Lecture 3 : Hypothesis testing and model-fitting

These lectures

- Lecture 1 : basic descriptive statistics
- Lecture 2 : searching for correlations
- Lecture 3 : hypothesis testing and model-fitting
- Lecture 4 : Bayesian inference

Lecture 3 : hypothesis testing and model-fitting

- Goodness-of-fit : chi-squared probability distribution
- What is meant by the p-value (probability value)?
- Parameter estimation
- Marginalization of parameters
- Confidence limits, skewed distributions
- Is adding another parameter justified by the data?

Objective

- When comparing data and models we are typically doing one of two things :

- Hypothesis testing : we have a set of N measurements $x_i \pm \sigma_i$ which a theorist says should have values μ_i. How probable is it that these measurements would have been obtained, if the theory is correct?

- Parameter estimation : we have a parameterized model which describes the data, such as $y = a\,x + b$, and we want to determine the best-fitting parameters and errors in those parameters

The chi-squared statistic

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \qquad \begin{array}{l} \text{Data } \; x_i \pm \sigma_i \\ \text{Model } \; \mu_i \end{array}$$

- The chi-squared statistic is a measure of the goodness-of-fit of the data to the model
- We penalize the statistic according to how many standard deviations each data point lies from the model
- If the data are numbers taken as part of a counting experiment we can use a Poisson error $\sigma_i^2 = \mu_i$
- [Small print : this equation assumes the data points are independent]

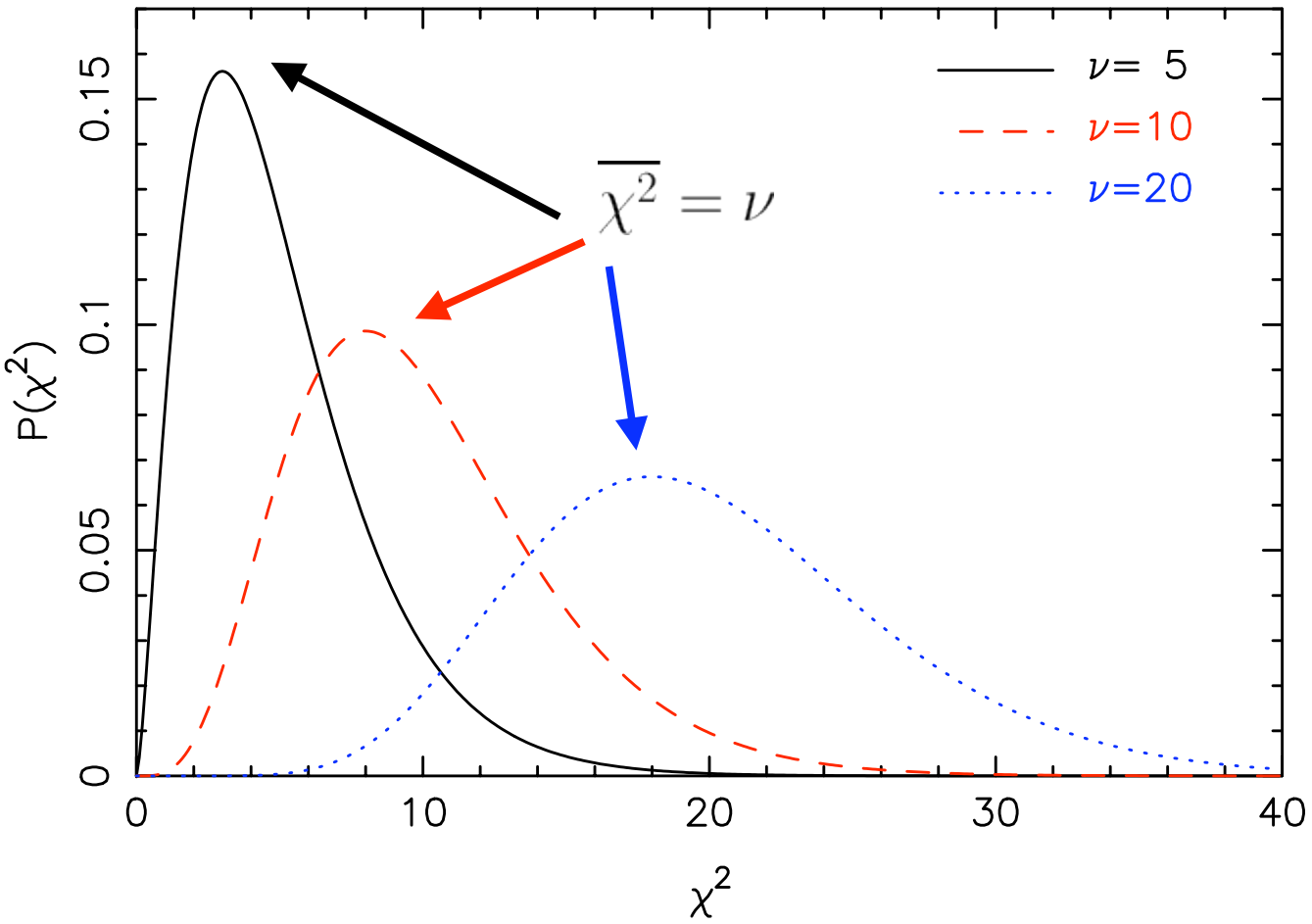
Chi-squared probability distribution

$$P(\chi^2) \propto (\chi^2)^{\frac{\nu-2}{2}} \exp(-\chi^2/2)$$

- Probability distribution if the model is correct
  [Small print : this assumes the variables are Gaussian-distributed]
- ν is number of degrees of freedom
- If the model has no free parameters then $\nu = N$
- If we are fitting a model with p free parameters, we can "force the model to exactly agree with p data points" and $\nu = N - p$

Chi-squared probability distribution

$$P(\chi^2) \propto \left(\chi^2\right)^{\frac{\nu-2}{2}} \exp\left(-\chi^2/2\right)$$



Chi-squared probability distribution

$$P(\chi^2) \propto \left(\chi^2\right)^{\frac{\nu-2}{2}} \exp\left(-\chi^2/2\right)$$

