# Lecture 7: Parameter Estimation and Uncertainty


D. Jason Koskinen

koskinen@nbi.ku.dk


*Advanced Methods in Applied Statistics*

*Feb - Apr 2018*

Photo by Howard Jackman

# Oral Presentation and Report

- By Feb. 26, everyone should have filled in the google spreadsheet noting which ‘paper’ you will present
- Wednesday March 7 the 1-2 page written report will be due by 16:00 CET
- Thursday March 8 will be the oral report

# Outline

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

*Material derived from T. Petersen, D. R. Grant, and G. Cowan

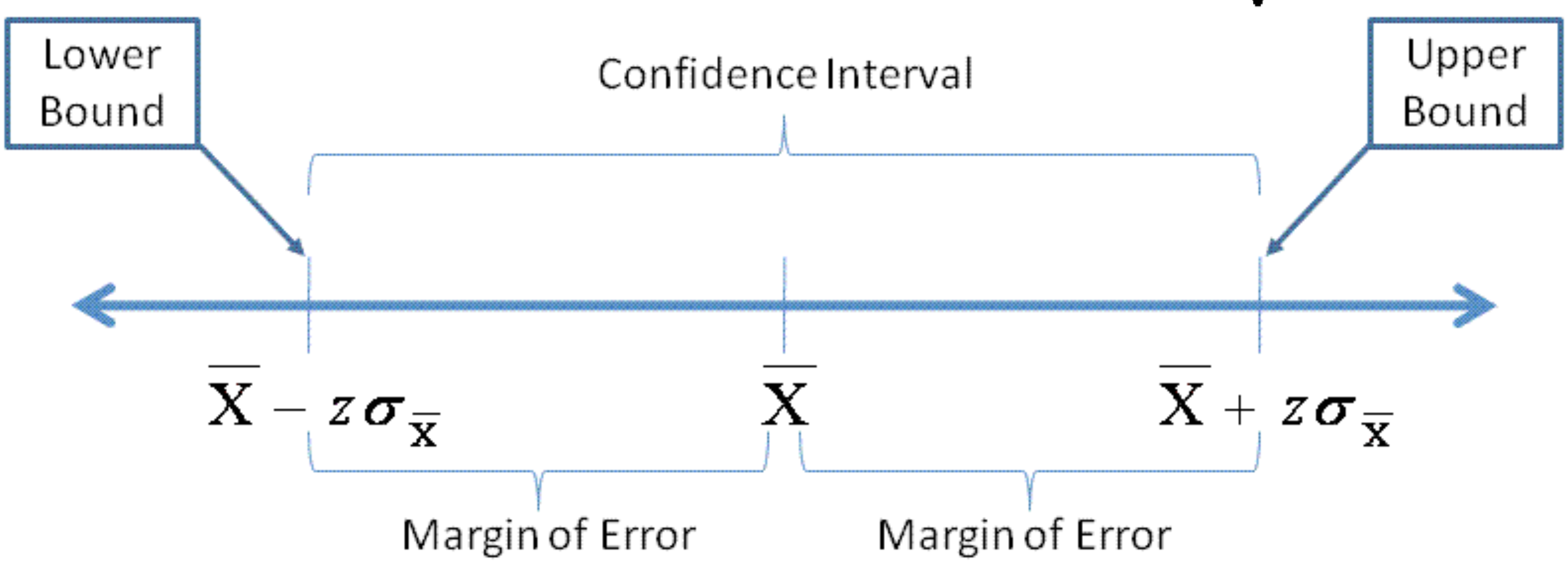
# Confidence intervals

*"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 σ or percent 50%, 90%, 95%, and 99%

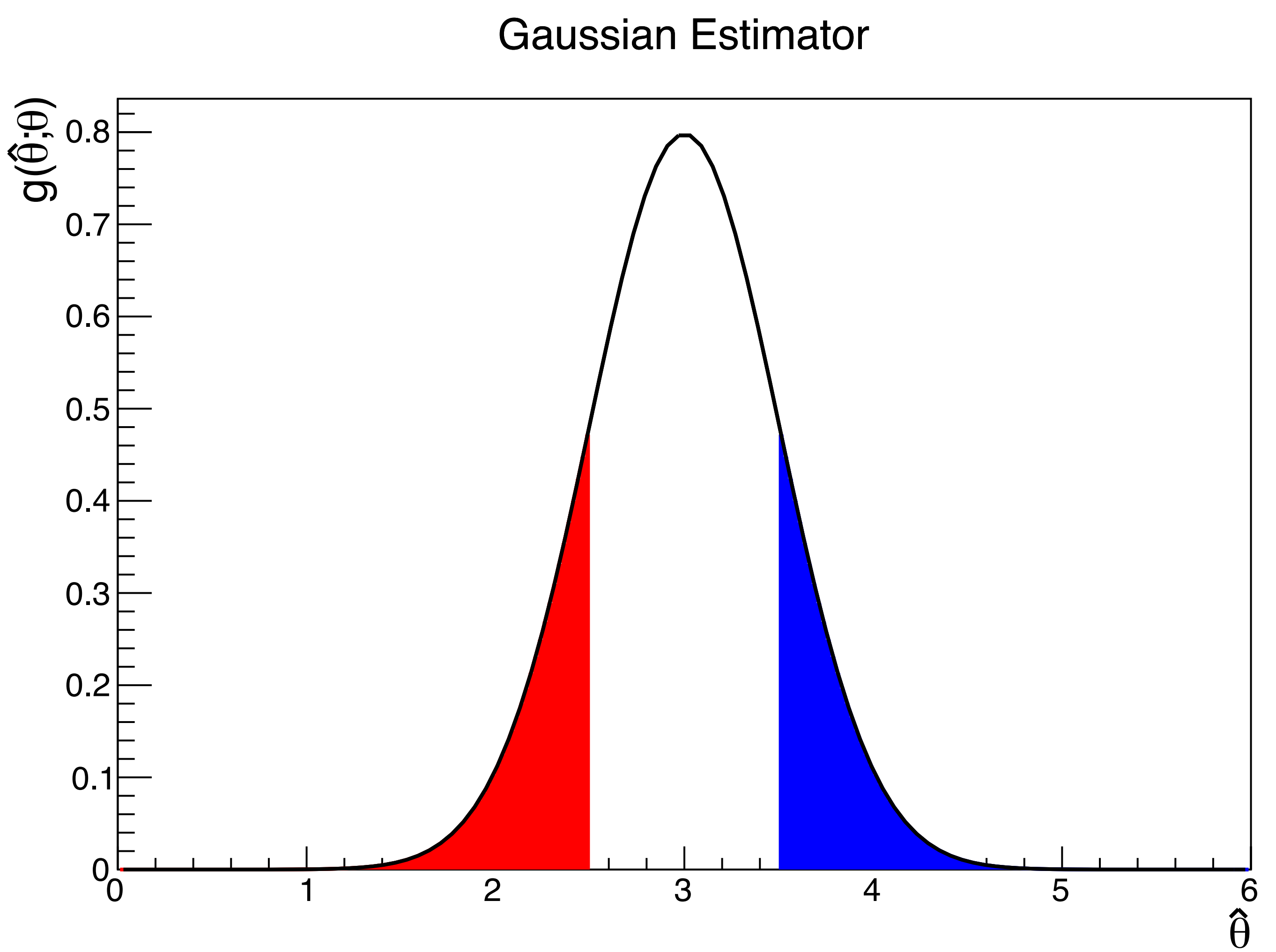
<u>There is a choice as follows:</u>

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



# Variance of Estimators - Gaussian Estimators

- Used for 1 or 2 parameters when the maximum likelihood 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 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 lnL is reduced by an amount 1/2. For a Gaussian likelihood function w/ **1** fit parameter:

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

$$\ln L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) = \ln L_{max} - \frac{1}{2}$$ 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 decrease in the ln(likelihood) of 1/2 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 $\chi^2$ distributions multiplied by 1/2, we often calculate the likelihoods as (-)2*LLH
- Translates to >1 parameters too, with the appropriate change in 2*LLH confidence values
  - 1 parameter, Δ(2LLH)=1 for 68.3% C.L.
  - 2 parameter, Δ(2LLH)=2.3 for 68.3% C.L.

