

# Lecture 5: Parameter Estimation and Uncertainty



D. Jason Koskinen  
[koskinen@nbi.ku.dk](mailto:koskinen@nbi.ku.dk)

*Advanced Methods in Applied Statistics*  
*Feb - Apr 2024*

# 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

# Outline

- Recap in 1D
- Extension to 2D
  - Likelihoods
  - Contours
  - Uncertainties

\*Some material from T. Petersen, D. R. Grant, and G. Cowan

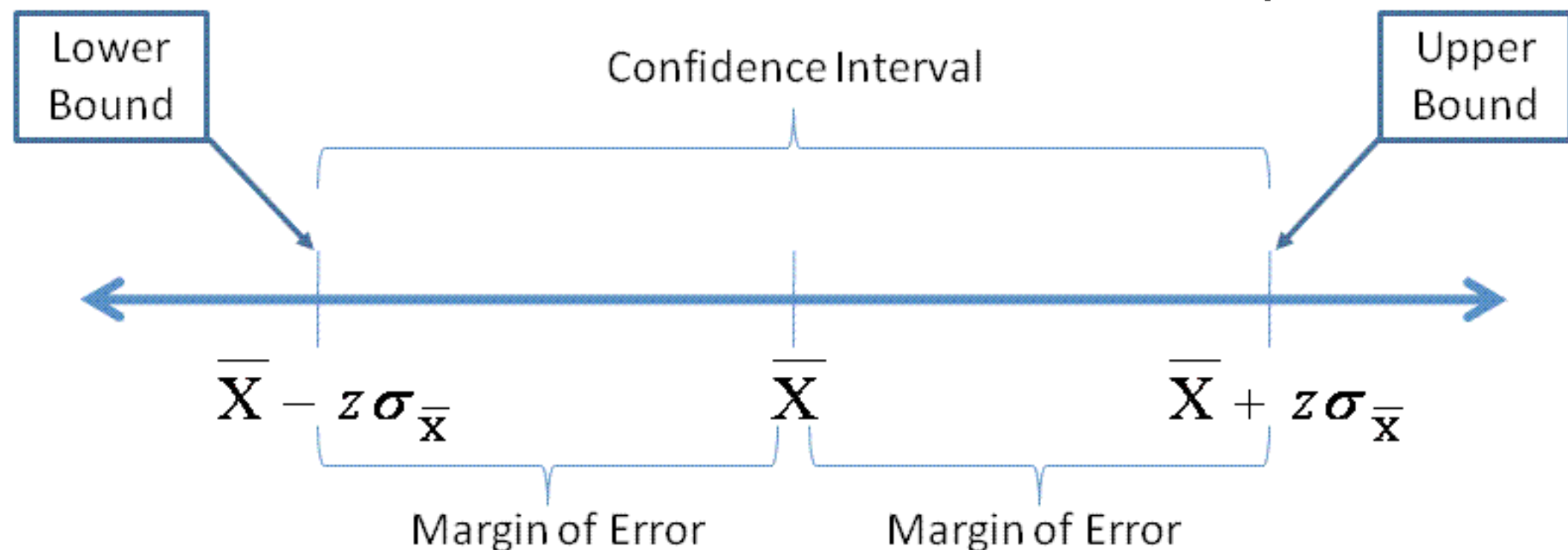
# Confidence intervals

*“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  $\sigma$  or percent 50%, 90%, 95%, and 99%

There is a choice as follows:

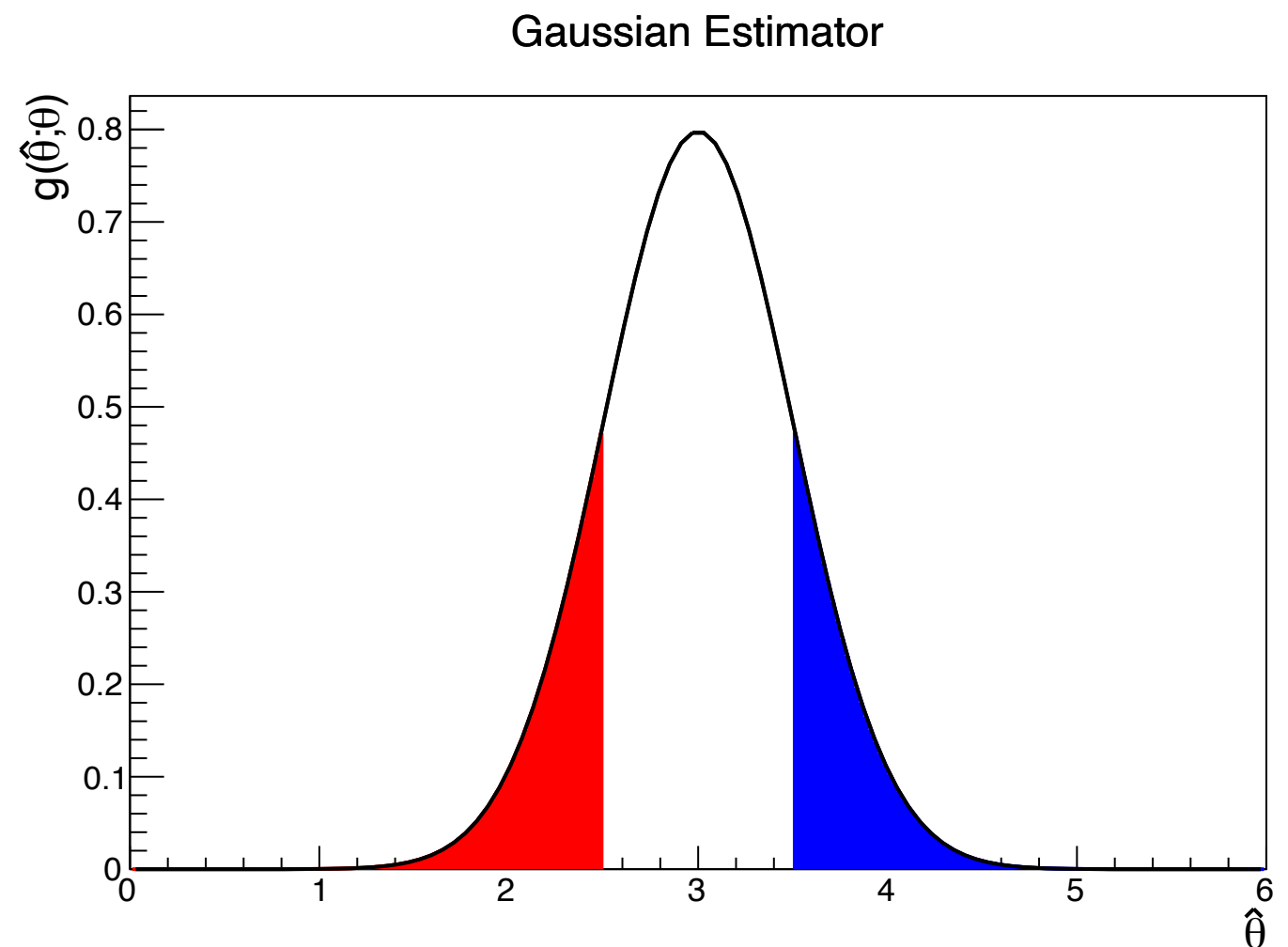
1. Require symmetric interval ( $x_+$  and  $x_-$  are equidistant from  $\mu$ ).
2. Require the shortest interval ( $x_+$  to  $x_-$  is a minimum).
3. Require a central interval (integral from  $x_-$  to  $\mu$  is the same as from  $\mu$  to  $x_+$ ).

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:

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

- First term is  $\ln L_{\max}$ , 2nd term is zero, third term can be used for information inequality (not covered here)

- For **1** parameter:

- Minimize, or scan, as a function of  $\theta$  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:

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

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

# 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:

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

- First (not
- For more information, see “Variance of ML Estimators” sections from “Statistical Data Analysis” ([https://www.sherrytowers.com/cowan\\_statistical\\_data\\_analysis.pdf](https://www.sherrytowers.com/cowan_statistical_data_analysis.pdf))

- For 1 parameter:
  - Minimize, or scan, as a function of  $\theta$  to get  $\hat{\theta}$
  - Uncertainty deduced from positions where  $\ln L$  is reduced by 0.5. For a Gaussian likelihood function w/ 1 fit parameter:

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

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



# $\ln(\text{Likelihood})$ and $2 \cdot \text{LLH}$

- A change of 1 standard deviation ( $\sigma$ ) in the maximum likelihood estimator (MLE) of the parameter  $\theta$  leads to a change in the  $\ln(\text{likelihood})$  value of 0.5 for a **gaussian distributed estimator**
  - Even for a non-gaussian MLE, the  $1\sigma$  region<sup>a</sup> defined as  $\text{LLH}-1/2$  can be an *okay* approximation
  - Because the regions<sup>a</sup> defined with  $\Delta\text{LLH}=1/2$  are consistent with common  $\chi^2$  distributions multiplied by 1/2, we often calculate the likelihoods as  $(-)\cdot 2 \cdot \text{LLH}$
- Translates to  $>1$  fit parameters too, with the appropriate change in  $2 \cdot \text{LLH}$  confidence values
  - 1 fit parameter,  $\Delta(2\text{LLH})=1$  for 68.3% C.L.
  - 2 fit parameter,  $\Delta(2\text{LLH})=2.3$  for 68.3% C.L.