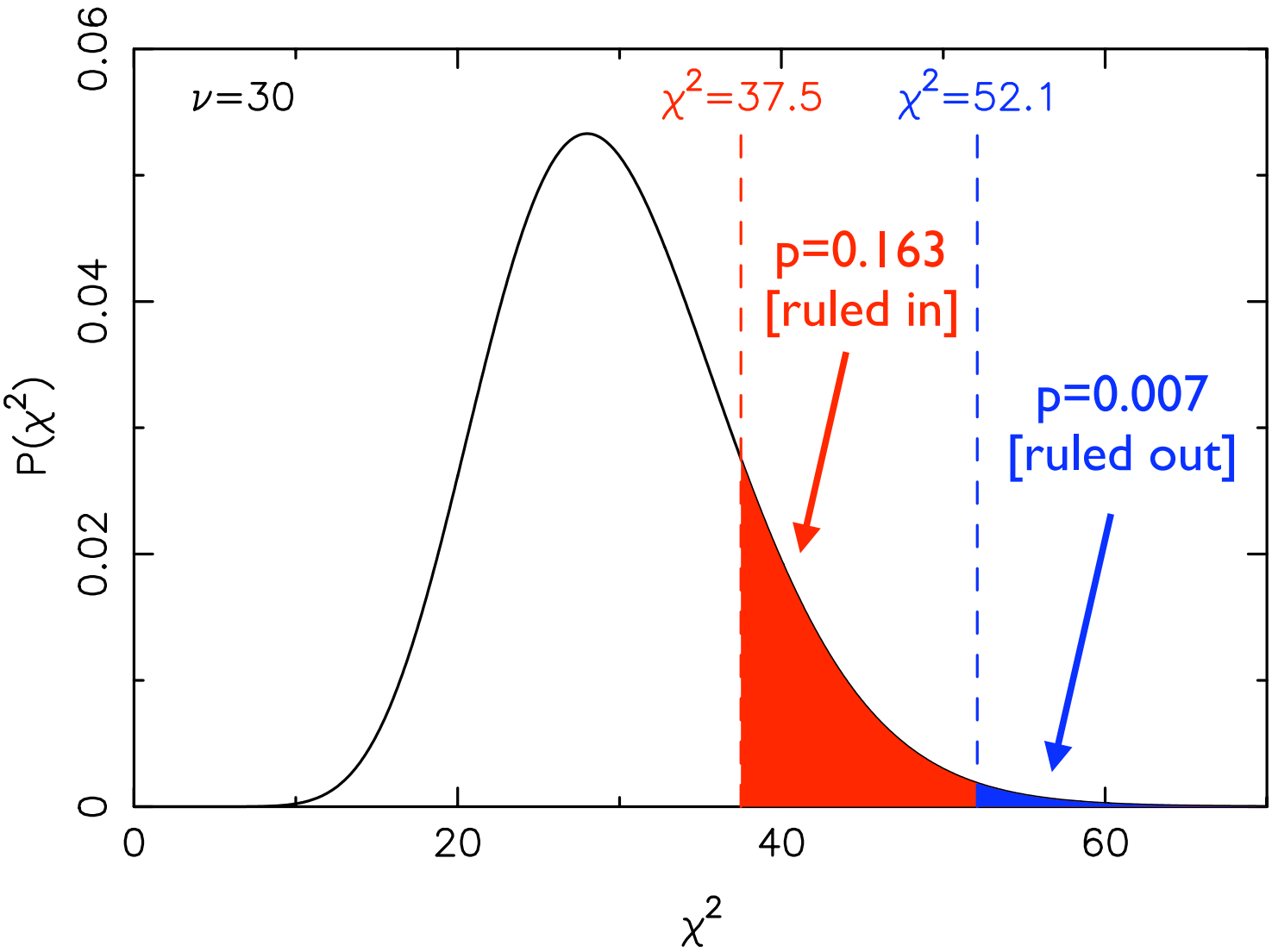
- Mean : $\overline{\chi^2} = \nu = N - p$
- Variance : $\mathrm{Var}(\chi^2) = 2\,\nu$
- If the model is correct we expect : $\chi^2 \sim \nu \pm \sqrt{2\nu}$
- Makes intuitive sense because each data point should lie about ~1-sigma from the model and hence contribute 1.0 to the chi-squared statistic

Hypothesis testing with chi-squared

- We ask the question : if the model is correct [our hypothesis], what is the probability that this value of chi-squared, or a larger one, could arise by chance
- This probability is called the p-value and may be calculated from the chi-squared distribution
- If the p-value is not low, then the data are consistent with being drawn from the model, which is "ruled in"
- If the p-value is low, then the data are not consistent with being drawn from the model. The model is "ruled out" in some sense

Hypothesis testing with chi-squared

- Example 1 on the problem set



Hypothesis testing with chi-squared

- Note that we are assuming the errors in the data are Gaussian and robust
- If the errors have been under-estimated then an improbably high value of chi-squared can be obtained
- If the errors have been over-estimated then an improbably low value of chi-squared can be obtained
- Since errors can sometimes be non-Gaussian or not robust, a model is typically only rejected for very low values of p such as 0.001

Hypothesis testing with chi-squared

- As a way of summarizing the model fit we can quote the reduced chi-squared χ^2/ν
- For a good fit $\chi^2/\nu \sim 1$ [because $\overline{\chi^2} = \nu$]
- However, the true probability of the data being consistent with the model depends on both (χ^2, ν)
- Do not just quote the reduced chi-squared

Hypothesis testing with chi-squared

- If variables are correlated, modify chi-squared equation

$$\chi^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} y_i \left(C^{-1}\right)_{ij} y_j \qquad y_i = x_i - \mu_i$$

- C is the covariance error matrix of the data

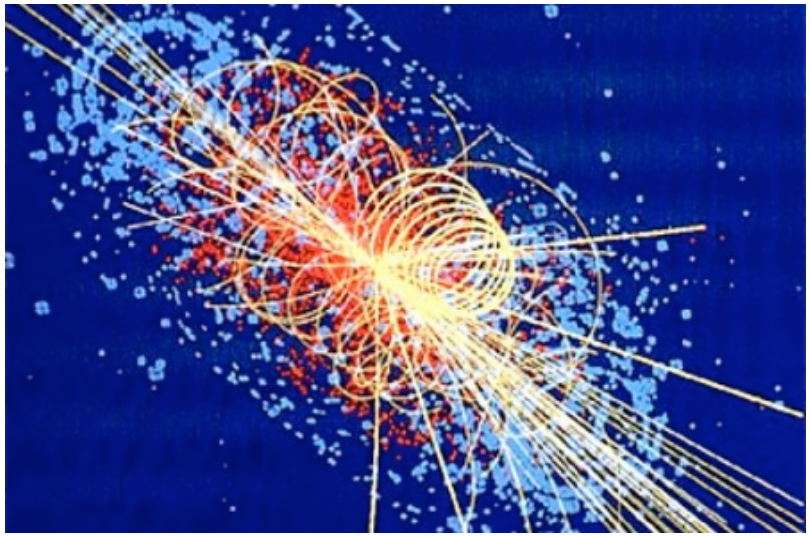
$$C_{ij} = < x_i x_j > - < x_i >< x_j >$$

j=i : $$C_{ii} = < x_i^2 > - < x_i >^2 = \mathrm{Var}(x_i)$$

- The number of degrees of freedom is unchanged for anything less than complete correlation!

Lies, damn lies and statistics

Example 5



"The new particle is very near to the 5-sigma level of significance - meaning that there is less than one in a million chance that their results are a statistical fluke" [The Independent, July 2012]

Why was this poor statistics? The p-value quoted by the LHC experiments is not the probability the Higgs particle doesn't exist. It is the probability of obtaining the measurement assuming the Higgs doesn't exist.

