Lecture 5: Parameter Estimation and Uncertainty

D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2025

Photo by Howard Jackman

Oral Presentation and Report

- Now would be a good to time to make sure you have:
  - Selected a topic
  - Selected a paper
  - Done some work on preparing the presentation and/or report

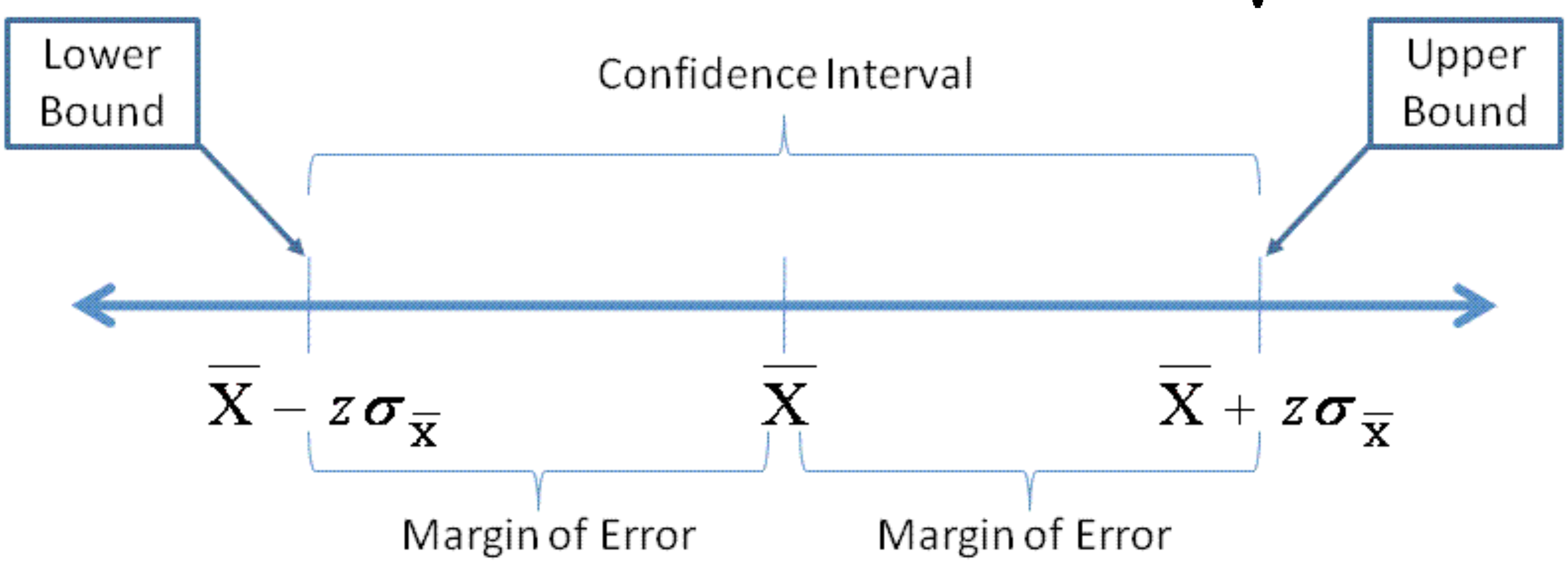
Confidence intervals

"Confidence intervals consist of a range of values (interval) that act as good estimates of the unknown population parameter."

It is thus a way of giving a range where the true parameter value probably is.

A very simple confidence interval for a Gaussian distribution can be constructed as:
(z denotes the number of sigmas wanted)

$$\bar{x} \pm z \frac{s}{\sqrt{n}}$$



Confidence intervals

Confidence intervals are constructed with a certain **confidence level C**, which is roughly speaking the fraction of times (for many experiments) to have the true parameter fall inside the interval:

$$Prob(x_- \leq x \leq x_+) = \int_{x_-}^{x_+} P(x)dx = C$$

Often, C is in terms of σ or percent 50%, 90%, 95%, and 99%

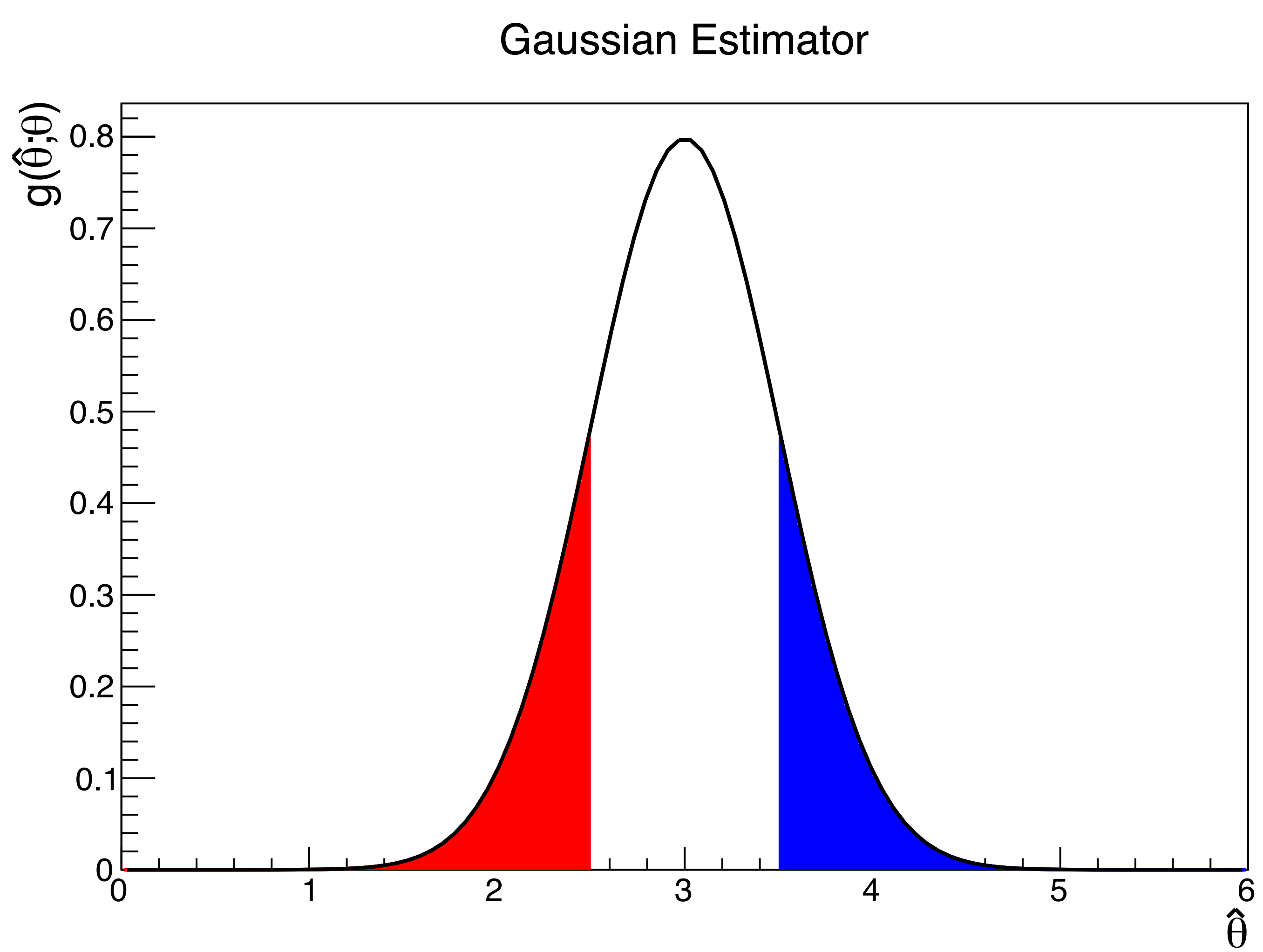
<u>There is a choice as follows:</u>

1. Require symmetric interval (x+ and x- are equidistant from μ).
2. Require the shortest interval (x+ to x- is a minimum).
3. Require a central interval (integral from x- to μ is the same as from μ to x+).

For the Gaussian, the three are equivalent!
Otherwise, 3) is usually used.

Confidence Intervals

- Confidence intervals are often denoted as C.L. or "Confidence Limits/Levels"
- Central limits are different than upper/lower limits
- We can establish uncertainties on our extracted best-fit parameters using likelihoods



Variance of Estimators - Gaussian Estimators

- Used for 1 or 2 parameters when the maximum likelihood estimate and variance cannot be found analytically. Expand $\ln$L 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 can used for information inequality (not covered here)

  - For **1** parameter:

    - Minimize, or scan, as a function of θ to get $\hat{\theta}$

    - Uncertainty deduced from positions where $LLH(\theta)$ is different from $LLH_{max}(\theta)$ by 0.5. For a Gaussian likelihood function w/ **1** fit parameter:

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

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

For N standard deviations

Variance of Estimators - Gaussian Estimators

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

For more information, see "Variance of ML Estimators" sections from "Statistical Data Analysis" (https://www.sherrytowers.com/cowan_statistical_data_analysis.pdf)

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

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

ln(Likelihood) and 2*LLH

- A change of 1 standard deviation (σ) in the maximum likelihood estimator (MLE) of the parameter θ leads to a change in the ln(likelihood) value of 0.5 for a **gaussian distributed estimator**
  - Even for a non-gaussian MLE, the 1σ region[a] defined as LLH-1/2 can be an *okay* approximation
  - Because the regions[a] defined with ΔLLH=1/2 are consistent with common χ^2 distributions multiplied by 1/2, we often calculate the likelihoods as (-)2*LLH
- Translates to >1 fit parameters too, with the appropriate change in 2*LLH confidence values
  - 1 fit parameter, Δ(2LLH)=1 for 68.3% C.L.
  - 2 fit parameter, Δ(2LLH)=2.3 for 68.3% C.L.

