Statistics and Probability

Bibliography

- Bayesian Data Analysis, Carlin, Stern and Rubin, CHAPMAN & HAA/CRC

- Bayesian Reasoning in Data Analysis, Giulio D'Agnostini, World Scientific.

- ICIC Data Analysis Workshop 2016, Alan Heavens Lectures.

- MACSS 2016, 2017, 2019 Lecture notes.

Statistics

- Descriptive
- Inferential:
  - Infer something from data set.
  - Test hypothesis.
  - Select a model or take decisions.

In cosmology and astrophysics, as well as in many areas, 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? What are the values of parameters involved in a model of COVID-19 positive cases.
* Test an hypothesis.	Is the CMB consistent with a scale free initial power spectrum of fluctuations, and with a gaussian distribution? Is COVID-19 evolution in a exponential or a decreasing state.
* Select a model.	¿Is General Relativity the correct and final theory, or modified theories works better? Is a SI, SIR, SEIR, or else a better model to describe COVID-19

Probability & Inference

Typical answer

- "The ratio of the number of favorable cases to the number of all cases"
- " The ratio of the number of times the event occurs in a test series to the total number of trials in the series"
- Frequentist approach

A subjective definition

- A formal definition would be: "The quality, state, or degree of something being supported by evidence strong enough make it likely though not certain to be true"
- A simple definition: "A measure of the degree of belief that an event will occur"
- Bayesian approach.

Probability Rules

$$0 < p(x) < 1$$

Probability of event x happens is coherent

$$p(x) + p(\sim x) = 1$$

Probability that event "x" happen, and probability of event x do not happen are complementary.

$$p(x, y) = p(x|y)p(y)$$

Product rule

$$p(x) = \sum_i p(x, y_i)$$

Probability that event "x" happen, given that y happened: Marginalization

$$p(x) = \int p(x, y)dy$$

In the continuous limit we change the sum by an integral.

BAYES THEOREM arise from the these rules.

Probability function (discrete variable)

To each possible value of x we associate a degree of belief.

$$f(x) = p(X = x)$$

f(x) must satisfy the Probability rules.

Define the Cumulative distribution function ,

$$F(x_k) \equiv P(\leq x_k) = \sum_{xi \leq x_k} f(x_i)$$

CDF

with properties:

$$F(-\infty) = 0$$

$$F(\infty) = 1,$$

Also define de mean, or expected value.

$$\mu = \bar{x} = E(x) = \sum_i x_i f(x_i)$$

In general: $E(g(x)) = \sum_i g(x_i) f(x_i)$ $\quad E(aX + b) = aE(X) + b$

- The standard deviation and Variance .

$$\sigma^2 \equiv Var(X) = \bar{X^2} - \bar{X}^2 \qquad \sigma = \sqrt(\sigma^2)$$

$$Var(aX + b) = a^2 Var(X),$$

- The mode, or the most probable value (for unimodal functions)

$$\left(\frac{df(x)}{dx} \right)_{x_m} = 0$$

Probability density function (continus variable)

The degree of believe of each value is quantified by the probability density function, pdf. $f(x)$

