Notes : P-Values

D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2025

Photo by Howard Jackman

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

Exercise #1

- It has been referenced before in this course that the interpretation of the reduced chi-squared changes based on the effective 'degrees of freedom'. For example, a reduced $\chi^2 = 1.2$ with 19 degrees of freedom will result in a reasonable p-value, wheres the same reduced $\chi^2 = 1.2$ with 19000 degrees of freedom will result in a bad p-value. We are going to illustrate that with two data sets on the course webpage

Exercise #1 (continued)

- In the two files (PVals1_1.txt and PVals1_2.txt) there are 24 data points, and 240 data points; respectively. Each data set has 2 columns (x, y) where x is the independent variable and y is the dependent variable y=f(x).
- For both datasets there is a constant expectation that f(x)=1.14, and the uncertainty on that expectation is $\sigma = 0.004$.
  - Calculate the χ^2 for each data set
  - Calculate the reduced χ^2 for each data set
  - Calculate the p-value for each data set

- Because there is no fitting involved to get the expectation, the degrees-of-freedom are equal to the number of data points, i.e. DoF = 24 and DoF=240.

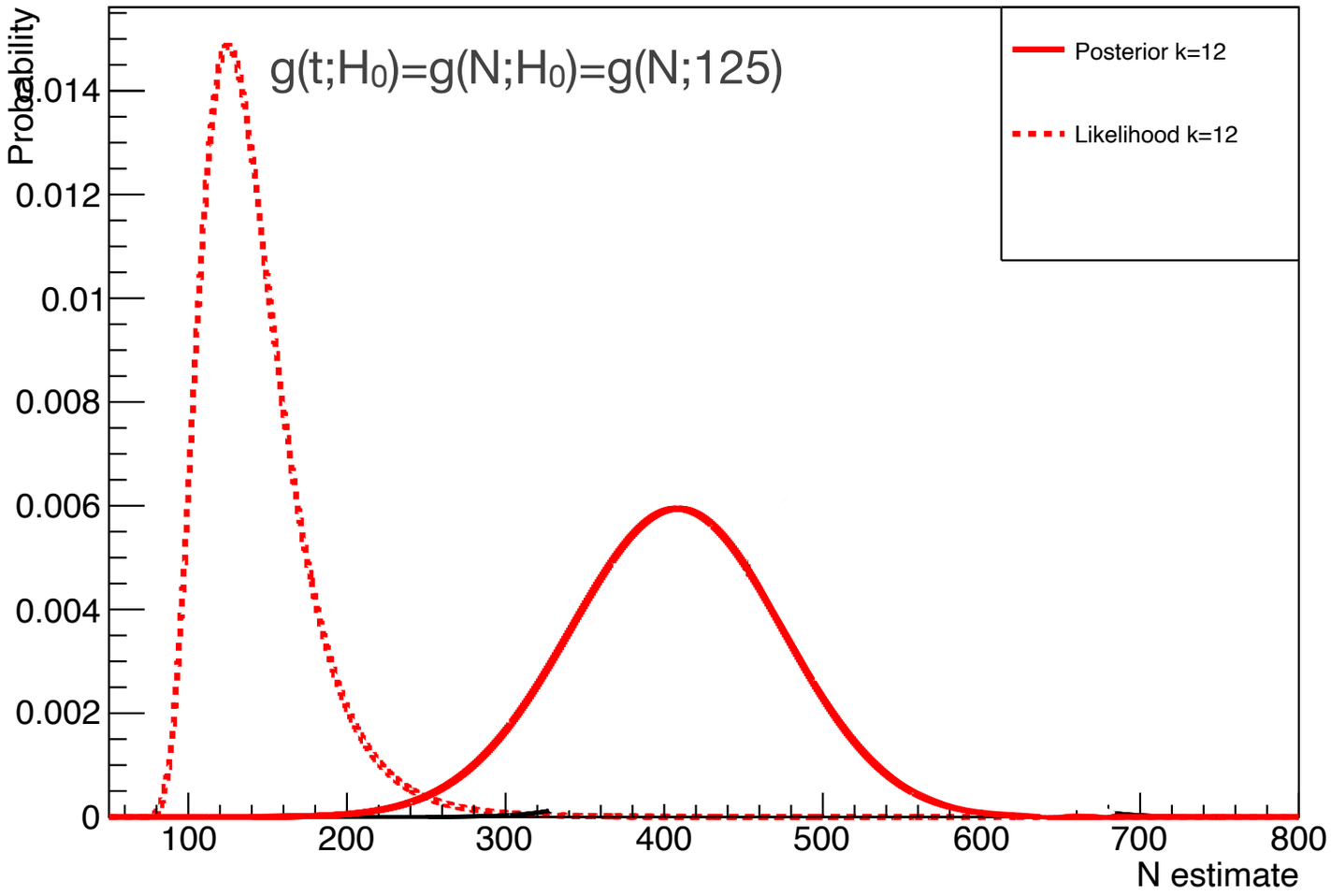
One-sided Test-Statistics

- Using a χ^2 or a likelihood as test-statistics is nice because they are always one-sided, e.g. a lower value of the likelihood is always worse than a higher likelihood.
