- ¿Why do we need a statistics and probability course?

In cosmology and astrophysics most of the problems consist of having a set of data from which we want to INFER something.

- Infer some parameter values.

  What is the value the parameters involved in the LCDM paradigm?

- Test an hypothesis.

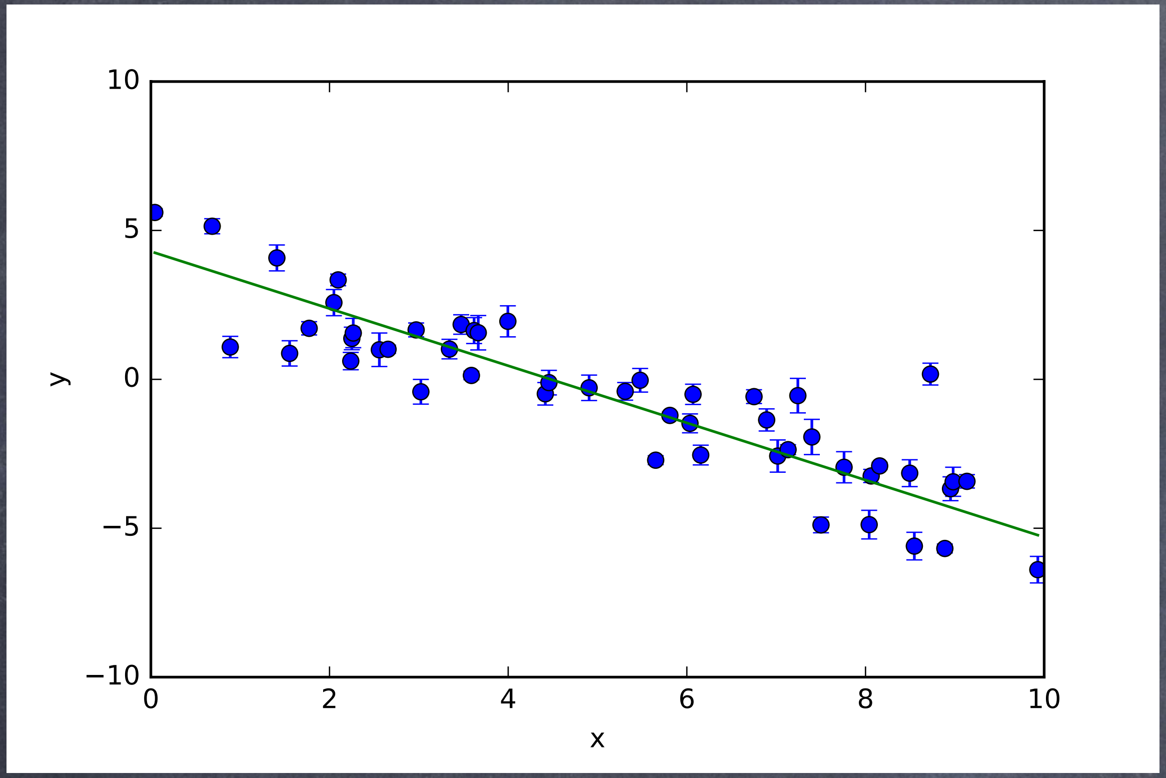
  Is the CMB consistent with a scale free initial power spectrum of fluctations, and with a gaussian distribution?

- Select a model.

  ¿Is General Relativity the correct and final theory, or modified theories works better?

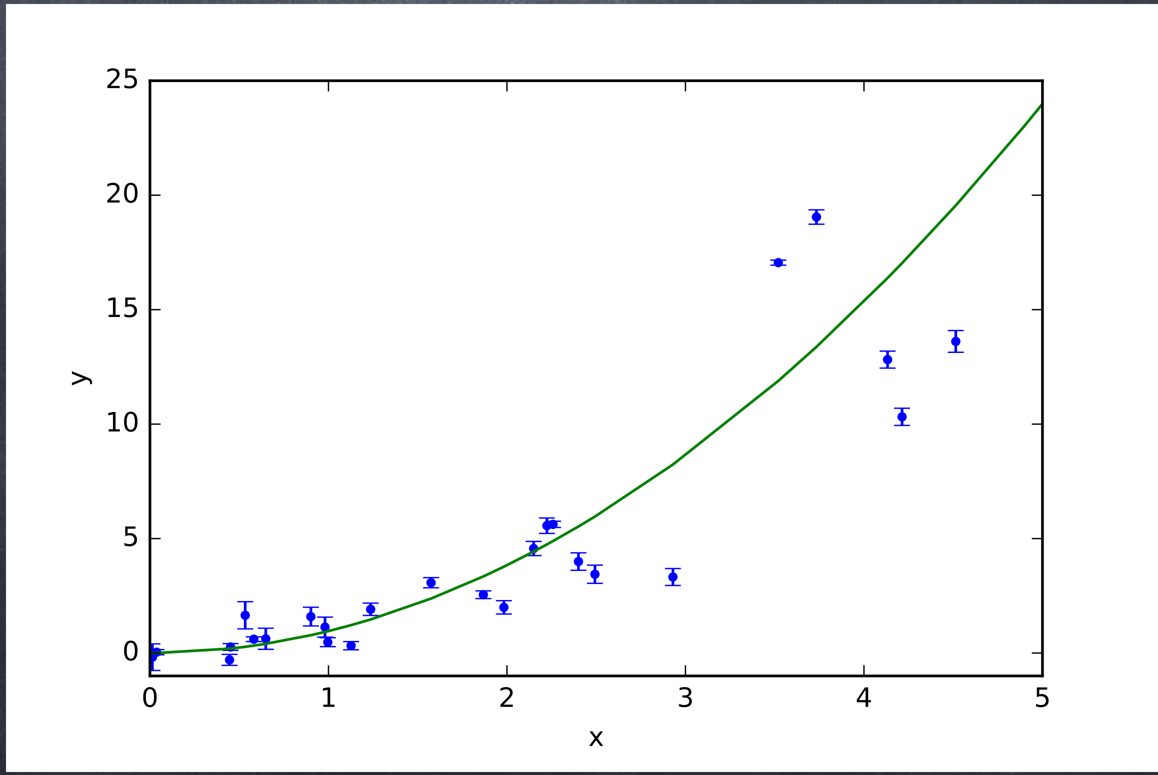
Parameter Estimation

- What do we do if we want to estimate the slope and y-intercept?



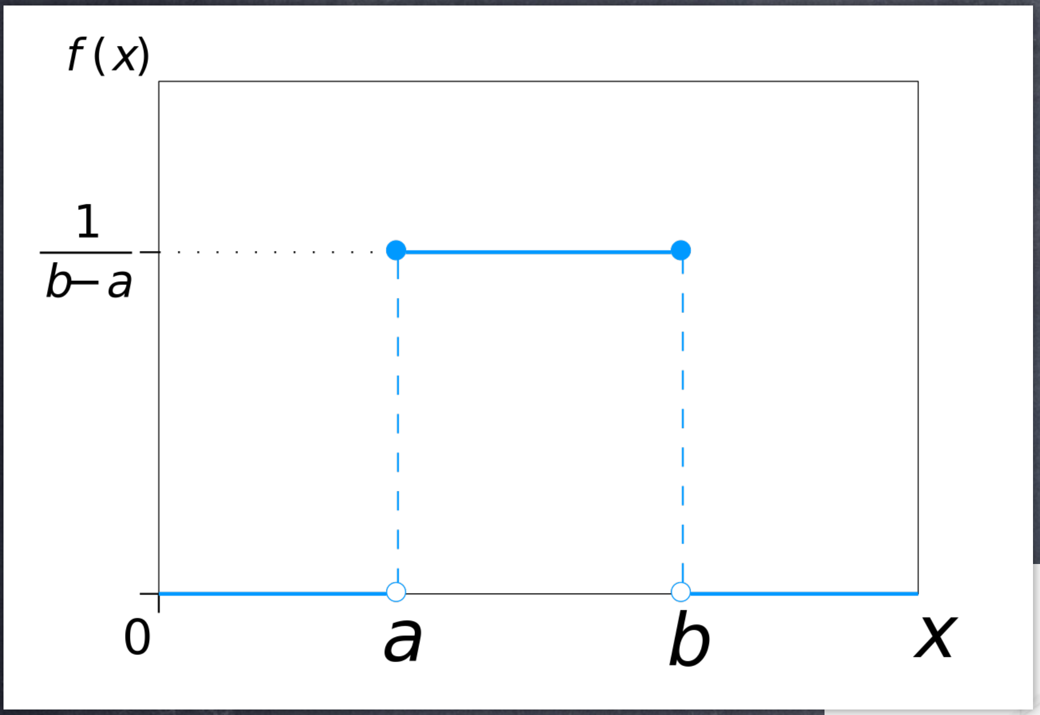
- Linear least square method

- What if data is not a straight line? And/or model is not linear, and/or if we have more than two free parameters? or more important what if I dont believe too much on the error bars…