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

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

$$E(X) = \int_{\infty}^{\infty} xf(x)dx$$

$$E(g(X)) = \int_{\infty}^{\infty} g(x)f(x)dx$$

$$\sigma^2 \equiv Var(X) = \bar{X}^2 - \bar{X}^2$$

Centra Limit Theorem

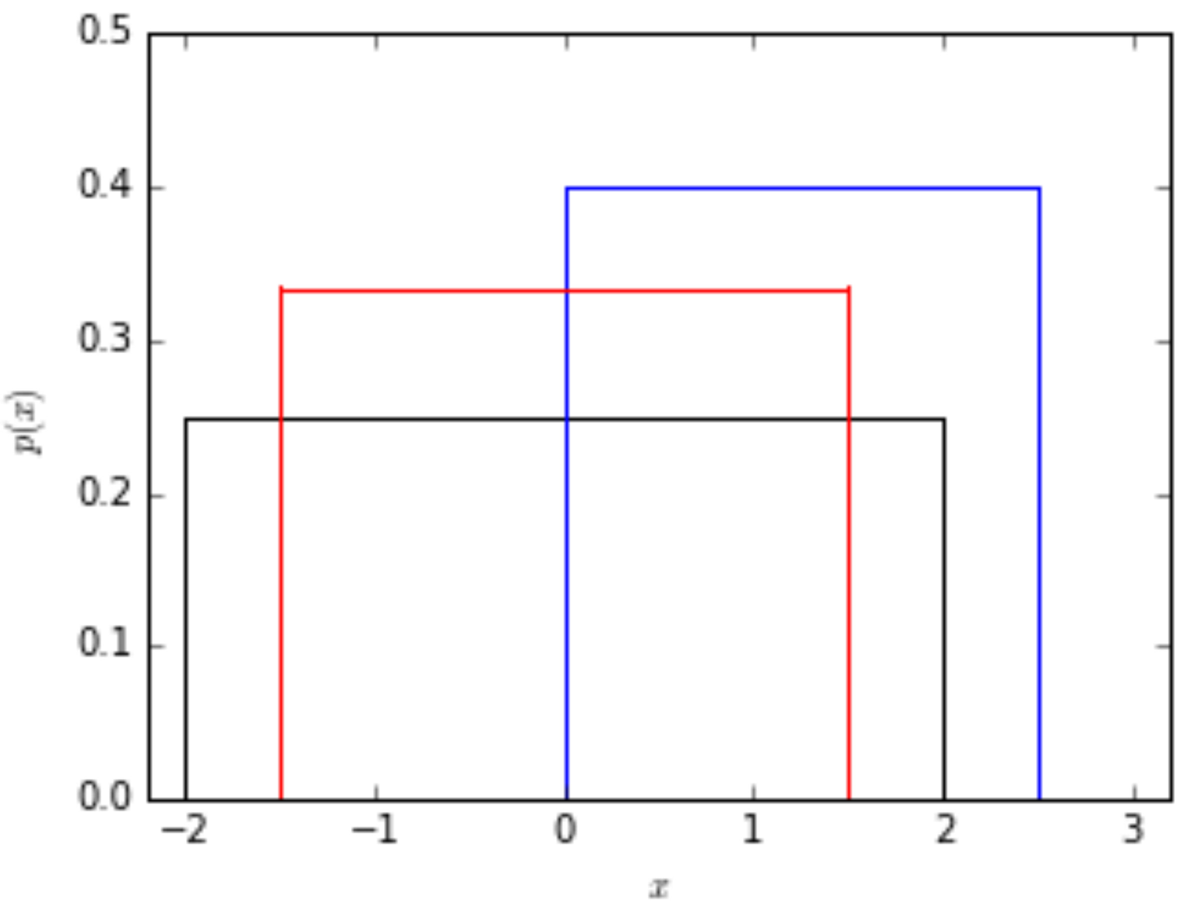
- The mean and variance of a linear combination of random variables is given by:

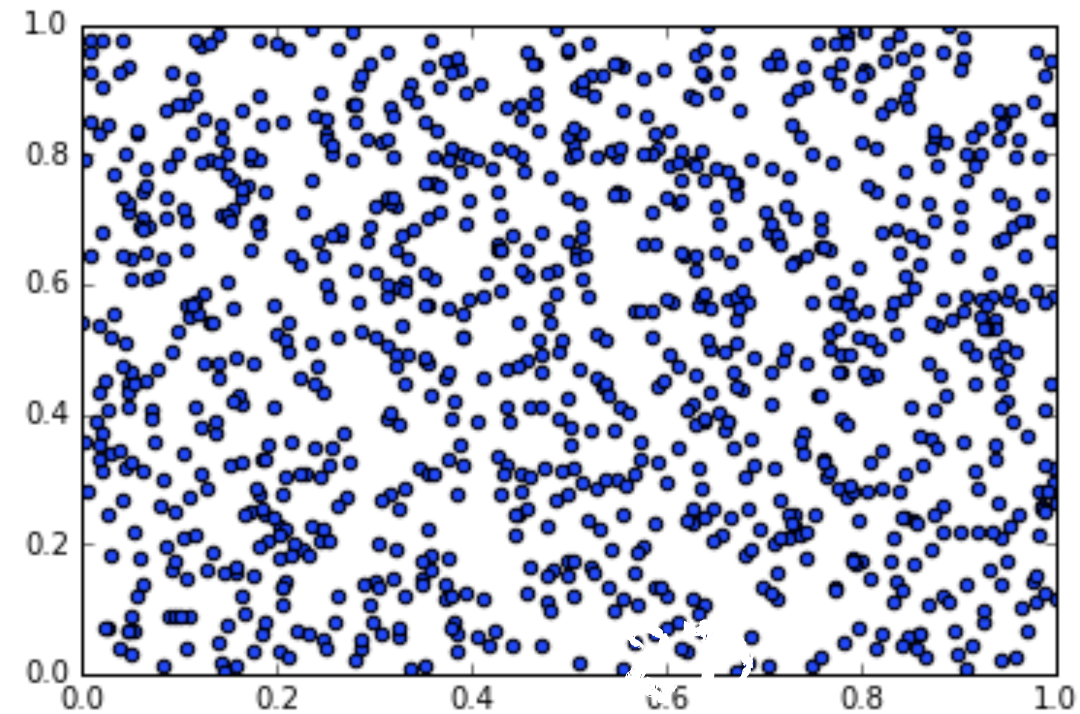
$$Y = \sum_{i=1}^{n} c_i X_i \qquad \sigma_Y^2 = \sum_{i=1}^{n} c_i^2 \sigma_i^2$$