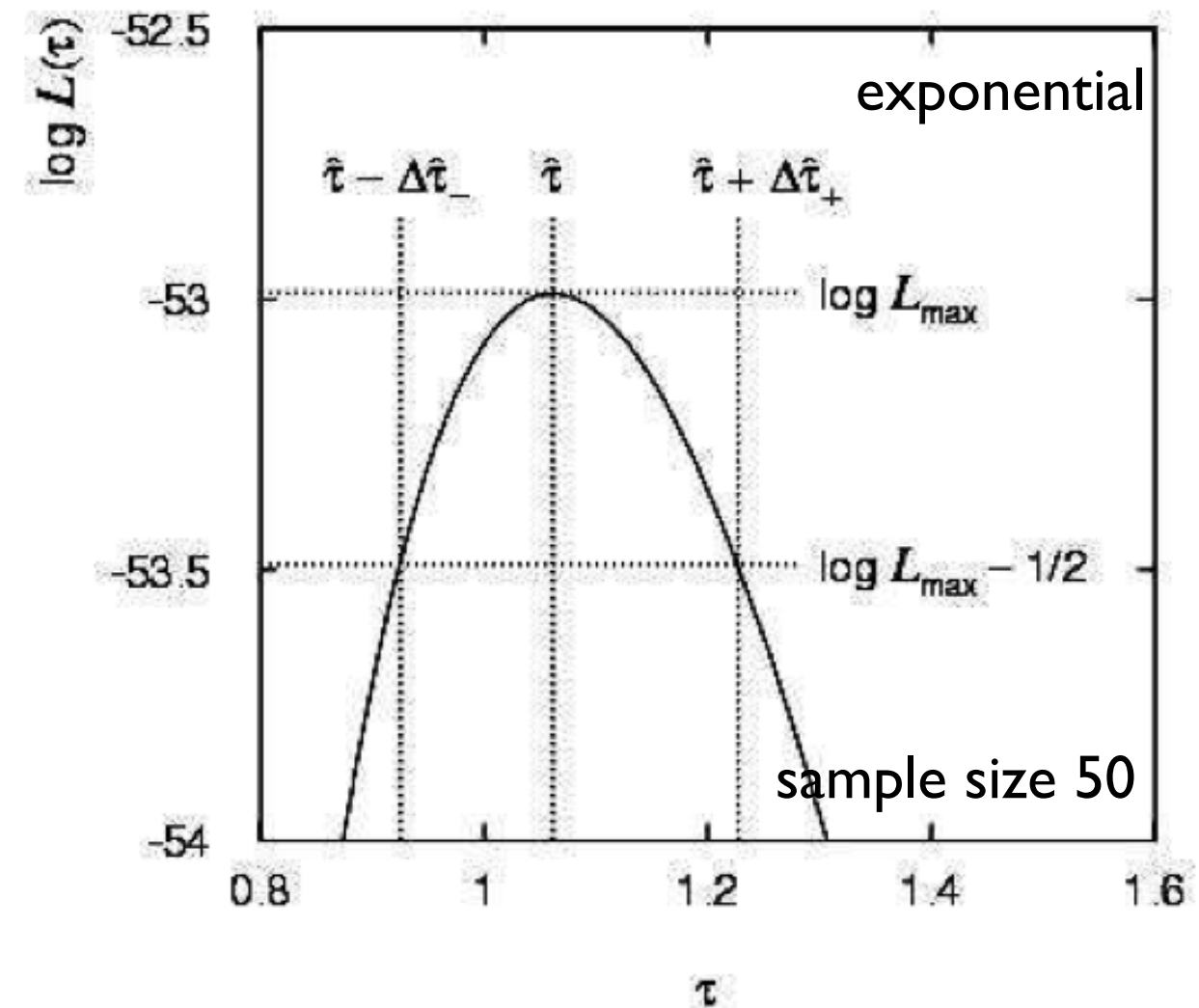
<sup>a</sup>for a distribution w/ 1 fit parameter

# Variance of Estimator

Likelihood is from Lecture 3 and is

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

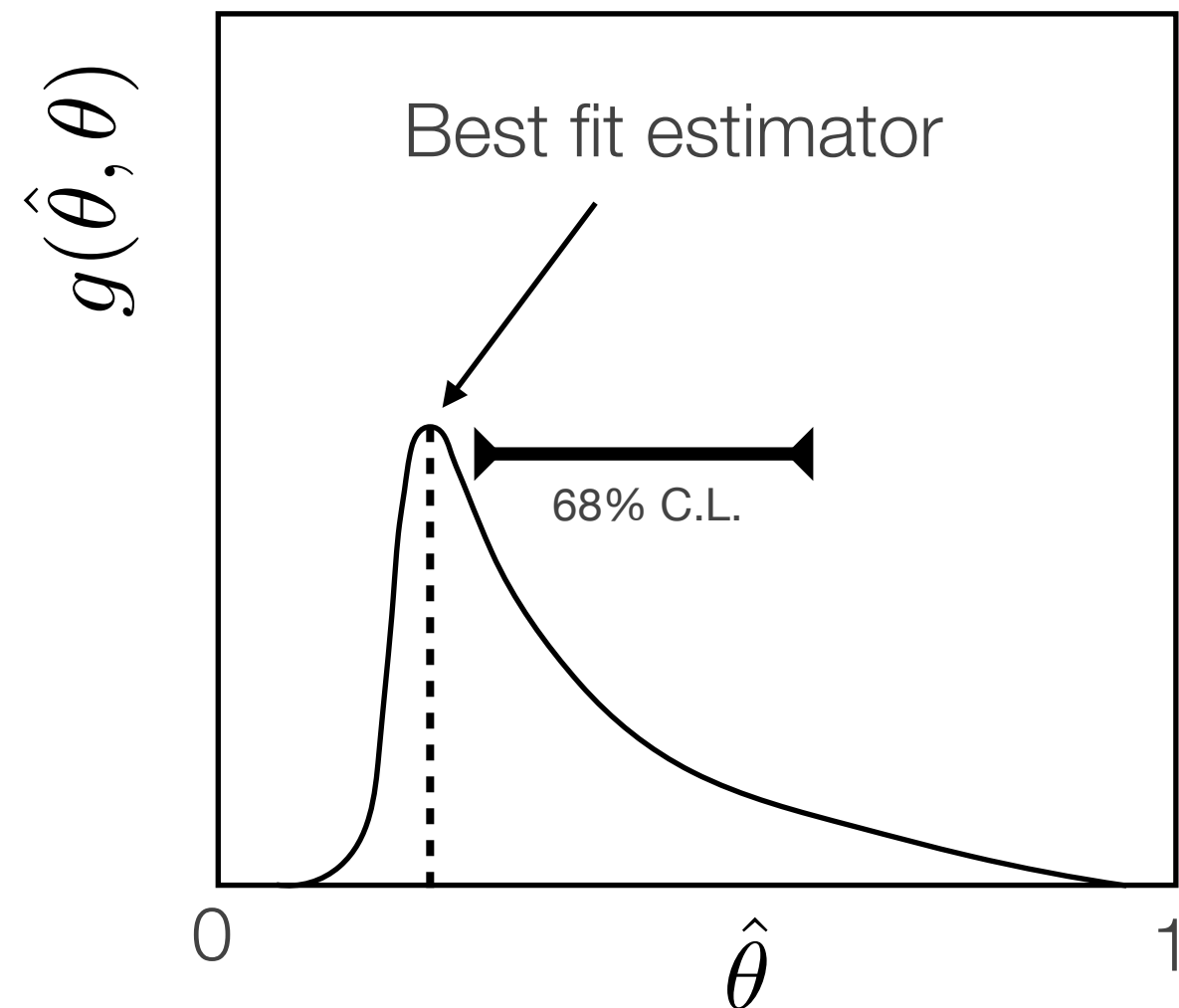
- First, we find the best-fit estimate of  $\tau$  via our LLH minimization to get  $\hat{\tau}_{best}$ 
  - Provides  $LLH(\hat{\tau}_{best}) = -53.0$
  - We could scan to get  $\hat{\tau}_{best}$ , but it won't be as precise or fast as a minimizer algorithm
- We only have 1 fit parameter, so from slide 7 we know that values of  $\hat{\tau}$  which cross  $LLH(\hat{\tau}_{best}) - 0.5$  are the  $1\sigma$  ranges, i.e. when the LLH equals -53.5



$$\begin{aligned}\hat{\tau} &= 1.062 \\ \Delta\hat{\tau}_- &= 0.137 \\ \Delta\hat{\tau}_+ &= 0.165 \\ \hat{\sigma}_{\hat{\tau}} &\approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15\end{aligned}$$

