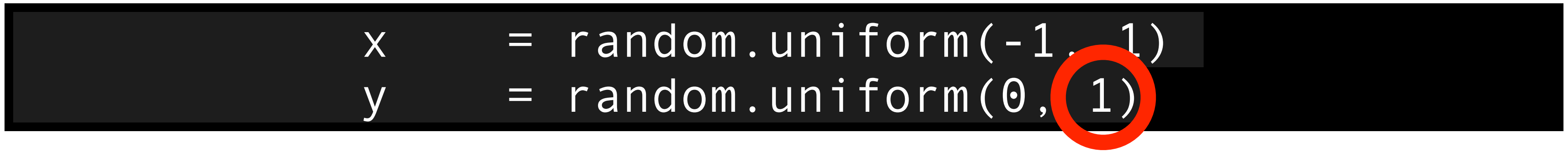
Only 1 of 800 pseudo-experiments had an upward fluctuation for x from 0.98 to 1.0. But, I expect ~1/2 of the pseudo-experiments to have an upward fluctuation.

Solution

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

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

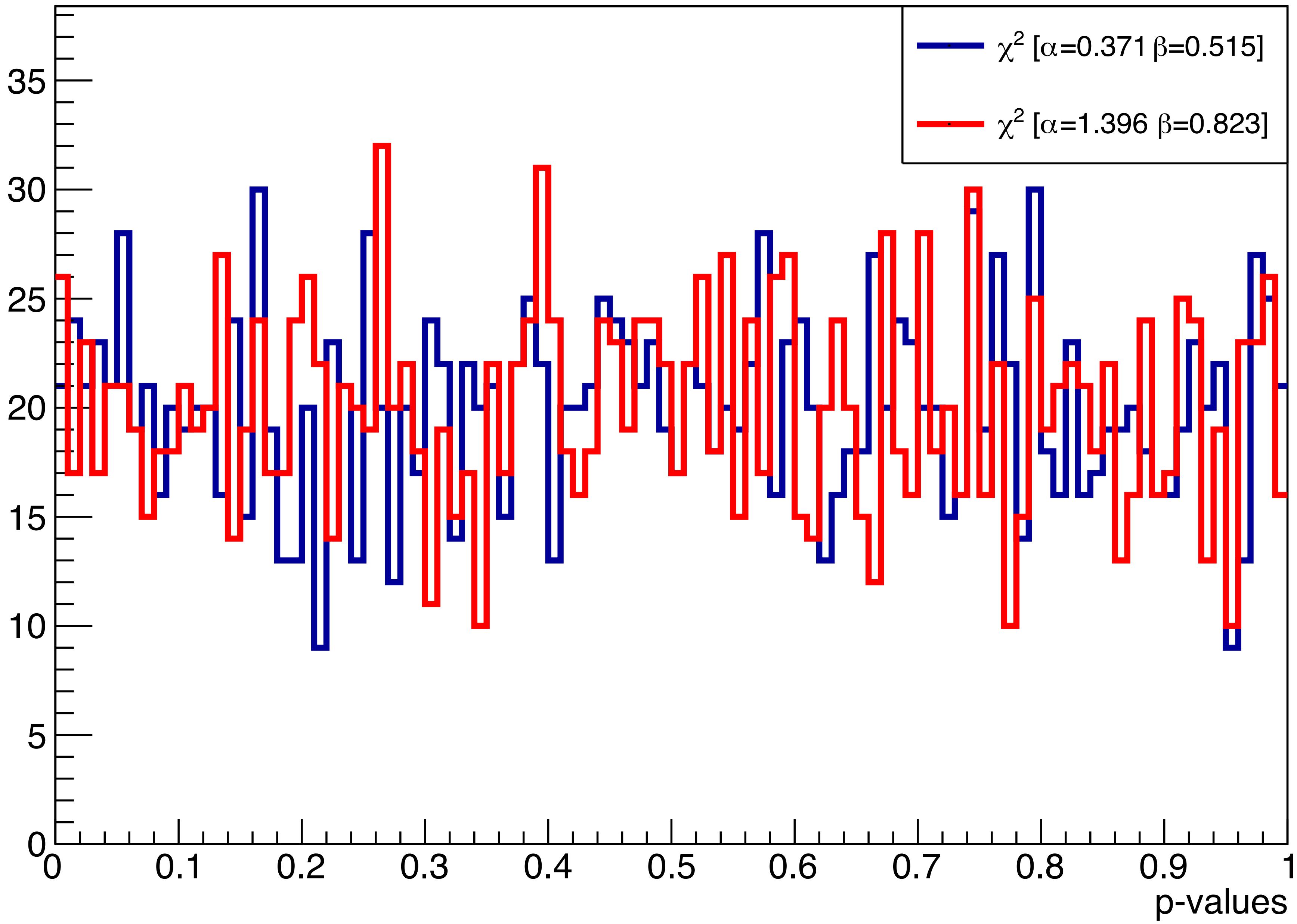
- What’s so special about x≈0.8?
    - Well, f(x=0.8; **α**=1.396, **β**=0.823)=1.039
    - The distribution is normalized to 1, but the instantaneous ‘probability’ goes above 1 in the range of ~0.8-1.
    - My accept/reject method of sampling the PDF goes from -1 to 1 in x, but only 0 to 1 in ‘y’.