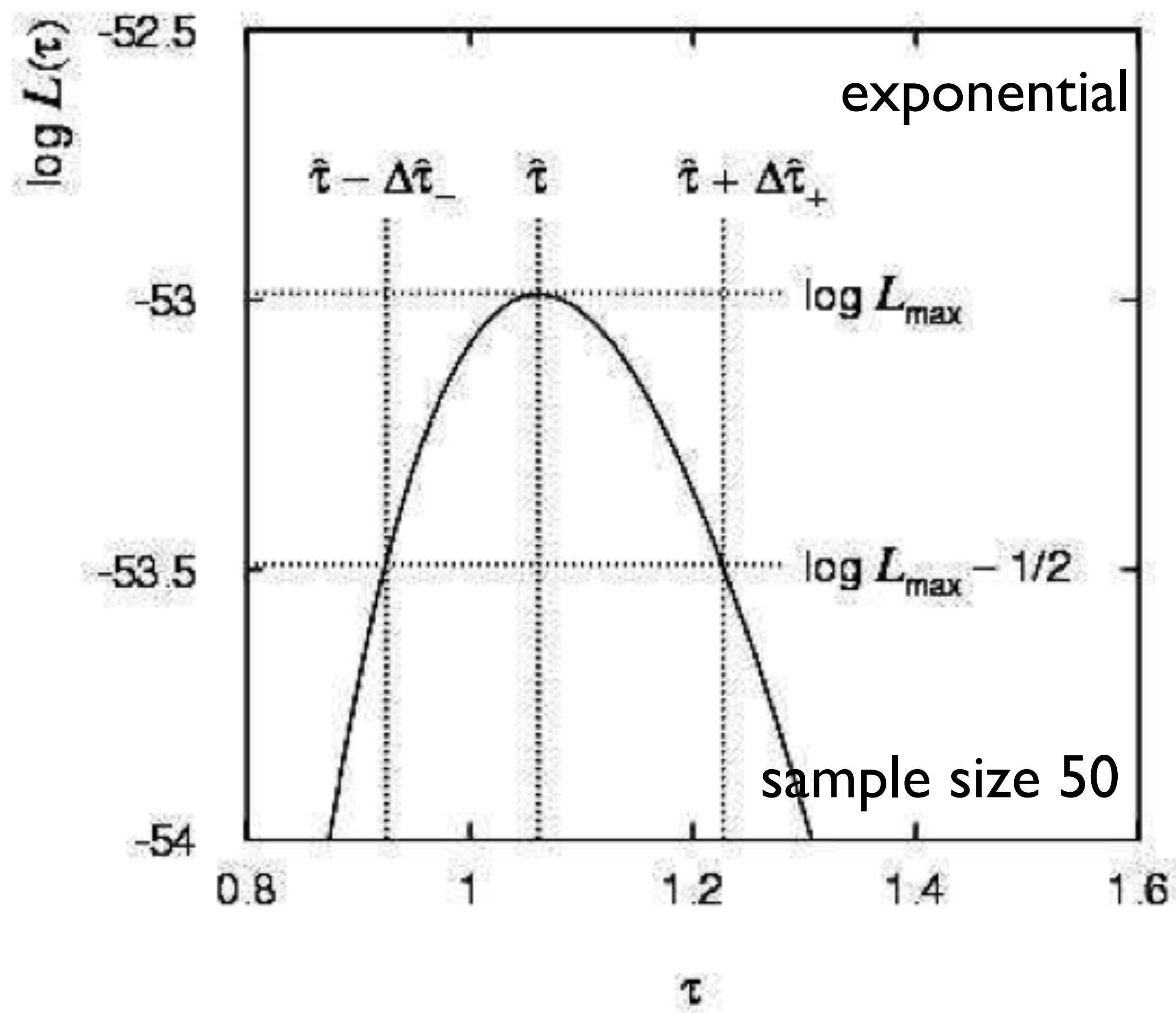
[a]for a distribution w/ 1 fit parameter

Variance of Estimator

Likelihood is from Lecture 3 and is

$$f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$$

- First, we find the best-fit estimate of τ via our LLH minimization to get $\hat{\tau}_{best}$
  - Provides LLH($\hat{\tau}_{best}$)=-53.0
  - To get $\hat{\tau}_{best}$, we can use a minimizer/maximizer fitting algorithm
- We only have 1 fit parameter, so from slide 7 we know that values of $\hat{\tau}$ which cross LLH($\hat{\tau}_{best}$)-0.5 are the 1σ ranges, i.e. when the LLH equals -53.5



$$\hat{\tau} = 1.062$$

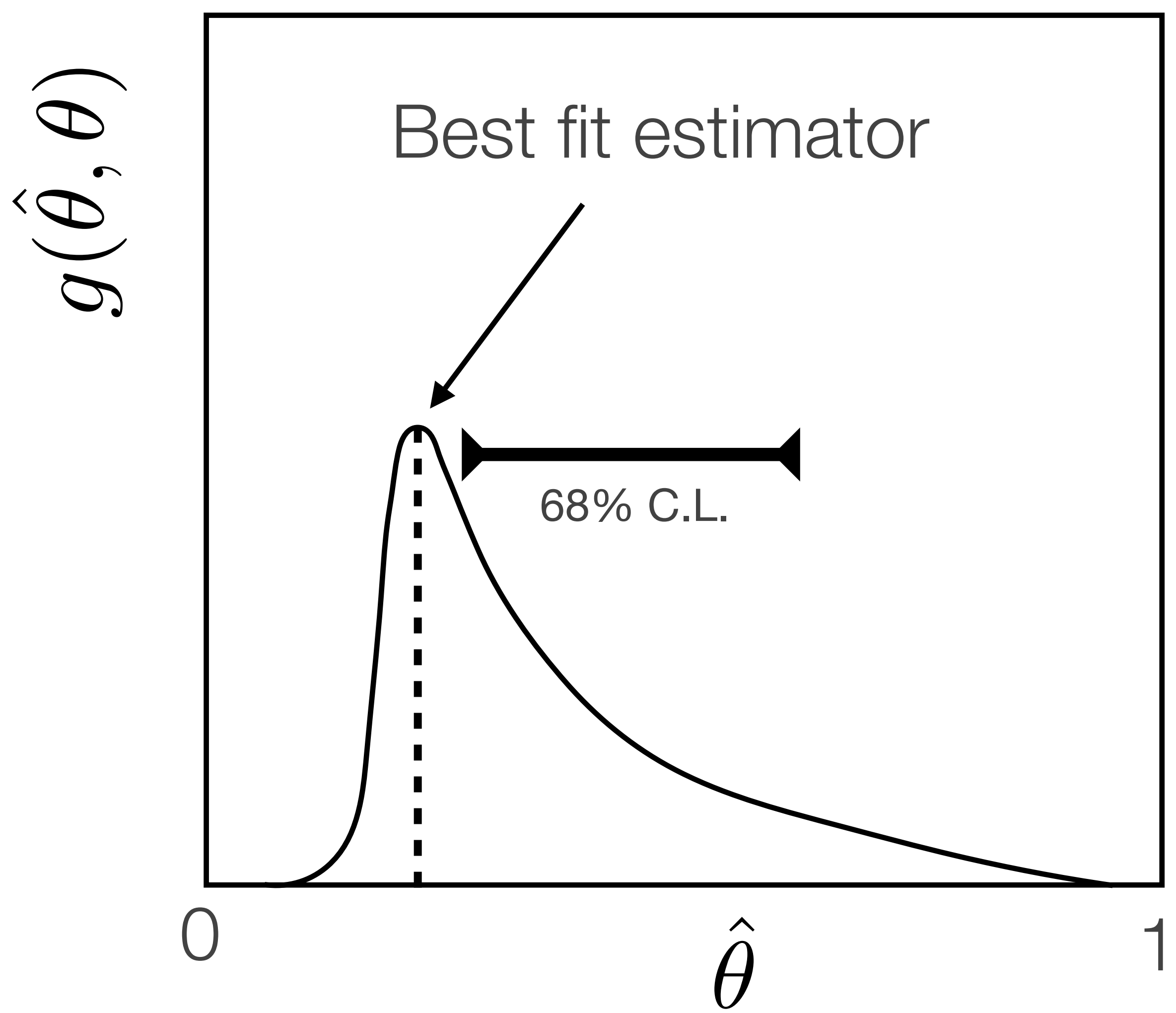
$$\Delta\hat{\tau}_{-} = 0.137$$

$$\Delta\hat{\tau}_{+} = 0.165$$

$$\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_{-} \approx \Delta\hat{\tau}_{+} \approx 0.15$$

Reporting Very Asymmetric Central Limits

- Central limits are often reported as $\hat{\theta} \pm \sigma_\theta$ or $\hat{\theta}^{+\sigma_1}_{-\sigma_2}$ if the error bars are asymmetric
- What happens when upper or lower range away from the best-fit value(s) does not have the right coverage? E.g. for 68% coverage, the lower 17% of the distribution includes the best fit point.
    - Quote the best-fit estimator of θ and the limit ranges separately. “Best fit is θ=0.21 and the 90% central confidence region is 0.17-0.77”