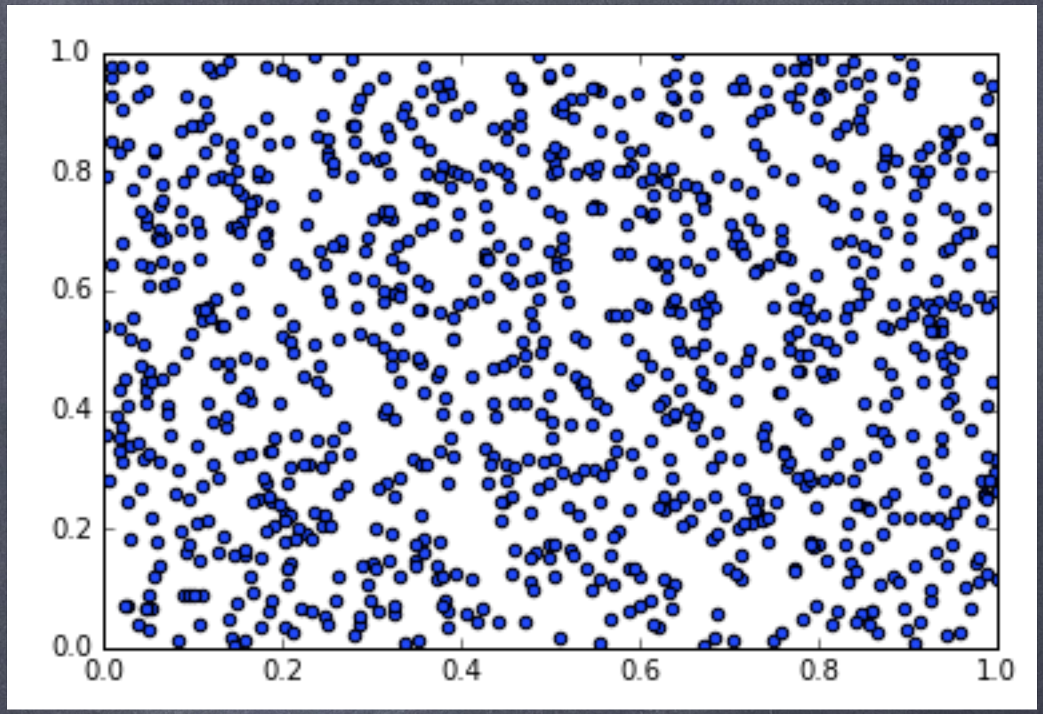


MonteCarlo uses probability distributions

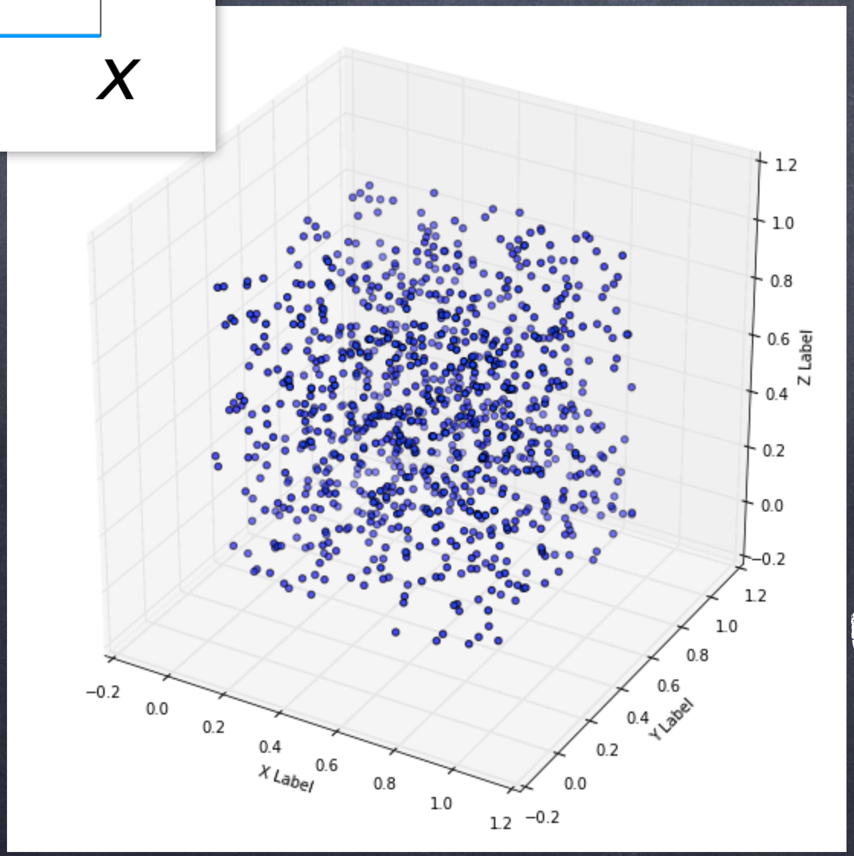
Probability distribution functions:
The basic one: Uniform



1D



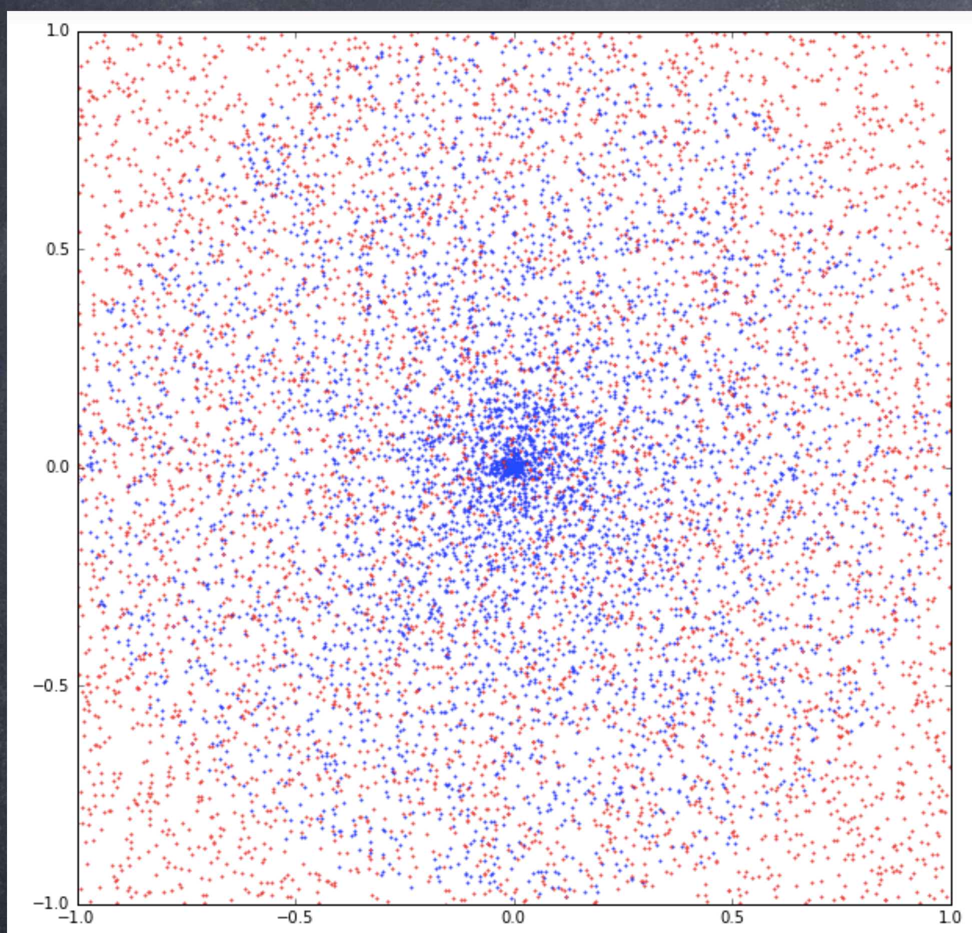
2D



3D

Python
numpy.random.rand()
numpy.random.uniform(a,b)

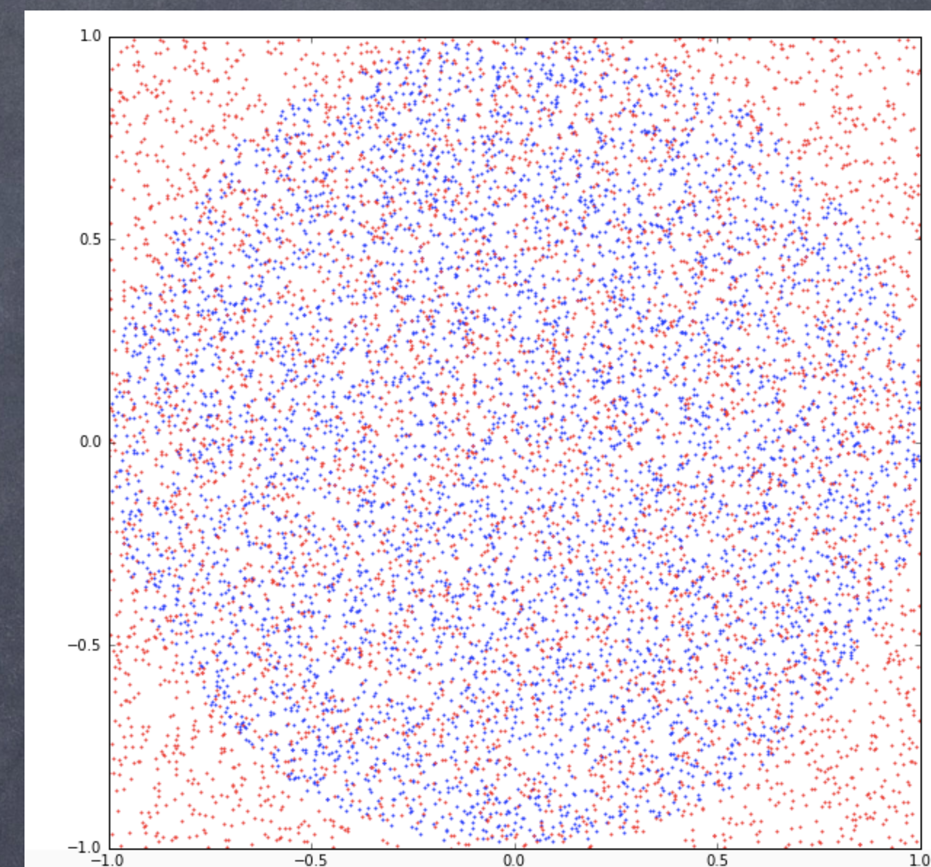
Draw samples from a specific distribution. Use the Uniform distribution (U). (Ej. 2D)



non-Uniform for r<1

1) generate r and theta, from U.