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

Fixed

- Changing the bounds on my accept/reject sampling fixes 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.



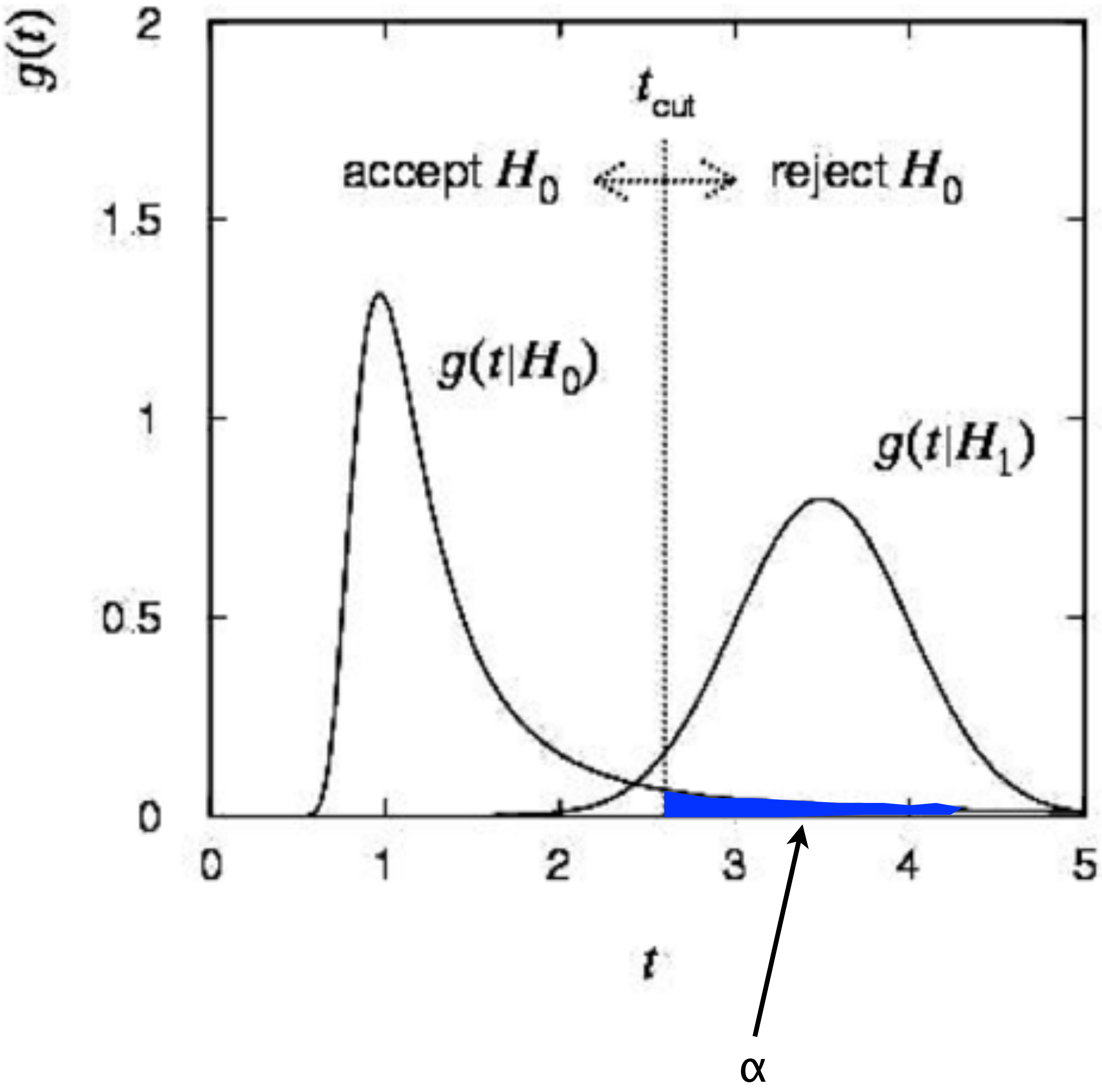