Hypothesis testing with chi-squared

- Suppose a chi-squared hypothesis test yields p=0.01
- This means : there is a 1% chance of obtaining a set of measurements at least this discrepant from the model, assuming the model is true. It does not mean :
- "the probability that the model is true is 1%"
- "the probability that the model is false is 99%"
- "if we reject the model there is a 1% chance that we would be mistaken"
- Frequentist statistics cannot assess the probability that the model itself is correct [see next lecture]

Hypothesis testing with chi-squared

- An issue : using the chi-squared statistic for hypothesis testing often involves binning of data
- For example, suppose we have a sample of galaxy luminosities. To compare the data with a Schechter function we would bin it into a luminosity function
- Warning : the binning of data loses information, can cause bias if the bin sizes are too large compared to changes in the function, and if the numbers in each bin is too small the probabilities can become non-Gaussian
- [As a rule of thumb, 80% of bins must have $N > 5$]

Parameter estimation

- A model typically contains free parameters. How do we determine the most likely values of these parameters and their error ranges?
- Suppose we are fitting a model with 2 free parameters (a,b)
- The most likely ("best-fitting") values of (a,b) are found by minimizing the chi-squared statistic
- The joint error distribution of (a,b) can be found by calculating the values of chi-squared over a grid of (a,b), where the grid spans a parameter range much wider than the eventual errors

Parameter estimation

- We plot 2D contours of constant $\chi^2 = \chi^2_{\rm min} + \Delta\chi^2$
- A joint confidence region for (a,b) can be defined by the zone which satisfies $\chi^2 < \chi^2_{\rm min} + \Delta\chi^2$
- The values of $\Delta\chi^2$ depend on the number of variables and confidence limits, e.g. for 2 variables :

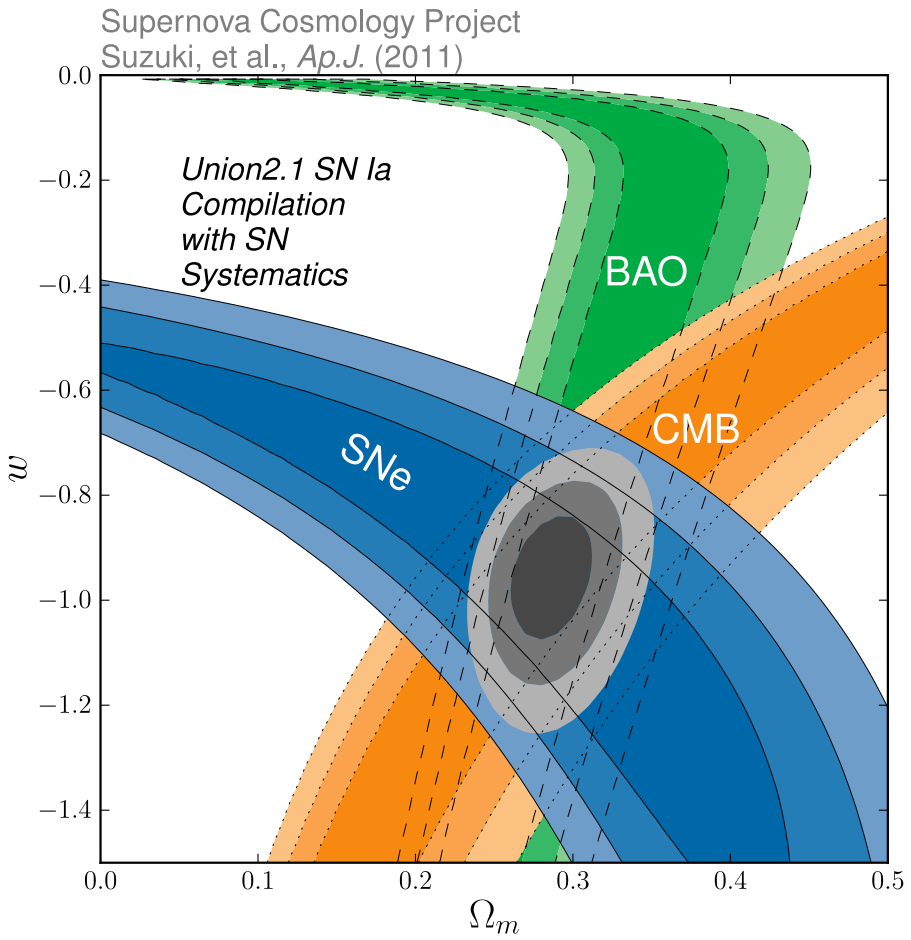
$$68\%\ \text{confidence}: \quad \Delta\chi^2 = 2.30$$
$$95\%\ \text{confidence}: \quad \Delta\chi^2 = 6.17$$

- [Small print : assumes the variables are Gaussian-distributed]

Parameter estimation

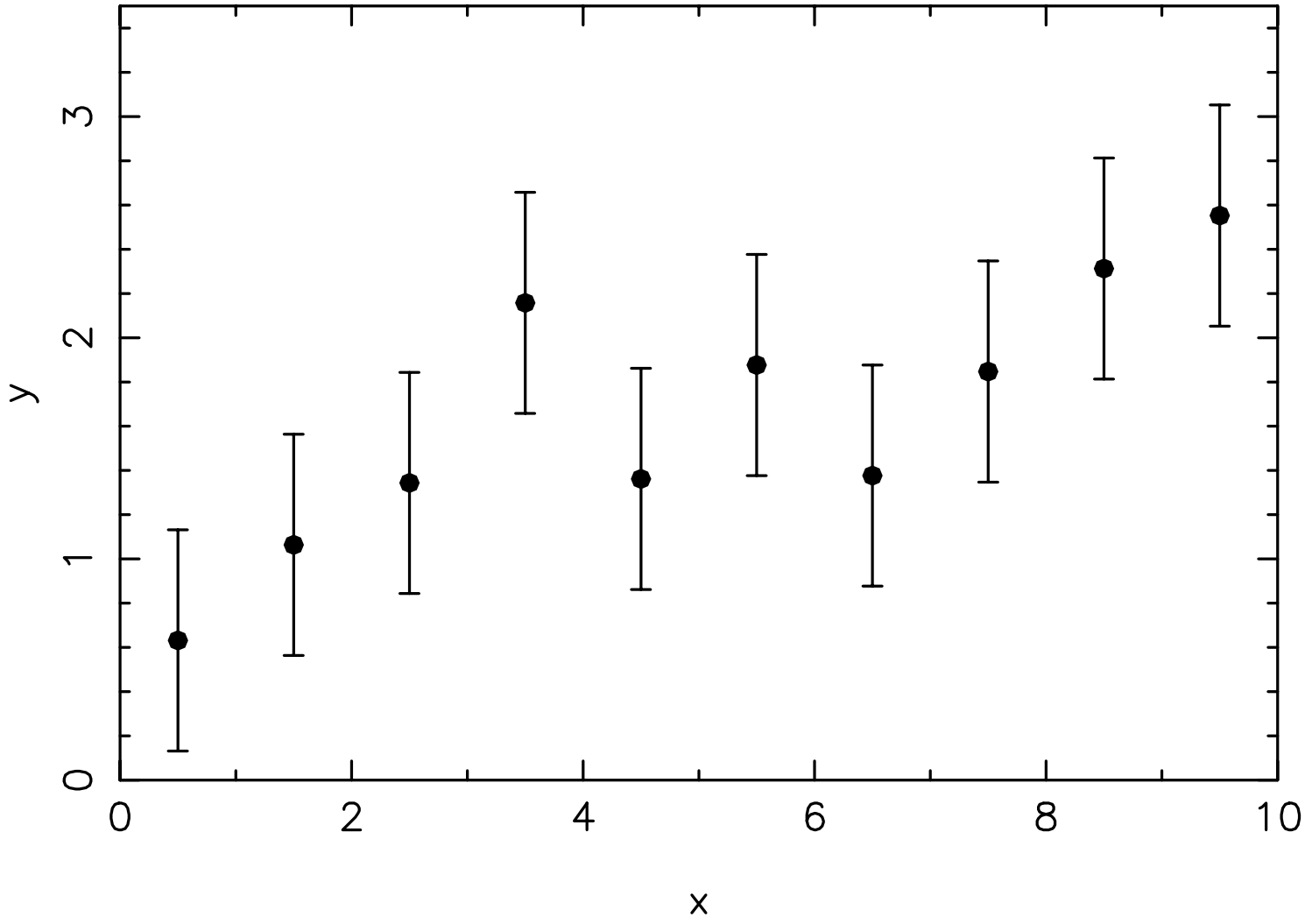
- Warning : Levenberg-Marquardt method (often used to minimize chi-squared for a non-linear model) returns an error in the parameters : treat cautiously