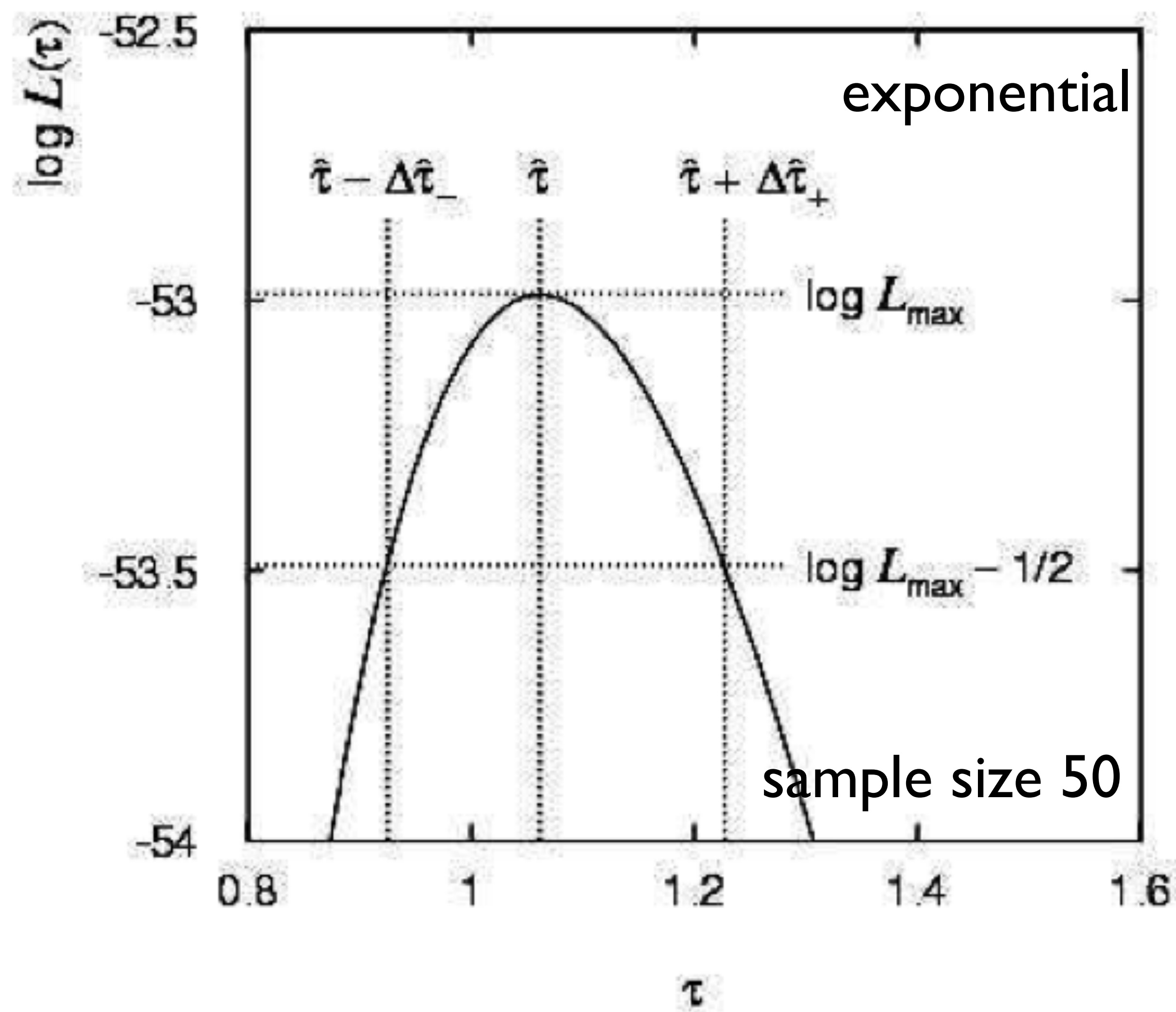
[a]for a distribution w/ 1 fit parameter

# Variance of Estimator

Likelihood is from Lecture 3 and is

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

- The formula can apply to non-Gaussian estimators, i.e. change variables to g(θ) which produces a Gaussian distribution. Likelihood distribution is invariant under parameter transformation.
- If the distribution of the estimated value of τ is asymmetric, as happens for small sample size, then an asymmetric interval about the most likely value may result