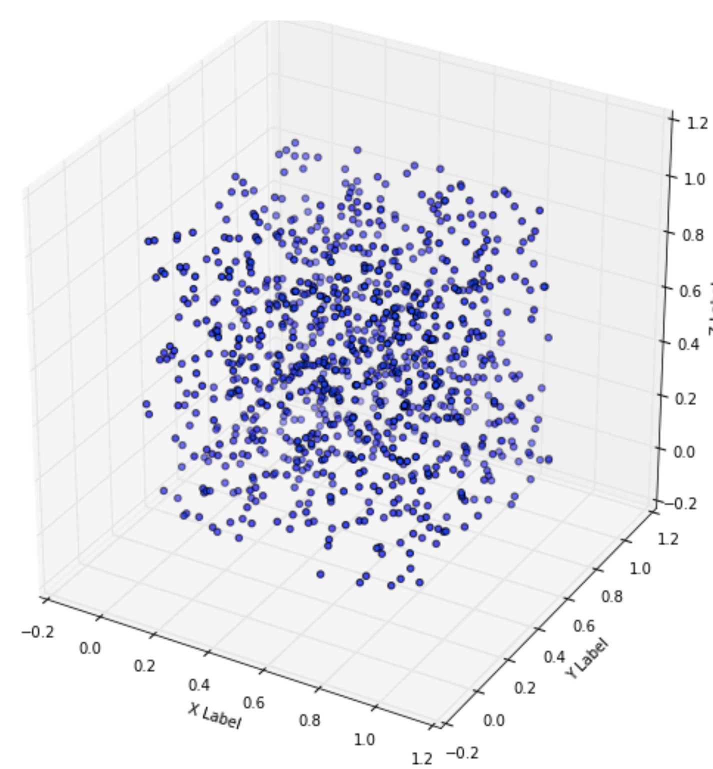
- CLT: The distribution of a linear combination Y will be approximately normal if the variables X_i are independent and \sigma_Y^2 is much larger than any single component c_i^2 \sigma_i^2 from a non-normally distributed X_i.

Probability distributions

Probability distribution functions: The basic one: Uniform







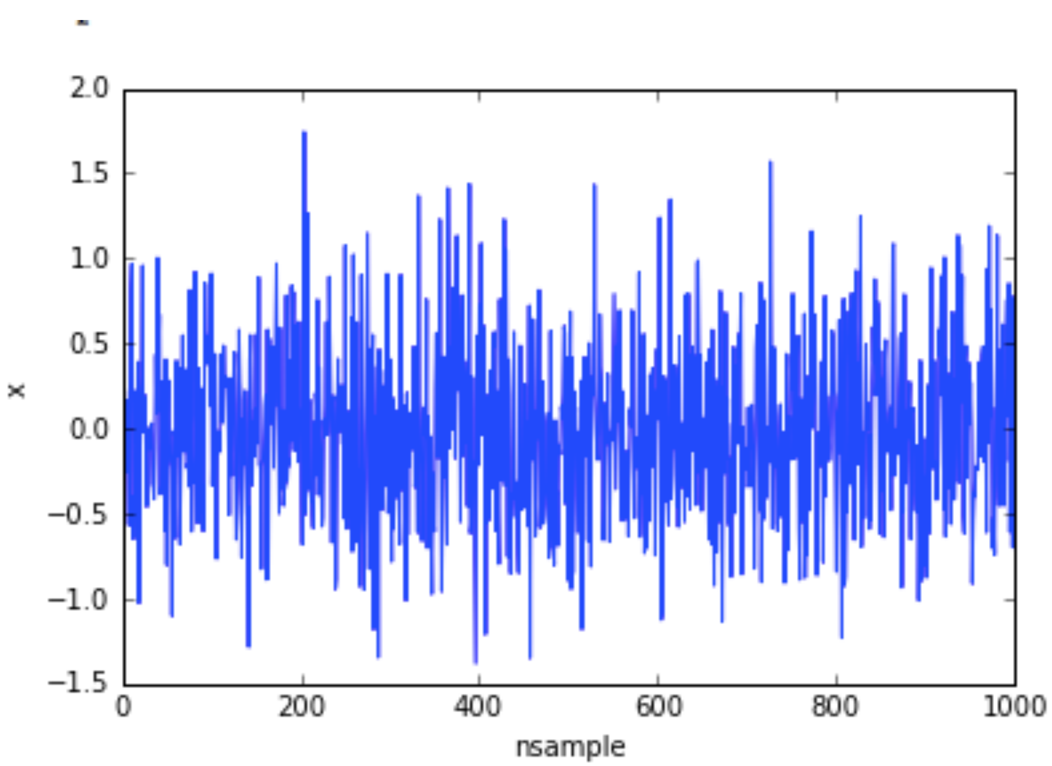
$$p(x \mid \mu, W) = \frac{1}{W} \text{for} \mid x - \mu \mid \leq \frac{W}{2}$$