- This error is based on an elliptical Gaussian approximation for the likelihood at the minimum



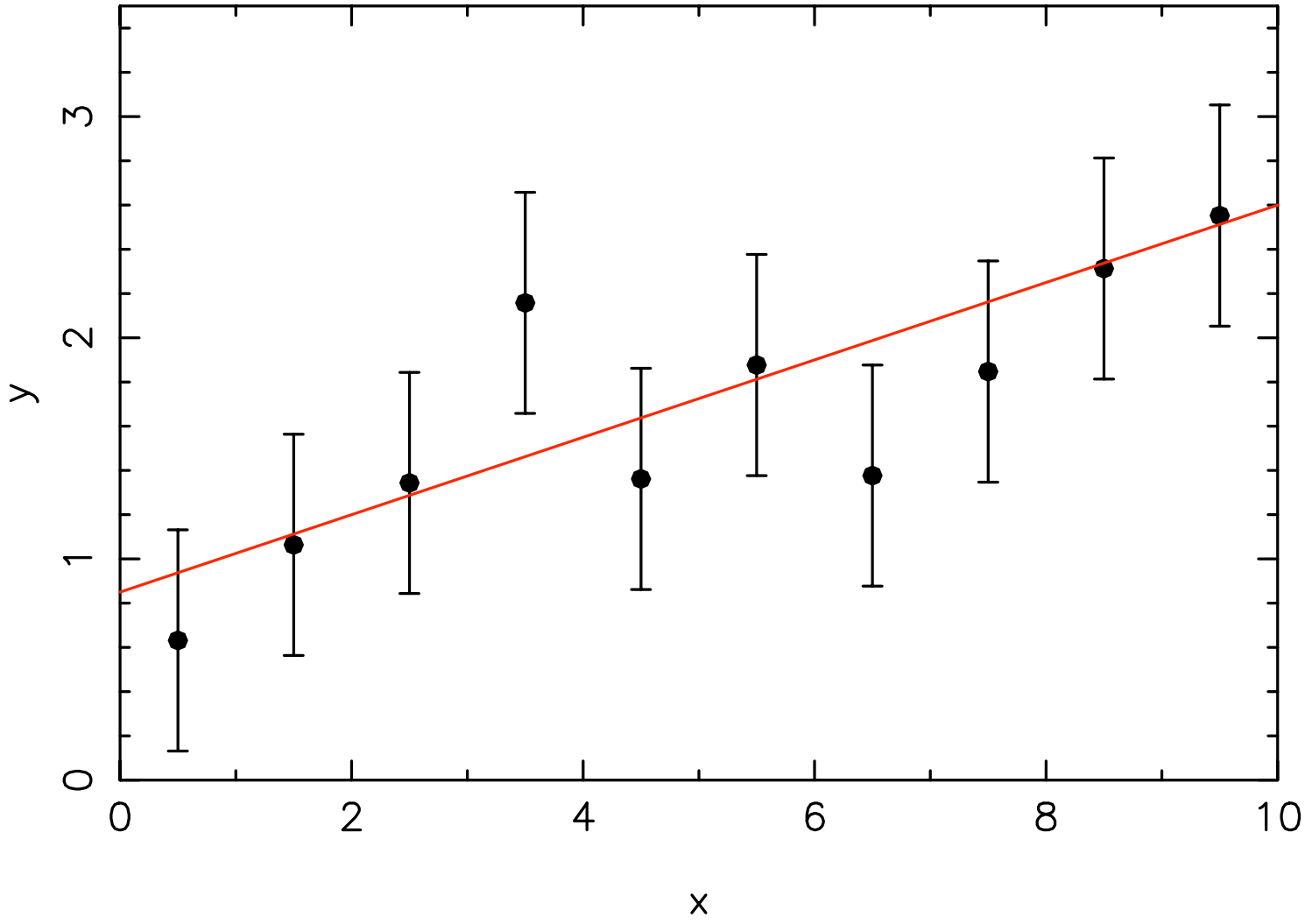
Parameter estimation

- Example 2 : fit model y = a x + b to this dataset :