!=

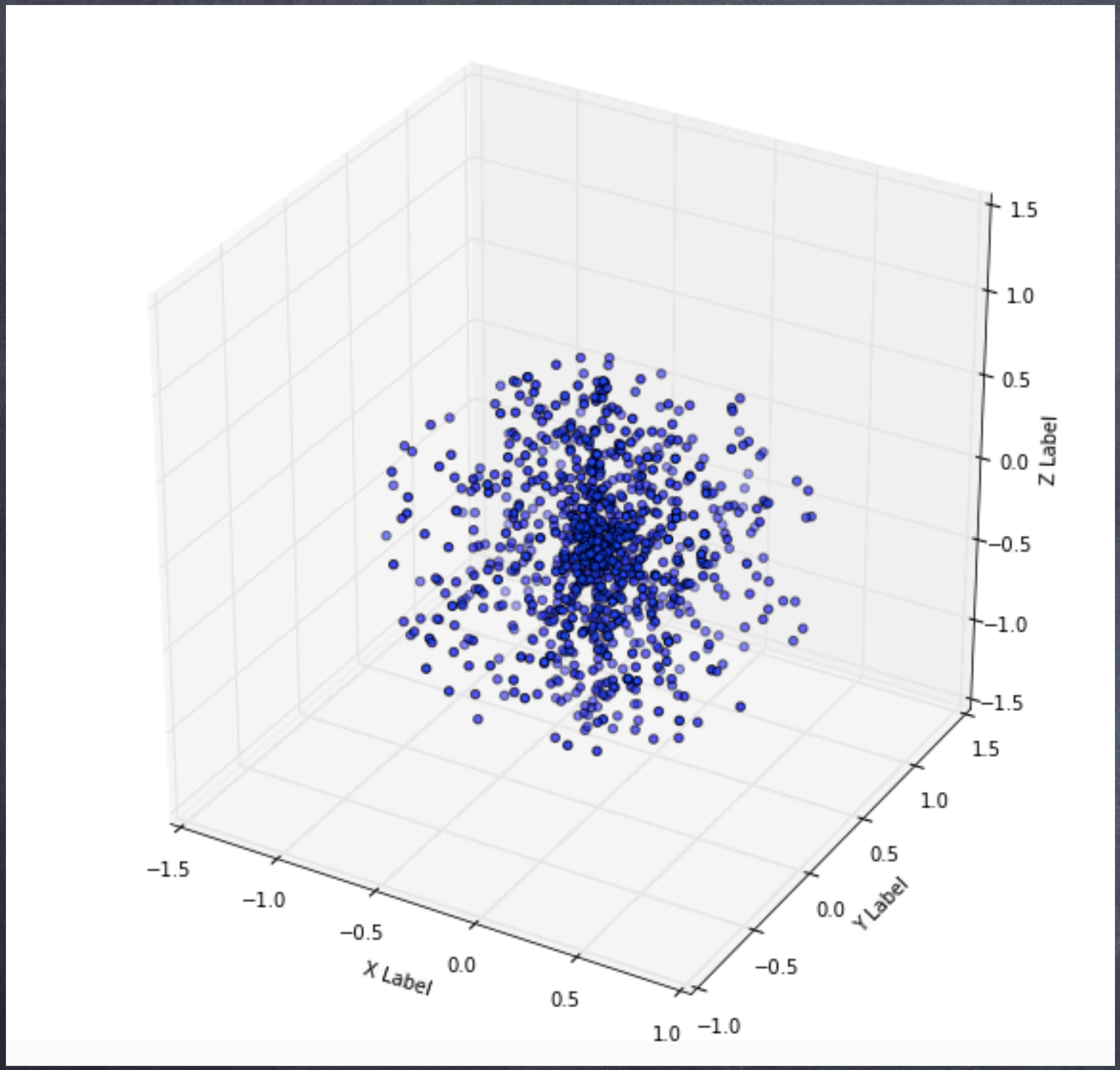


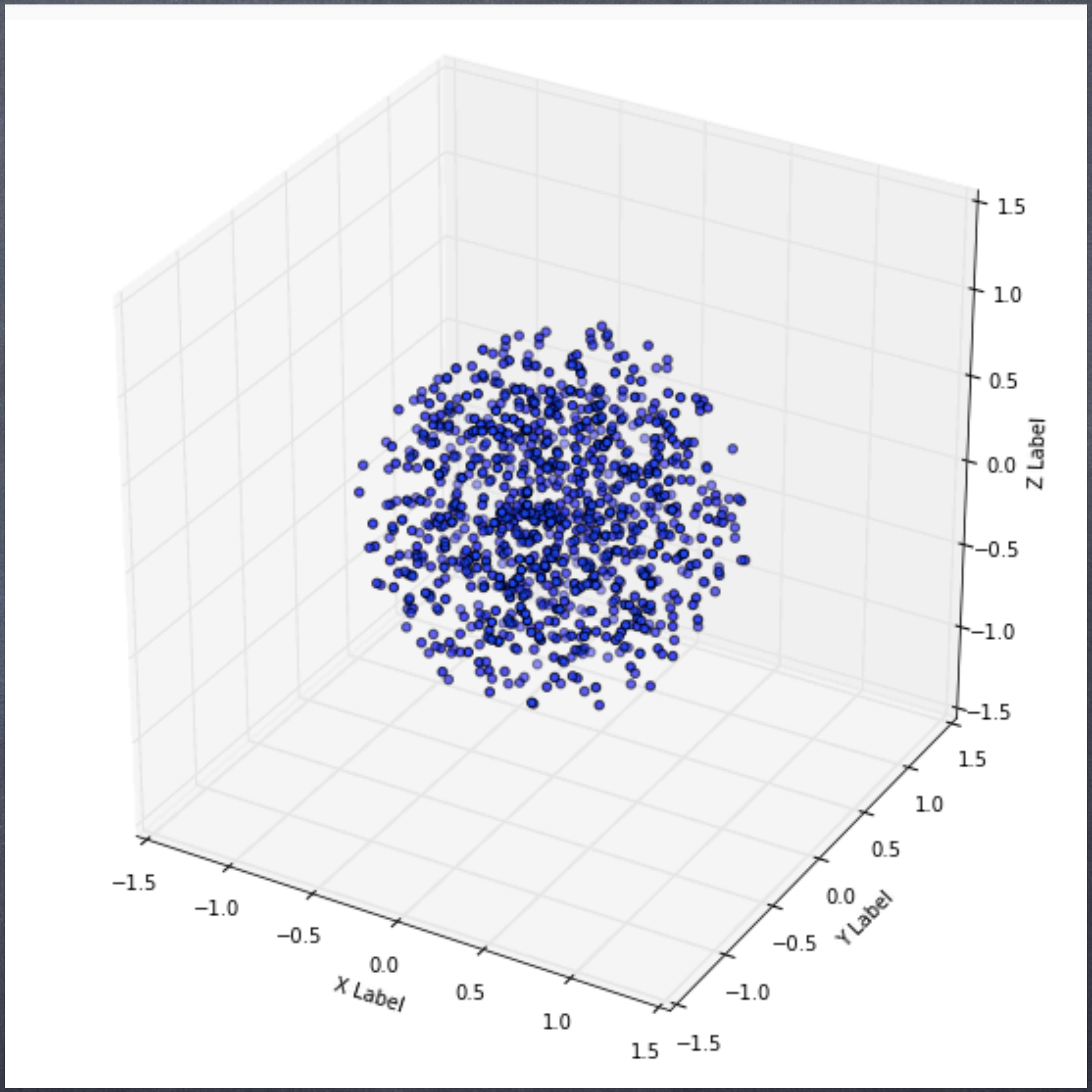
Uniform for r<1

1) generate x,y from U so that r=sqry(x^2+y^2)<=1

2) generate sqrt[r] and theta, from U.

Draw samples from a specific distribution. Use the Uniform distribution (U).