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

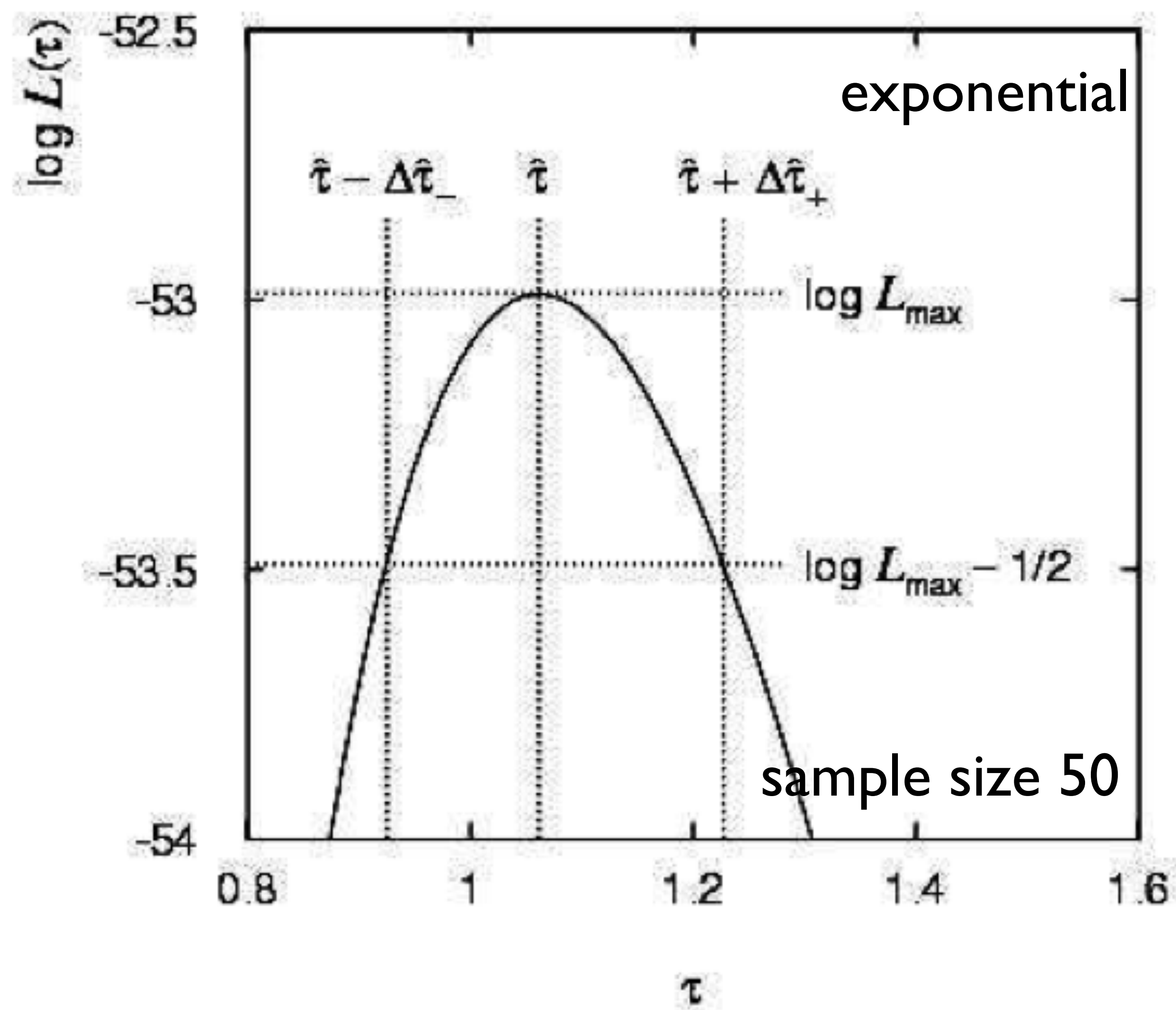
$$\Delta\hat{\tau}_{-} = 0.137$$

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

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

# Variance of Estimator

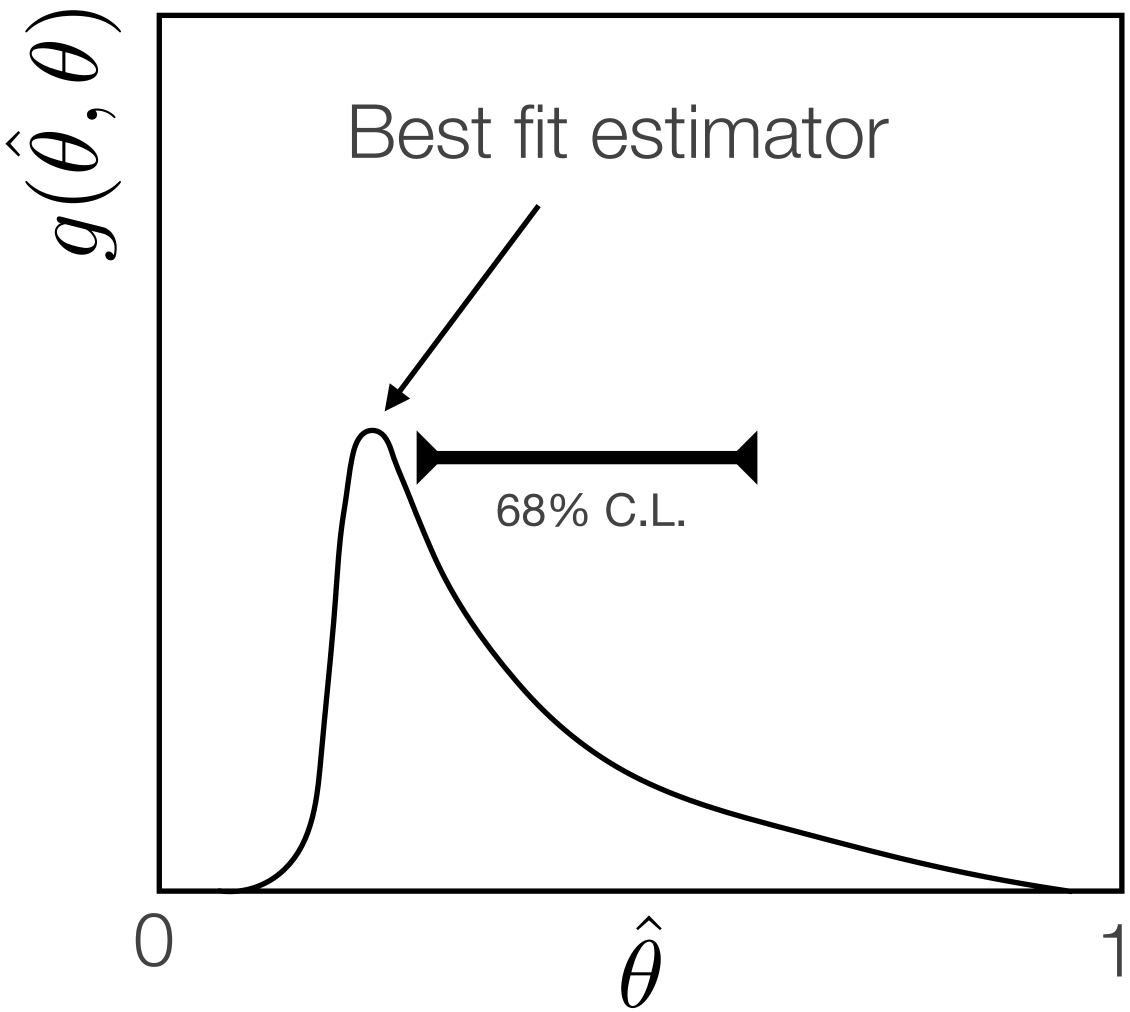
- First, we find the best-fit estimate of $\tau$ via our LLH minimization to get $\hat{\tau}_{best}$
    - Provides LLH($\hat{\tau}_{best}$)=-53.0
    - We could scan to get $\hat{\tau}_{best}$, but it won't be as precise or fast as the minimizer
- We only have 1 fit parameter, so from slide 7 we know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




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

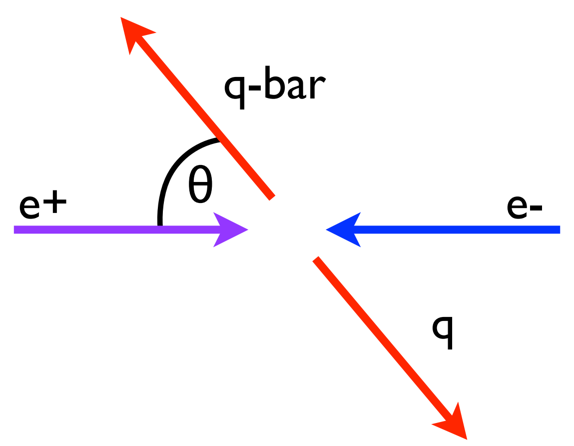
# Reporting Very Asymmetric Central Limits

- Central limits are often reported as $\hat{\theta} \pm \sigma_\theta$ or $\hat{\theta}^{+\sigma_{\theta_1}}_{-\sigma_{\theta_2}}$ 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"



# Variance of Estimators - Graphical Method

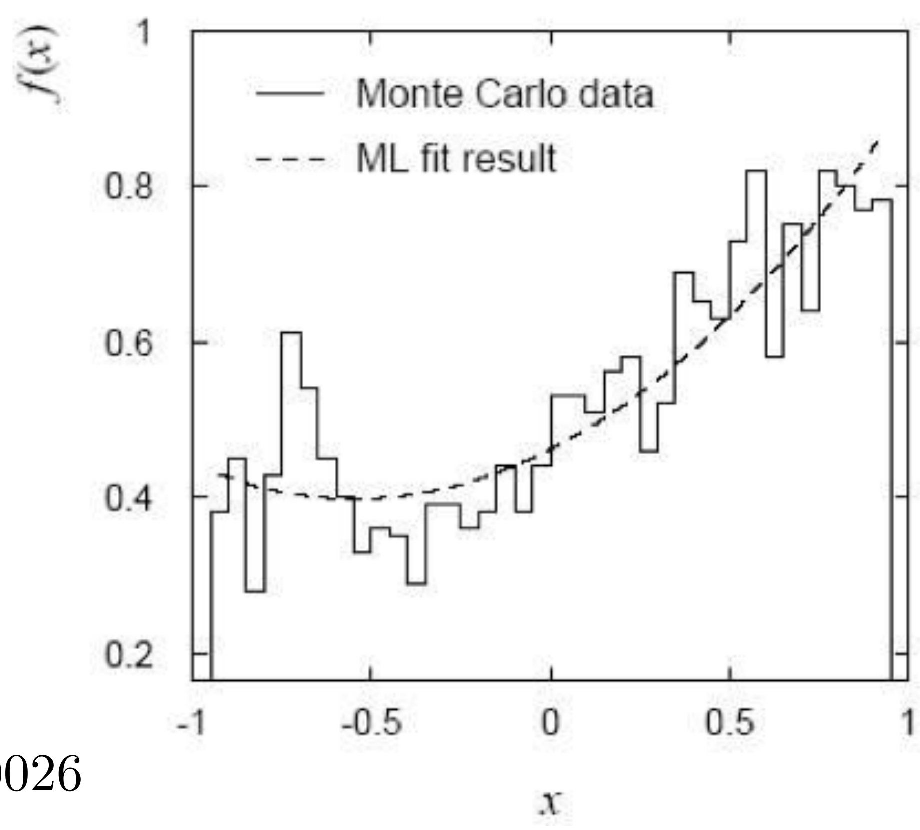
- Consider an example from scattering with an angular distribution given by $x = cos\theta$




- if $x_{min} < x < x_{max}$ then the PDF needs to be normalized:

$$f(x;\alpha,\beta) = \frac{1+\alpha x+\beta x^2}{2+2\beta/3} \qquad \int_{x_{min}}^{x_{max}} f(x;\alpha,\beta)dx = 1$$

- Take the specific example where **α**=0.5 and **β**=0.5 for 2000 points where -0.95 ≤ x ≤ 0.95




- The maximum may be found numerically, giving values $\alpha = 0.508,\ \beta = 0.47$ for the plotted data

- The statistical errors can be estimated by numerically solving the 2nd derivative (shown here for completeness)

$$(\hat{V^{-1}})_{ij} = -\frac{\partial^2 \ln L}{\partial\theta_i\partial\theta_j}|_{\vec{\theta}=\hat{\vec{\theta}}} \qquad \hat{\sigma}_{\hat{\alpha}} = 0.052,\ \hat{\sigma}_{\hat{\beta}} = 0.11,\ cov[\hat{\alpha},\hat{\beta}] = 0.0026$$

# Exercise #1

- Before we use the LLH values to determine the uncertainties for $\alpha$ and $\beta$, let's do it via Monte Carlo first
- Similar to the exercises 2-3 from Lecture 3, the theoretical prediction:

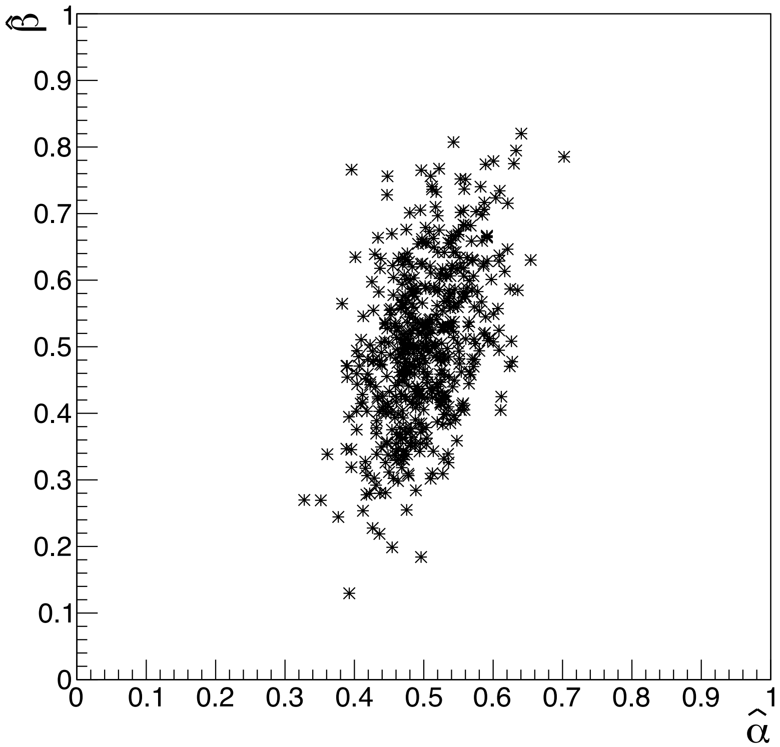
$$f(x;\alpha,\beta) = 1 + \alpha x + \beta x^2$$

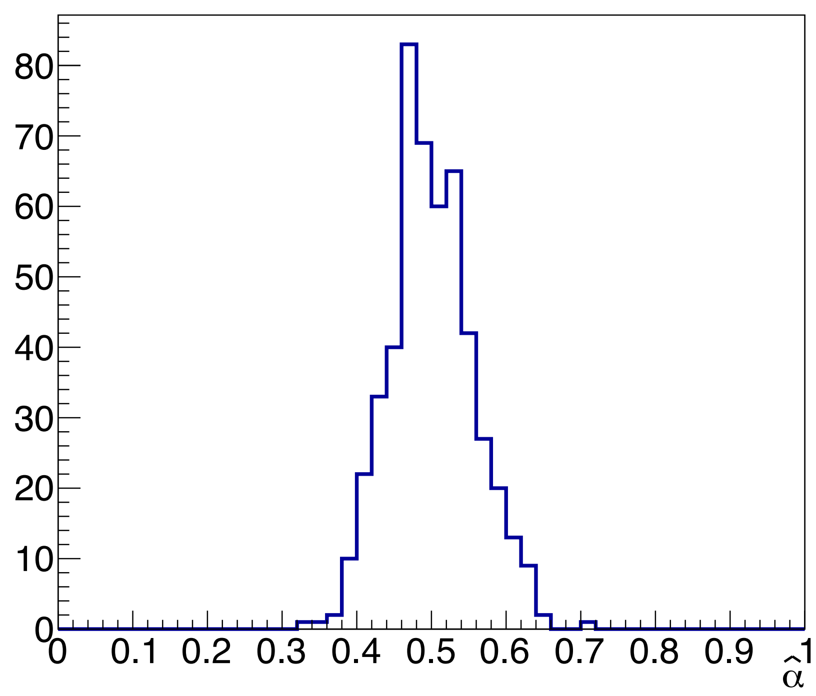
  - For $\alpha$=0.5 and $\beta$=0.5, generate 2000 Monte Carlo data points using the above function transformed into a PDF over the range $-0.95 \leq x \leq 0.95$
    - Remember to normalize the function properly to convert it to a proper PDF
    - Fit the MLE parameters $\hat{\alpha}$ and $\hat{\beta}$ using a minimizer/maximizer
    - Repeat 100 to 500 times plotting the distributions of $\hat{\alpha}$ and $\hat{\beta}$ as well as $\hat{\alpha}$ vs. $\hat{\beta}$

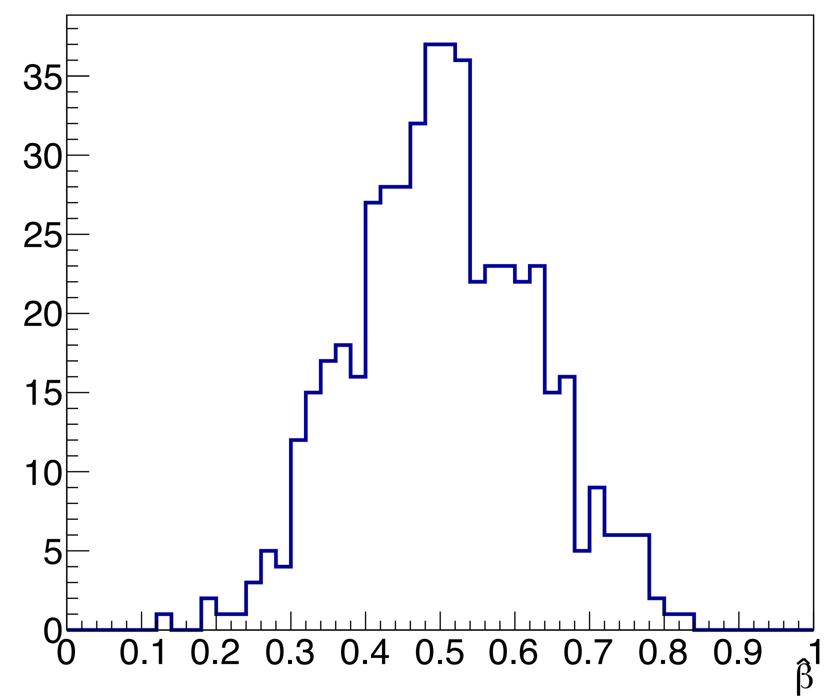
# Exercise #1

- Shown are 500 Monte Carlo pseudo-experiments
- The estimates average to approximately the true values, the variances are close to initial estimates from slide 8 and the estimator distributions are approximately Gaussian

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



# Comments

- 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, the uncertainty is accurate even if the Central Limit Theorem is invalid for your data/parameters
  - 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

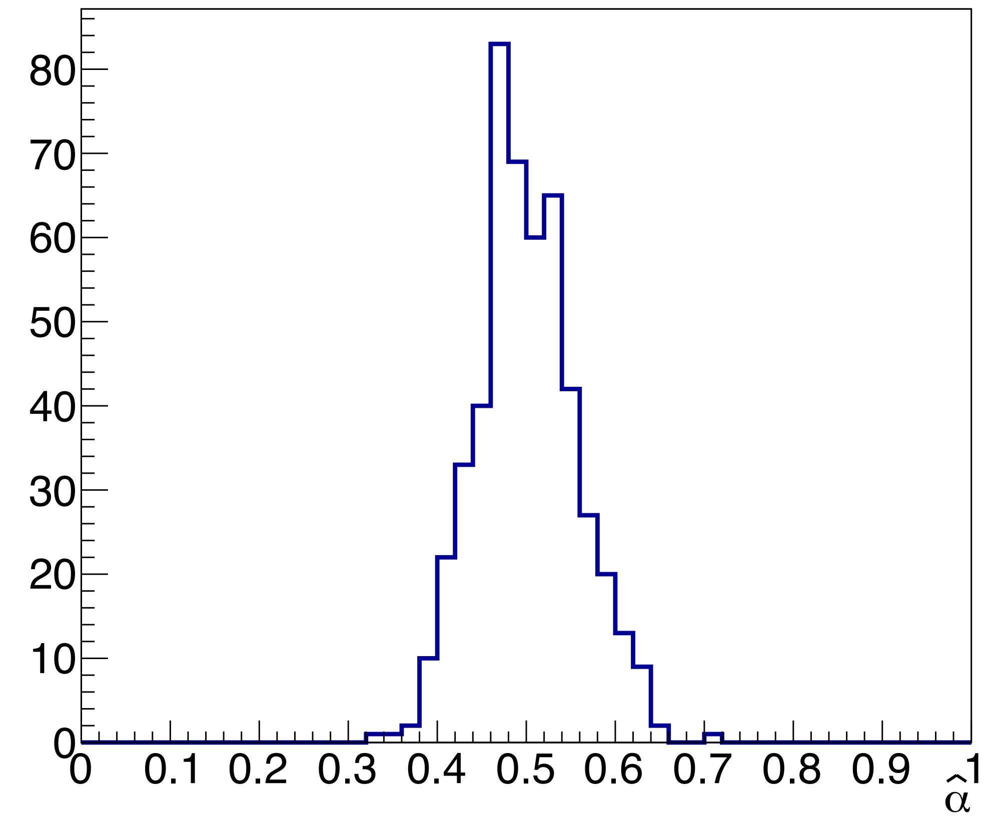
# Brute Force

- If we either did not know, or did not trust, that our estimator(s) are nicely analytic PDFs (gaussian, binomial, poisson, etc.) 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 ±1σ central interval