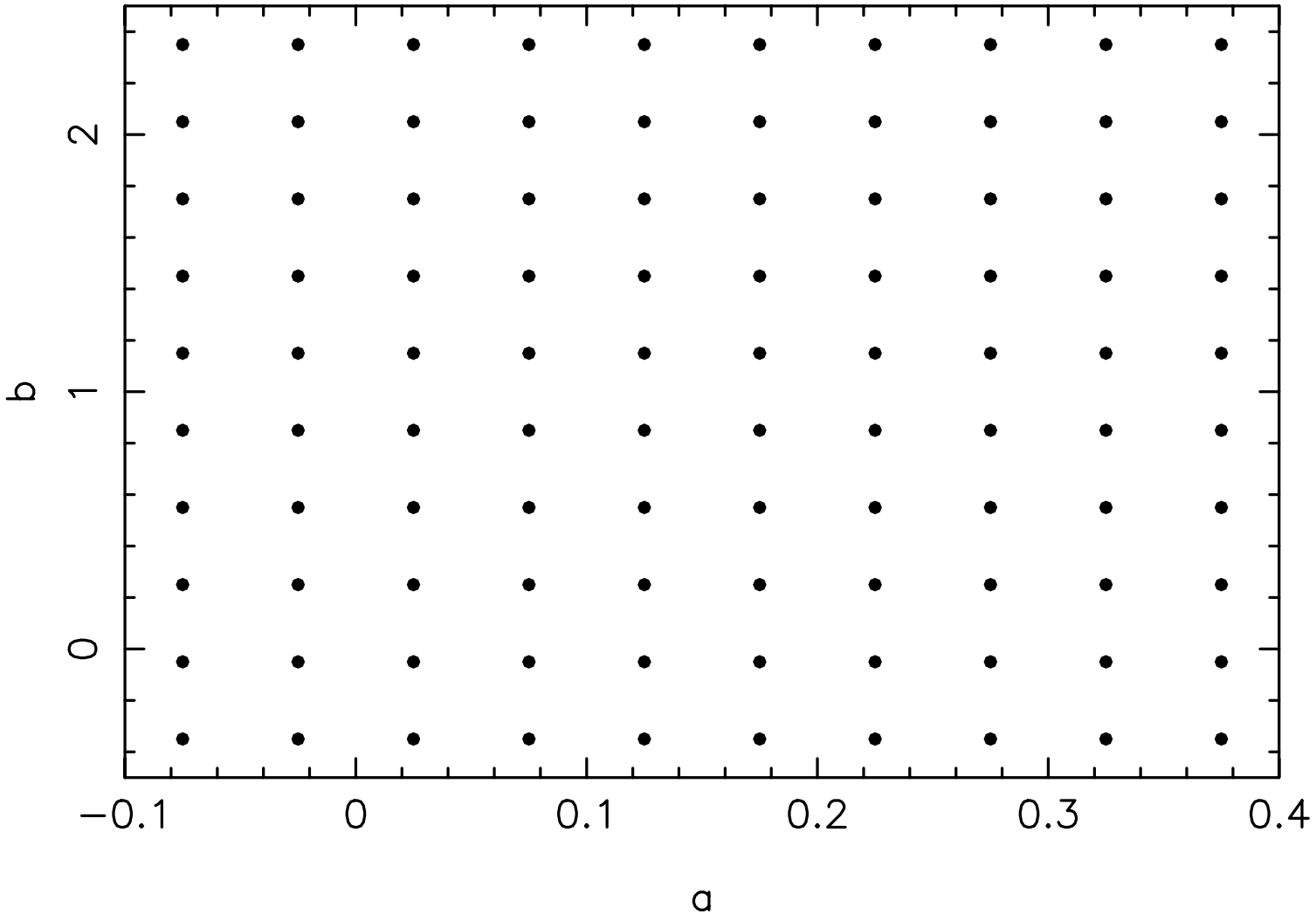
Parameter estimation

- $a_{\rm best} = 0.16 \quad b_{\rm best} = 0.83 \quad \chi^2_{\rm min} = 4.28 \quad \nu = 8 \quad {\rm Prob} = 0.83$