Statistical Tests - Decision Boundary

- 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

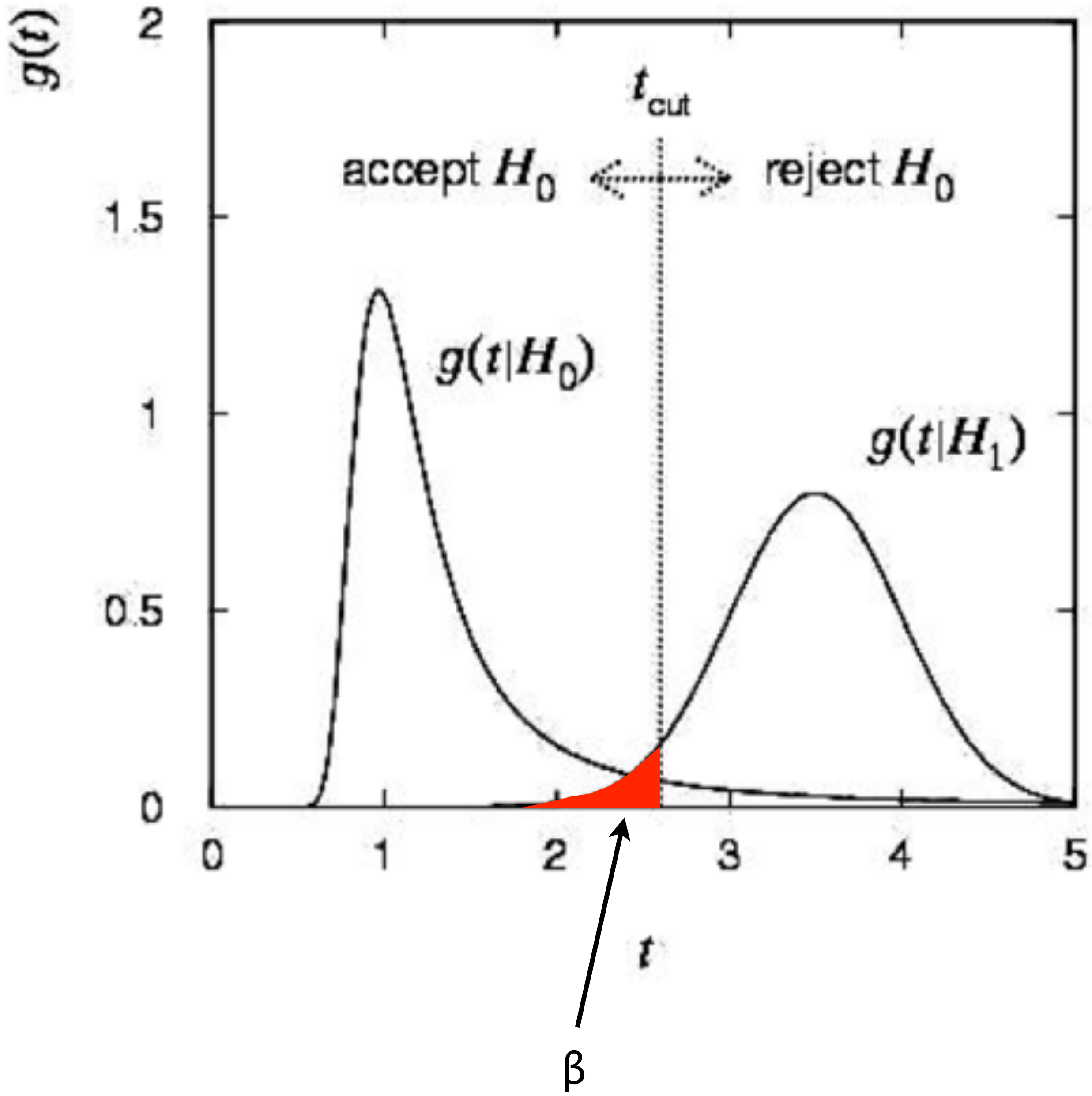
- Consider now the alternate hypothesis.