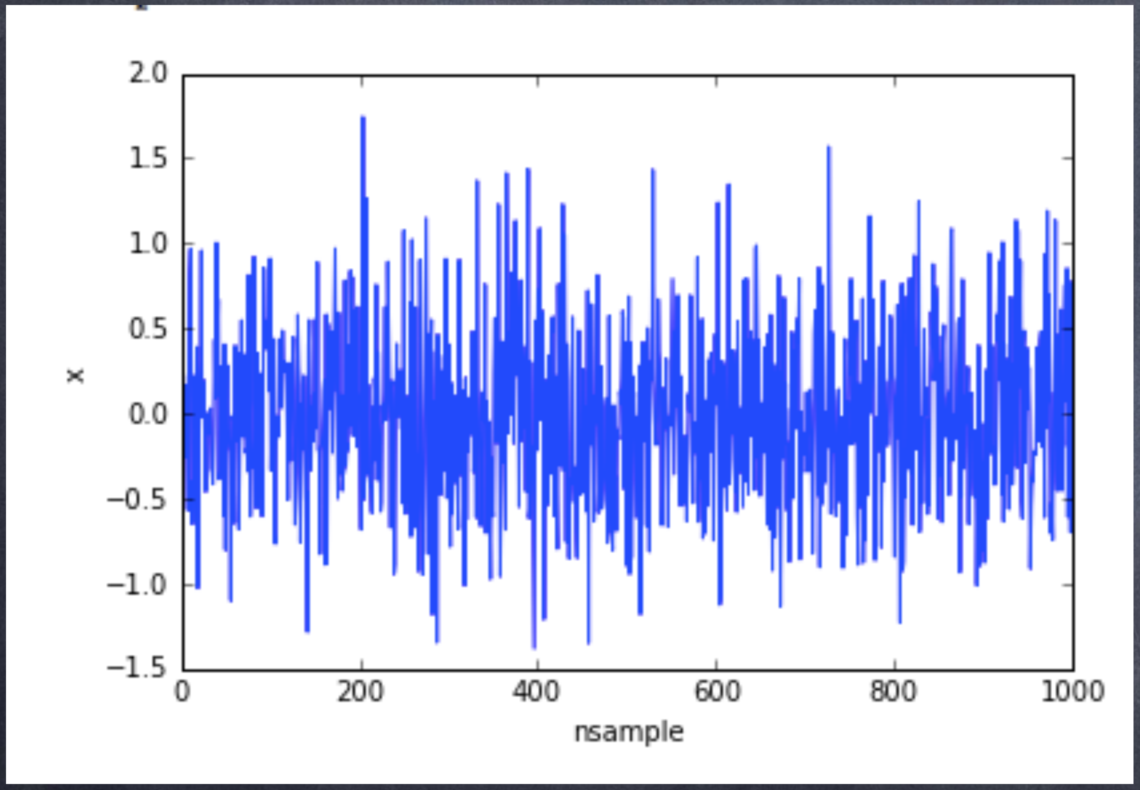


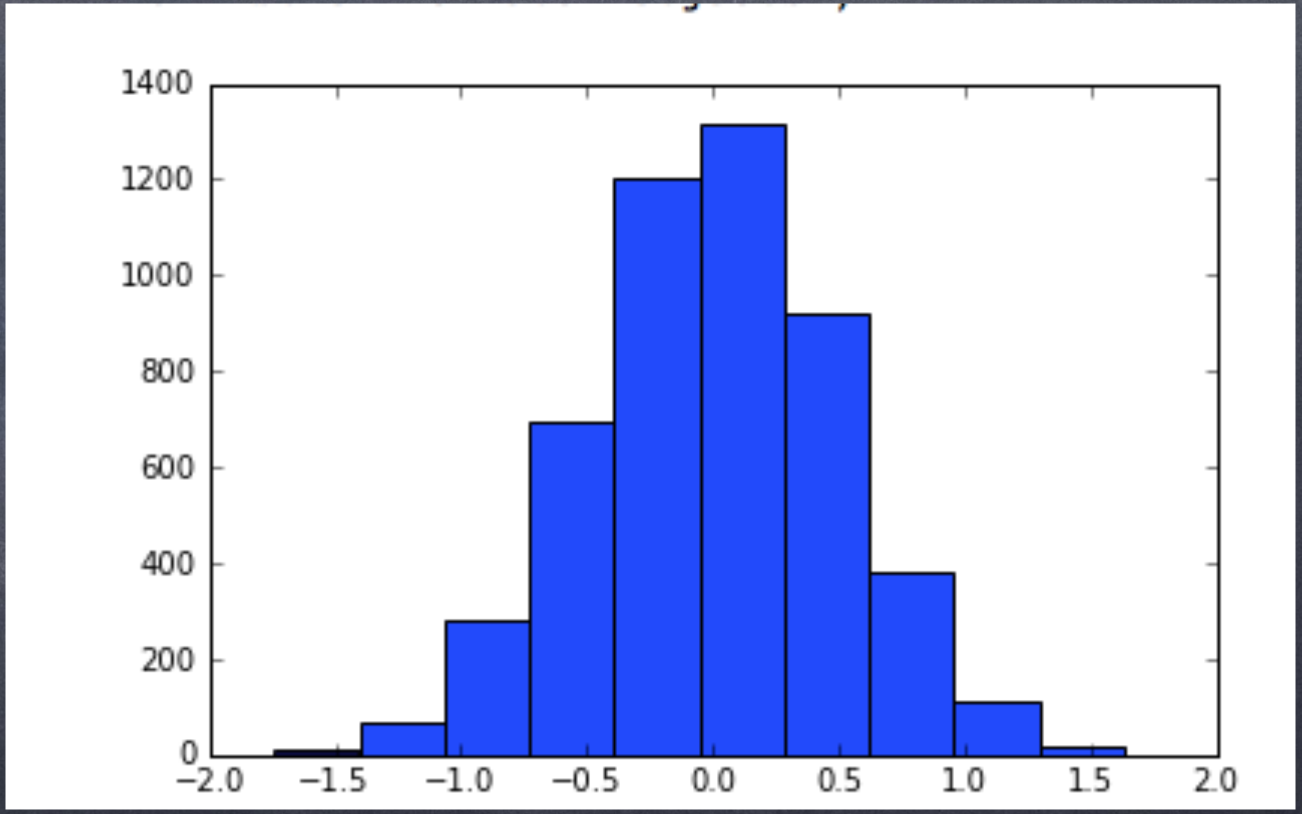
1) generate r and theta, phi from U.

1) generate x,y from U r=sqry(x^2+y^2+z^2)<=1
2) generate sqrt[r], theta and z=cos(phi) from U.

Probability distribution functions:

The next basic one: Gaussian

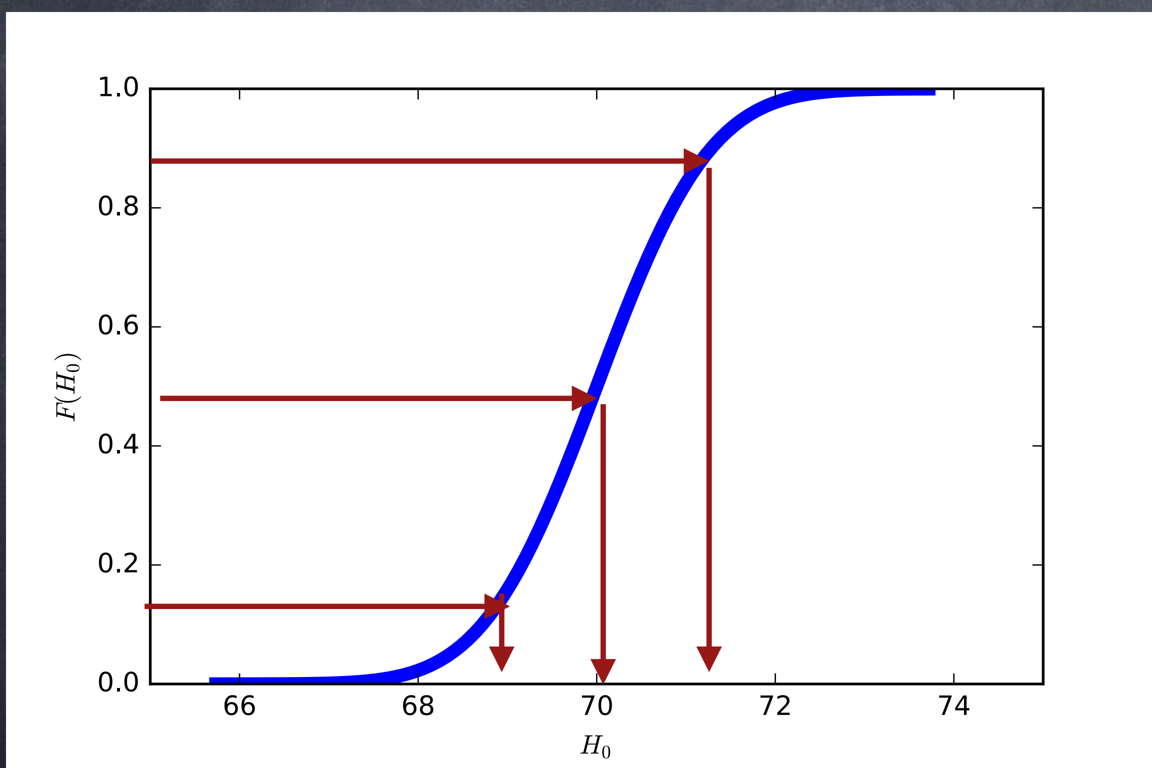


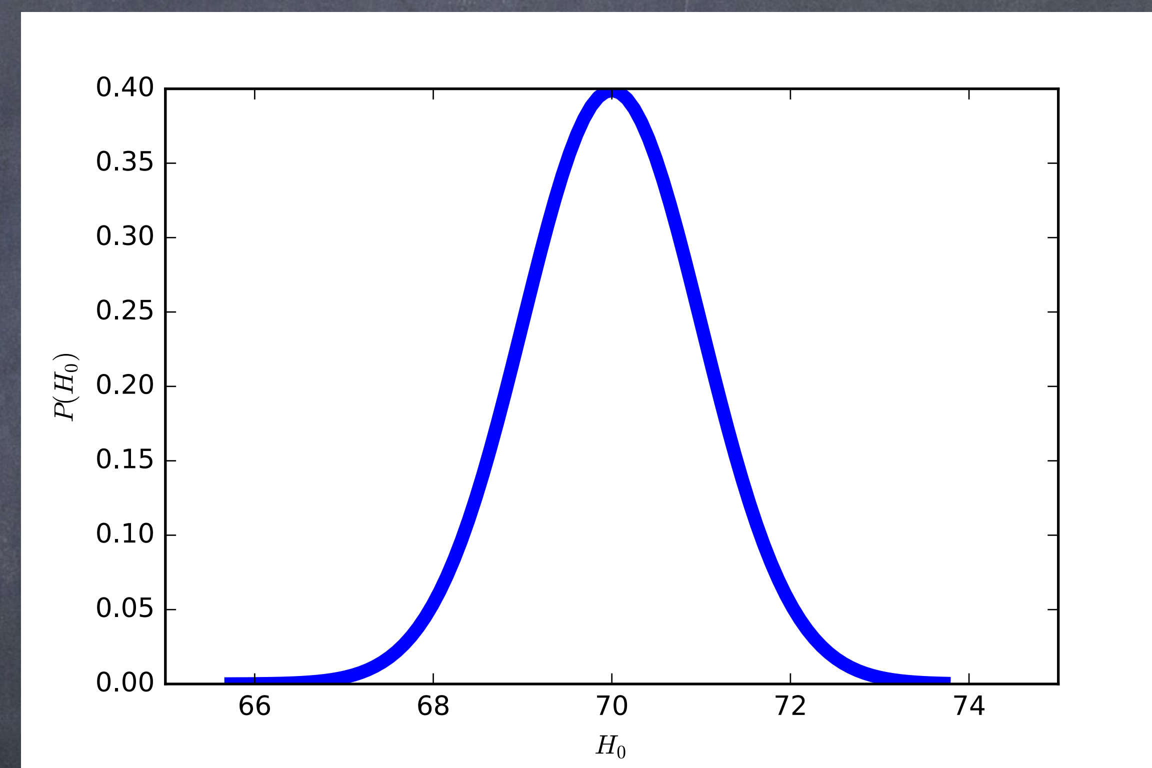


How to sample a PDF

- Depending on the programing lenguaje you are using it can be more or less difficult. But simple method is by using the CDF.

- Depending on the programing lenguaje you are using it can be more or less difficult. But a simple method is by using the CDF.



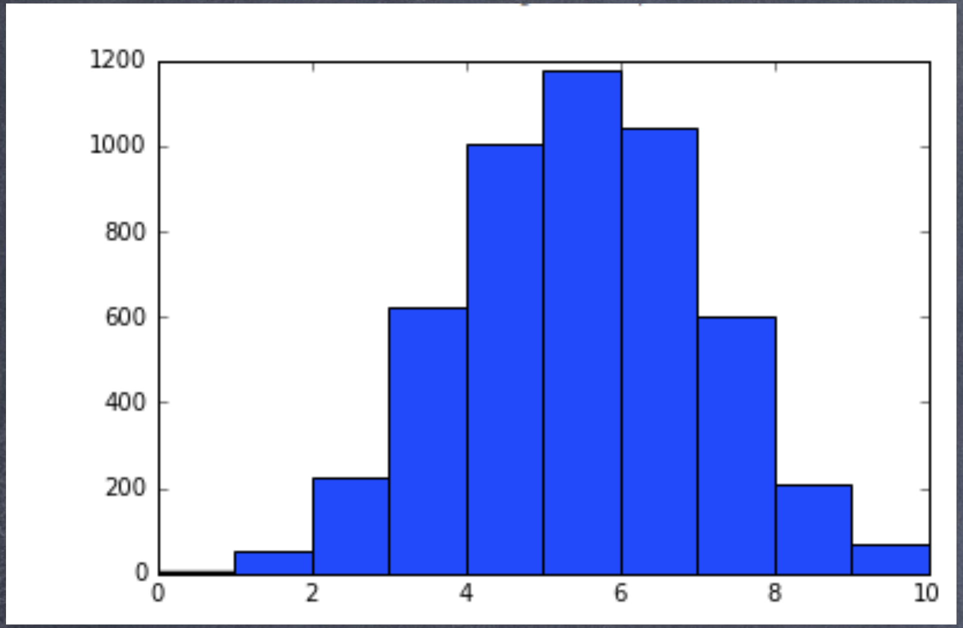


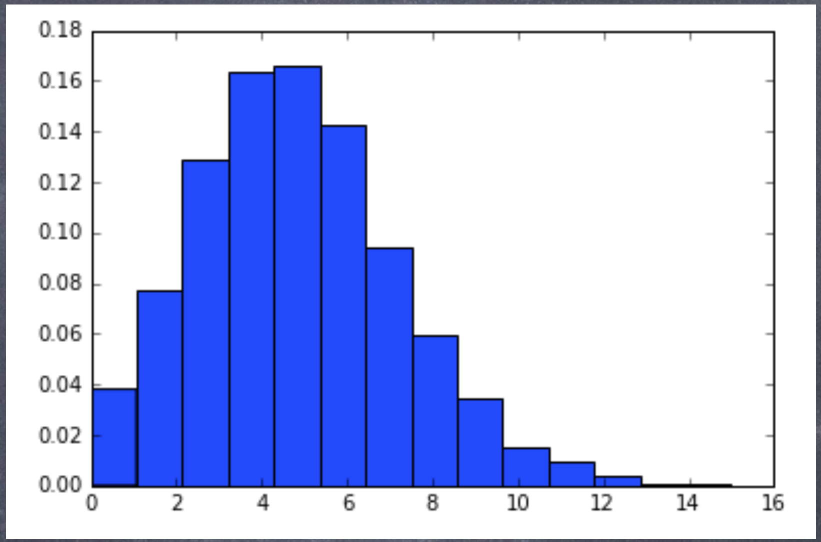
$$F(x) = \int_{-\infty}^{x} f(x')dx'$$

Choose a random number between 0 and 1, and use the CDF to assign the corresponding value of H0. Generate as many as you want, and make the histogram of H0 to verify you did it right.

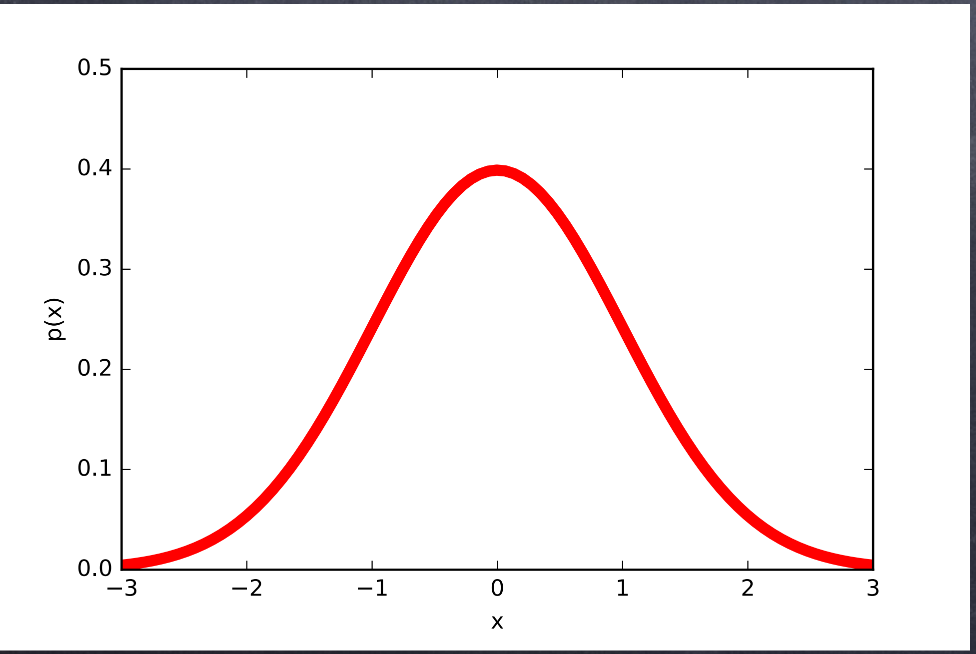
Probability distribution

- Binomial, Poisson and χ^2 distribution can be approximated, for large numbers, by a Gaussian distribution. Other