# Reporting Very Asymmetric Central Limits

- Central limits are often reported as  $\hat{\theta} \pm \sigma_{\theta}$  or  $\hat{\theta}_{-\sigma_2}^{+\sigma_1}$  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  $\theta$  and the limit ranges separately.  
"Best fit is  $\theta=0.21$  and the 90% central confidence region is 0.17-0.77"



# Exercise #0.5

- There is a data file on the course webpage for Exercise which has 2000 data points. We will assume that the unnormalized function describing the data is:

$$f(x; \beta) = 1 + 0.47x + \beta x^2$$

- Using the In-likelihood difference approach ( $\Delta LLH$ ) how would you:
  - Estimate the best-fit value of  $\hat{\beta}$ ?
  - Estimate the uncertainties on  $\hat{\beta}$ , i.e.  $\sigma_{\hat{\beta}}$ ?
  - Make sketches, diagrams, or pseudo-code to illustrate to yourself and others how you would approach establishing  $\hat{\beta}$  and  $\sigma_{\hat{\beta}}$  using the  $\Delta LLH$  approach

# Exercise #1

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

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

- For data that has unknown values of  $\alpha$  and  $\beta$  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.
  - There are 2000 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  $-0.95 \leq x \leq 0.95$ .
  - Remember to **normalize** the function properly to convert it to a proper PDF

# Exercise #1 (cont.)

- Fit the maximum likelihood estimate (MLE) parameters  $\hat{\alpha}_{data}$  and  $\hat{\beta}_{data}$  from the data files using a minimizer/maximizer
- To get an idea of what the distribution of  $\hat{\alpha}_{data}$  and  $\hat{\beta}_{data}$  look like we will generate a certain number "N" of pseudo-trials, fit  $\hat{\alpha}_{pseudo-trial,i}$  and  $\hat{\beta}_{pseudo-trial,i}$  for each "i" independent and identically distributed pseudo-trial, and then plot the "N" outcomes
  - Each pseudo-trial has 2000 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'



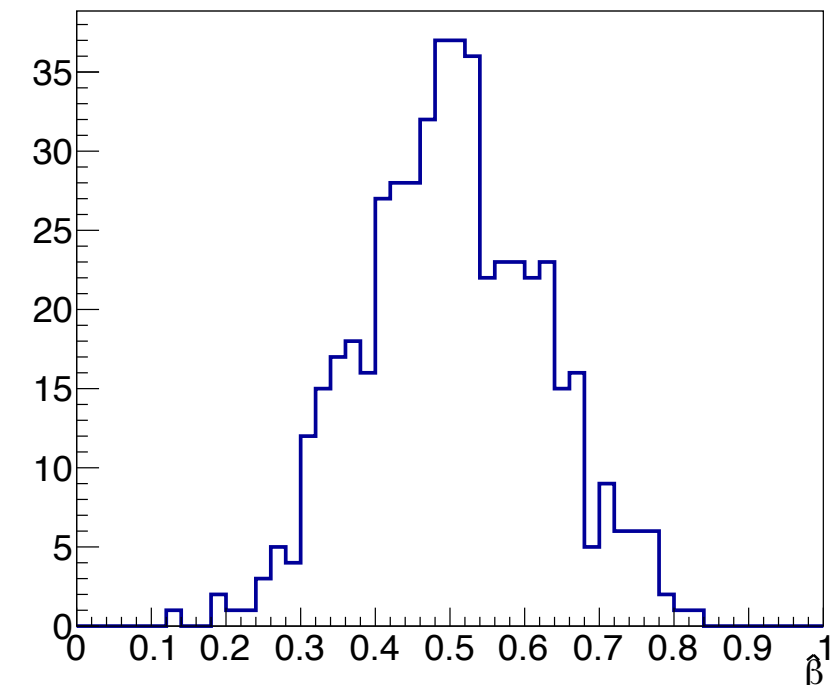
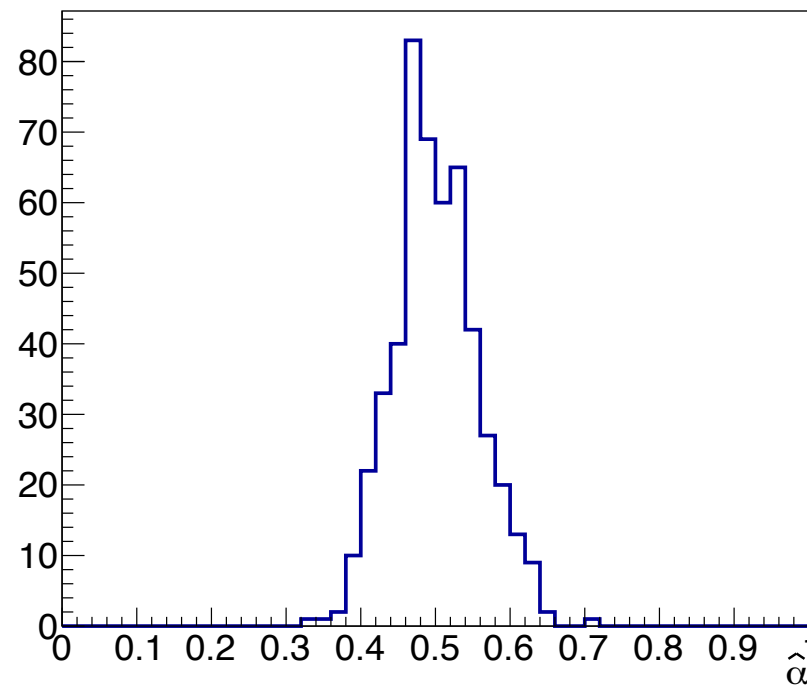
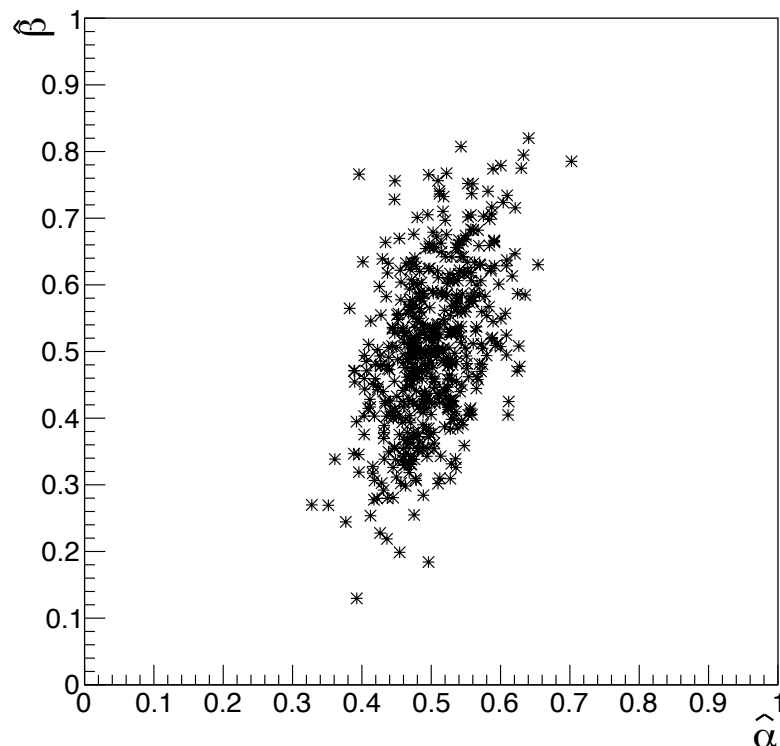
# Exercise #1

\*values shown here are NOT the same as what is used to generate the data file on the webpage

- Shown are 500 Monte Carlo pseudo-experiments
- The estimates average to approximately the best-fit values, the variances are close to initial estimates from earlier slides and the estimator distributions are approximately Gaussian
- But there is a much better way to estimate uncertainties than just assuming that the MC sample distributions of  $\hat{\alpha}$  and  $\hat{\beta}$  are Gaussian

$$\begin{aligned}\bar{\hat{\alpha}} &= 0.5005 \\ \hat{\alpha}_{RMS} &= 0.0557 \\ \bar{\hat{\beta}} &= 0.5044 \\ \hat{\beta}_{RMS} &= 0.1197\end{aligned}$$

RMSE = Root Mean Squared Error, i.e.  $\sqrt{\text{variance}}$



# Comments

- After finding the best-fit values via  $\ln(\text{likelihood})$  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