Exercise #1

- Similar to the exercises 2-3 from Lecture 3, we will use the theoretical function:

$$f(x : \beta) = 1 + 0.65x + \beta x^2$$

- For data that has an unknown β, we want to get the best-fit value of $\hat{\beta}$ from the data as well as the 1σ uncertainty $\sigma_{\hat{\beta}}$.
    - There are 3013 data points in a file for Exercise 1 on the course webpage. The data points come from the above function transformed into a PDF over the range -0.95 ≤ x ≤ 0.95.
    - Remember to **normalize** the function properly to convert it to a proper PDF

Exercise #1 - LLH Approach

- Normalize the function for all possible values of β
  - This looks a lot like a previously used function $f(x\,|\,\alpha,\beta) = 1 + \alpha x + \beta x^2$, but with now with an unchanging value $\alpha = 0.65$
- Have a function that can calculate the negative natural log-likelihood (LLH) using the data set with 3013 entries
- Have a minimizer get the best-fit value of β, i.e. $\hat{\beta}$
  - Be able to also get the negative LLH value at the best-fit, i.e. LLH($\hat{\beta}$)
- Since this is a 1-parameter fit, the 1σ uncertainty is then the value(s) of β where | (-LLH($\hat{\beta}$)) - (-LLH(β)) |=0.5.
  - There should be two values of β— i.e. $\beta_{+\sigma}$ and $\beta_{-\sigma}$—that satisfy the above because -LLH($\hat{\beta}$) is the minimum of the LLH landscape
  - If the two values $\beta_{+\sigma}$ and $\beta_{-\sigma}$ are equidistant from $\hat{\beta}$, then the uncertainty is $\pm\sigma_\beta = |\hat{\beta} - \beta_{\pm\sigma}|$
  - If the two values $\beta_{+\sigma}$ and $\beta_{-\sigma}$ are not equidistant from $\hat{\beta}$, then the uncertainty is $+\sigma_\beta = |\hat{\beta} - \beta_{+\sigma}|$ and $-\sigma_\beta = |\hat{\beta} - \beta_{-\sigma}|$

Likelihoods for Uncertainties

- Using the log-likelihood difference (ΔLLH) between the best-fit point to construct uncertainty regions is fast
- Requires some features that are not always satisfied
  - Properly known as 'Wilks theorem'
  - Expects that estimator distributions are gaussian, e.g. repeat measurements of β will be (mostly) gaussian
  - In Lecture 7 next week, we will cover in more detail the foundation of why the ΔLLH can be used to construct intervals

- As a cross-check, or in situations where the Wilk's theorem is violated (either knowingly or unknowingly), there is an extremely robust way to calculate uncertainties… parametric bootstrapping

Robust

- After finding the best-fit values via $\ln$(likelihood) maximization/minimization from data, one of **THE** best and most robust calculations for the *parameter uncertainties* is to run numerous pseudo-experiments using the best-fit values for the Monte Carlo 'true' values and find out the spread in pseudo-experiment best-fit values
  - MLEs don't have to be gaussian. Thus, a Monte Carlo based uncertainty is accurate even if the Central Limit Theorem is invalid for your data/parameters
  - The routine of 'Monte Carlo plus fitting' will take care of many parameter correlations
  - The problem is that it can be slow and gets exponentially slower with each dimension for multi-dimensional scenarios

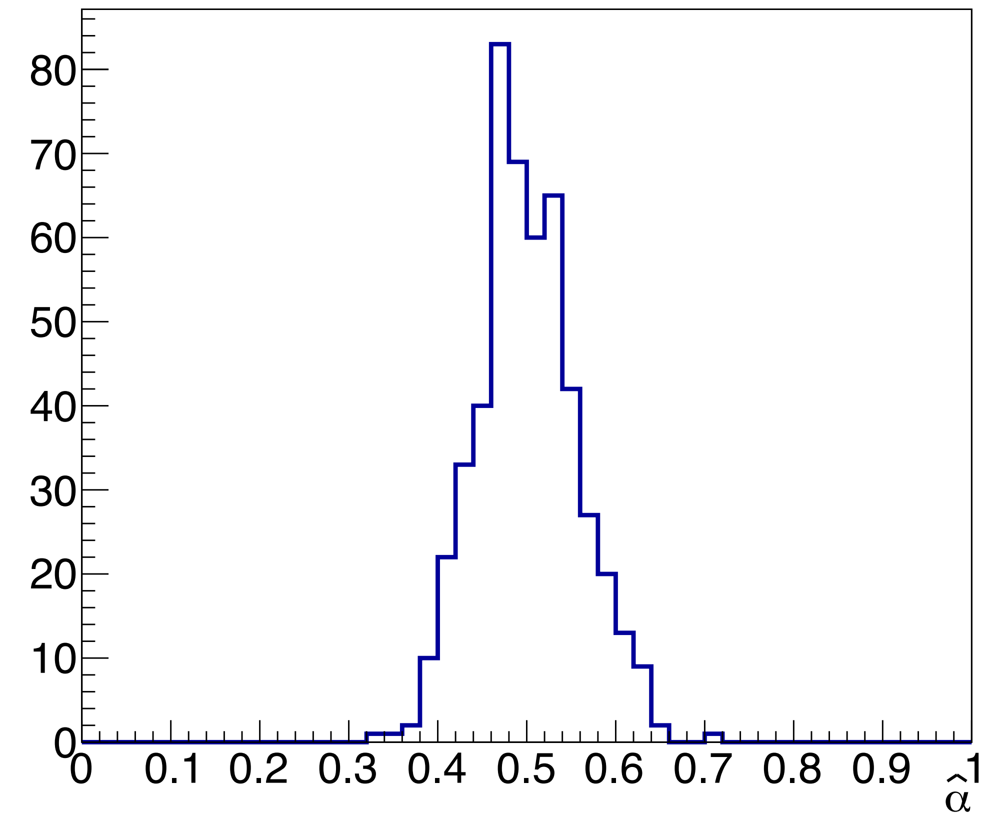
Brute Force

- If we either did not know, or did not trust, that our estimator(s) are a nicely analytic PDF (gaussian) we can use pseudo-experiments to establish the uncertainty on our best-fit values
  - Sample new pseudo-experiment data from the PDF with injected values of $\hat{\alpha}_{obs}$ and $\hat{\beta}_{obs}$ that were found as the best-fit values
  - Fit each pseudo-experiment
  - Repeat
  - Integrate ensuing estimator PDF

To get ±1σ central interval

$$\frac{100\% - 68.27\%}{2} = \int_{-\infty}^{C_-} g(\hat{\alpha}; \hat{\alpha}_{obs}) d\hat{\alpha}$$

$$\frac{100\% - 68.27\%}{2} = \int_{C_+}^{\infty} g(\hat{\alpha}; \hat{\alpha}_{obs}) d\hat{\alpha}$$

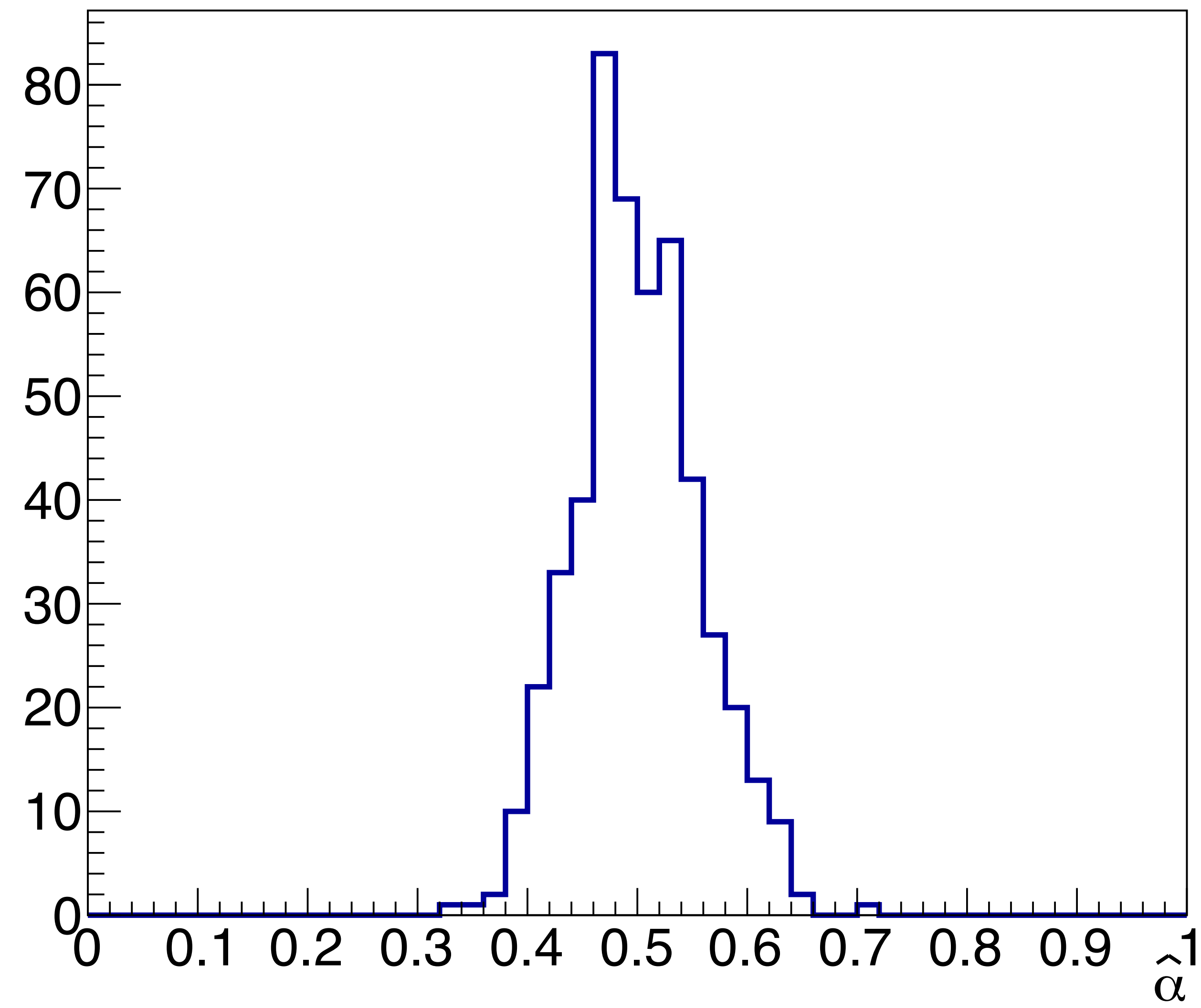


Brute Force

- For the Monte Carlo brute force method, i.e. "parametric bootstrapping", the lower value for the confidence interval is set at C_{-} and the upper value for the confidence interval is set at C_{+}, and we are calculating for a 1σ C.L., i.e. 68.27%