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

$$\frac{100\% - 68.2\%}{2} = \int_{-\infty}^{C_-} g(\hat{\alpha}; \alpha\hat{_{obs}}) d\hat{\alpha}$$



# Brute Force cont.

- The previous method is known as a parametric bootstrap
  - Overkill for the previous example
  - Useful for estimators which are complicated
- 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 100 or 500 values from the pseudo-experiments, i.e. parametric bootstrapping
    - Find the central 1σ confidence interval(s) for $\hat{\alpha}$ as well as $\hat{\beta}$ using bootstrapping
- Repeat, but now:
    - **Fix** α=0.5, and only fit for β, i.e. α is now a constant
    - What is the new 1σ central confidence interval for $\hat{\beta}$?
- Repeat with a new angular distributions range of the $-0.9 \leq x \leq 0.85$
    - Again, **fix** α=0.5
    - 2000 Monte Carlo 'data' points

# Good?

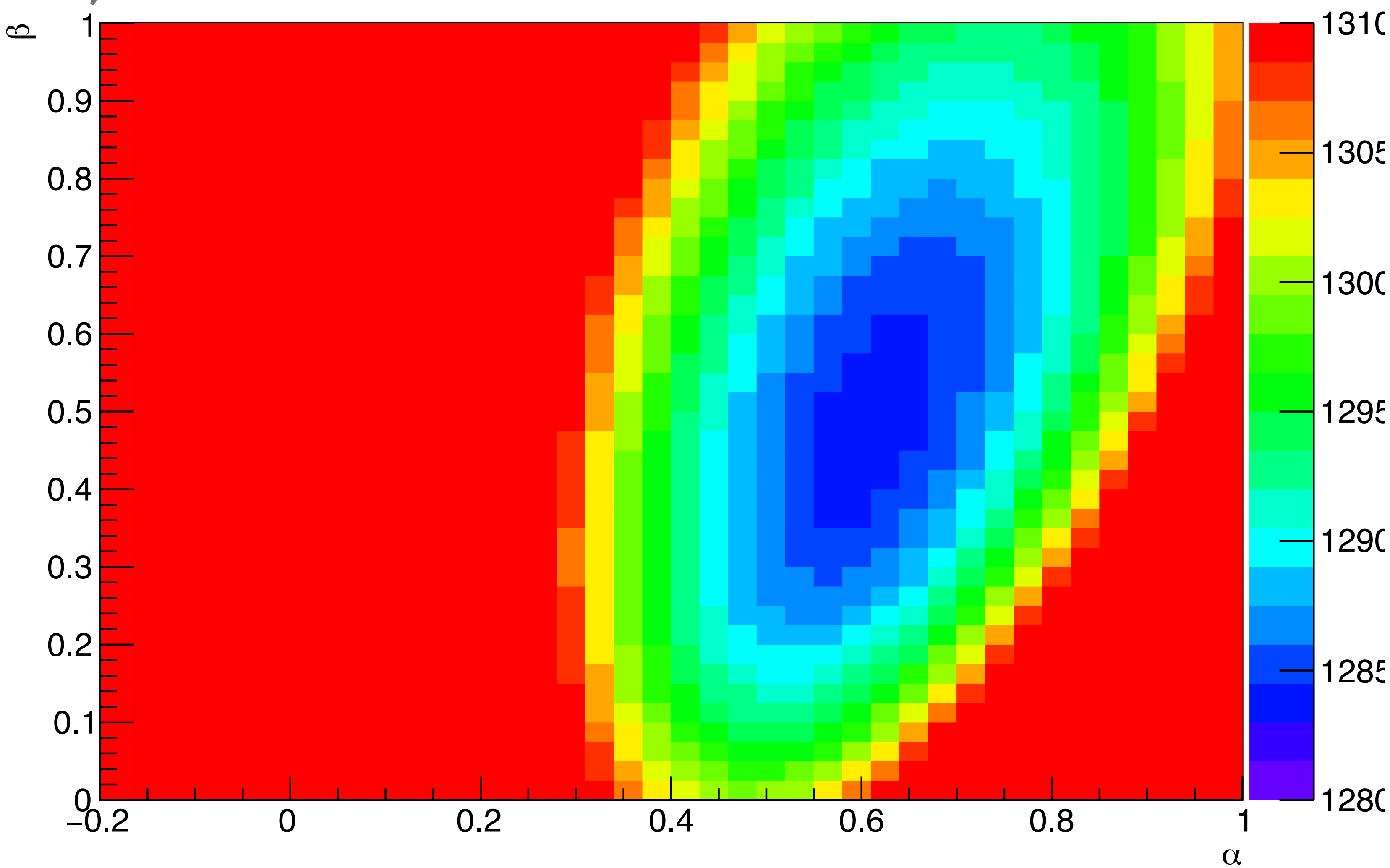
- The LLH minimization will give the best-fit values and often the uncertainty on the estimators. But, likelihood fits do not tell whether the data and the prediction agree
  - Remember that the likelihood has a form (PDF) that is provided by **<u>you</u>** and may not be correct
  - The PDF may be okay,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

# Multi-parameter

- Getting back to LLH confidence intervals
- In one dimension fairly straightforward
  - Confidence intervals, i.e. uncertainty, can be deduced from the LLH difference(s) to the best-fit point(s)
  - 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

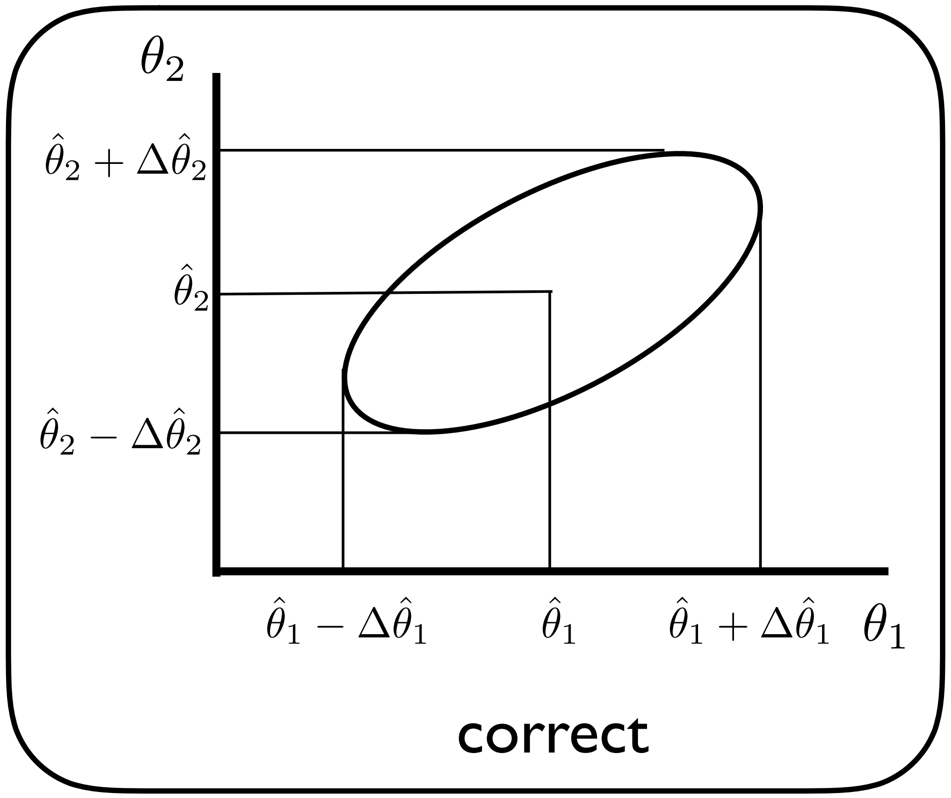
- For 2 dimensions, i.e. 2-parameter fits, we can produce likelihood landscapes. In 3 dimensions a surface, and in 3+ dimensions a likelihood hypersurface.
- The contours are then lines of with a constant value of likelihood or ln(likelihood)