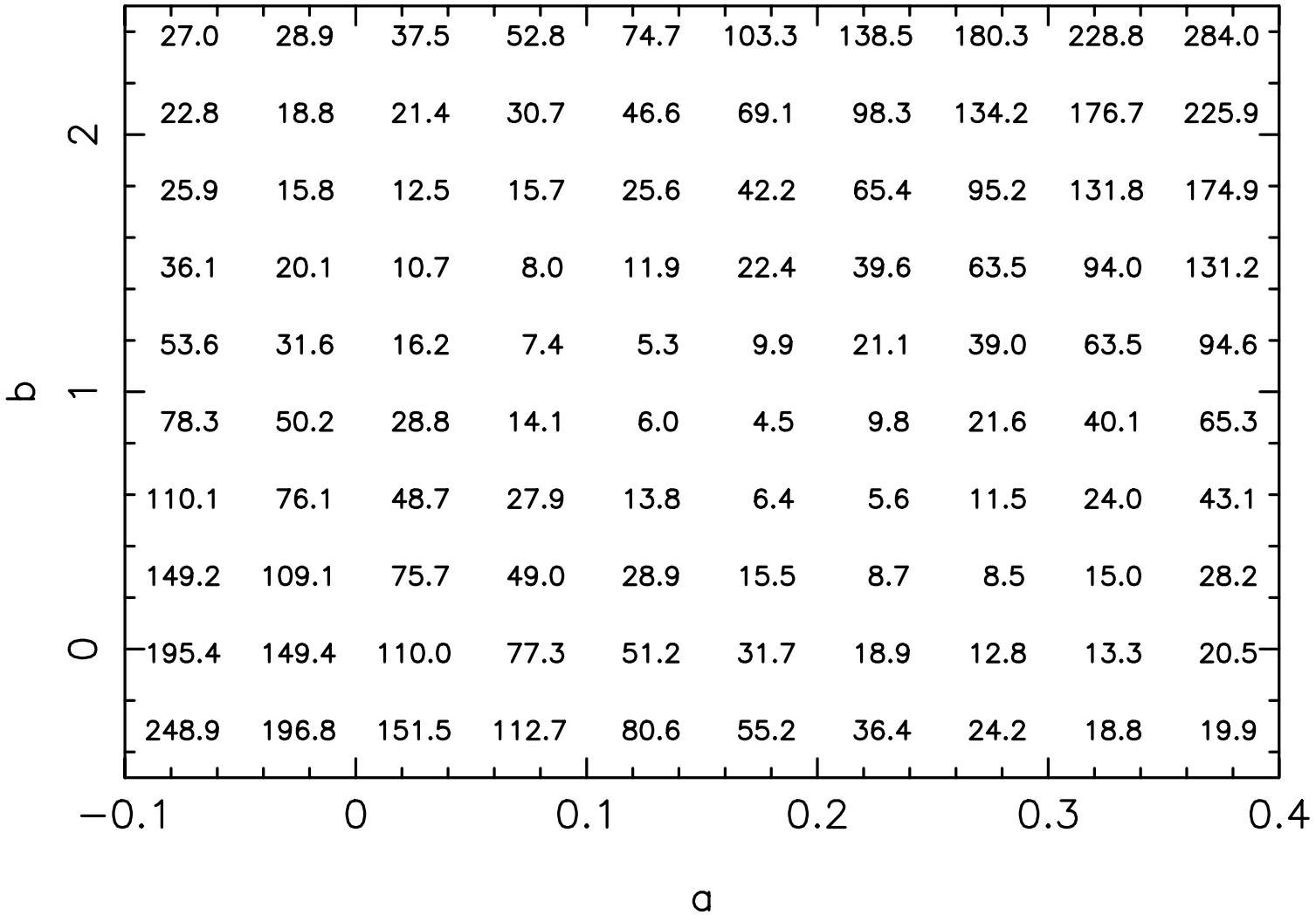
Parameter estimation

- Determine chi-squared for a grid of parameters



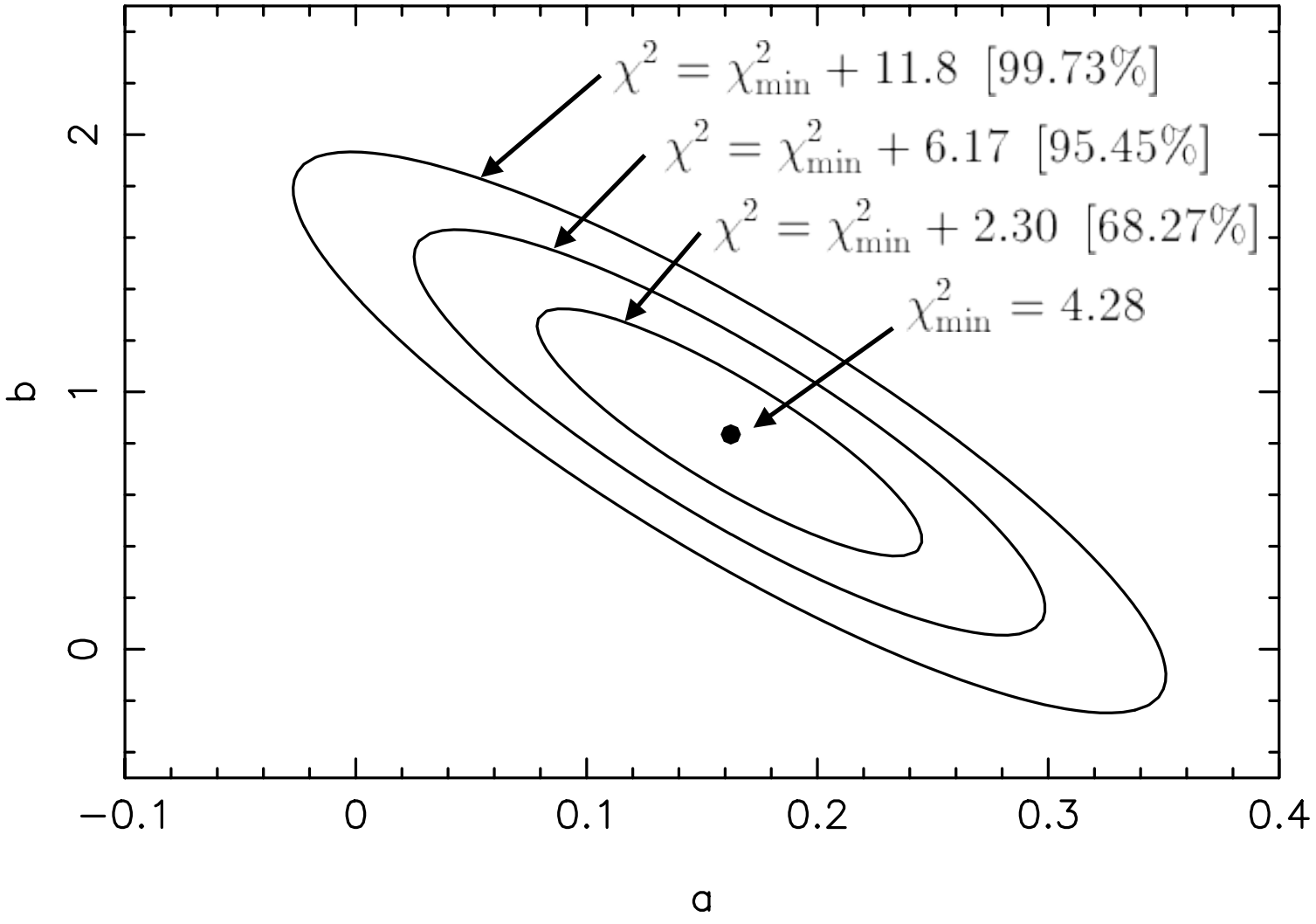
Parameter estimation

- Determine chi-squared for a grid of parameters



Parameter estimation

- Contours of constant chi-squared



Marginalization of parameters

- What is the probability distribution for parameter a, considering all possible values of parameter b? [This is known as marginalization of parameter b]

- For Gaussian variables: $\text{Likelihood} \propto \exp\left(-\chi^2/2\right)$

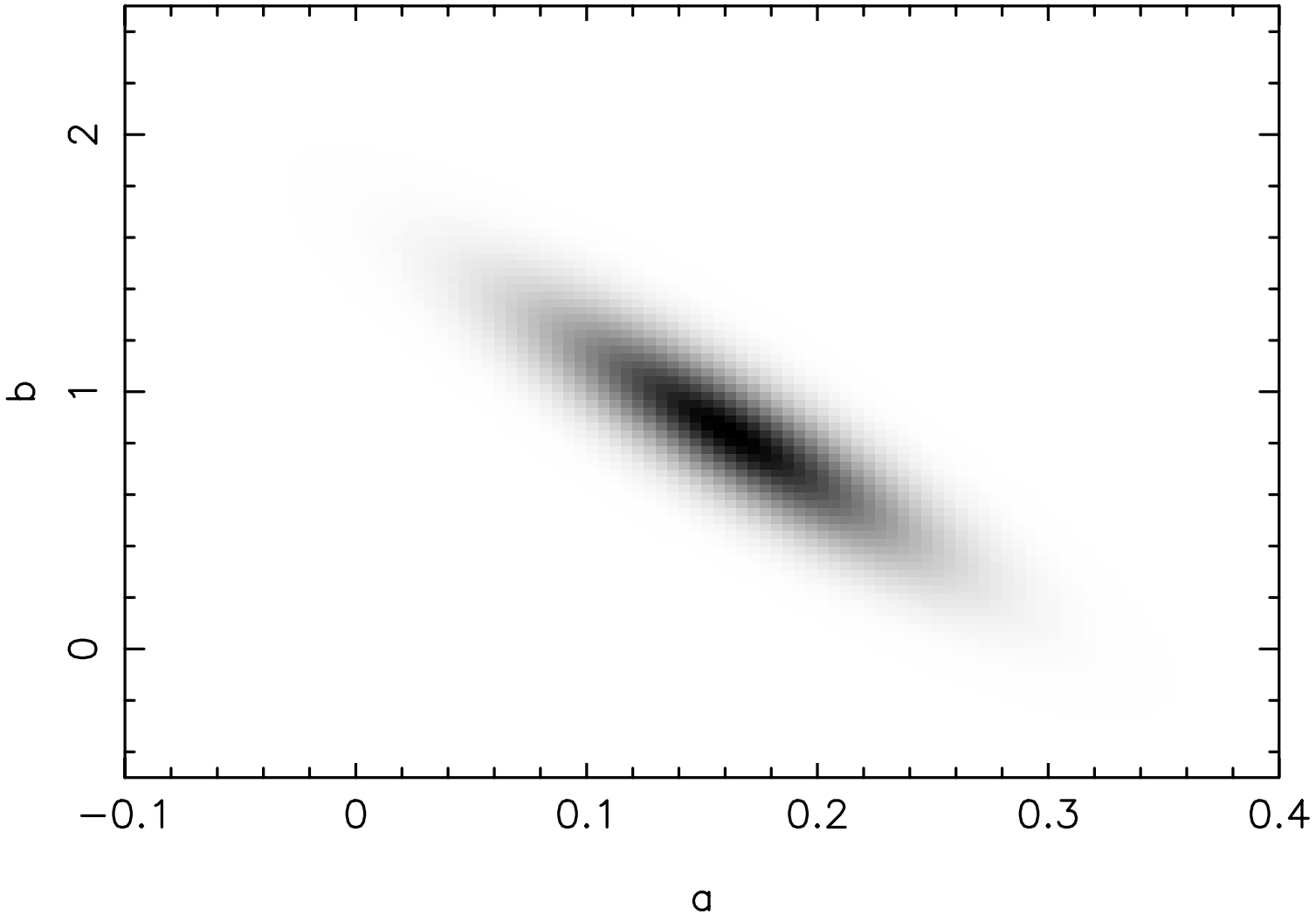
- Convert the 2D chi-squared grid into a 2D probability grid $P_{2\text{D}}(a,b) \propto \exp\left(-\chi^2/2\right)$

- Normalize the grid $\sum_{a,b} P_{2\text{D}}(a,b) = 1$

- Produce the marginalized probability distribution for one parameter by summing $P_{1\text{D}}(a) = \sum_{b} P_{2\text{D}}(a,b)$