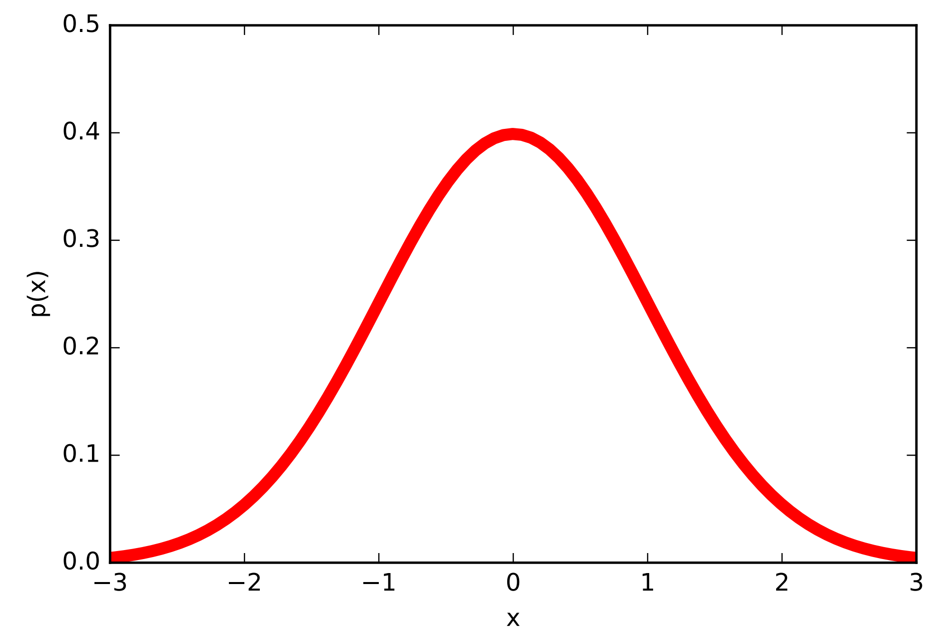
$$W = b - a$$

1D

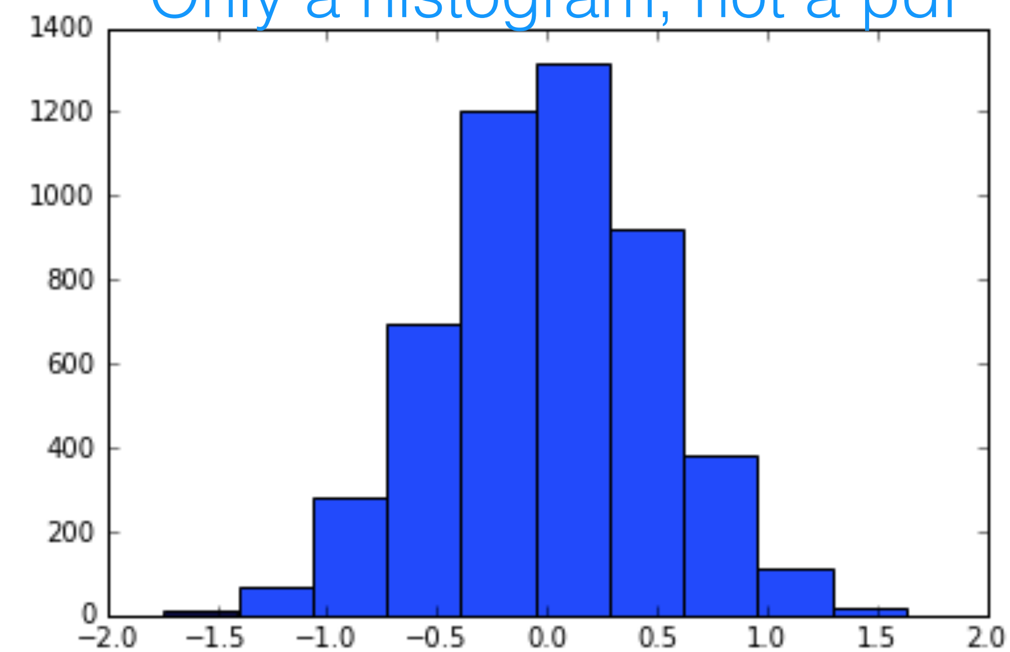
Probability distribution functions:

The next basic one: Gaussian/Normal





Only a histogram, not a pdf



$$p(x\,|\,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

Properties

- The convolution of two gaussian distribution is gaussian.

- Ej. $\mu_c = \mu_0 + b$ and $\sigma_c = \sqrt{\sigma_0^2 + \sigma_e^2}$ Where mu_o and sigma_o defines the distribution of some quantity we want to measure, and b and sigma_e defines de error distribution.