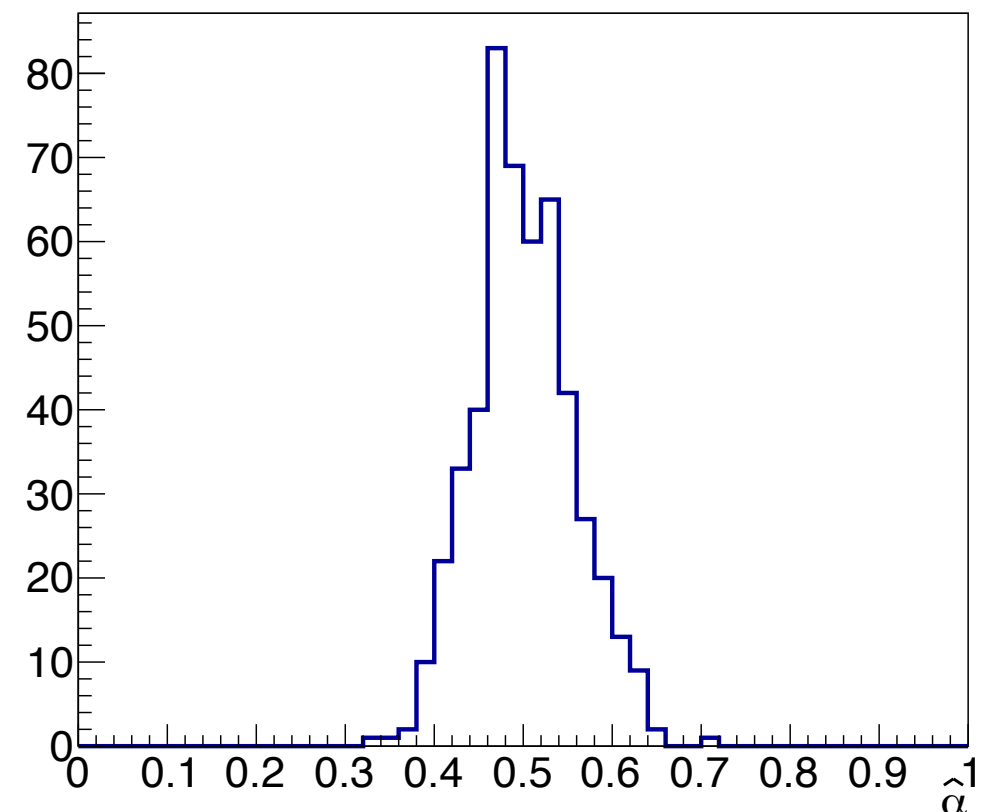
# Brute Force

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

To get  $\pm 1\sigma$  central interval

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

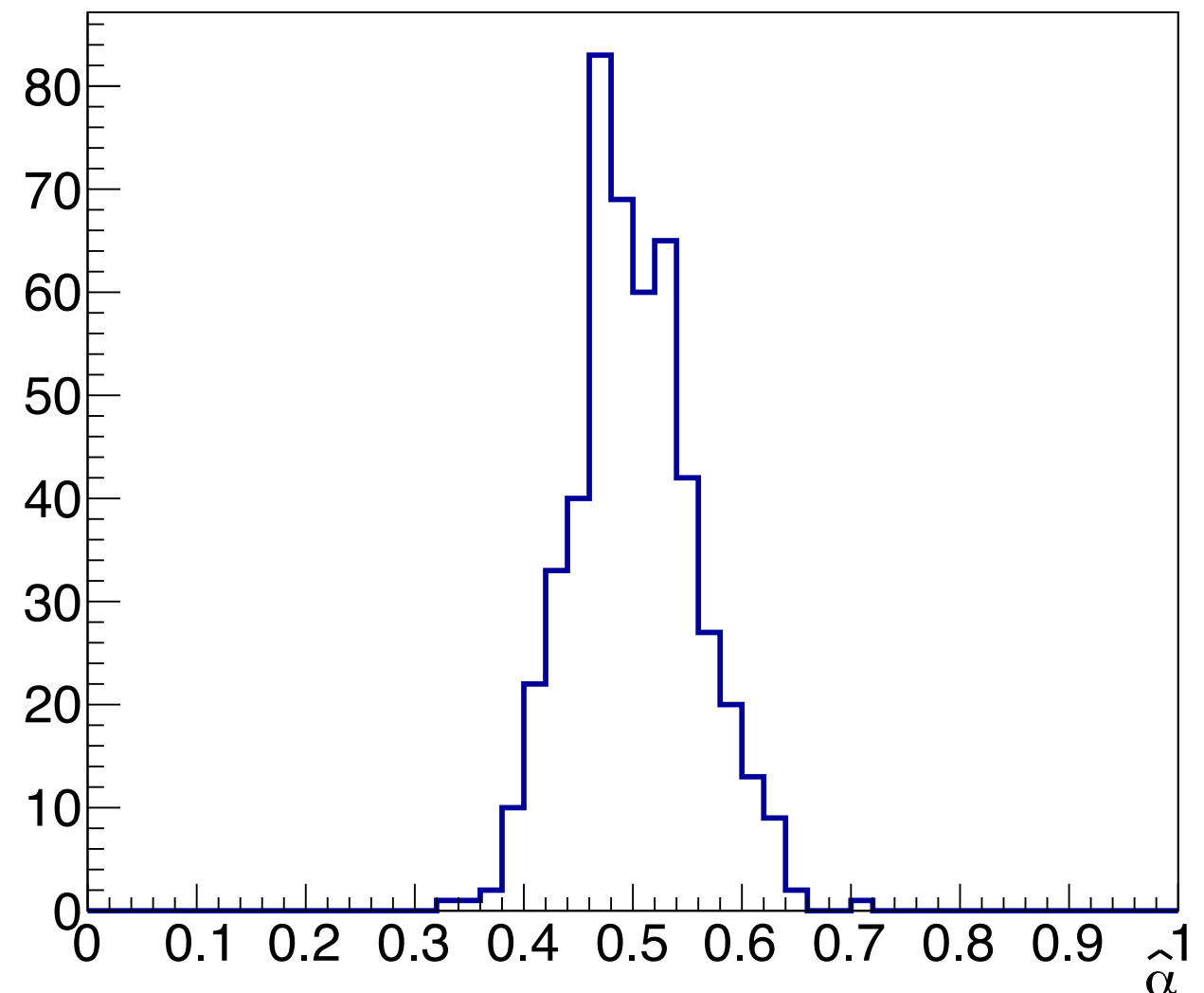
$$\frac{100\% - 68.27\%}{2} = \int_{C_+}^{\infty} g(\hat{\alpha}; \hat{\alpha}_{obs}) d\hat{\alpha}$$



# 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  $1\sigma$  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.

- The previous 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

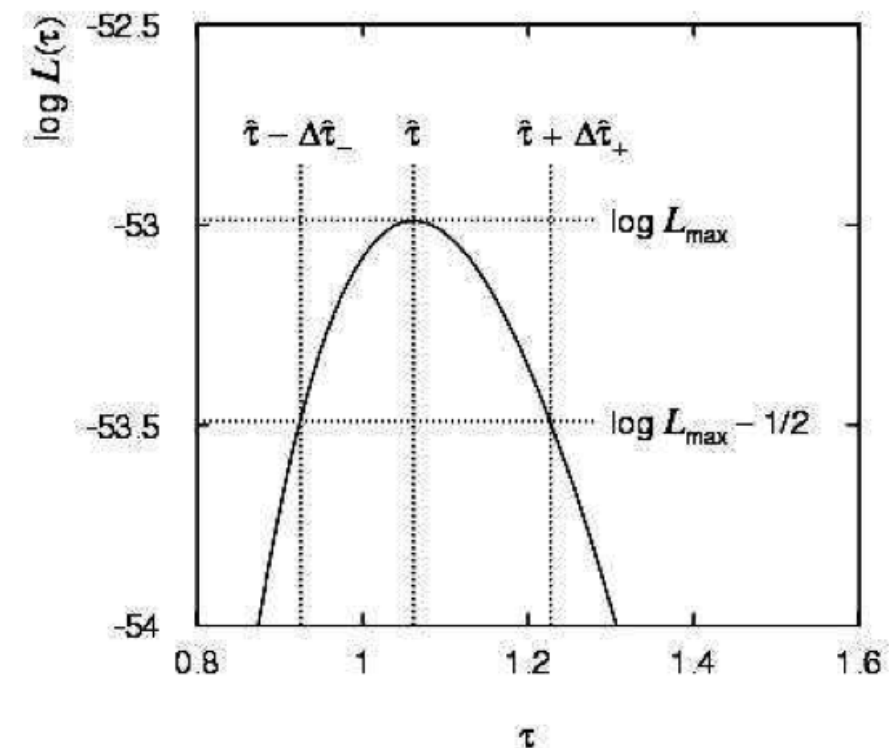
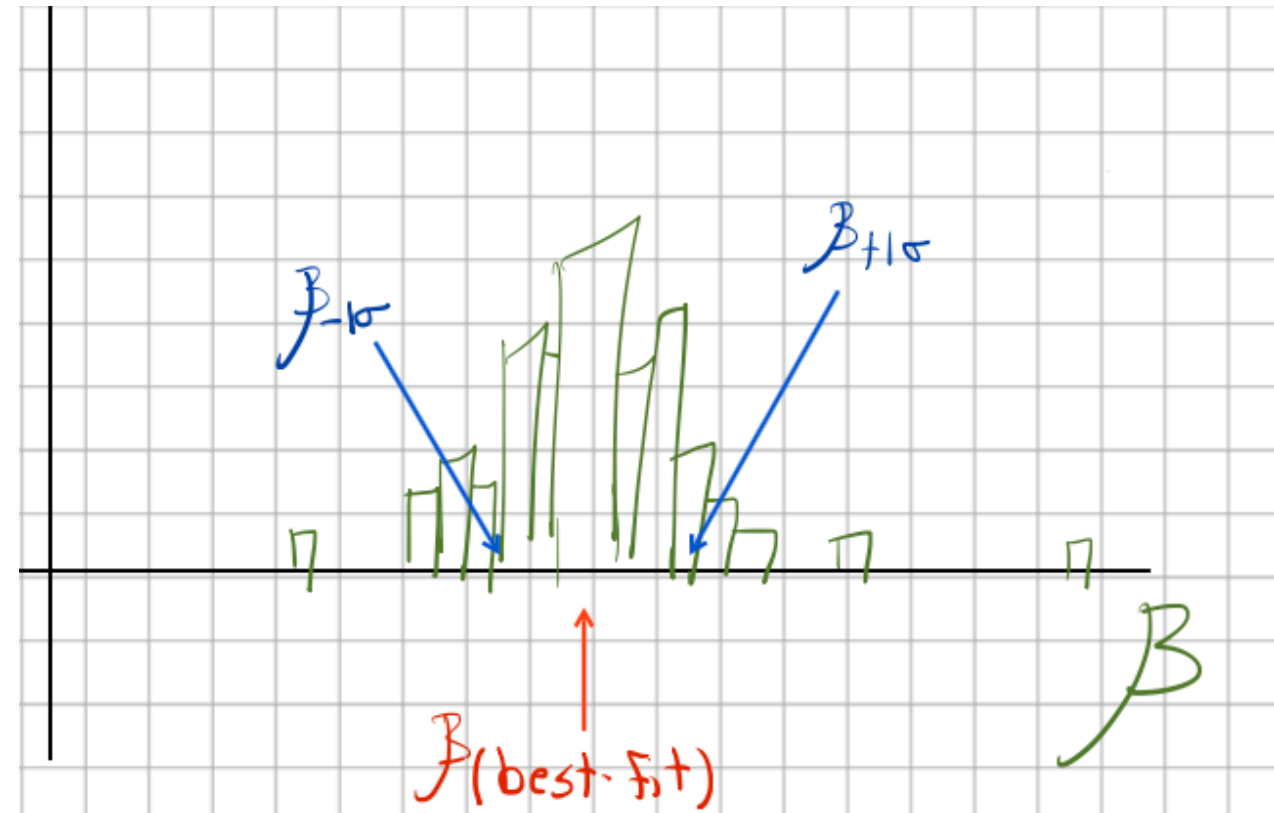
# Exercise 1b