$$\frac{100\% - 68.27\%}{2} = \int_{-\infty}^{C_-} g(\hat{\alpha};\hat{\alpha}_{obs})d\hat{\alpha}$$

$$\frac{100\% - 68.27\%}{2} = \int_{C_+}^{\infty} g(\hat{\alpha};\hat{\alpha}_{obs})d\hat{\alpha}$$



Brute Force cont.

- This method is known as a **parametric bootstrap**
  - Overkill for the previous example
  - Useful for estimators which are complicated
  - Useful for when you want to ***ensure*** your uncertainties and confidence intervals are accurate

- Finding the uncertainty using the integration of the tails works for bayesian posteriors in same way as for likelihoods

Exercise #2a

- We will use the theoretical prediction:

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

- For data that has unknown values of α and β we want to get an idea of the best-fit values of $\hat{\alpha}$ and $\hat{\beta}$ from the data as well as the uncertainties.
    - Remember to **normalize** the function properly to convert it to a proper PDF
    - Same data set as in Exercise 1

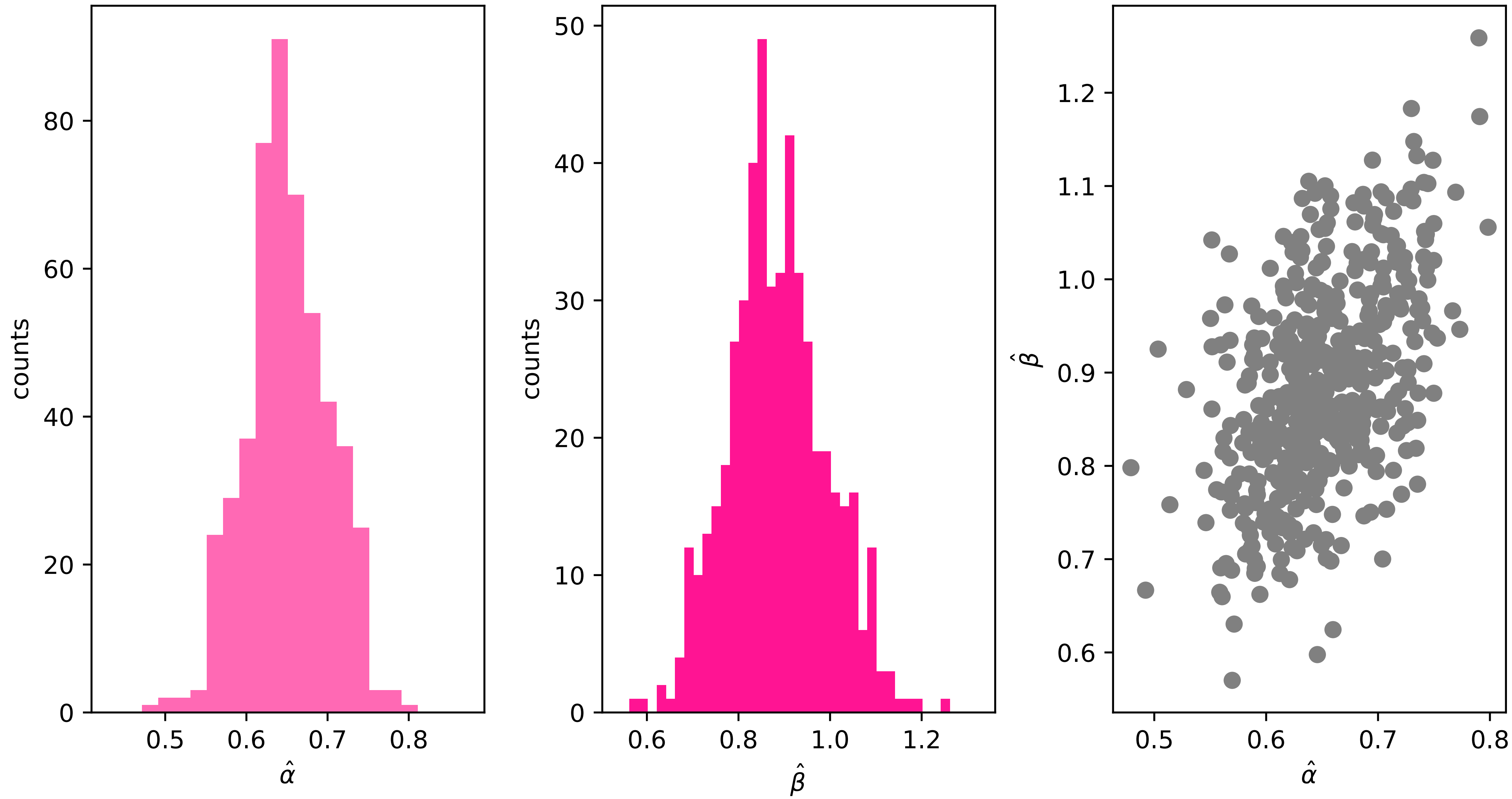
- Fit the maximum likelihood estimate (MLE) parameters $\hat{\alpha}_{data}$ and $\hat{\beta}_{data}$ from the data files using a minimizer/maximizer

Exercise #2b

- To get an idea of what the distribution of $\hat{\alpha}_{data}$ and $\hat{\beta}_{data}$ look like we will generate a "N" of pseudo-trials data sets, fit $\hat{\alpha}_{pseudo-trial,i}$ and $\hat{\beta}_{pseudo-trial,i}$ for each "i" independent and identically distributed pseudo-trial data set, and then plot the "N" outcomes
    - Each pseudo-trial has 3013 Monte Carlo data points
    - Generate N=500 pseudo-trials
    - Plot a 1D histogram of all $\hat{\alpha}_{pseudo-trial,i}$, a 1D histogram of all $\hat{\beta}_{pseudo-trial,i}$, and a 2D scatter-plot of $\hat{\beta}_{pseudo-trial,i}$ versus $\hat{\alpha}_{pseudo-trial,i}$
    - 'pseudo-trials' are also known as 'pseudo-experiments'

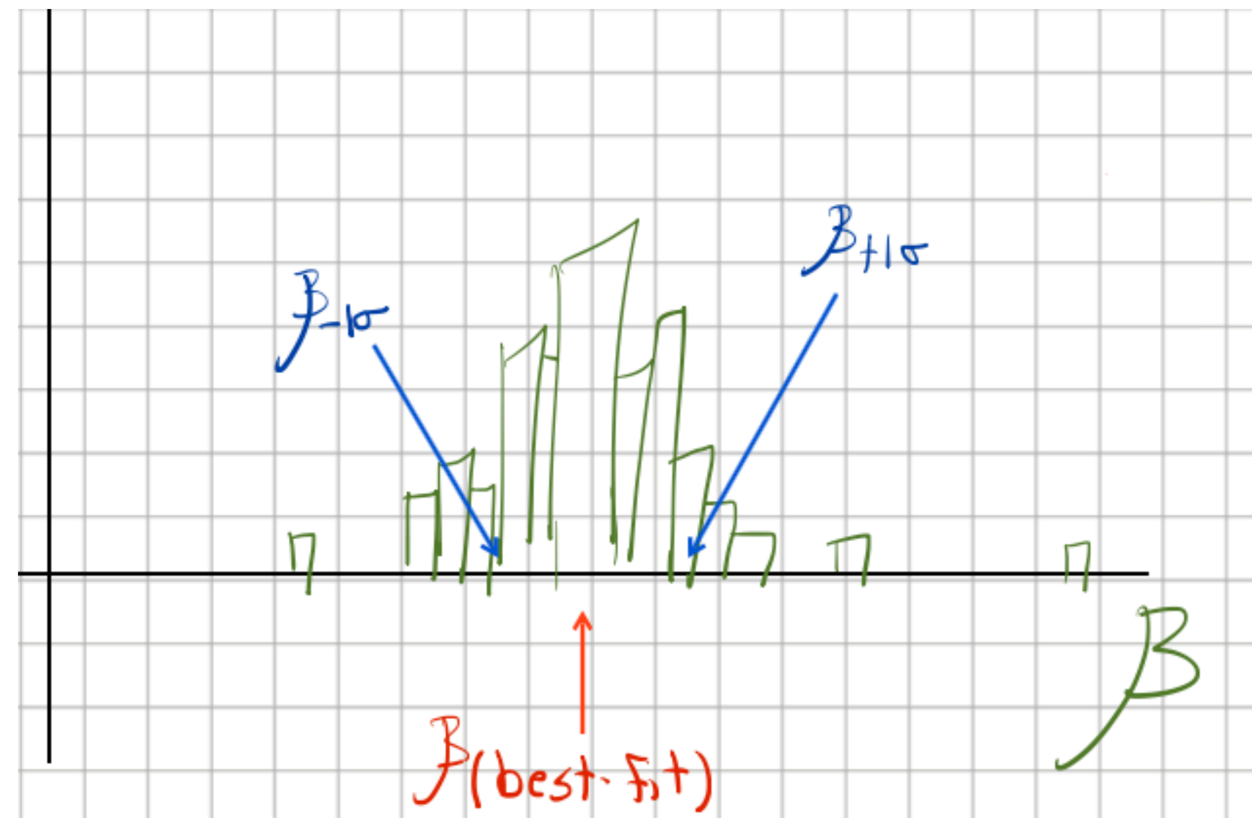
Exercise #2b (cont.)