Convolution

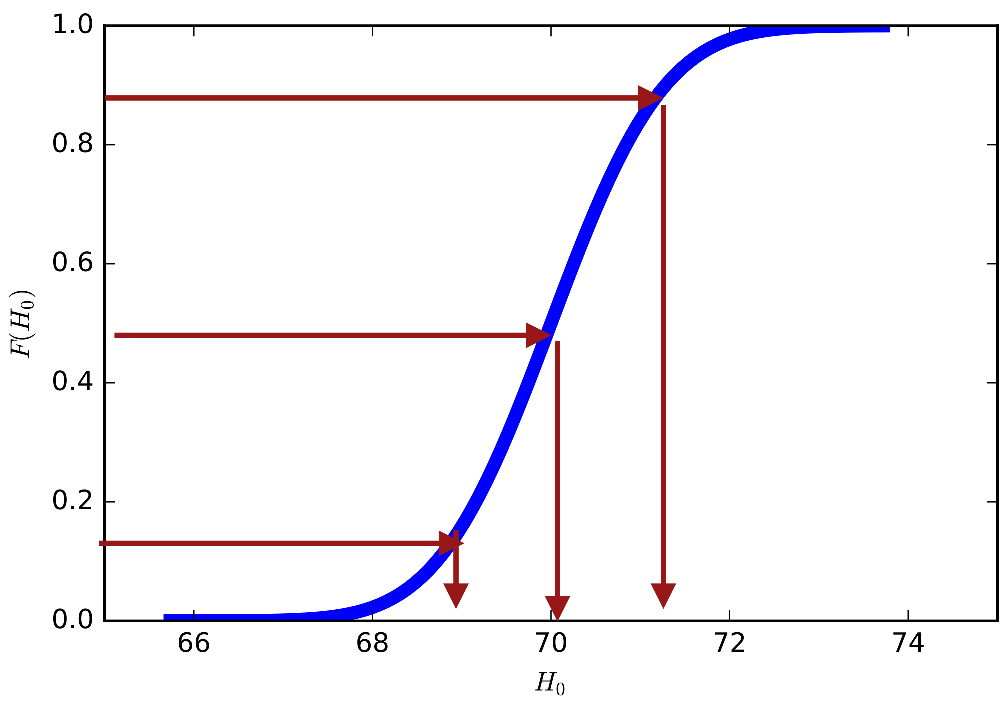
$$(f \star g)(x) = \int_{-\infty}^{\infty} f(x')g(x - x')dx' = \int_{-\infty}^{\infty} f(x - x')g(x')dx'$$

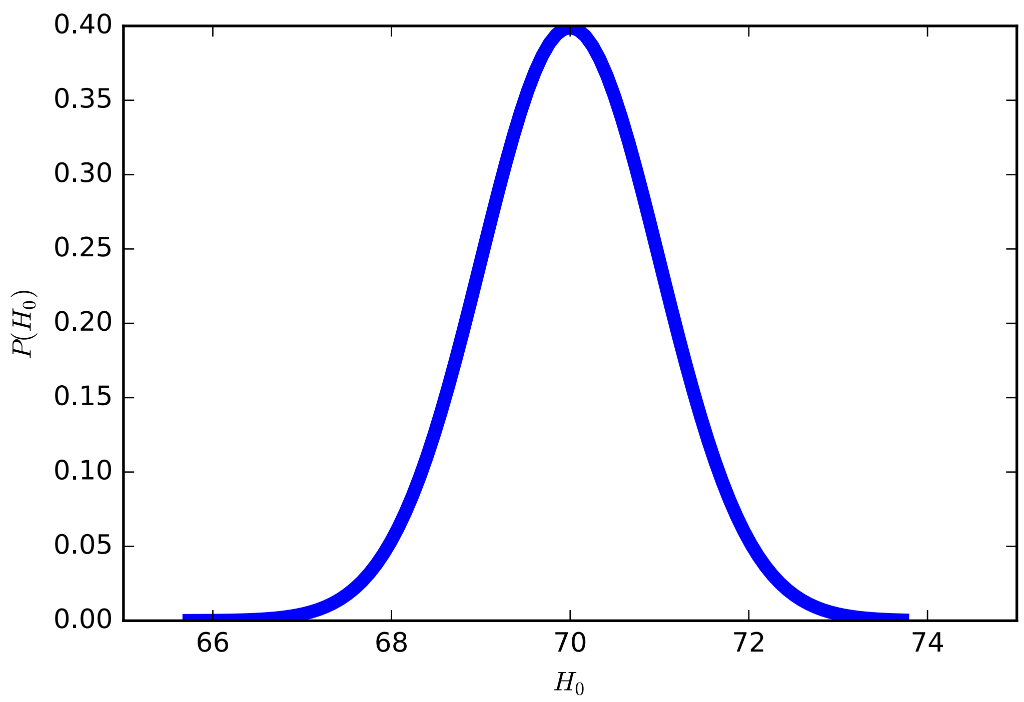
- Fourier transform of a Gaussian is a Gaussian.

- Central limit: the mean of samples drawn from almost any distribution will follow a Gaussian.