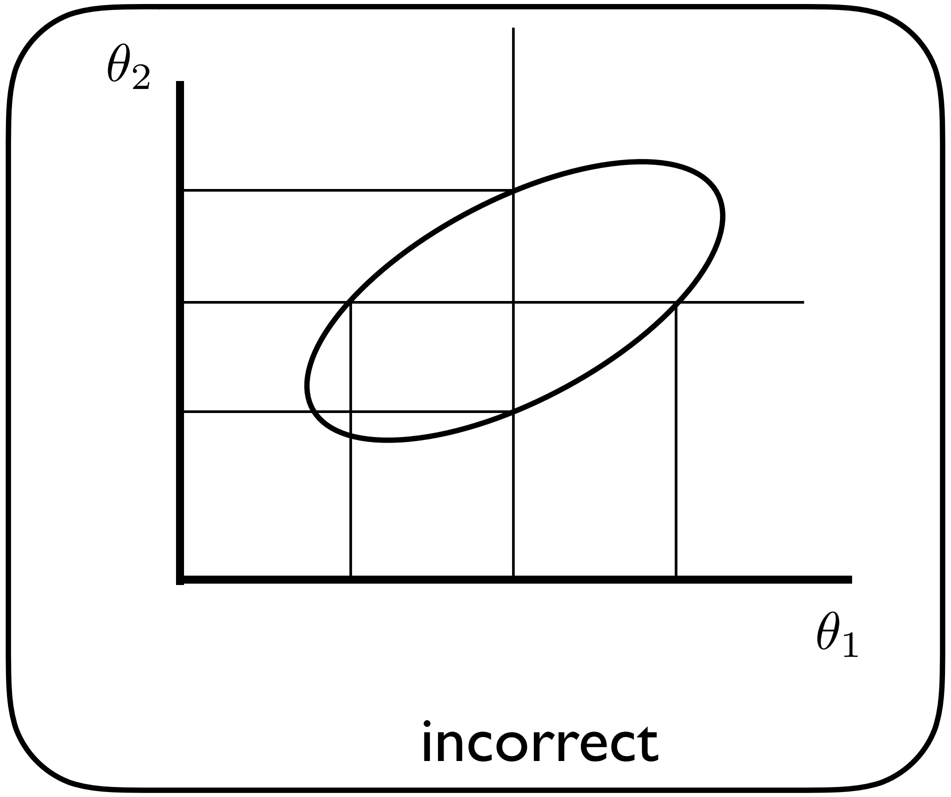



*LLH landscape is from Lecture 3

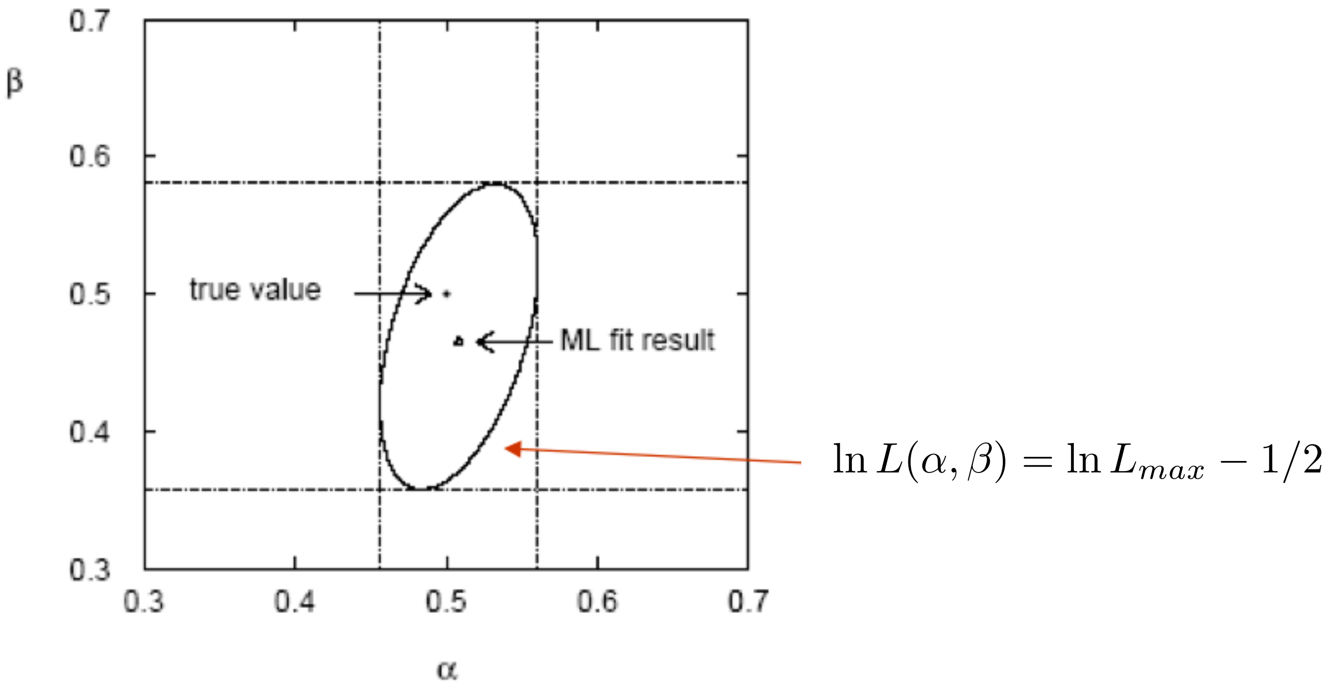
# Variance of Estimators - Graphical Method