- Define the error of the second kind as β as a probability to accept the null hypothesis but the true hypothesis was not the null but 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) = max. Aim for α and β small as possible.



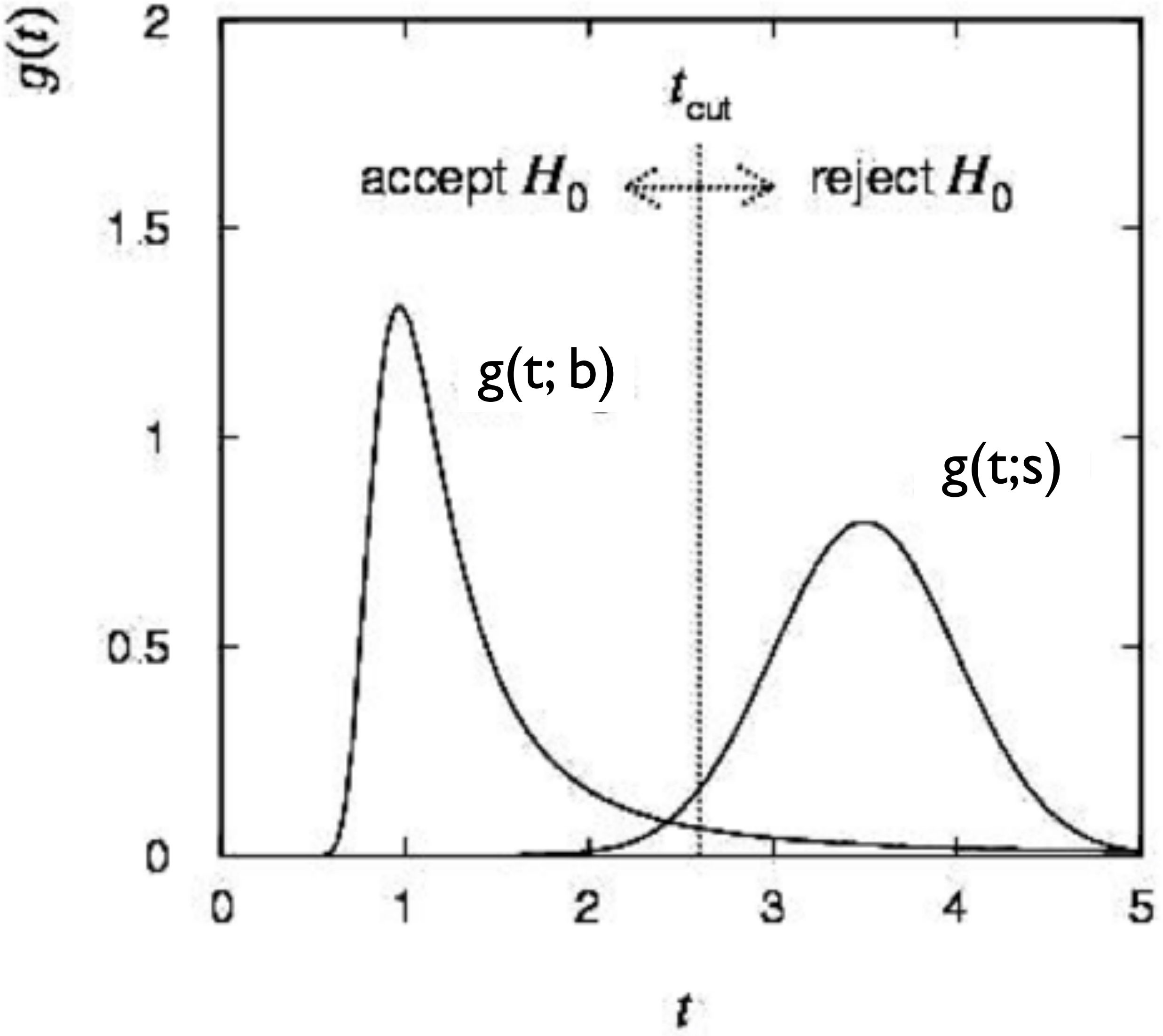
Statistical Tests - Signal & Background

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

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

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

Statistical Tests - Likelihood Ratio

- Likelihood Ratio Test
  - A test based on the Neyman-Pearson acceptance region for the vector of test statistics, t, is a test using an one dimensional statistic given by the ratio:

$$r = \frac{g(t; H_0)}{g(t; H_1)}$$