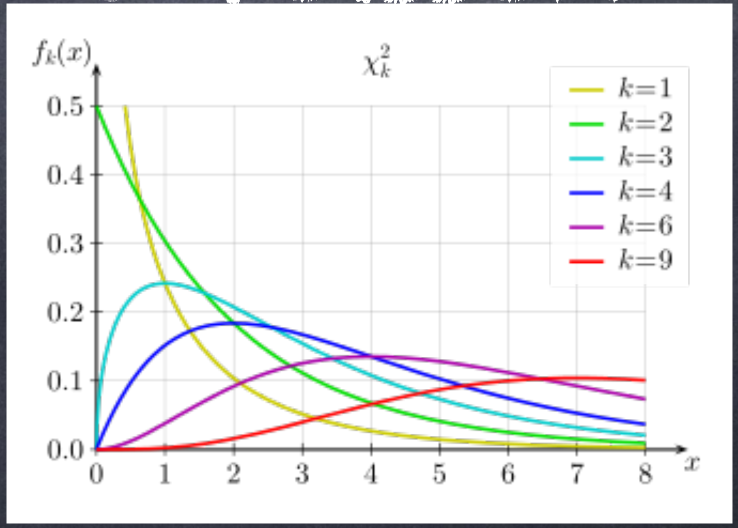




- Poisson



- Gaussian



- χ^2 distribution

Parameter Estimation. Level 0

- Least square method.

$$
\begin{aligned}
a &= \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i\, y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\
&= \frac{\bar{y}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x} \sum_{i=1}^{n} x_i\, y_i}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2} \\
b &= \frac{n \sum_{i=1}^{n} x_i\, y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\
&= \frac{\left(\sum_{i=1}^{n} x_i\, y_i\right) - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}
\end{aligned}
$$

Where this comes from?

Minimize the residual

$$R^2 \equiv \sum \left[y_i - f(x_i, a_1, a_2, ..., a_n)\right]^2$$

assuming:

- A linear function f=ax+b.

- Errors are Gaussian and uncorrelated.

Minimization implies:

$$\frac{\partial R^2}{\partial a_i} = 0$$

Parameter Estimation and optimization

$$\chi^2$$

$$\chi^2 = \sum (y_i - y(x_i, \theta))^2 / \sigma_{y_i}^2 \qquad \frac{\partial \chi^2}{\partial \theta} = 0$$

θ : free parameters

σ_{y_i} : variance ony_i

Chisq minimization becomes difficult (sometimes imposible) when the number of parameters increases…

We can do the minimization of Chi^2 with MonteCarlo sampling.

Least square, and minimum Chi^2 methods are just special cases of Statistical Inference. This Chi^2 is a gaussian distribution if data points are independent, and errors are also gaussian.

Likelihood

- The probability , under the assumption of a model/theory, to observe the data as was actually obtained.

$$\mathcal{L} \longrightarrow P(\mathrm{Data}, \mathrm{Model})$$

- For data that can be thought as samples of a sequence of normal random variables that have a mean and a variance, the likelihood is Gaussian

$$\mathcal{L} \propto \prod_{i}^{n} \frac{1}{2\pi\sigma_i^2} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma_i^2}\right)$$

mu will be the expected mean given our model

- A minimization of the Chi-square correspond to the maximization of the likelihood.

Gaussian Likelihood

$$-\ln(\mathcal{L}(\vec{x},\vec{y}|\vec{\theta})) \propto \frac{1}{2}\sum_i \left(\frac{(y_i - \lambda(x_i,\vec{\theta}))^2}{\sigma_i^2}\right)$$

In terms of data points and parameters.
Lambda is our model for y_i

How do we maximize the likelihood if there 2,3, or more parameters...?

How do we maximize the likelihood if we have a complex model?

What if the likelihood is not Gaussian...?

MonteCarlo Markov Chain

Draw random samples and accept them or reject them according to the likelihood.

If the likelihood of a new sample is higher than the previous one we accept the sample and save it. new—>old

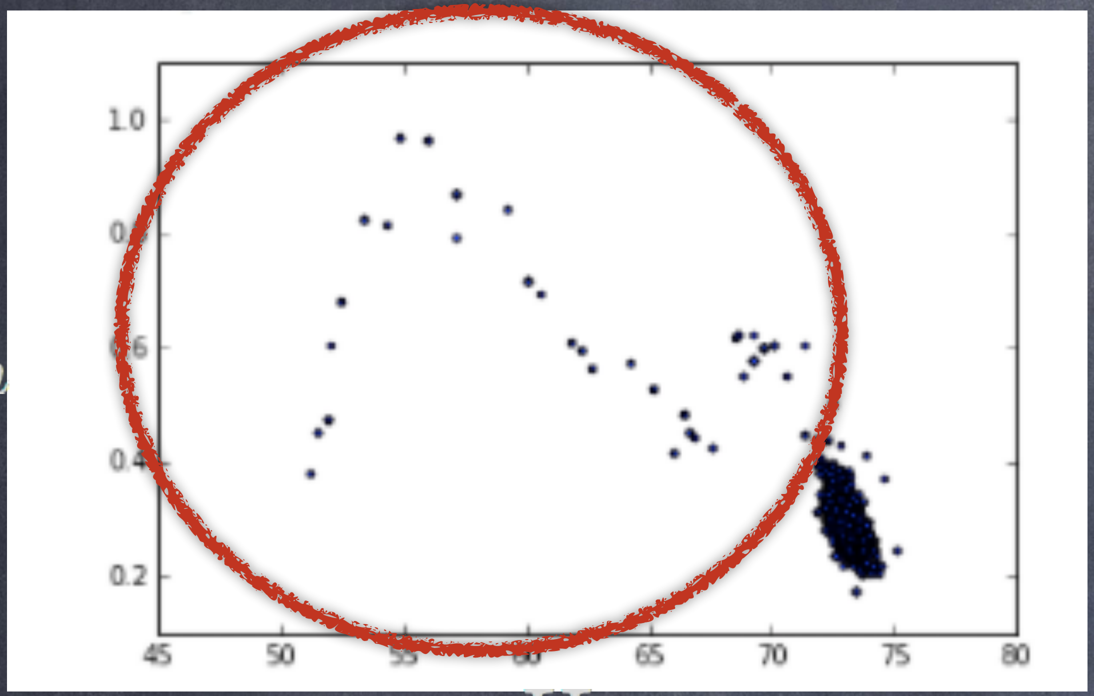
If the likelihood of a new sample is lower than the previous one, we then draw a random number U(0-1), if the ratio of likelihoods (new/old) is larger than such number then we accept it, not otherwise.

Draw a new sample in de vicinity of the previous one, and start again….

After many steps, look at the resultant distribution (the chains) of parameters, i.e., the likelihood..

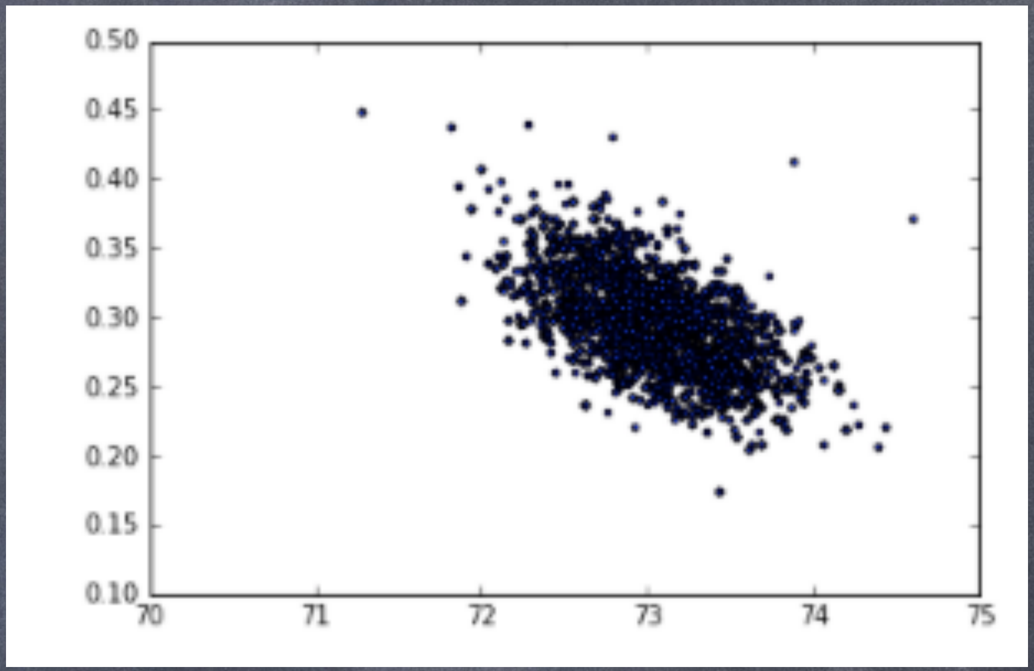
Look at the burning period and the convergence….