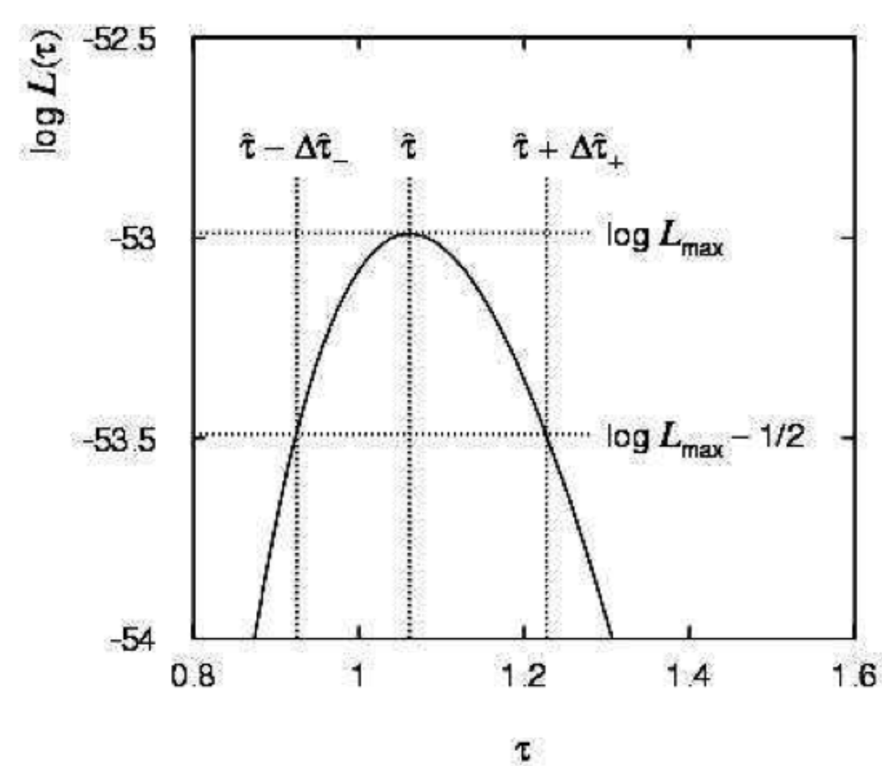
- Shown are 500 Monte Carlo pseudo-experiments
- Parametric bootstrapping can establish the confidence intervals even when the estimator distribution isn't gaussian distributed
  - DO NOT FIT THE PSEUDOEXPERIMENT DISTRIBUTION AS IF IT IS GAUSSIAN



Uncertainty from Bootstrapping vs. Likelihood

- The uncertainty estimate from bootstrapping: uses multiple Monte Carlo generated samples (using the best-fit from the original data sample) and the best-fit values of those MC samples to build a distribution. The 'width' of the ensuing fit values from the Monte Carlo constitutes the uncertainties.
- The uncertainty estimate from likelihood(s): get the best-fit of a parameter. Establish the value of the parameter where the LLH difference to the best-fit point is equal to the critical value for the number of fit parameters.
    - See critical values on later slides, or find chi-square tables online for a more complete list





Good?

- The LLH minimization will give the best-fit values and often the uncertainty on the estimators. But, likelihood fits do not tell whether the data and the prediction agree
  - Remember that the likelihood has a form (PDF) that is provided by **<u>you</u>** and may not be correct
  - The PDF may be okay, but there may be some measurement systematic uncertainty that is unknown or at least unaccounted for which creates disagreement between the data and the best-fit prediction
  - Likelihood *ratios* between two hypotheses are a good way to exclude models, and we'll cover hypothesis testing next week

Quick Note

- The term and definition of '1σ' can represent both the gaussian 1σ, and the 68.27% confidence interval.
  - For a Gaussian distribution the 1σ in the equation $\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ and 68.27% confidence intervals are the same thing
  - For non-gaussian distributions, 1σ is interpreted as <u>*only*</u> the 68.27% confidence interval or uncertainty, and does not imply that the original C.I. or uncertainty is gaussian.

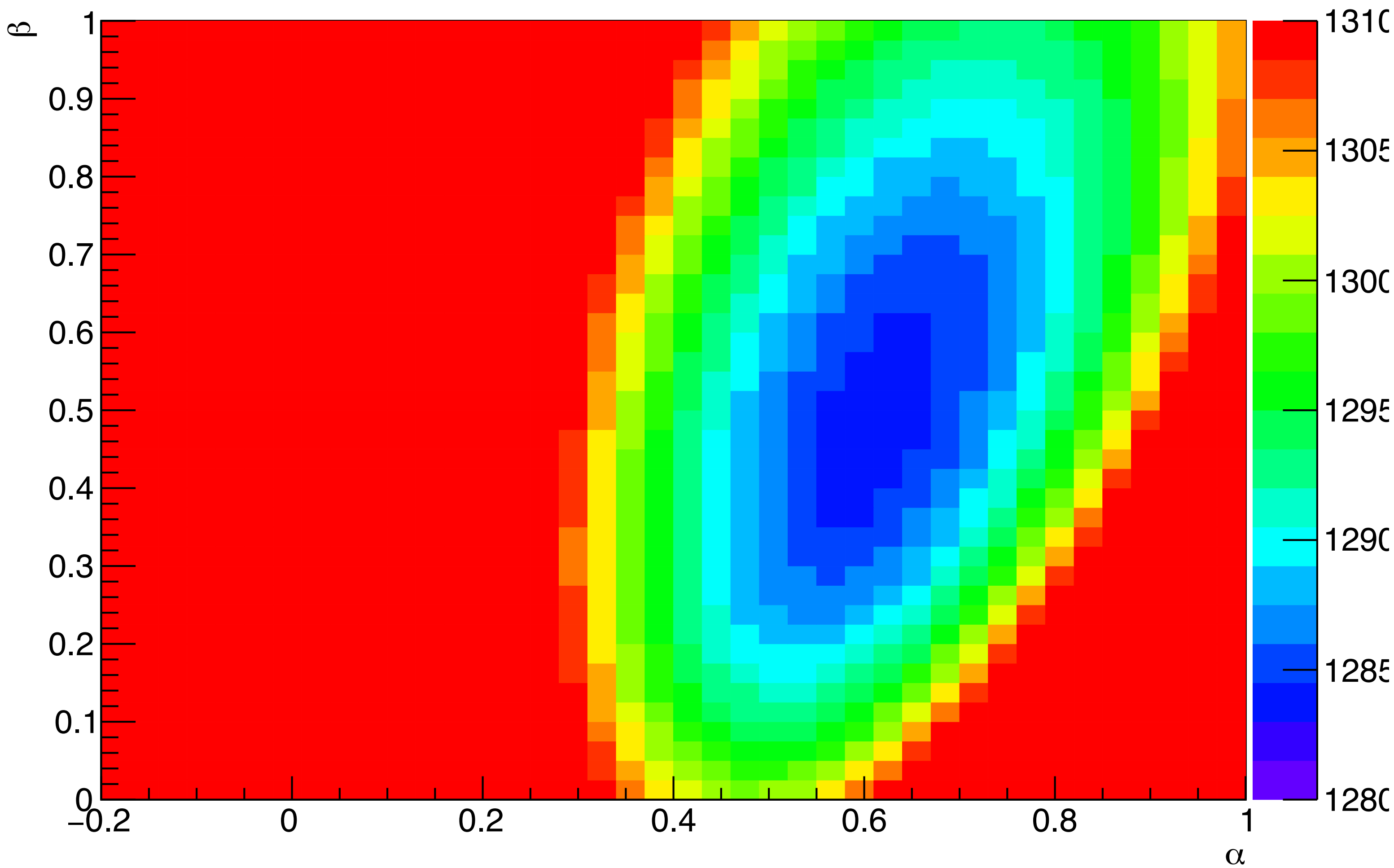
- For asymmetric uncertainties, the 68.27% confidence interval is still commonly used.
  - From the abstract of A gravitational-wave standard siren measurement of the Hubble constant : "We determine the Hubble constant to be $70^{+12.0}_{-8.0} km\ s^{-1}\ Mpc^{-1}$ (maximum a posteriori and 68% credible interval)."
  - From the abstract of A measurement of Hubble's Constant using Fast Radio Bursts "…our best-fitting value of H_0 is calculated to be $73^{+12}_{-8} km\ s^{-1}\ Mpc^{-1}$"