- They are also nice because they are always 1-dimensional values, and likelihood & χ^2 always return single values;
  - regardless of how many data points are in the data sample,
  - regardless of how many dimensions/variables comprise the data, e.g. the data can be in only x-space (x), or 3-dimensional space (x, y, z), or N-dimension space (x, y, z,…,Ω)
  - regardless of whether the PDF has lots of parameters, e.g. $f(x\,|\,\alpha) \propto 1 + \alpha x$ — versus— $f(x\,|\,\alpha, \beta, \gamma\,\rho) \propto 1 + \alpha x + \beta x^3 - \gamma x^6 + \rho x^7$
- These two traits are helpful when calculating p-values because it makes it easier to calculate the probability of "…observing a test statistic as compatible to, or less compatible with, the observed data"

P-Value in Action

- For this example we consider N to be the test statistic (t=N), the *maximum a posteriori* (MAP) value of 409 is the alternate hypothesis (H_1), and value of 125 is the null hypothesis (H_0). We want to know "Assuming the null hypothesis, what is the chance that we would get a value of 409, *or greater than 409*?"
- We assume H_0 to be true, and $g(t;H_0)$ 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

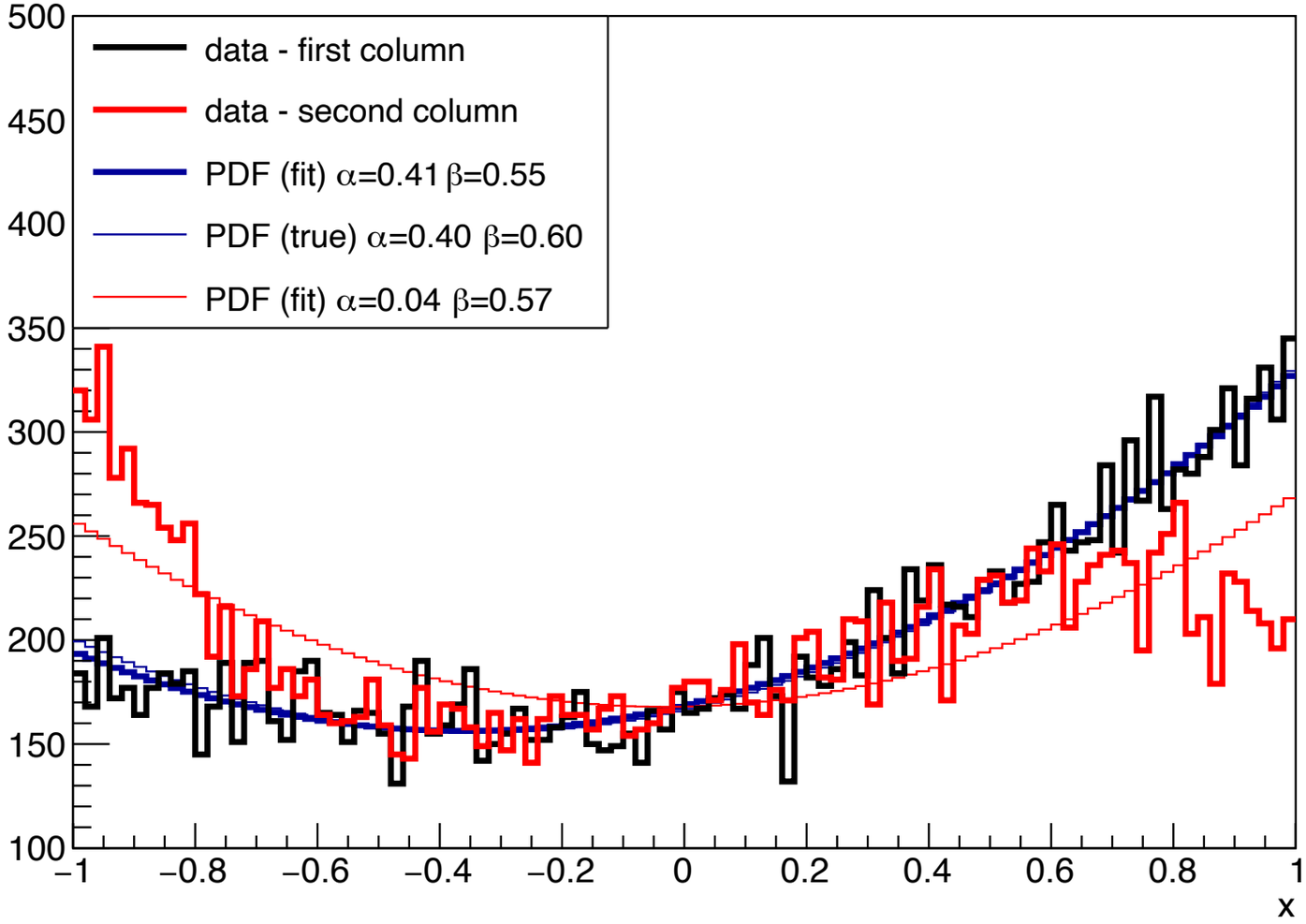
- There is a file posted on the class webpage from "Parameter Estimation and Confidence Intervals" lecture which has two columns of x numbers (not x and y, only x for 2 pseudo-experiments) corresponding to x over the range $-1 \leq x \leq 1$
- Using the function:

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

- Find the best-fit for the unknown α and β for each data set
- Find the uncertainties σ_α and σ_β for each data set
- Plot the 2D contours for the 50%, 90%, and 95% confidence intervals
- [Optional] Using a chi-squared test statistic, calculate the 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 it was an optional exercise to calculate the goodness-of-fit. The p-value from a chi-squared distribution was the suggestion.
  - Visually, the previous plot of the x data from the first column looks to agree with the PDF using the 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):
(120.8030913720248, 0.05120506553561213 9)
data 2 (chi-square, p-value):
(384.85801188036919, 6.3385429186073 07e-36)
```

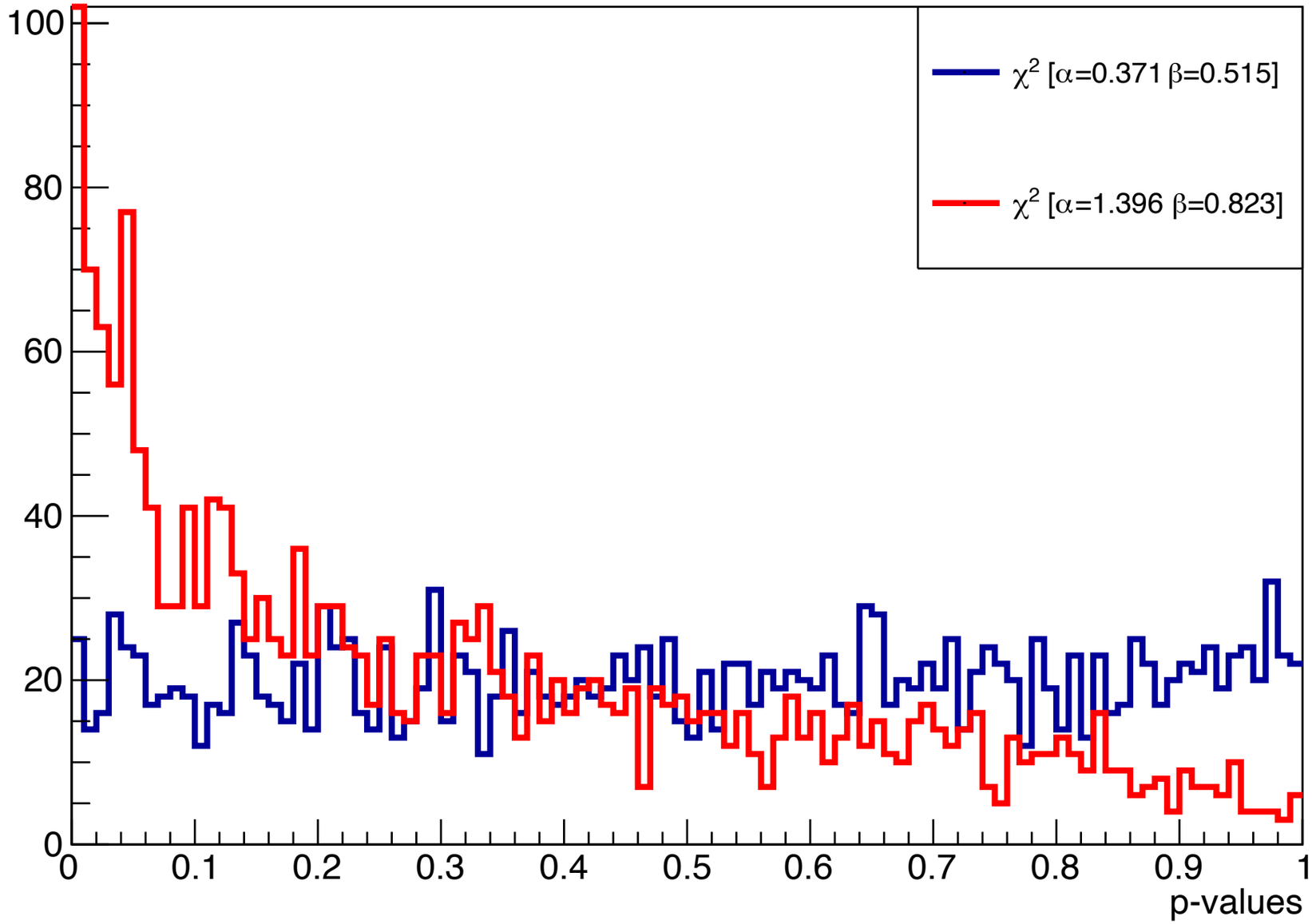
*from a binned histogram

Funny Thing

- A previous 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 $(1+\alpha x+\beta x^2)/(2+\beta/3)$

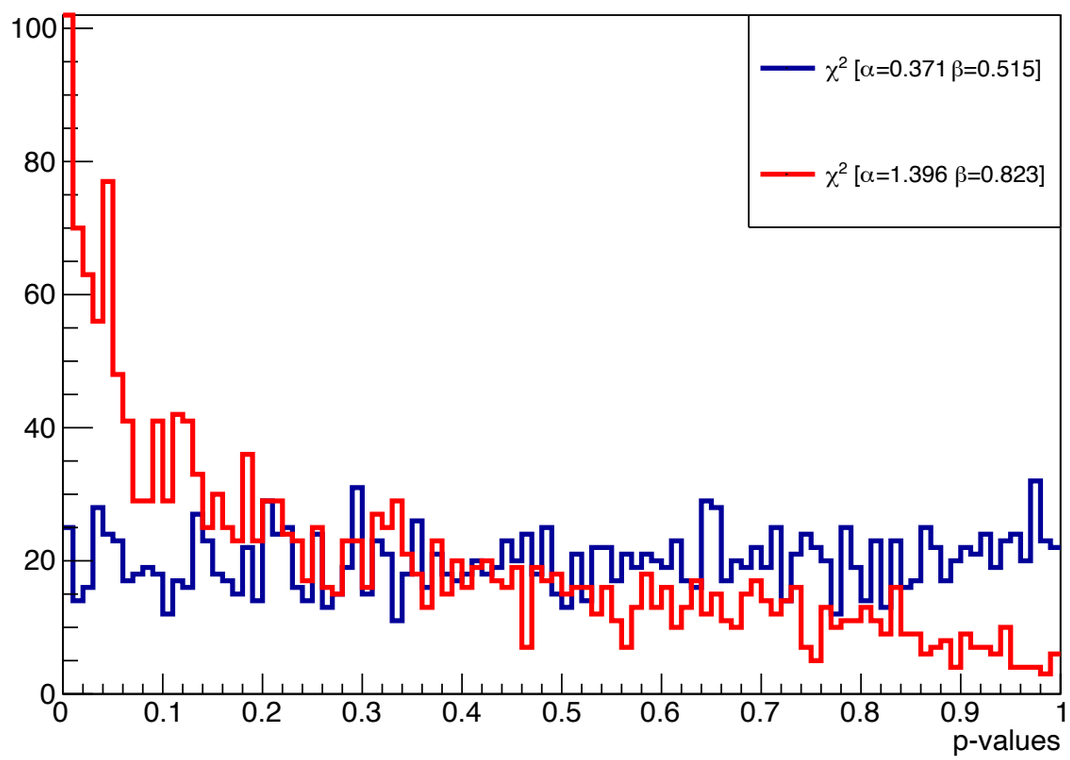


*Different file than what is posted online

Results - Discussion

- The blue line was where the hypothesis (underlying PDF) matched the pseudo-data for many trials. When I created statistically independent datasets using the same underlying PDF, then calculated the p-value using the χ^2, I get a p-value distribution for the 800 pseudo-experiments that is flat. So blue looks good.