  - This is known as the likelihood ratio with an acceptance region r > k.   The best test statistic, in terms of maximum power, for a given significance level is given by the Likelihood Ratio.
  - If r is approximately one, then it is likely the null hypothesis is true and if approximately 0 then it is unlikely the hypothesis is true.
  - For a large sample, one can use the asymptotic behavior for likelihood ratios such that if the null hypothesis imposes n constraints then -2ln(r) is distributed as a chi-square with n degrees of freedom.
  - The PDFs are often determined by Monte Carlo simulation or calibration data (independent samples).

Maximum Likelihood Ratio

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

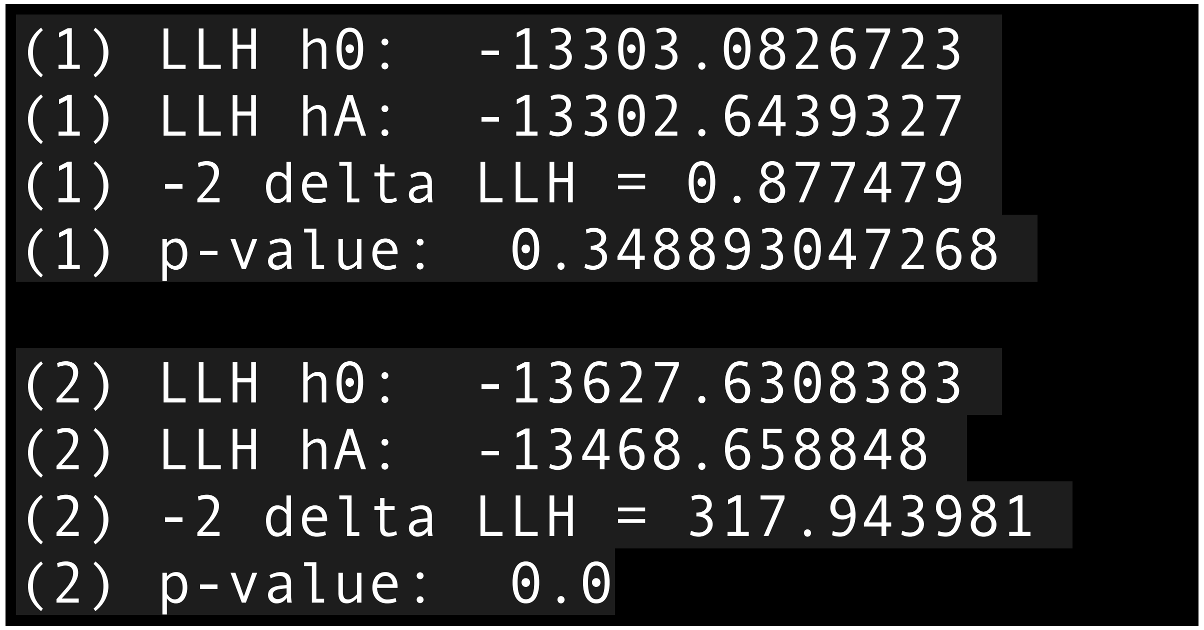
$$\Lambda(\theta, x_{obs}) = -2\ln\frac{\mathcal{L}(\theta_0|x_{obs})}{\mathcal{L}(\hat{\theta}|x_{obs})}$$