How to sample a PDF

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



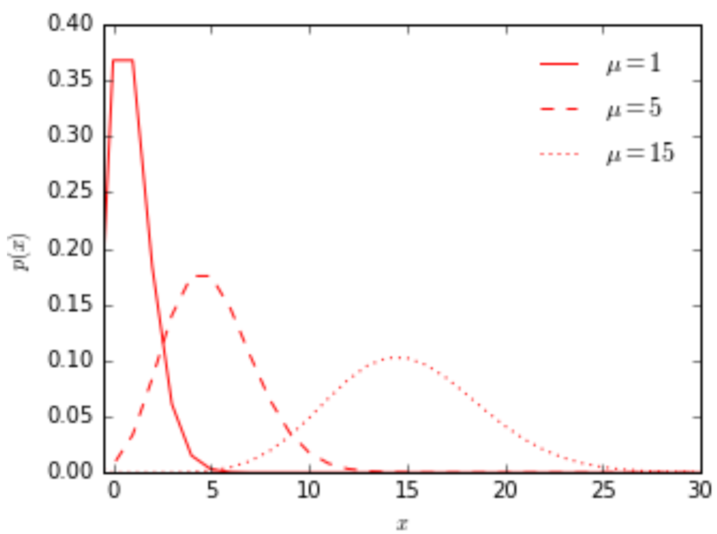


$$F(x) = \int_{-\infty}^{x} f(x')dx'$$

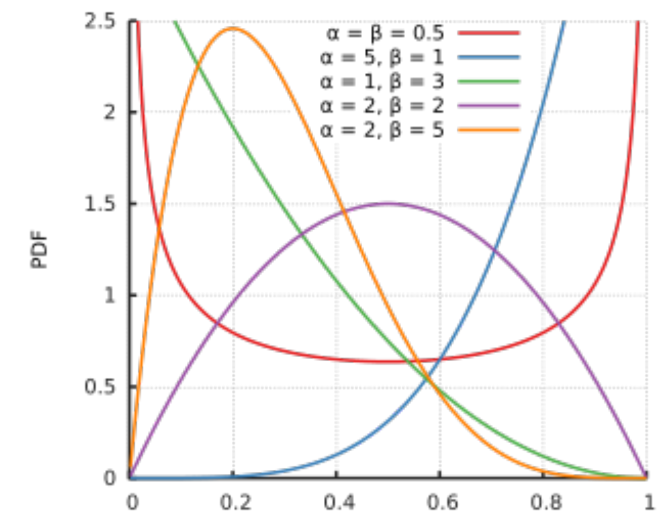
Homework: Find the cumulative distribution function for the Gaussian Distribution, and reproduce the plots. Choose a random number between 0 and 1, and use the CDF to assign the corresponding value of H0. Generate as many as you want, and make the histogram of H0 to verify you did it right. Use a mean of 70 and a sigma=2.

Other Probability distribution

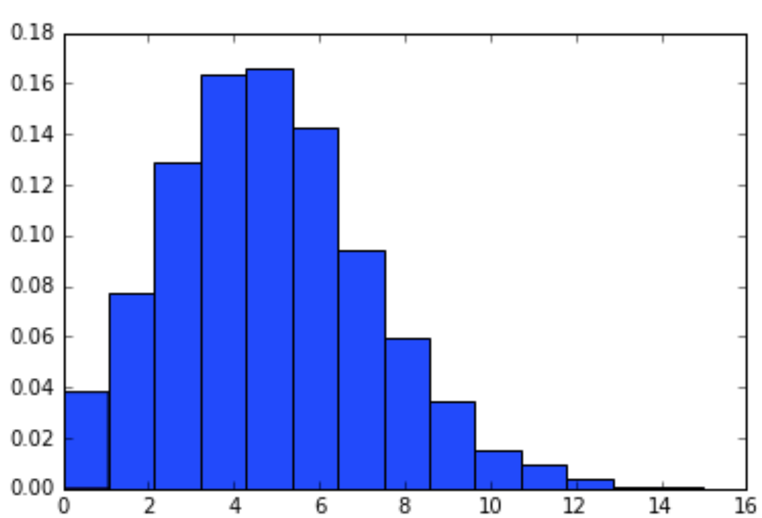
Binomial, Poisson and χ^2 distribution can be approximated, for large numbers, by a Gaussian distribution. Other distributions