- Continuing from Exercise 1 and using the same procedure for the 500 values from the pseudo-experiments, i.e. parametric bootstrapping
  - Find the central  $1\sigma$  confidence interval(s) for  $\hat{\alpha}$  as well as  $\hat{\beta}$  using bootstrapping
- Repeat, but now:
  - **Fix**  $\alpha=0.65$ , and only fit for  $\beta$ , i.e.  $\alpha$  is now a constant
  - What is the new  $1\sigma$  central confidence interval for  $\hat{\beta}$ ?
- Repeat with a new range of the  $-0.9 \leq x \leq 0.85$ 
  - Again, **fix**  $\alpha=0.65$
  - 2000 Monte Carlo 'data' points



# 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



# Exercise 1c

- Estimate the uncertainty only from the log-likelihood difference ( $\Delta LLH$ ), no parametric bootstrapping
  - Use the same data and function from the earlier exercises and the updated range of  $-0.9 \leq x \leq 0.85$
  - **Fix**  $\alpha=0.65$ , i.e.  $\alpha$  is not a fit parameter and never changes.
  - Since  $\alpha$  is fixed, the function  $f(x; \alpha, \beta)$  is a 1 parameter equation, and the PDF of  $f(x; \alpha, \beta)$  is also only dependent on 1 parameter. So the  $1\sigma$  uncertainty is where  $|\ln(\mathcal{L}(x; \alpha, \beta_{best-fit})) - \ln(\mathcal{L}(x; \alpha, \beta_\sigma))| = 0.5$ , and  $\sigma_\beta = \beta_{best-fit} - \beta_\sigma$
- [optional] Check to see if  $\sigma_\beta$  is asymmetric, i.e.  $+\sigma_\beta \neq -\sigma_\beta$ , for this problem when using the likelihood prescription to estimate the uncertainty.

# Good?

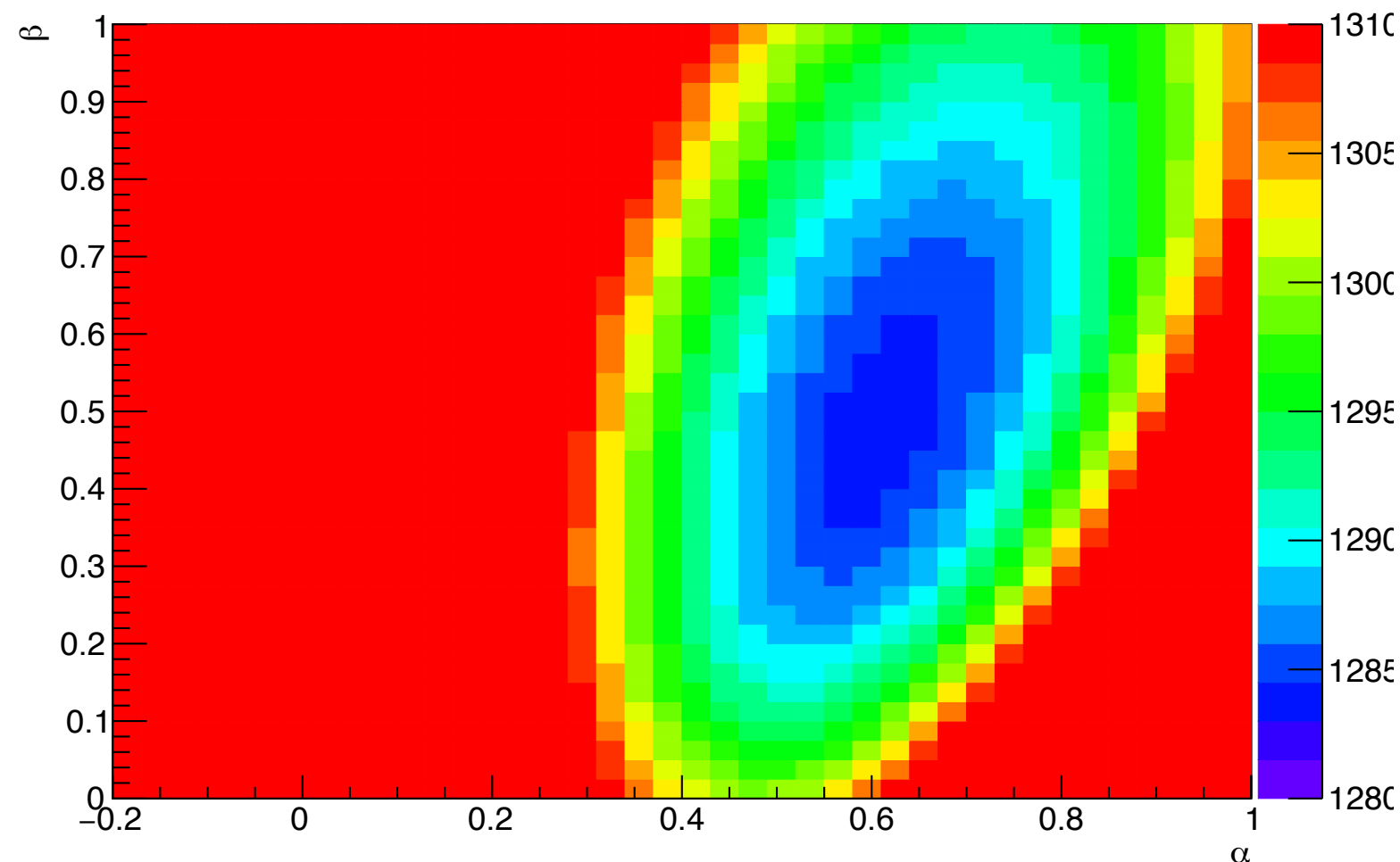
- The LLH minimization will give the best-fit values and often the uncertainty on the estimators. But, likelihood fits do not tell whether the data and the prediction agree
  - Remember that the likelihood has a form (PDF) that is provided by you and may not be correct
  - The PDF may be okay, but there 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

# 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
  - Often produce 2D contours of  $\hat{\theta}$  vs.  $\hat{\phi}$
  - There are some common mistakes to avoid