- Two Parameter Contours



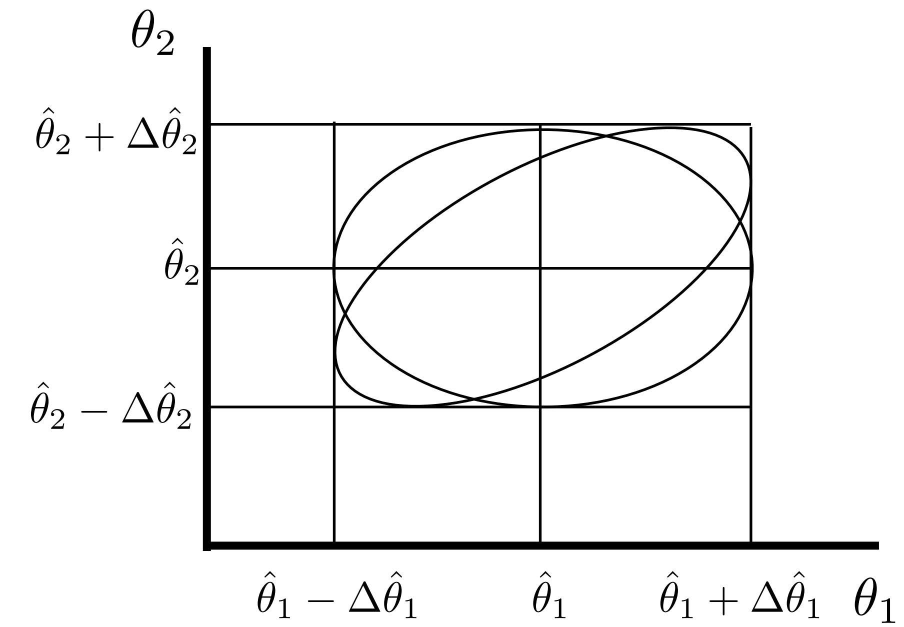




- Tangent lines to the contours give the standard deviations



# Variance of Estimators - Graphical Method

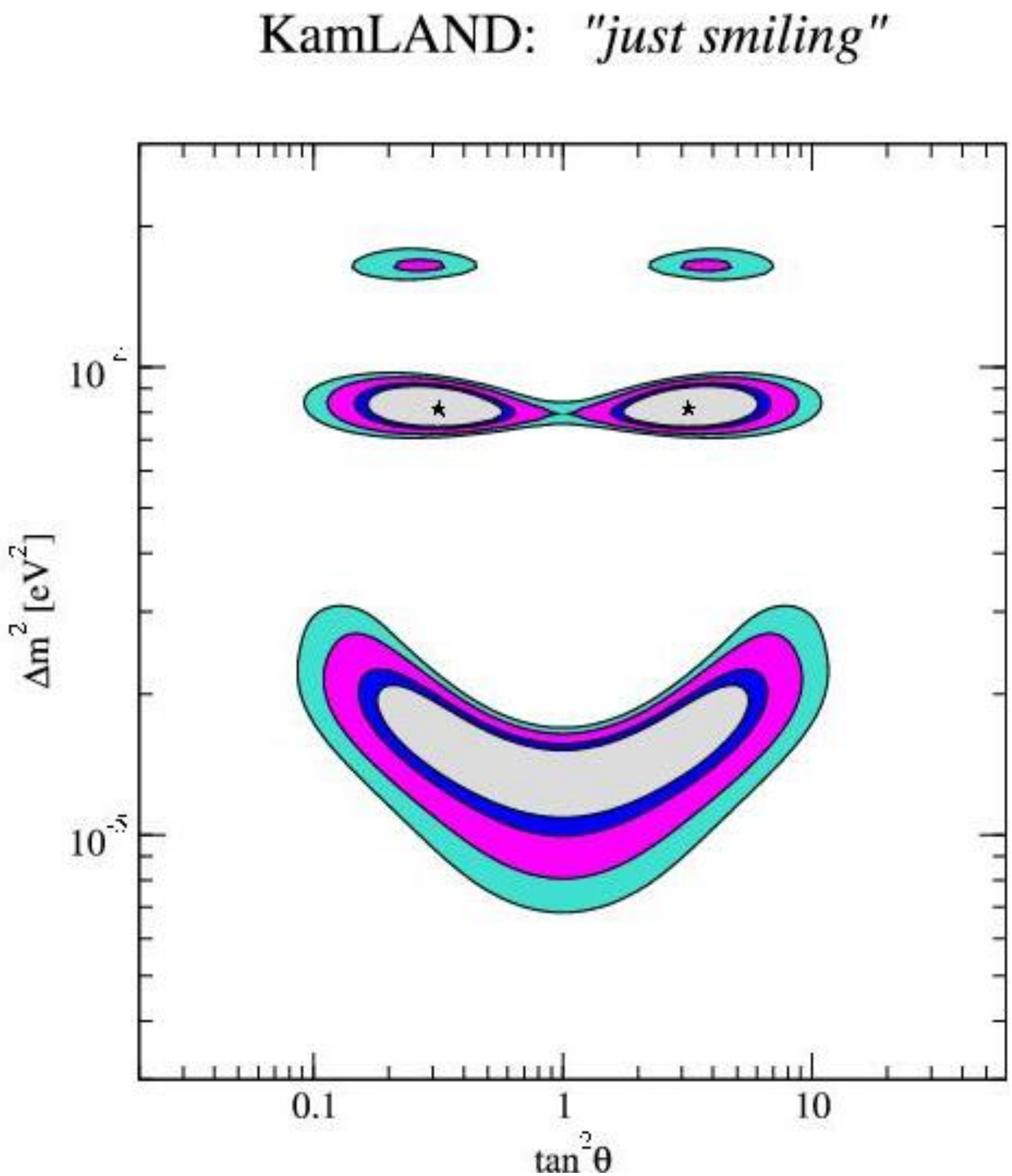
- When the correct, tangential, method is used then the uncertainties are not dependent on the correlation of the variables.
- The probability the ellipses of constant $\ln L = \ln L_{max} - a$ contains the true point, $\theta_1$ and $\theta_2$ , is:




**correct**

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

# Best Result Plot?



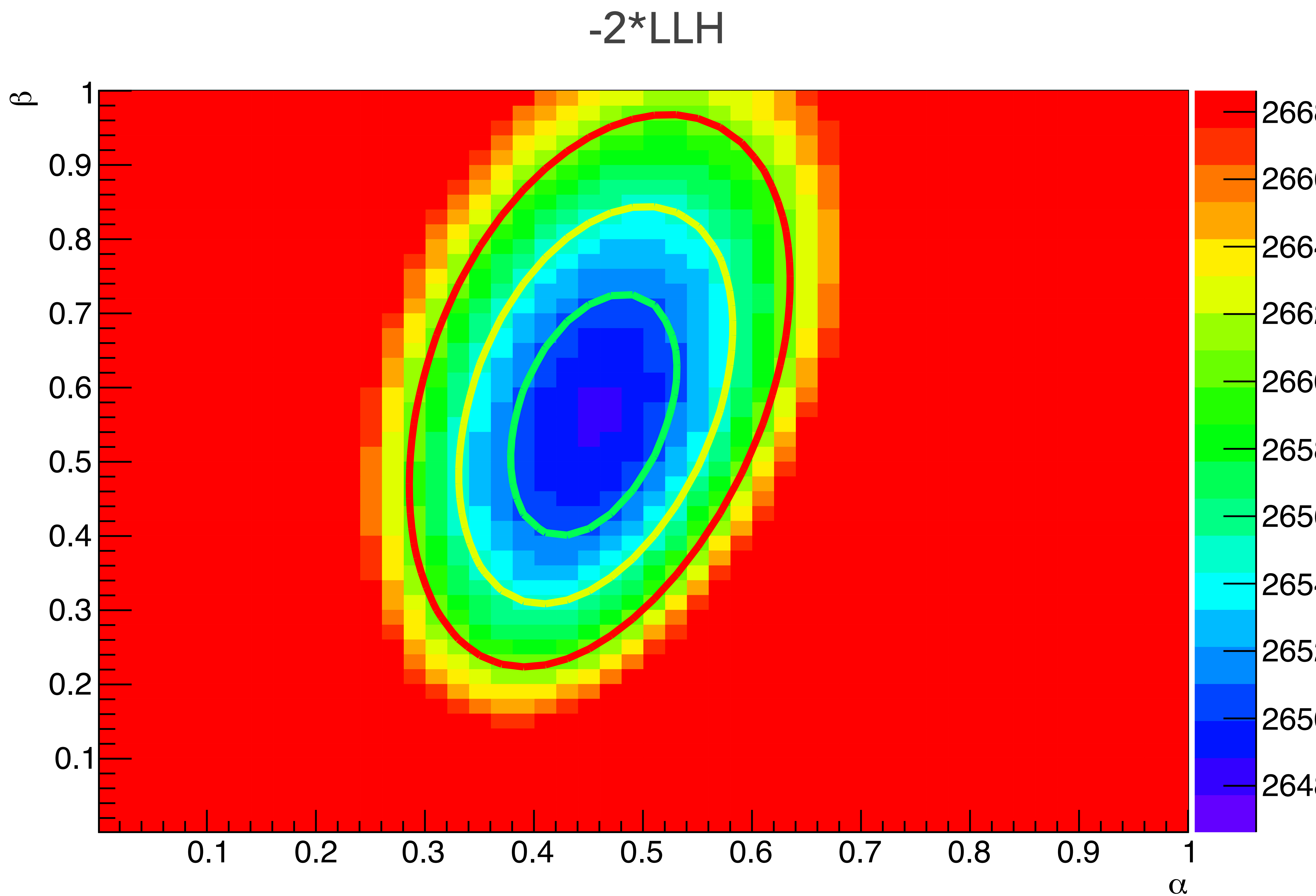
# Variance/Uncertainty - Using LLH Values