- When the hypothesis **does not match** the data, then I get biased values to lower p-values. This is ALSO a good thing, because if the hypothesis for the underlying PDF does not match the data, then I **should** get a low p-value.



Solution

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

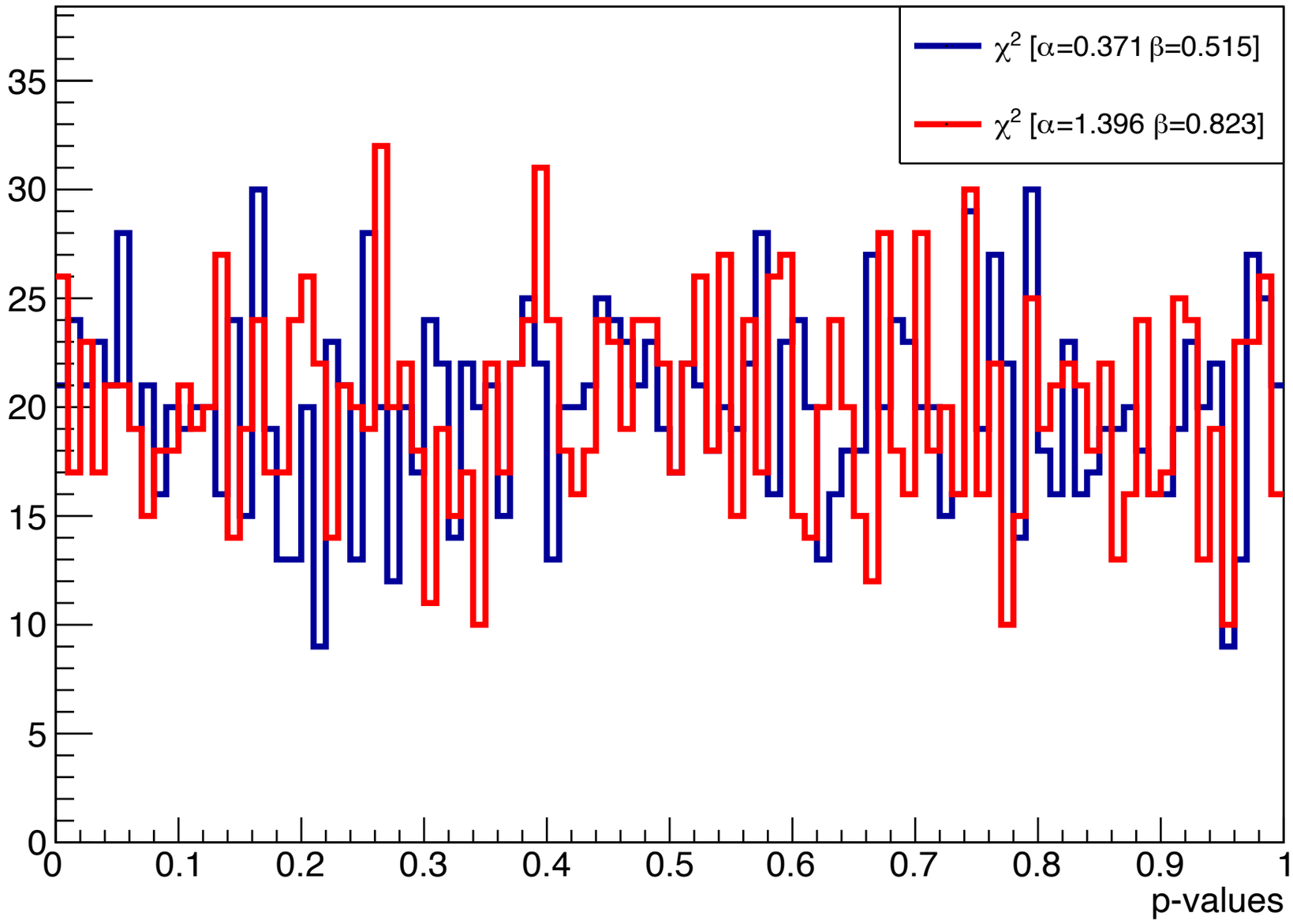
$$f(x;\alpha,\beta) = \frac{1+\alpha x+\beta x^2}{2+2\beta/3}$$

- What's so special about x≈0.8?
  - Well, $f(x=0.8\,|\,\alpha=1.396,\ \beta=0.823)=1.039$. The distribution is normalized to 1, but the instantaneous probability density goes above 1 in the 'x' 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.