Marginalization of parameters

- Greyscale of probability



Marginalization of parameters

- Correlation coefficient of a and b

$$\rho = \frac{< a\, b > - \mu_a\, \mu_b}{\sigma_a\, \sigma_b}$$

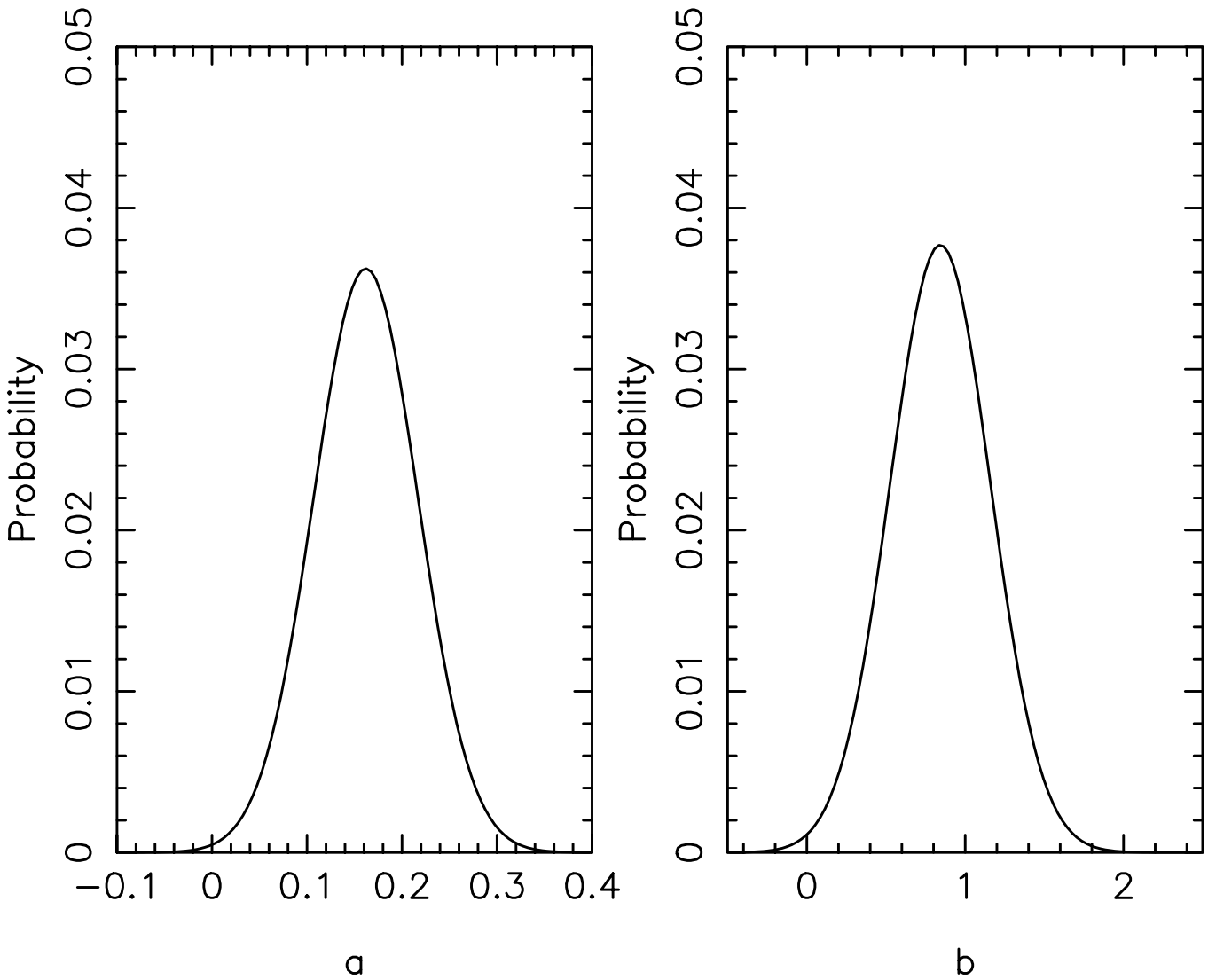
$$< a\, b >= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P_{\rm 2D}(a, b)\, da\, db$$

$$\rho = -0.87$$

Strong anti-correlation !

Marginalization of parameters

- 1D probability distributions

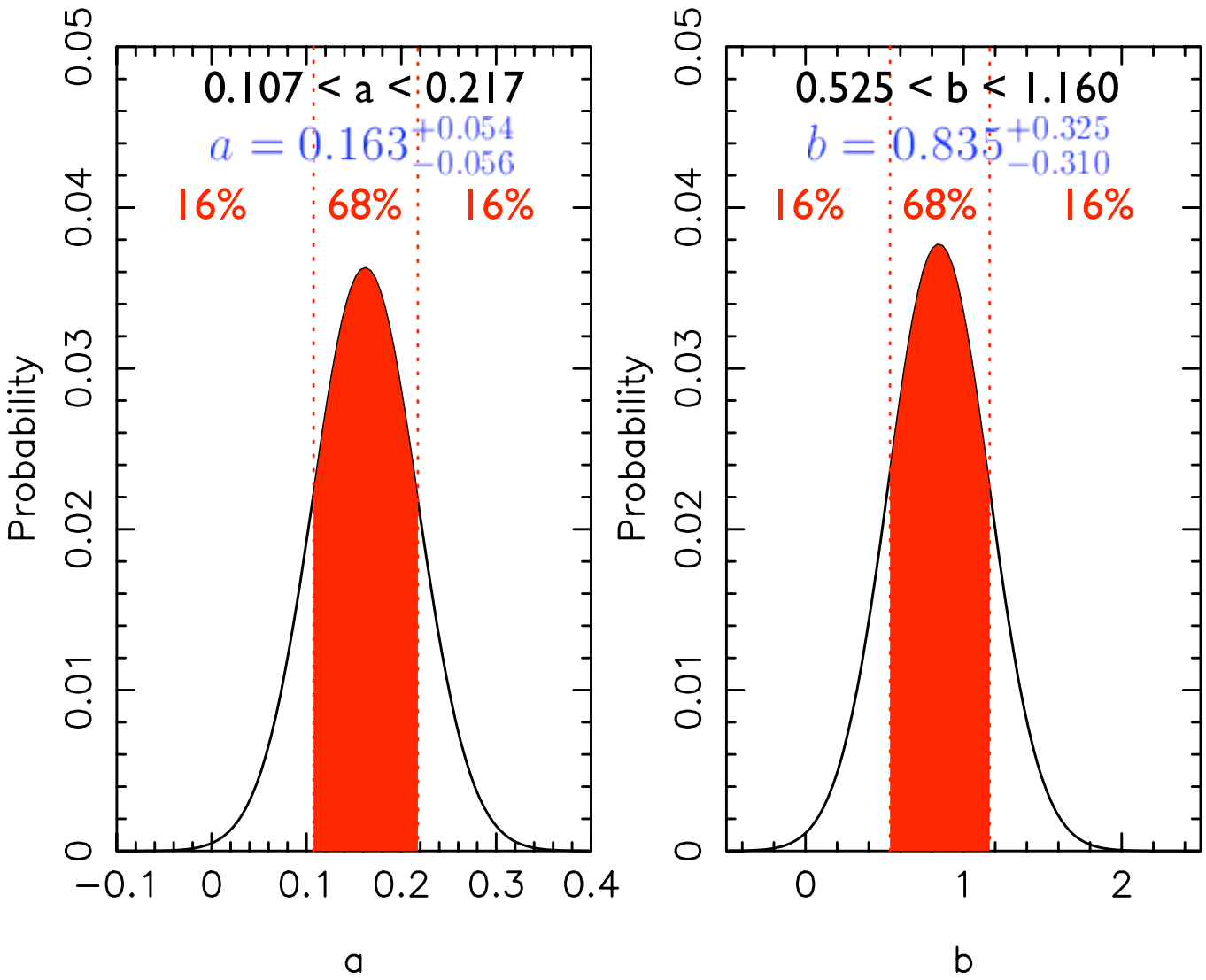


Marginalization of parameters

- We can use the 1D probability distribution to determine a confidence interval for the parameter
- Mean : $\mu_a = \int_{-\infty}^{\infty} a\, P_{\rm 1D}(a)\, da$
- Standard deviation : $\sigma_a^2 = \int_{-\infty}^{\infty} (a - \mu_a)^2\, P_{\rm 1D}(a)\, da$
- Only if the probability distribution is Gaussian is the mean equal to the best-fitting value and the standard deviation equal to the 68% confidence limit
- For a general probability distribution should determine the confidence interval by integration