Walker



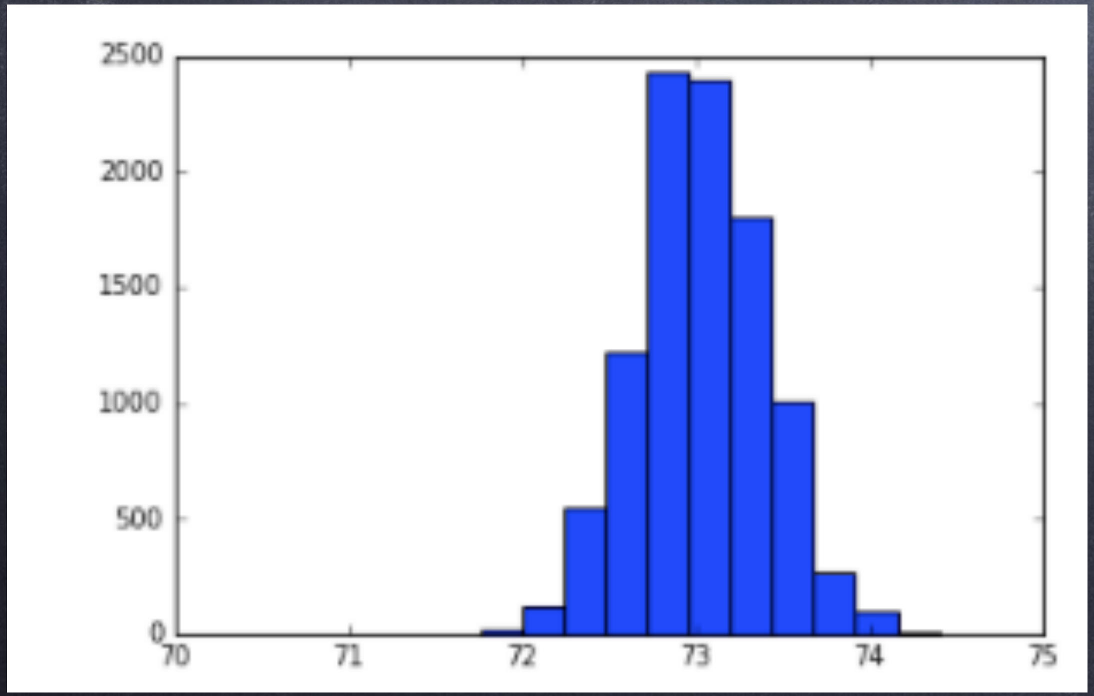
Ω_m

H_0

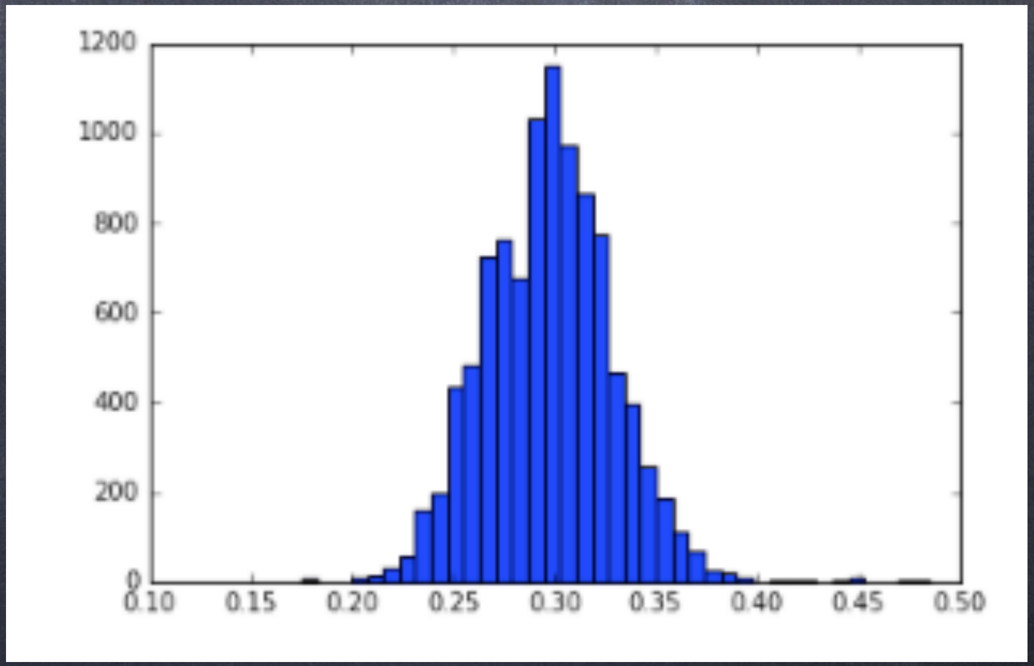


Ω_m

H_0

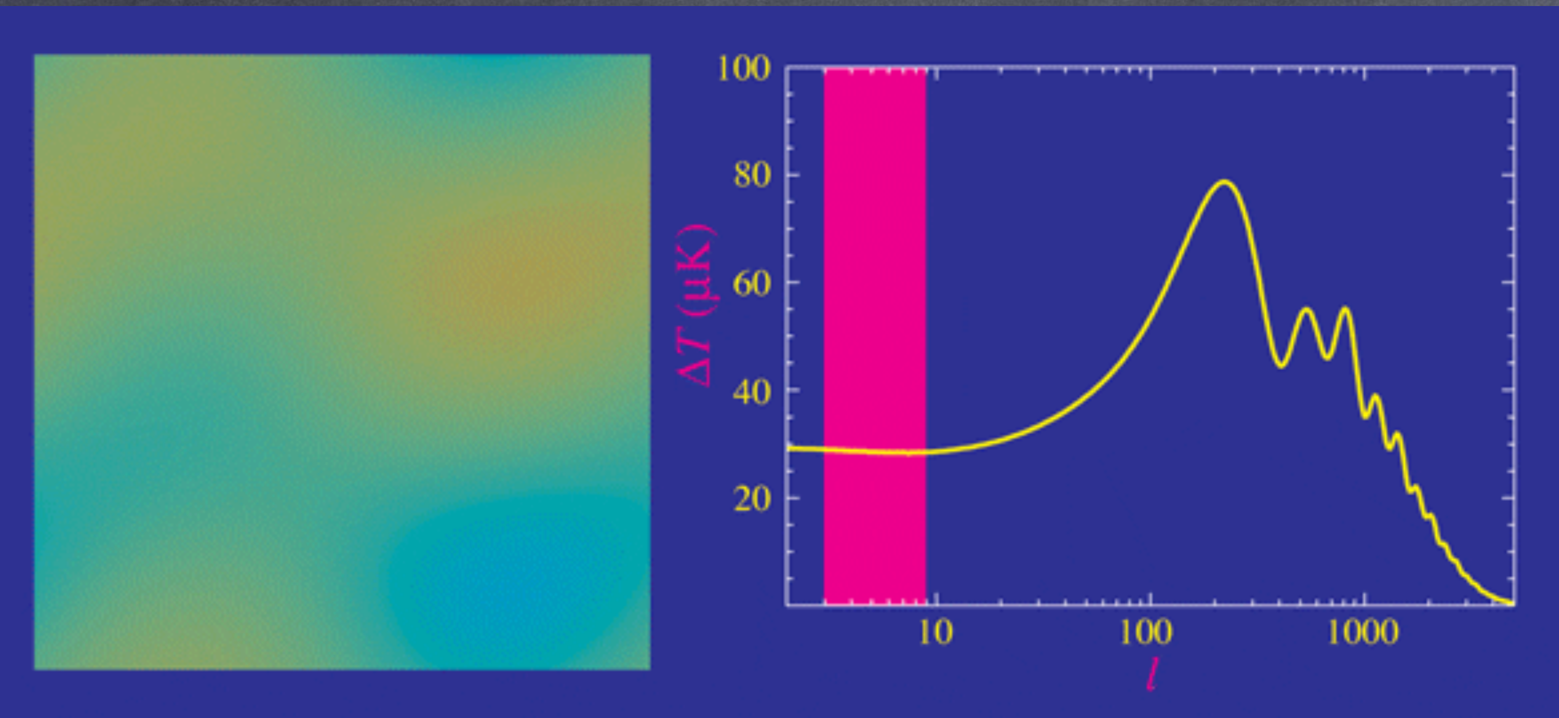


H_0



Ω_m

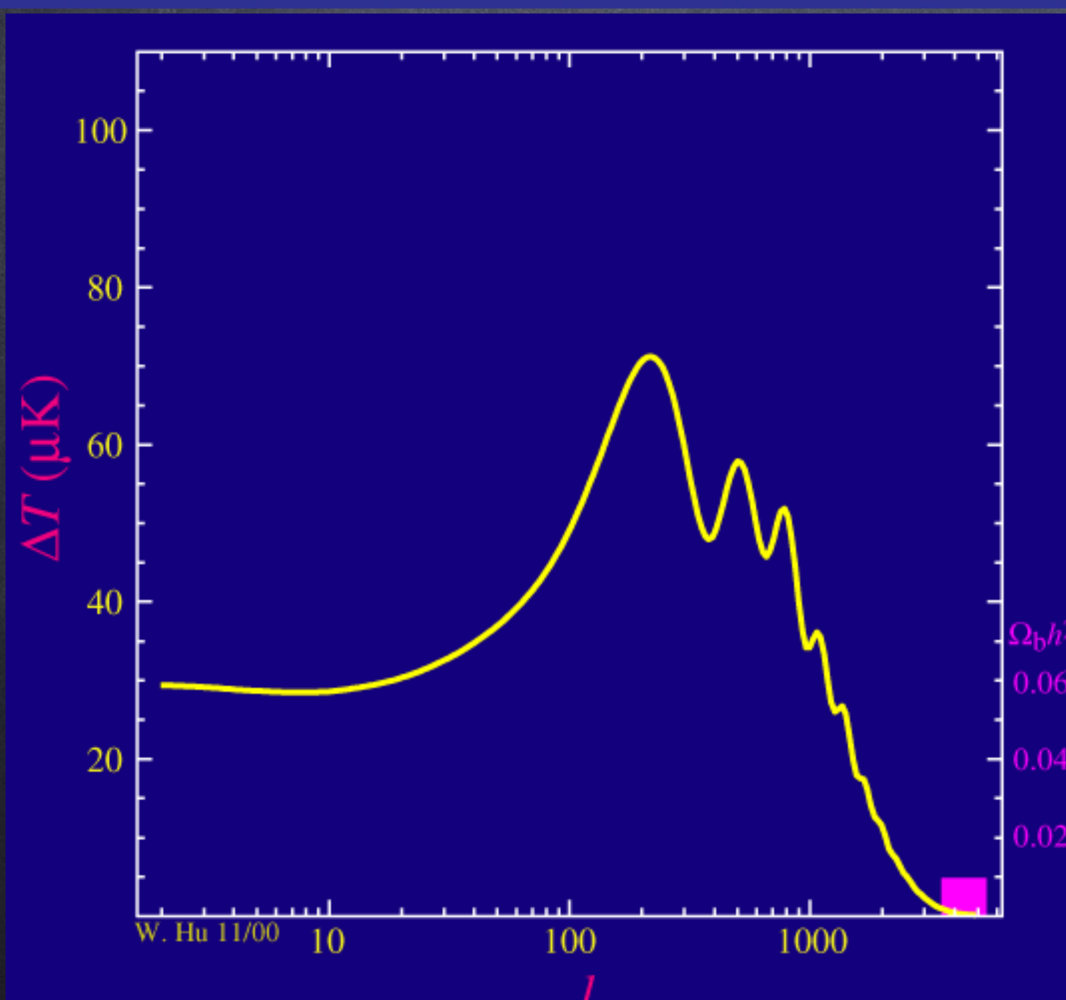
Ej. CMB DATA and Cosmological model