# Likelihood Contour/Surface

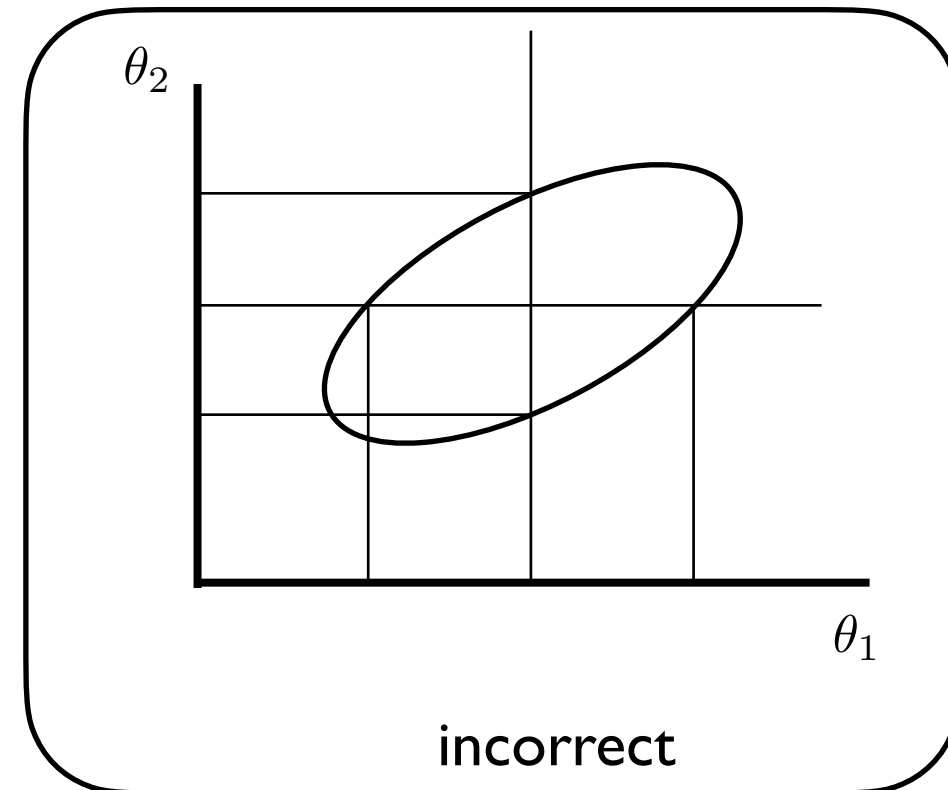
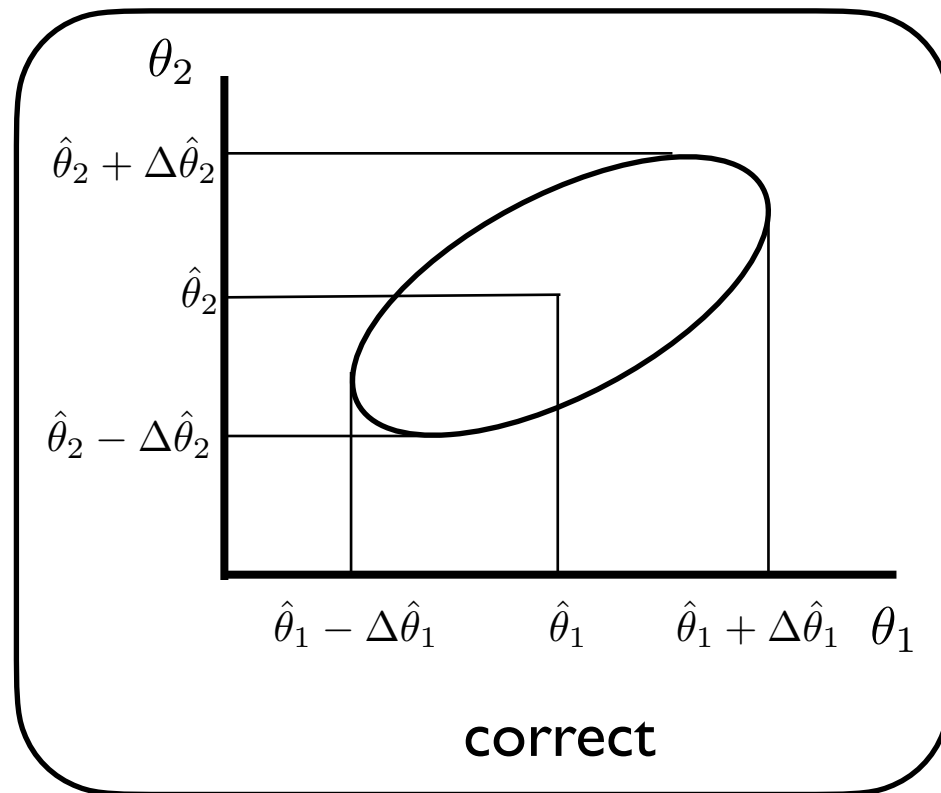
- For 2 dimensions, i.e. 2-parameter fits, we can produce likelihood landscapes. In 3 dimensions a surface, and in 3+ dimensions a likelihood hypersurface.
- The contours are then lines of with a constant value of likelihood or  $\ln(\text{likelihood})$



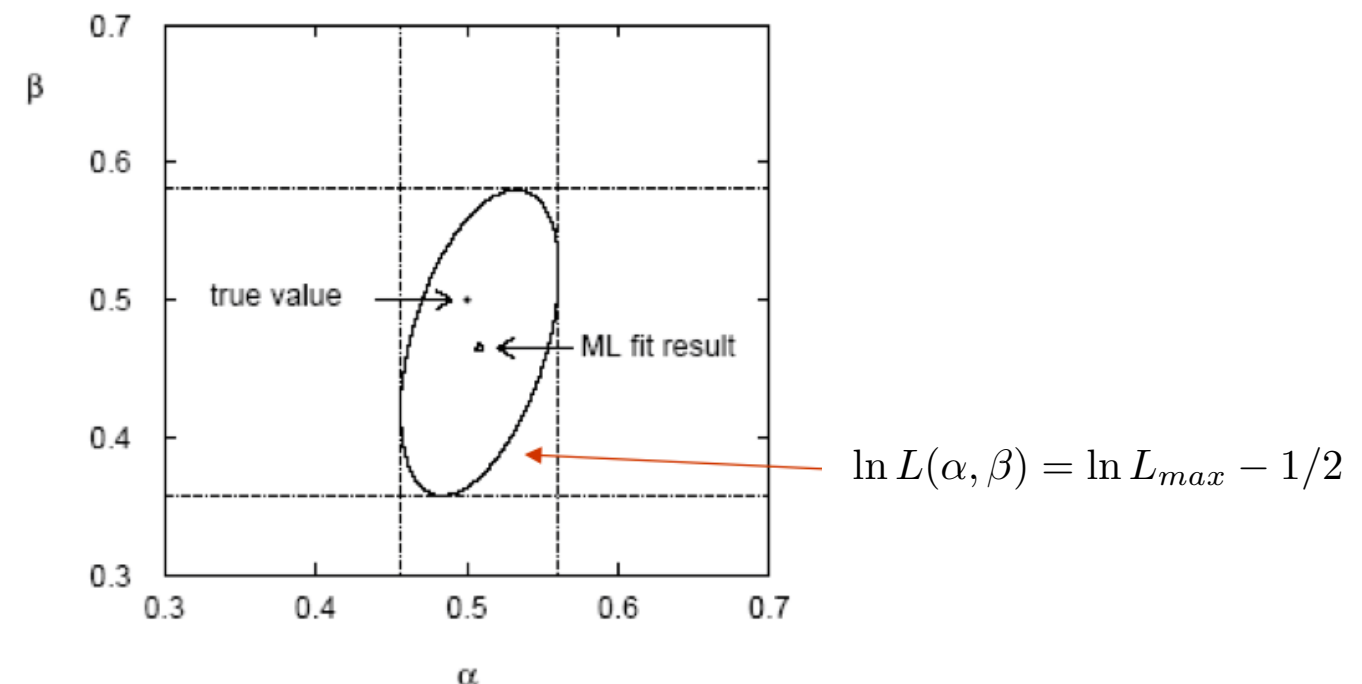
\*LLH landscape is from  
Lecture 3

# Variance of Estimators - Graphical Method

- Two Parameter Contours



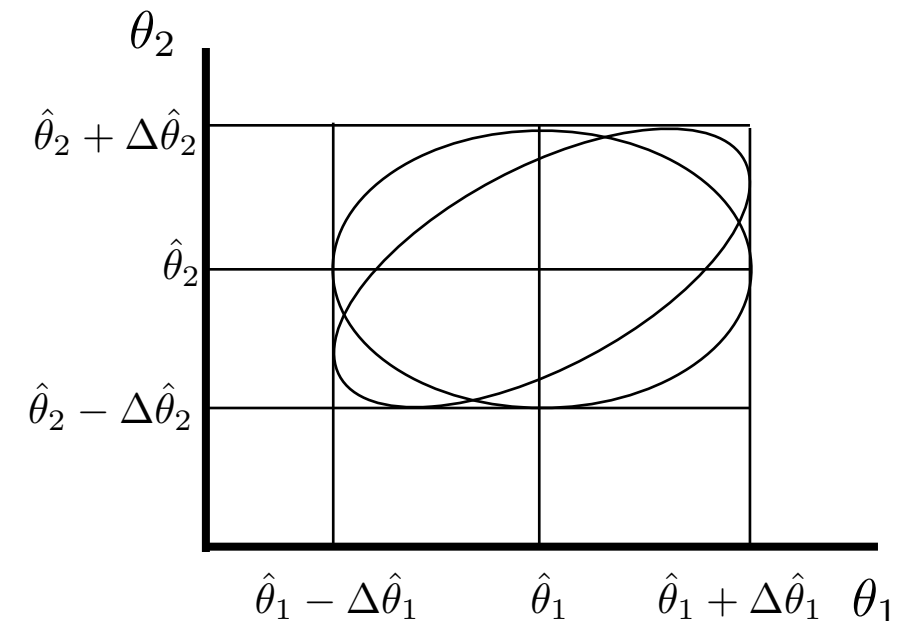
- Tangent lines to the contours give the standard deviations