- 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

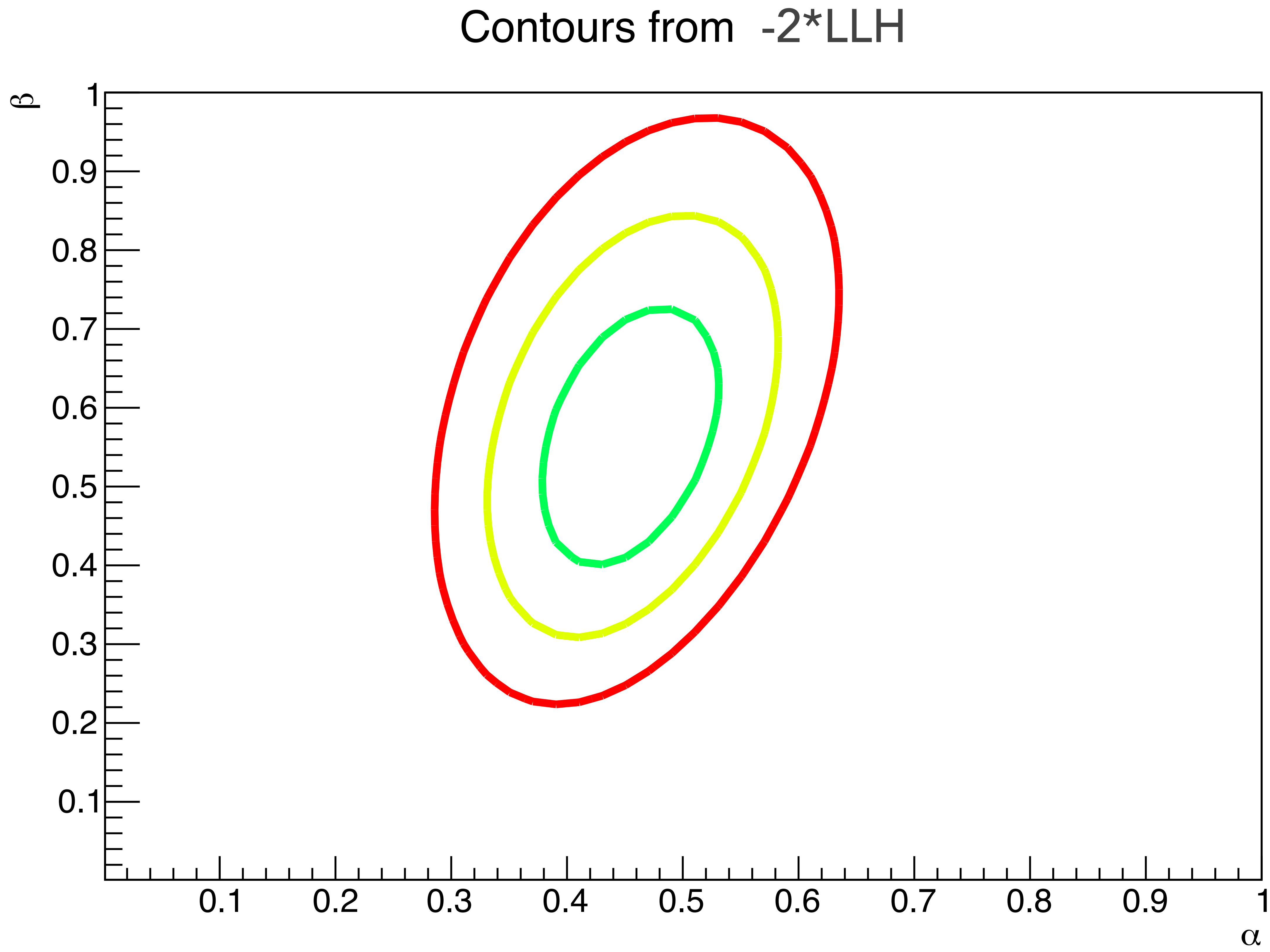
# Exercise #2

- Using the same function and **$\alpha$**=0.5 and **$\beta$**=0.5 as Exercise #1, find the MLE values for a single Monte Carlo sample w/ 2000 points
- Plot the contours related to the 1$\sigma$, 2$\sigma$, and 3$\sigma$ confidence regions
  - Remember that this function has 2 fit parameters
  - Because of different random number generators, your result is likely to vary from mine
- Calculate a goodness-of-fit
  - For a quick calculation a reduced chi-square might be enough, but it is better to quote the goodness-of-fit, i.e. p-value assuming gaussian estimator w/ a fixed **$\alpha$** and/or **$\beta$**
  - E.g. use a reduced chi-squared and convert to a goodness-of-fit value

# Contours on Top of the LLH Space



# Just the Contours

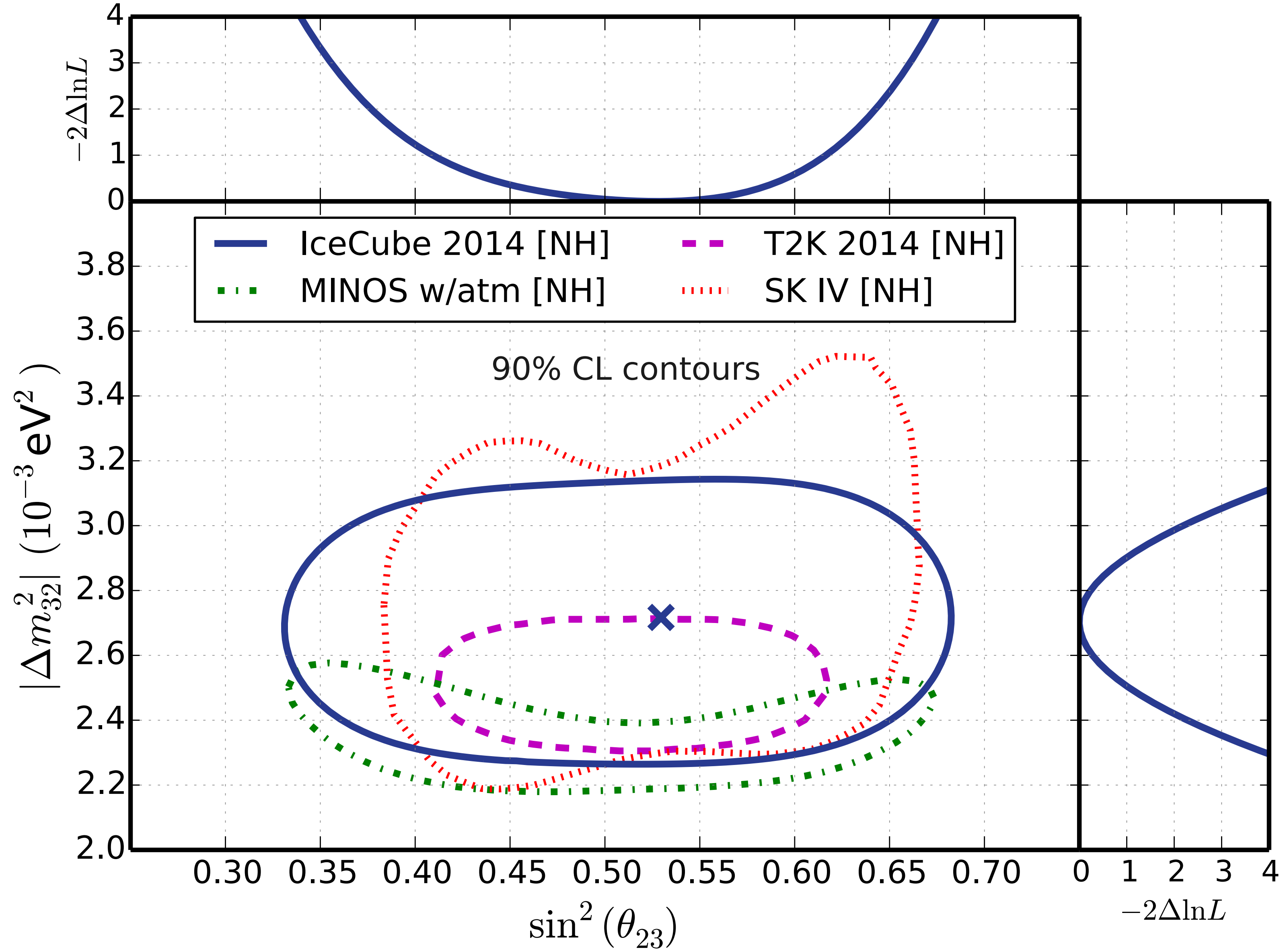


# Real Data

- 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 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 $-1 \leq x \leq 1$
- Using the function:

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

  - Find the best-fit for the unknown **α** and **β**
  - 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 exercise #1, show that Monte Carlo re-sampling produces similar uncertainties as the ΔLLH prescription for the 3D hyper-ellipse
  - 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

# Extra Extra

- Use Markov Chain to get the likelihood minimum and then use the LLH (or -2*LLH) values to get the uncertainties.
    - Is the MCMC quicker to converge to the 'best-fit' than using your LLH minimizer?
    - The Markov Chain estimator (maximum a posteriori - MAP) has a precision on the variance of $\mathcal{O}(1/n)$ for n simulation points, i.e. you can't get 99.9% interval without at least 1000 MCMC 'steps' after convergence. With a flat prior and using the 3-dimensional function the variance with an MCMC posterior distribution, do the best-fit values and uncertainties match what you get for the ΔLLH approach
        - Use the same 2000 data points for consistency from a single pseudo-experiment
        - Flat prior does not impact the $\mathcal{O}(1/n)$ variance, but just makes it easier to compare to the results already derived using the ΔLLH formulation for uncertainty