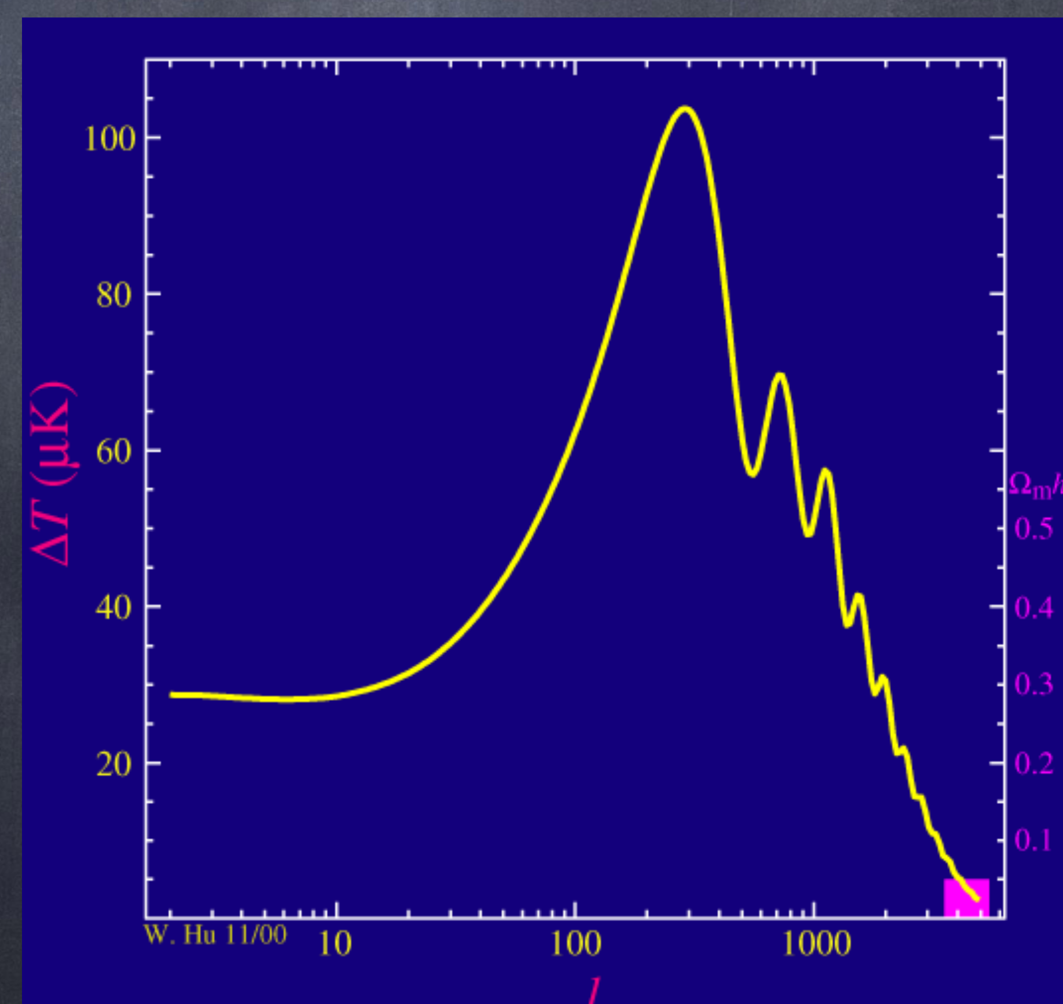


6 cosmological parameters

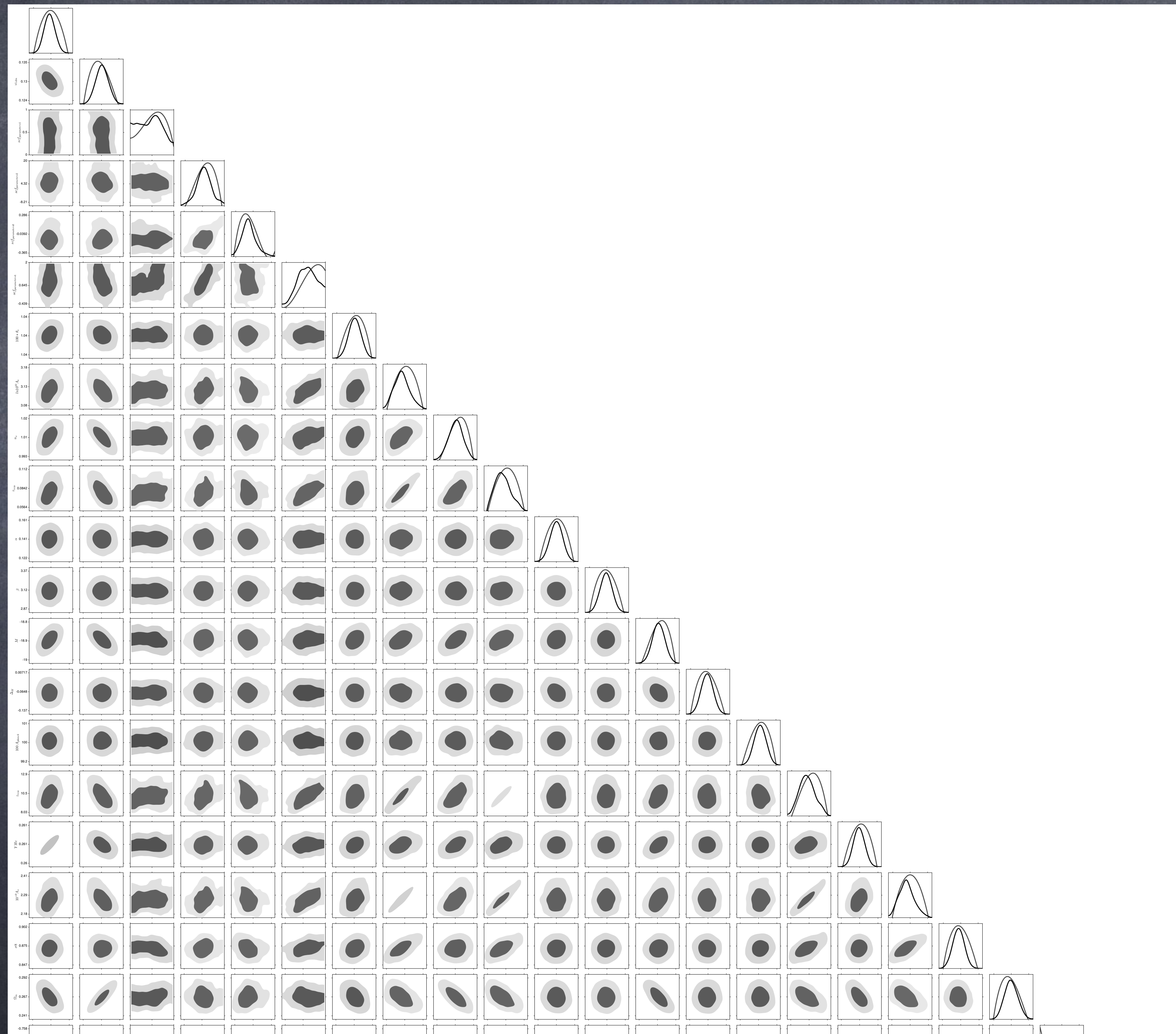
+

many more nuisance parameters





Ej. CMB DATA and Cosmological model



Ej. CMB DATA and Cosmological model