# 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,  $\theta_1$  and  $\theta_2$ , is:



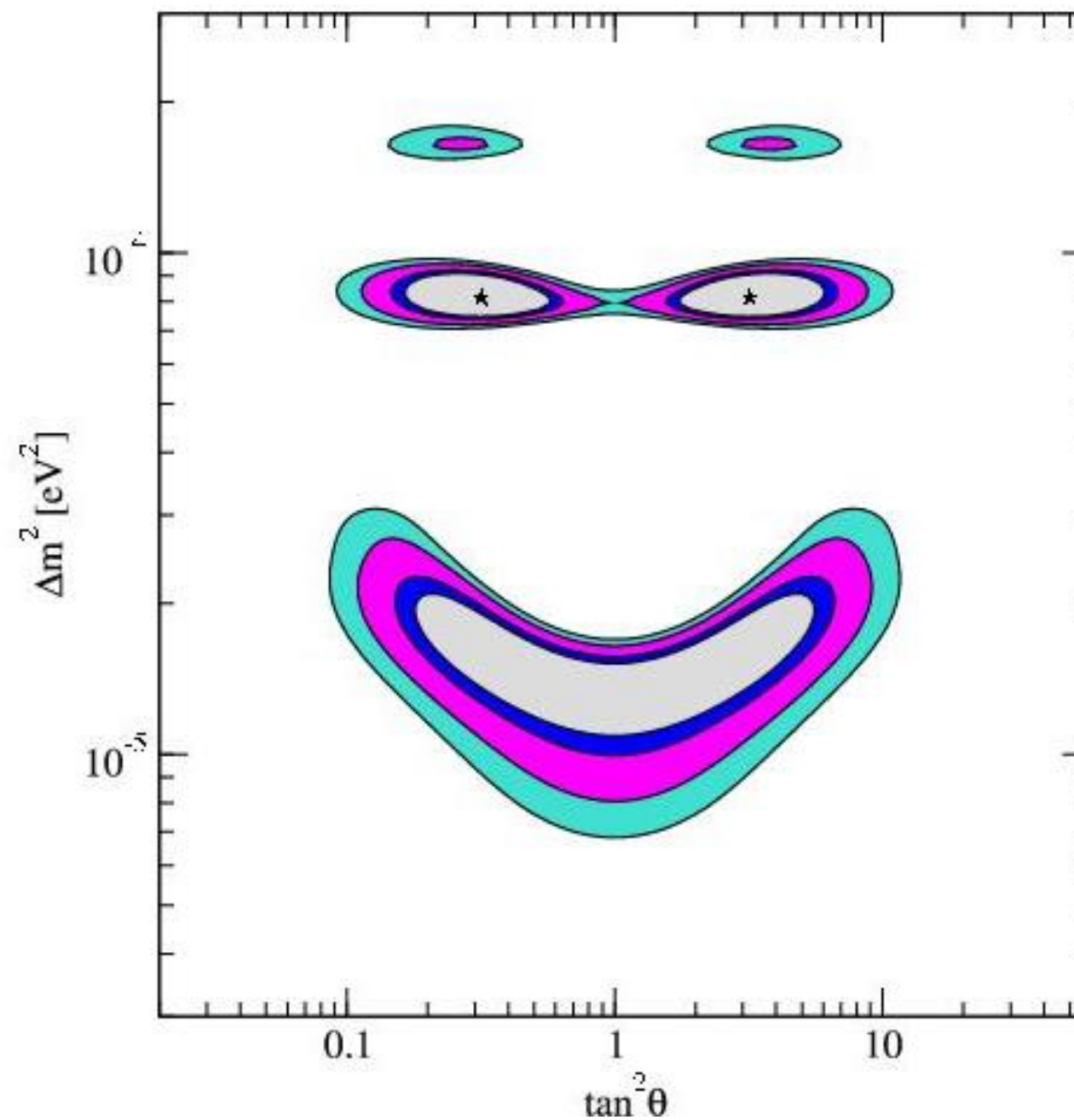
correct

a (1 DoF)	a (2 DoF)	$\sigma$
0.5	1.15	1
2.0	3.09	2
4.5	5.92	3

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

# Best Result Plot?

KamLAND: *"just smiling"*



# Variance/Uncertainty - Using LLH Values

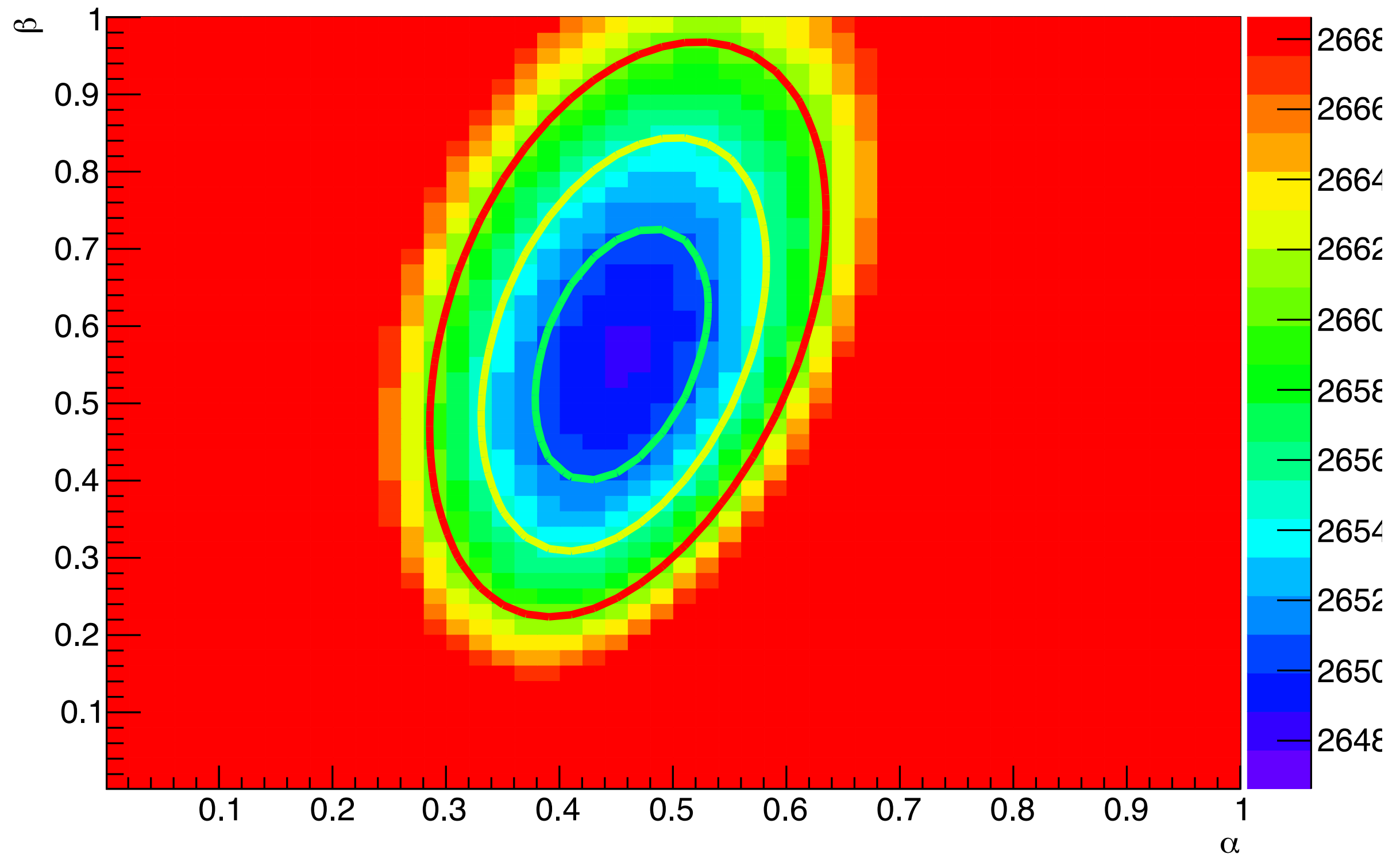
- The LLH (or  $-2 \times \text{LLH}$ ) landscape provides the necessary information to construct 2+ dimensional confidence intervals
  - Provided the respective MLEs are gaussian or well-approximated as gaussian the intervals are 'easy' to calculate
  - For non-gaussian MLEs — which is not uncommon — 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 #2