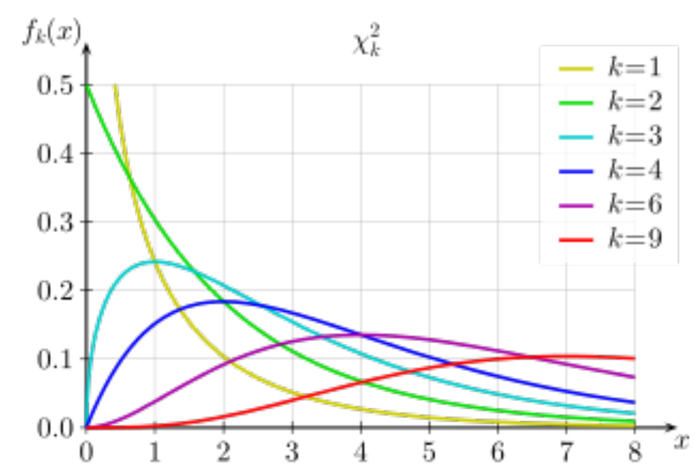
Poisson



- Beta distribution



Binomial



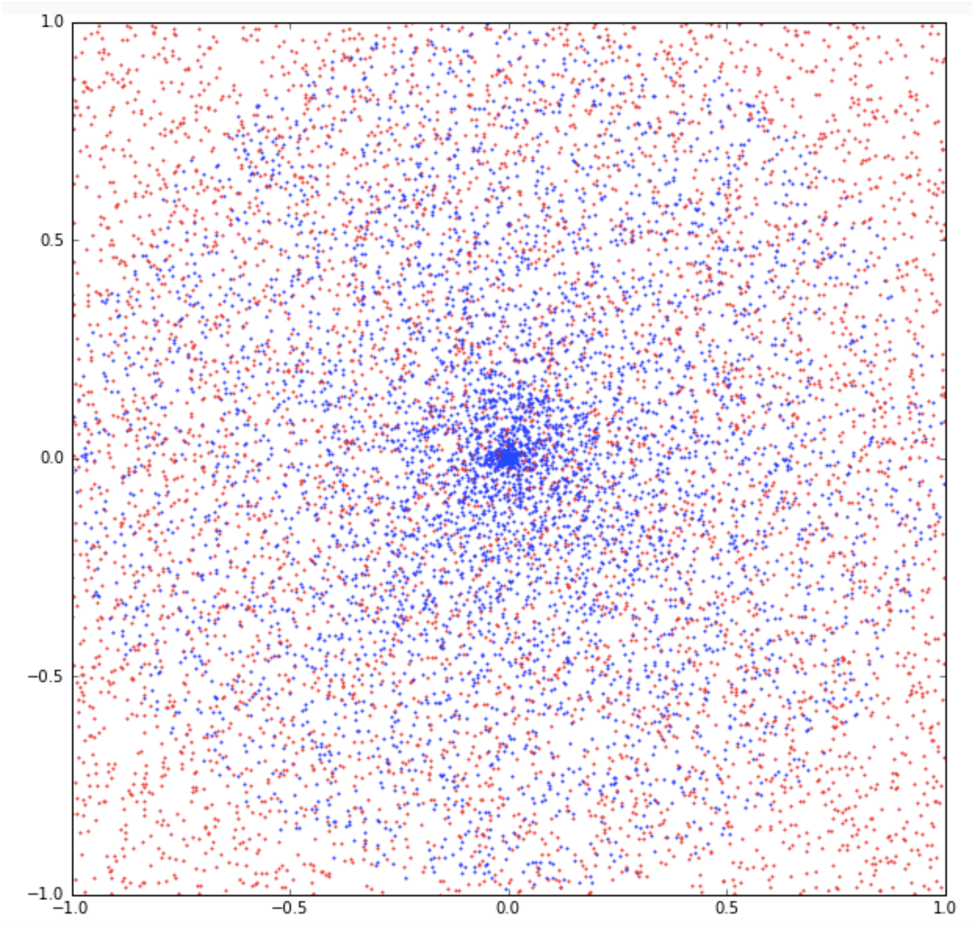
- χ^2 distribution

Homework

- For the Poisson and Binomial, etc... distributions mentioned, find the CDF mean and standard deviation. Plot both the PDF and CDF, for some different values of mean and sigma.
- Investigate about other useful distribution functions.

Draw samples from a specific distribution. Transformation of variables.

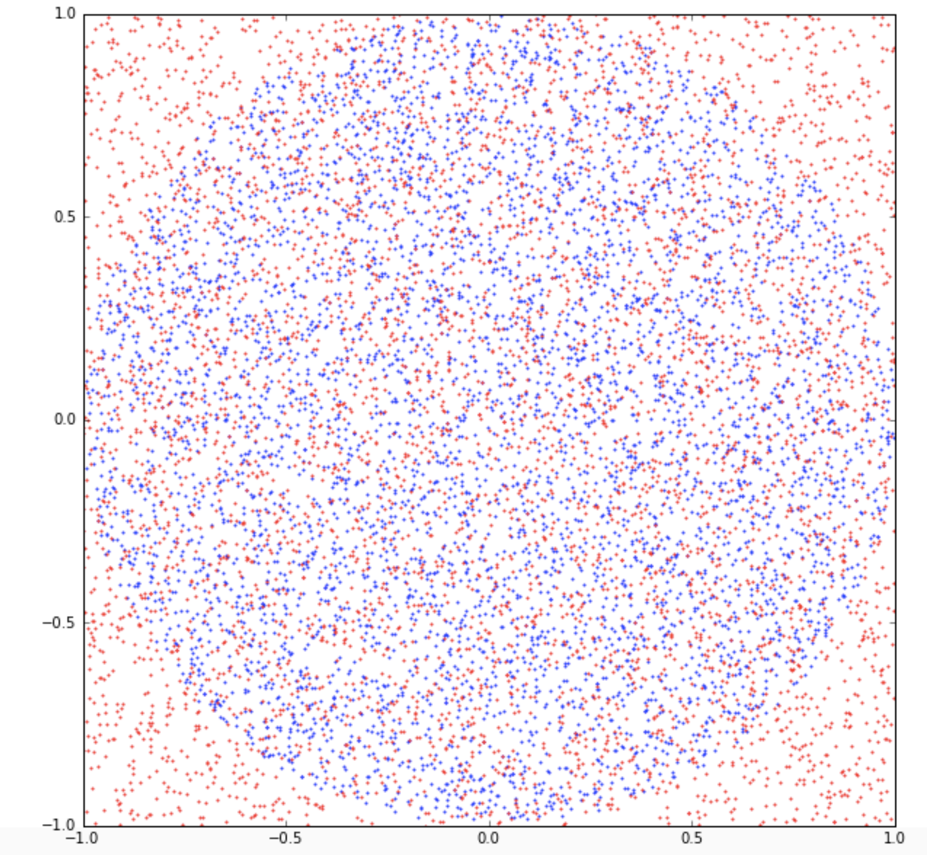
(Ej. 2D)