Marginalization of parameters

- 1D probability distributions



Marginalization of parameters

- 68% confidence interval $a_{bot} < a < a_{top}$

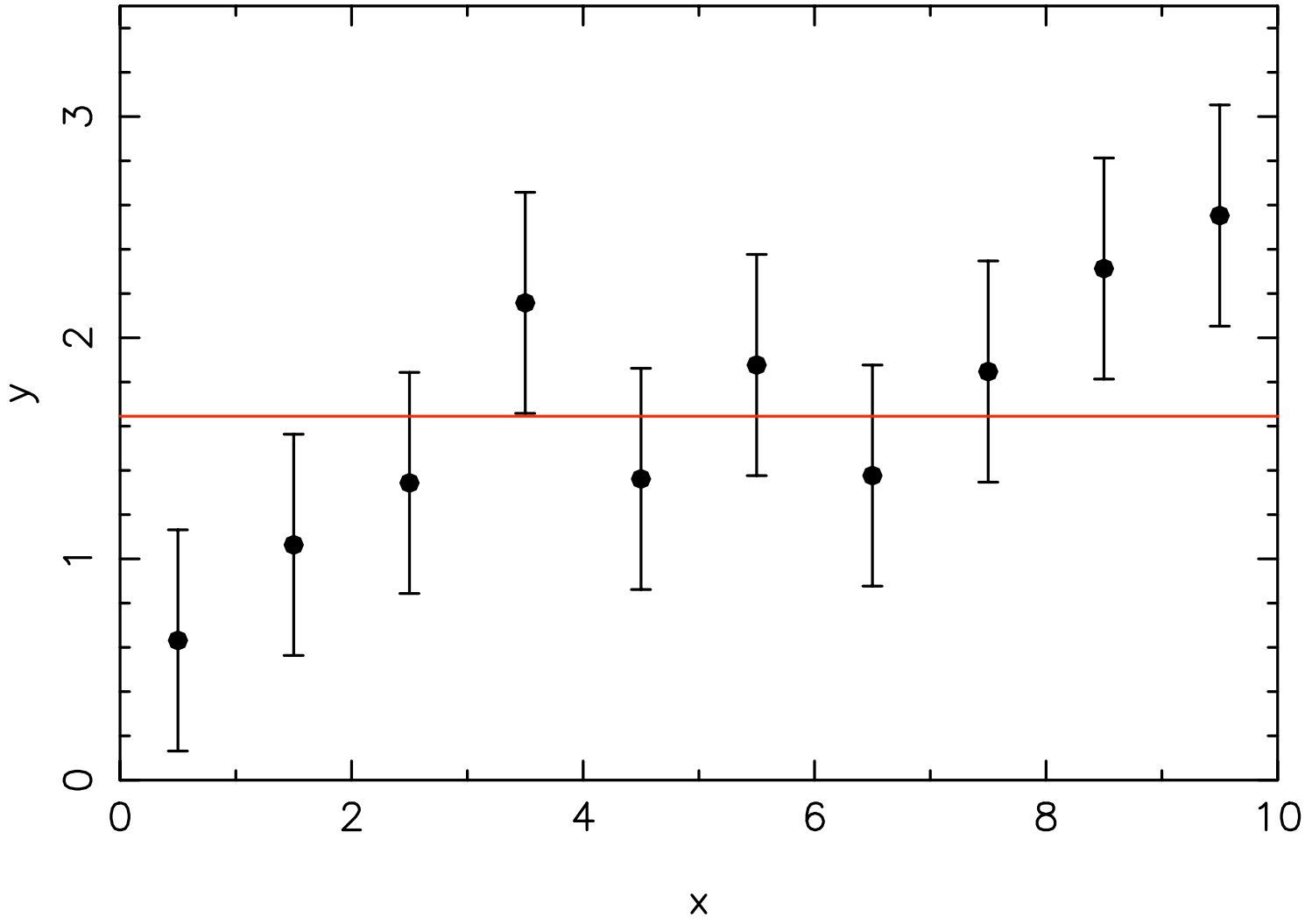
$$\int_{-\infty}^{a_{\rm bot}} P_{\rm 1D}(a)\,da = 0.16$$

$$\int_{a_{\rm bot}}^{a_{\rm top}} P_{\rm 1D}(a)\,da = 0.68$$

$$\int_{a_{\rm top}}^{\infty} P_{\rm 1D}(a)\,da = 0.16$$

Is adding another parameter justified?

- Fit model y = b : $b_{\rm best} = 1.65 \;\; \chi^2_{\rm min} = 12.94 \;\; \nu = 9 \;\; {\rm Prob} = 0.17$



Is adding another parameter justified?

- Model y = a x + b : $\chi^2 = 4.28 \quad N = 10 \quad p = 2 \quad \nu = 8$

- Model y = b : $\chi^2 = 12.94 \quad N = 10 \quad p = 1 \quad \nu = 9$

- Both models provide an acceptable fit to the data. Adding one parameter has produced an improvement in chi-squared of 8.66. Which model do we select?

Is adding another parameter justified?

- As a rule of thumb, the model with the minimum reduced chi-squared is usually the preferred one
- More rigorous (1) : create many Monte Carlo realizations of the dataset, and ask how often the model with the extra parameter is preferred
- More rigorous (2) : can use Akaike information criterion
- [Small print : also see Bayesian information criteria]

Is adding another parameter justified?

- Minimizing the Akaike information criterion allows selection between models with differing numbers of parameters
- If p = number of parameters, N = number of bins

$$AIC = \chi^2 + 2p + \frac{2p(p+1)}{N-p-1}$$

Penalty for parameters

Correction for sample size

- Model y = a x + b : AIC = 9.99 [preferred]
- Model y = b : AIC = 15.44

Errors in both co-ordinates

- Suppose we are fitting y = a x + b to data with errors in both co-ordinates. One solution is to modify the function we are minimizing :

$$\chi^2(a,b) = \sum_{i=1}^{N} \frac{(y_i - a\,x_i - b)^2}{\sigma_{y,i}^2 + a^2\,\sigma_{x,i}^2}$$

- Note 1 : errors in (a,b) may be obtained by bootstrap resampling (see previous lecture)
- Note 2 : this procedure is not symmetric - it minimizes the deviation in y, not necessarily x
- [Small print : rigorous solution uses maximum likelihood]