- Using the same function as Exercise #1, find the MLE values for the data in the file over the range  $-0.9 \leq x \leq 0.85$
- Plot the uncertainty contours related to the  $1\sigma$ ,  $2\sigma$ , and  $3\sigma$  confidence regions from the  $\Delta LLH$ 
  - Remember that this function has 2 fit parameters
  - Because of different random number generators, your result is likely to vary from mine

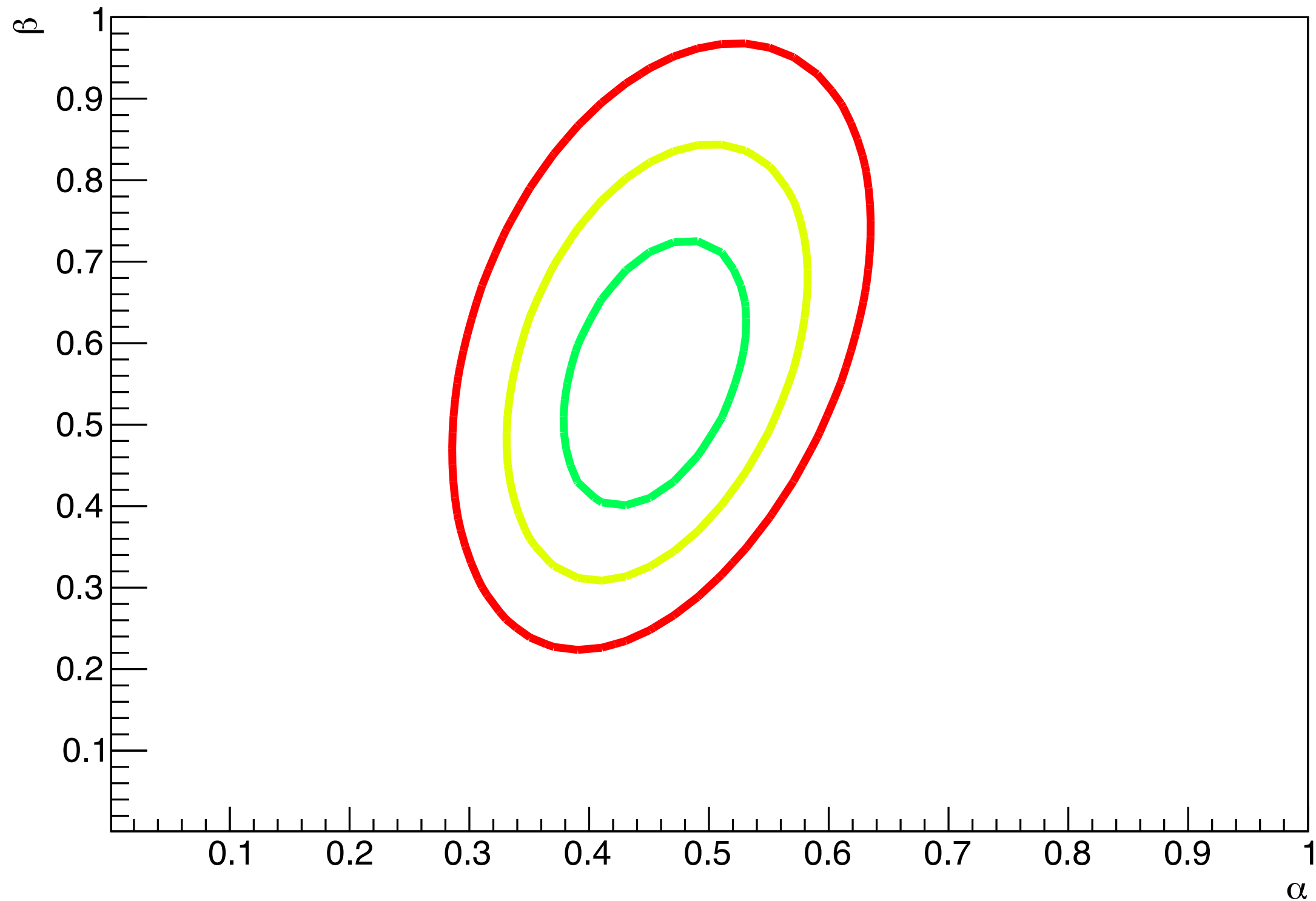
# Contours on Top of the LLH Space

$-2*LLH$



# Just the Contours

Contours from  $-2*LLH$



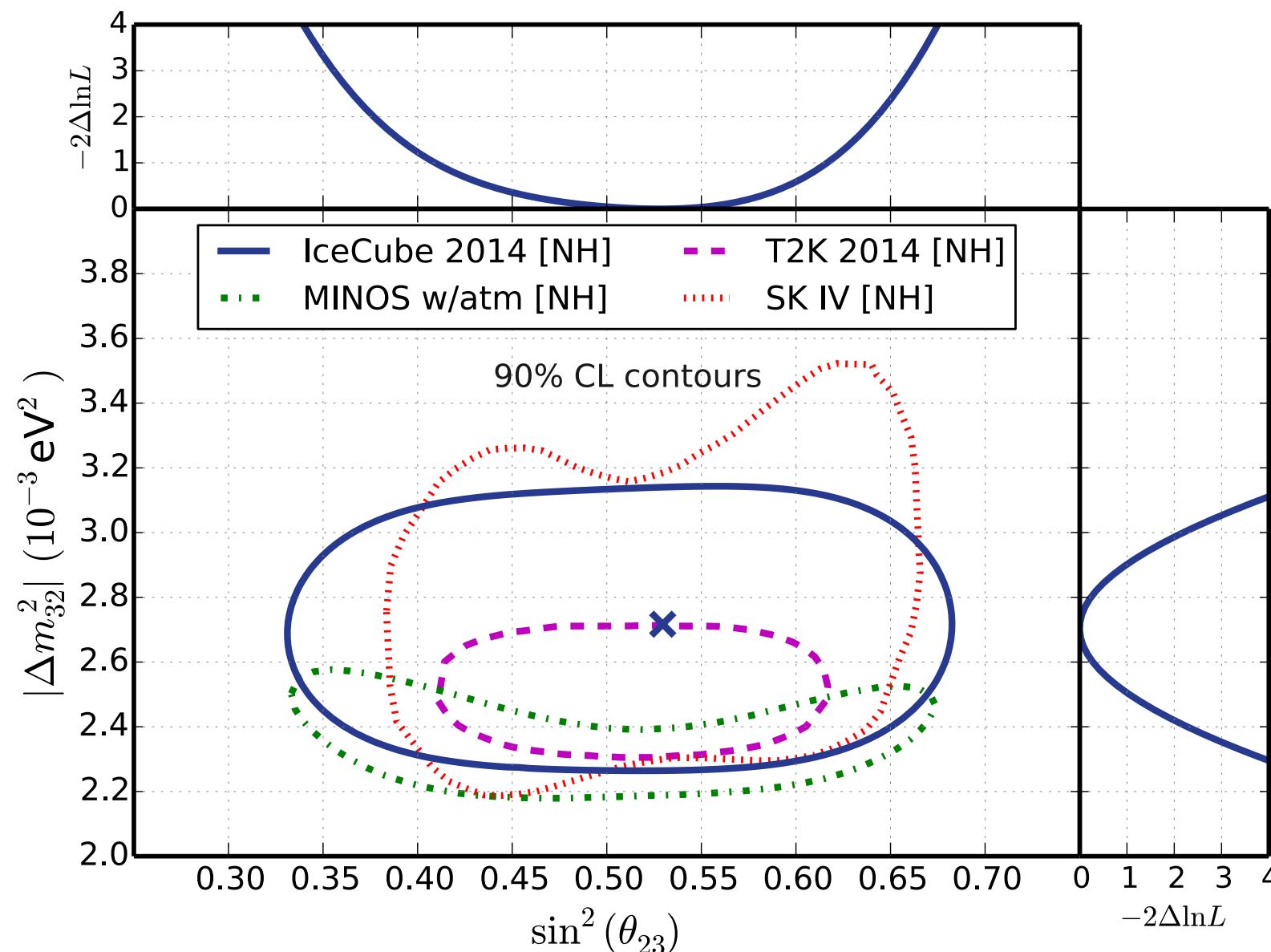


# Real Data

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

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

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



Remember, even though they are 1D projections the  $\Delta\text{LLH}$  conversion to  $\sigma$  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 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  $\alpha$  and  $\beta$
- [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  $\alpha=0.5$ ,  $\beta=0.5$ , and  $\gamma=0.9$  generate 2000 Monte Carlo data points using the function transformed into a PDF over the range  $-1 \leq x \leq 1$

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

- Find the best-fit values and uncertainties on  $\alpha$ ,  $\beta$ , and  $\gamma$
- Similar to exercise #1, show that Monte Carlo re-sampling produces similar uncertainties as the  $\Delta\text{LLH}$  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