  - 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 one (or more) of the θ parameters 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 ratio follows the χ^2 distribution,
  - N_{DOF} = difference in dimensionality between the models
  - Also requires that Wilk's Theorem is satisfied (more later)

Exercise #1

- From the file posted on the class webpage for “Class 8”, use the $\ln$-likelihood ratio and calculate the p-value of each data set using 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 $f(x;\alpha,\beta,\gamma) = 1 + \alpha x + \beta x^2 - \gamma x^5$
  - These hypotheses satisfy Wilk’s Theorem



```
(1) LLH h0:  -13303.0826723
(1) LLH hA:  -13302.6439327
(1) -2 delta LLH = 0.877479
(1) p-value:  0.348893047268

(2) LLH h0:  -13627.6308383
(2) LLH hA:  -13468.658848
(2) -2 delta LLH = 317.943981
(2) p-value:  0.0
```

Variance of Estimators - Graphical Method

- Used for 1 or 2 parameters when the ML estimate and variance cannot be found analytically. Expand lnL about its maximum via a Taylor series:

$$\ln L(\theta) = lnL(\hat{\theta}) + (\frac{\partial \ln L}{\partial \theta})_{\theta=\hat{\theta}}(\theta - \hat{\theta}) + \frac{1}{2!}(\frac{\partial^2 \ln L}{\partial \theta^2})_{\theta=\hat{\theta}}(\theta - \hat{\theta})^2 + ...$$

- First term is $\ln L_{max}$, 2nd term is zero, third term is used for information inequality.

From last Lecture

osition where L is largest. Sometimes there is more than one peak — take the highest.

- Uncertainty deduced from positions where lnL is reduced by an amount 1/2. For a Gaussian Likelihood function:

$$\boxed{\ln L(\theta) = \ln L_{max} - \frac{(\theta - \hat{\theta})^2}{2\hat{\sigma}^2_{\hat{\theta}}}}$$

$$\ln L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) = \ln L_{max} - \frac{1}{2} \quad \text{or} \quad \ln L(\hat{\theta} \pm N\hat{\sigma}_{\hat{\theta}}) = \ln L_{max} - \frac{N^2}{2}$$

For N standard deviations

Wilk’s Theorem… Kinda

- The expression we derived earlier can be rewritten as a ratio of likelihood and $\ln$-likelihoods that are χ^2 distributed:

$$\ln L(\theta) = \ln L_{max} - \frac{(\theta - \hat{\theta})^2}{2\hat{\sigma}^2_{\hat{\theta}}} \quad \Longrightarrow \quad \Lambda(\theta, x_{obs}) = -2\ln\frac{\mathcal{L}(\theta_0|x_{obs})}{\mathcal{L}(\hat{\theta}|x_{obs})}$$

- 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
- Find journal article for the oral presentation
- Nice link about quickly interpreting distributions of p-values
    - http://varianceexplained.org/statistics/interpreting-pvalue-histogram/