Multi-parameter

- Getting back to LLH confidence intervals
- In one dimension, they are fairly straightforward to use
  - Confidence intervals, i.e. uncertainty, can be deduced from the LLH difference(s) to the best-fit point
  - Brute force option is rarely a bad choice, and parametric bootstrapping is nice
- Both strategies work in multi-dimensions too
  - It is an excellent habit to produce 2D contours of $\hat{\theta}$ vs. $\hat{\phi}$
  - There are some common mistakes to avoid

Likelihood Contour/Surface

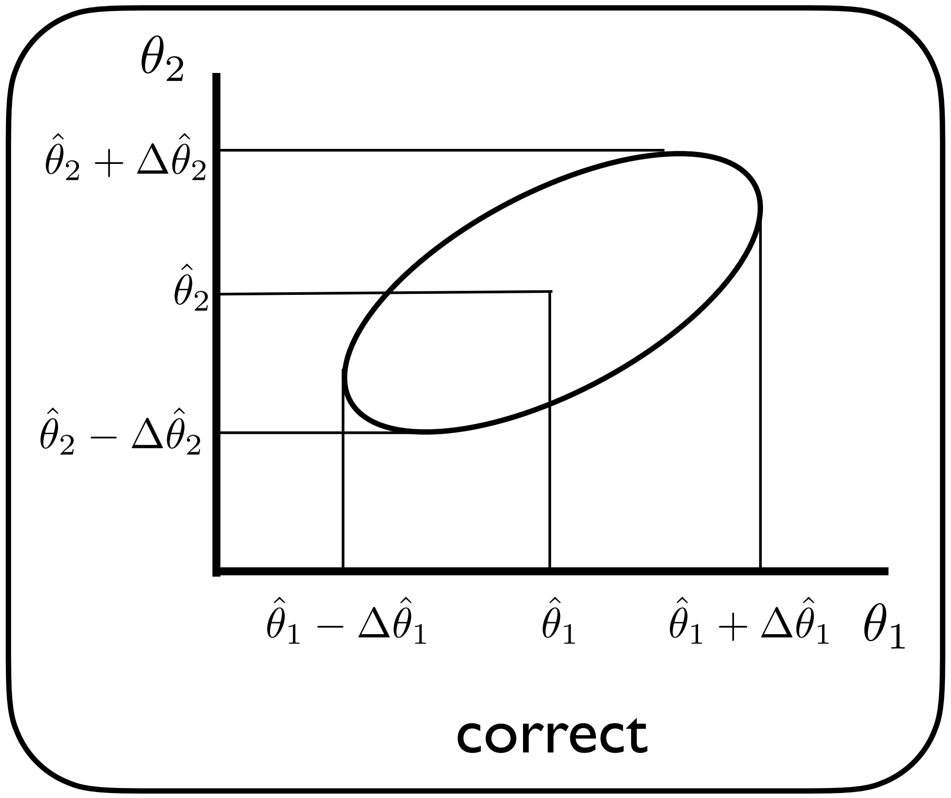
- For 2 dimensions, i.e. 2-parameter fits, we can produce likelihood landscapes. In 3 dimensions a surface, and in 3+ dimensions a likelihood hypersurface.
- The contours are then lines of with a constant value of likelihood or ln(likelihood)

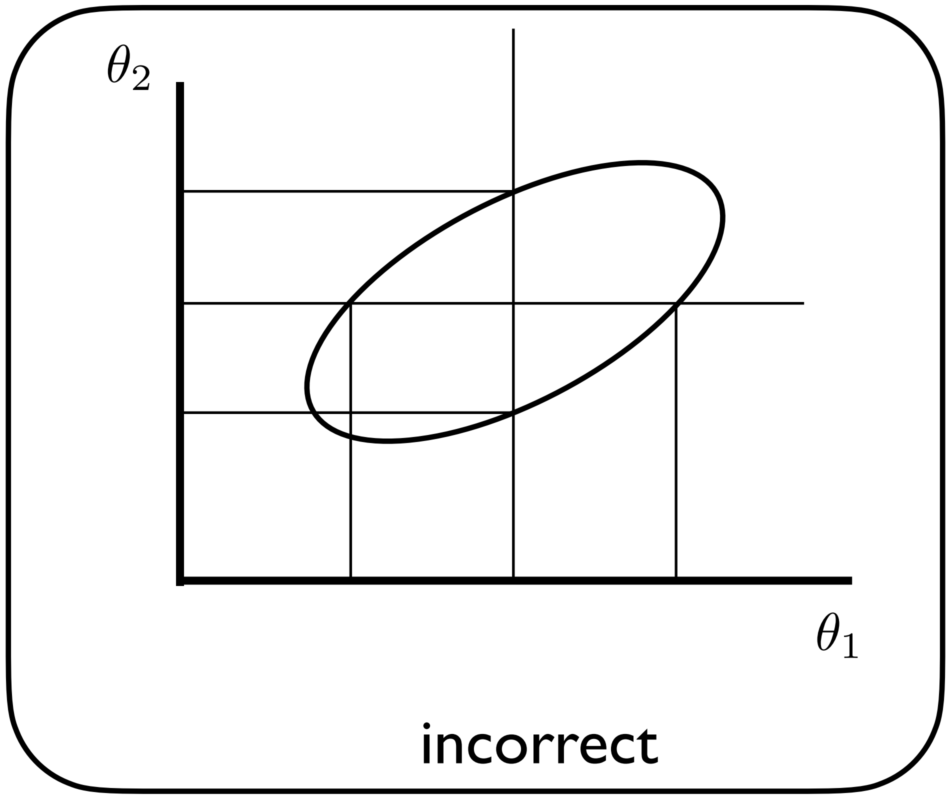


*LLH landscape is from Lecture 3

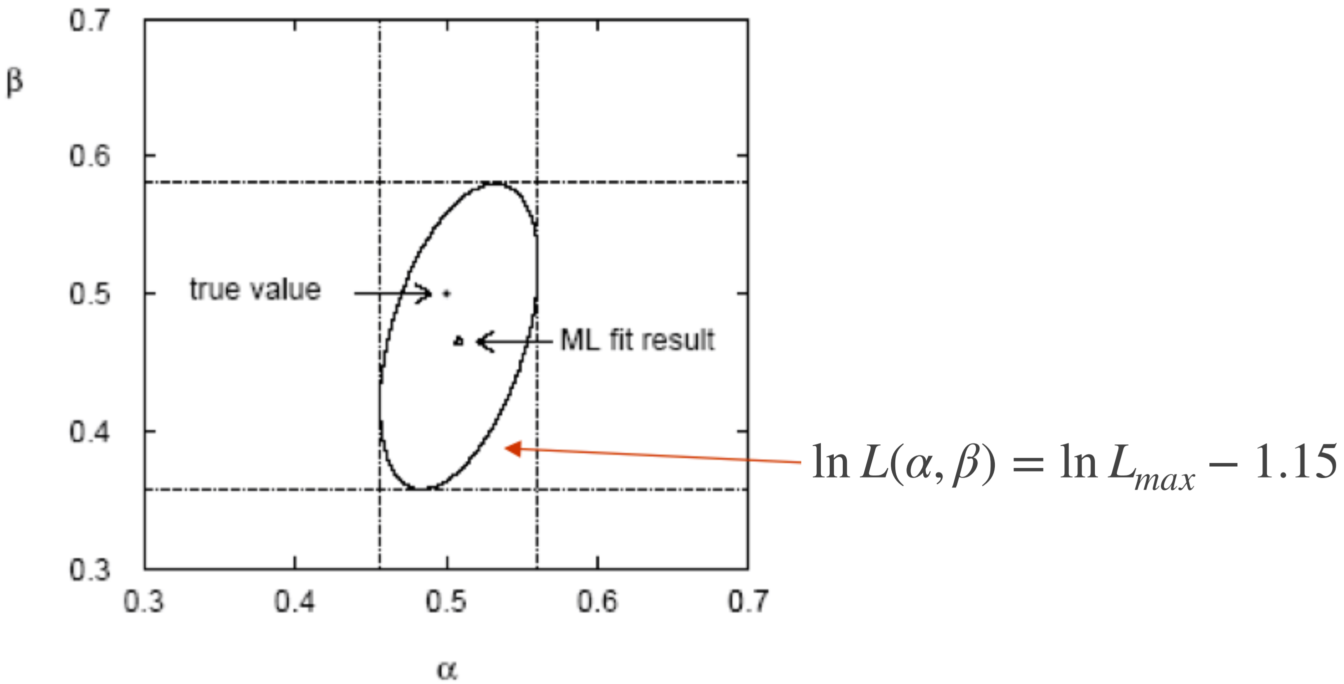
Variance of Estimators - Graphical Method