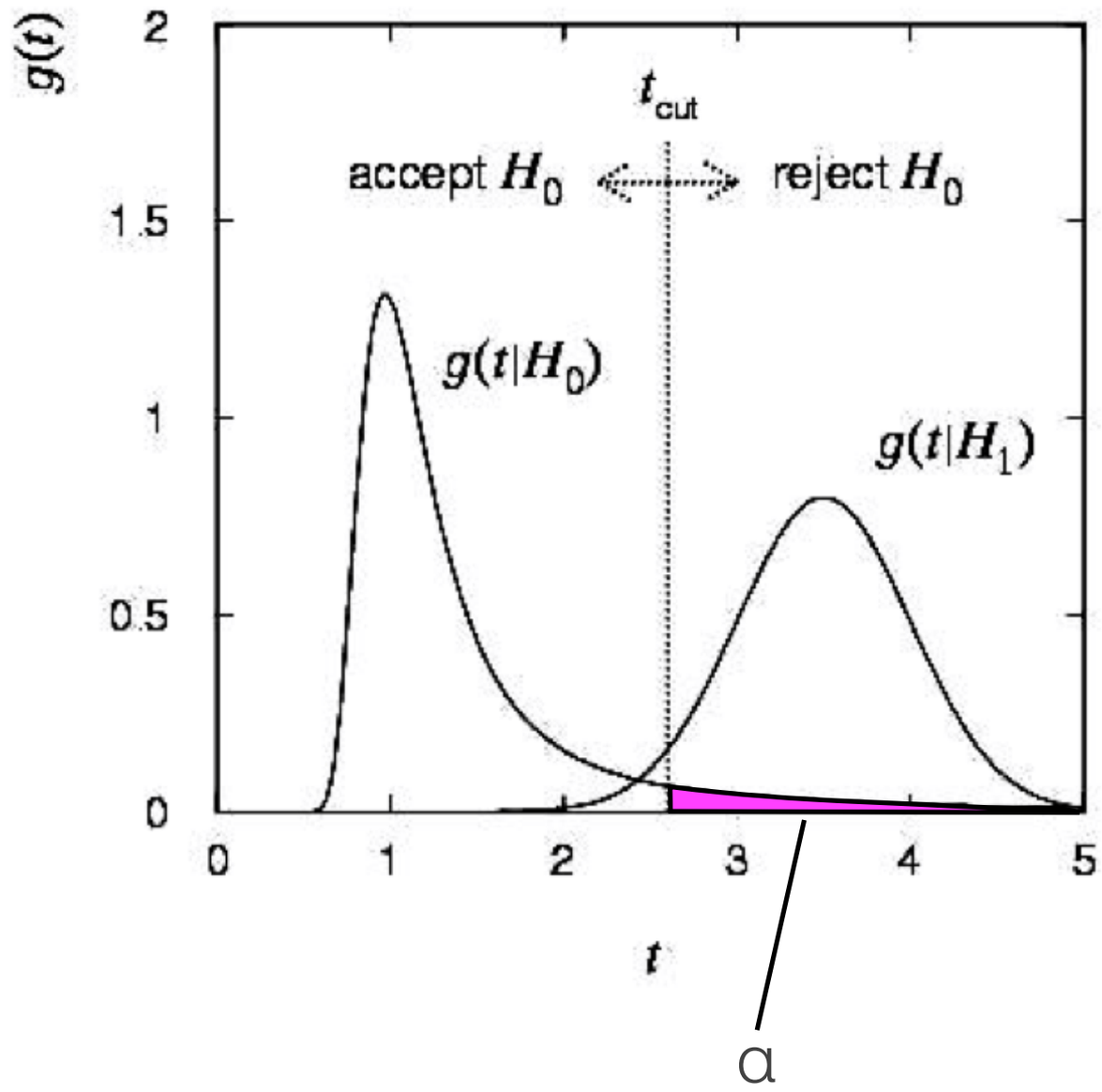


Statistical Tests - Decision Boundary

- 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



Maximum Likelihood Ratio

- A 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 variable, i.e. gaussian, 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)

Maximum Likelihood Ratio

- The test-statistic is the natural-log of the ratio

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

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

$$= -2 * \Delta LLH$$

- The test-statistic is **NOT** the ratio of the natural logs

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