- Two Parameter Contours



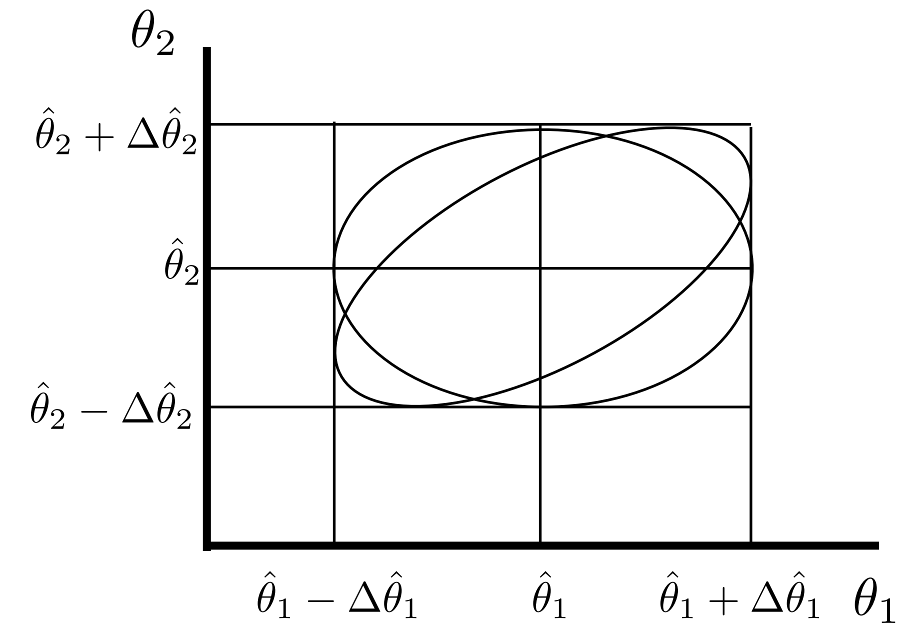


- Tangent lines to the contours give the uncertainties



Variance of Estimators - Graphical Method

- When the correct, tangential, method is used and 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:

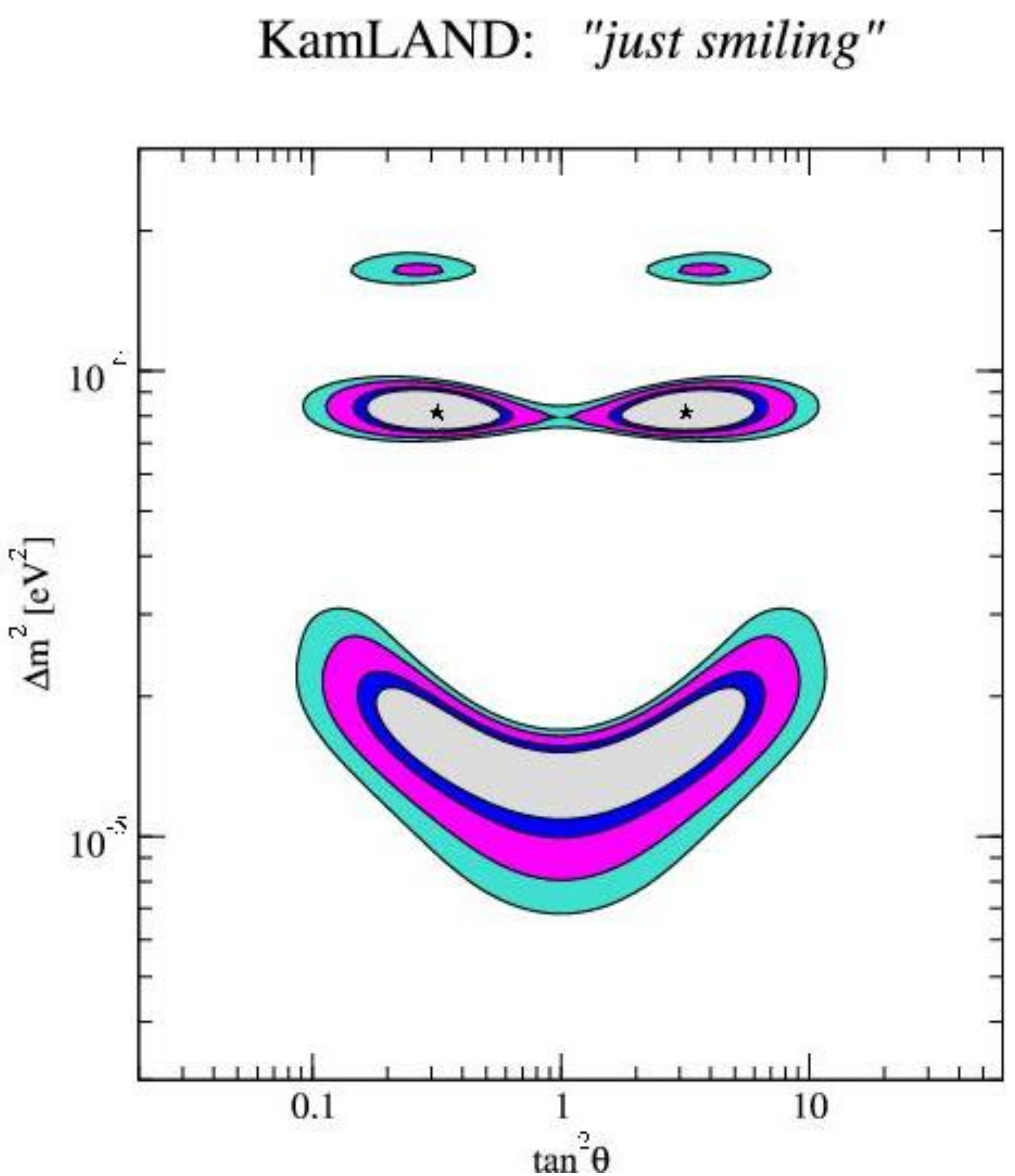


correct

a (1 DoF)	a (2	σ	C.I.
0.5	1.15	1	68.27%
2.0	3.09	2	95.4%
4.5	5.92	3	99.73%

*DoF = Degree of freedom. Here it equates to the number of fit parameters in the likelihood.

Multiple Localized Confidence Intervals — “Islands”



Variance/Uncertainty - Using LLH Values

- The LLH (or -2*LLH) landscape provides the necessary information to construct 2+ dimensional confidence intervals
  - Provided the respective MLEs are gaussian or well-approximated as gaussian the intervals are ‘easy’ to calculate
  - For non-gaussian MLEs — which is not uncommon — a more rigorous approach is needed, e.g. parametric bootstrapping
- Some minimization programs will return the uncertainty on the parameter(s) after finding the best-fit values
  - The .migrad() call in iminuit
  - It is possible to write your own code to do this as well

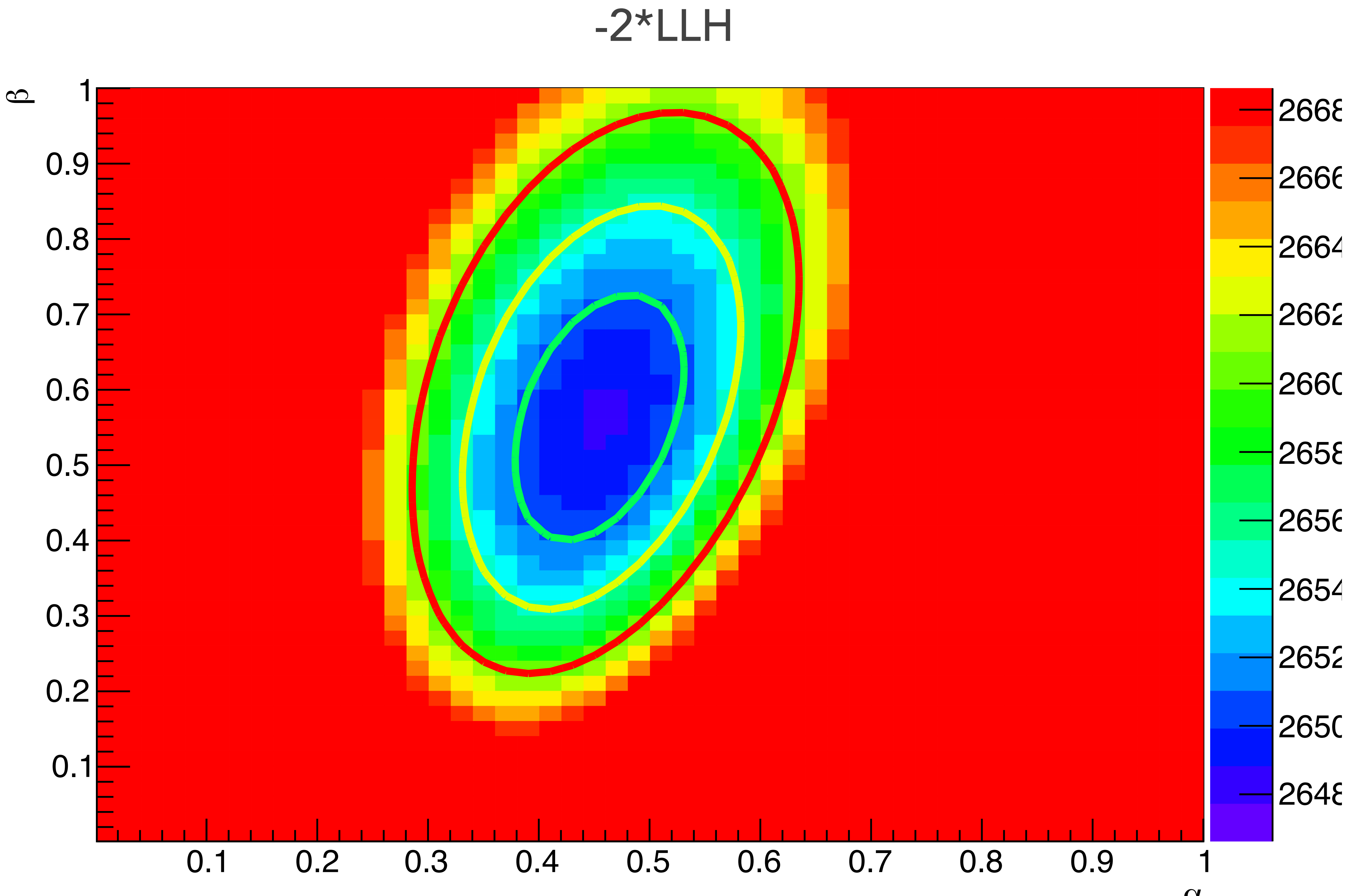
Exercise #2d

- Using the same function as Exercise #2a, find the MLE values for the data in the file using the function

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

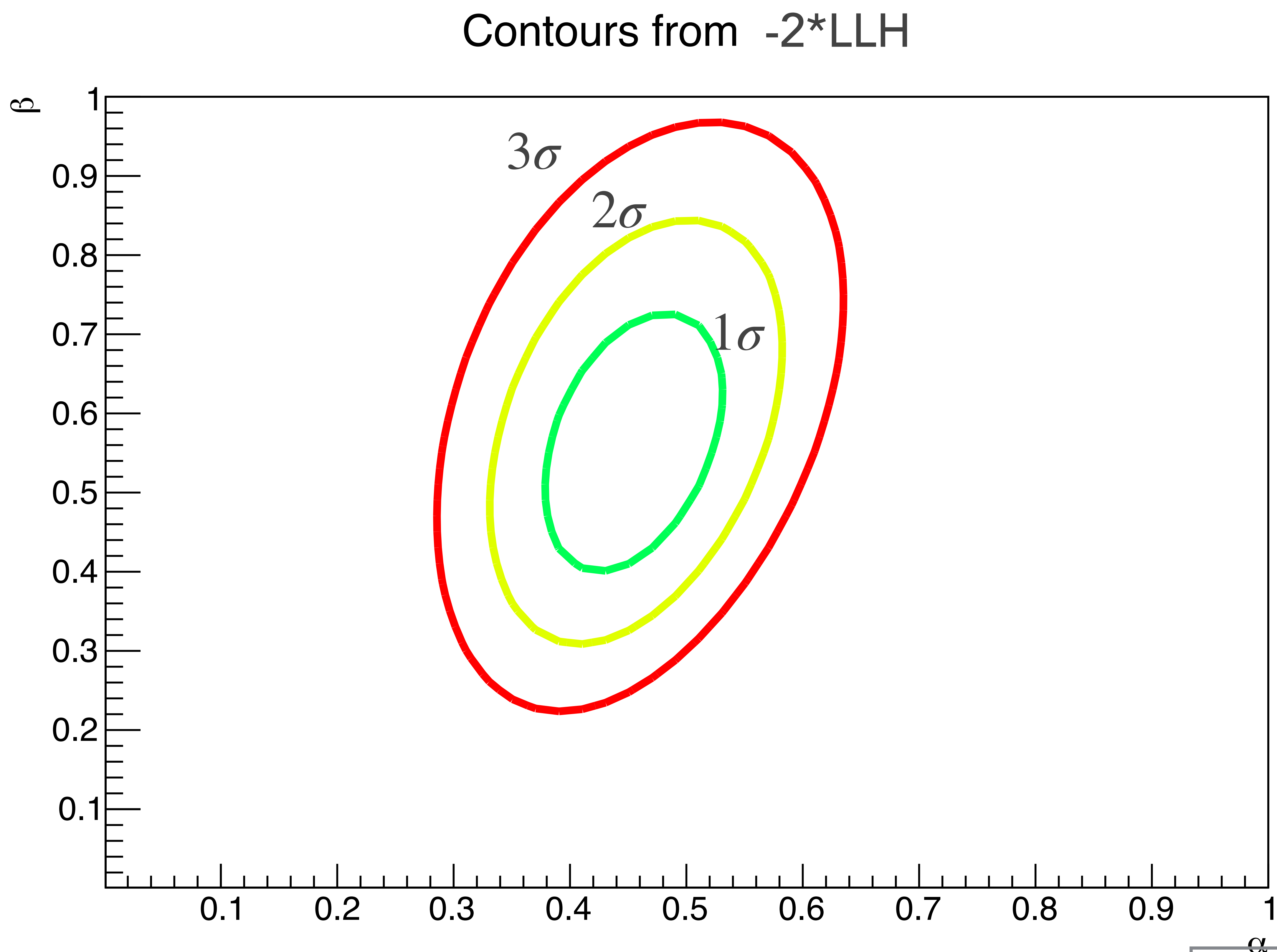
- Plot the uncertainty contours related to the 1σ, 2σ, and 3σ confidence regions from the ΔLLH
    - Remember that this function has 2 fit parameters

Contours on Top of the LLH Space



*from different data set

Just the Contours



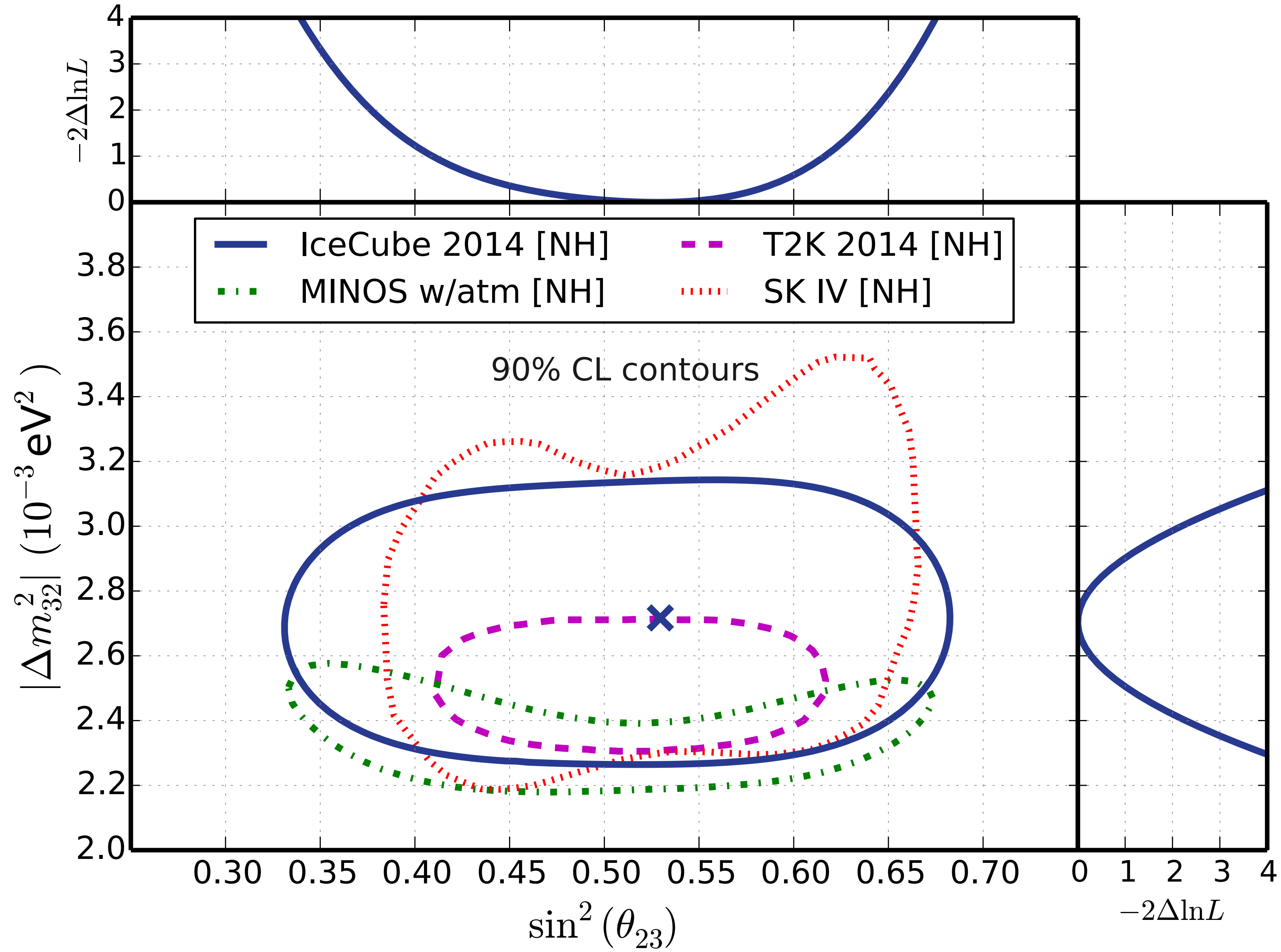
*from different data set

Real Data

- 1D projections of the 2D contour in order to give the best-fit values and their uncertainties

$$\sin^2\theta_{23} = 0.53^{+0.09}_{-0.12}$$

$$\Delta m^2_{32} = 2.72^{+0.19}_{-0.20} \times 10^{-3}\mathrm{eV}^2$$



Remember, even though they are 1D projections the ΔLLH conversion to **σ** must use the degrees-of-freedom from the actual fitting routine

*arXiv:1410.7227

Exercise #3

- There is a file posted on the class webpage which has two columns of x numbers (not x and y, *just* x for 2 data sets) 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 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

Extra

- Use a 3-dimensional function for **α**=0.5, **β**=0.5, and **ɣ**=0.9 generate 2000 Monte Carlo data points using the function transformed into a PDF over the range -1 ≤ x ≤ 1

$$f(x;\alpha,\beta,\gamma) = 1 + \alpha x + \beta x^2 + \gamma x^5$$

- Find the best-fit values and uncertainties on **α**, **β**, and **ɣ**
- Similar to Exercises #2, show that Monte Carlo re-sampling produces similar uncertainties as the ΔLLH prescription for the 3D hypersurface
    - In 3D, are 500 Monte Carlo pseudo-experiments enough?
    - Are 2000 Monte Carlo data points per pseudo-experiment enough?
    - Write a profiler to project the 2D contour onto 1D, properly