non-Uniform for r<1

1) generate r and theta, from U.

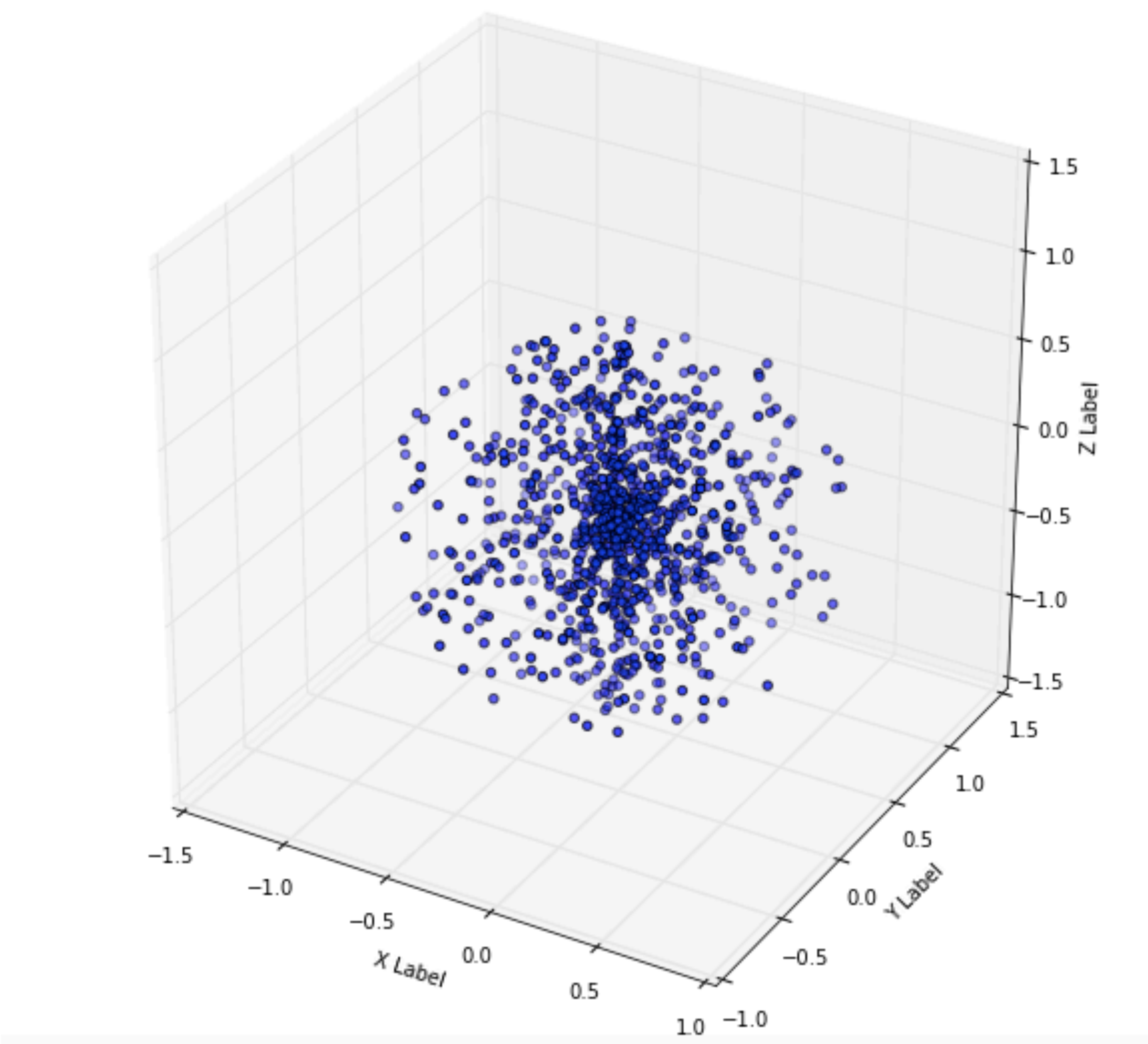
!=

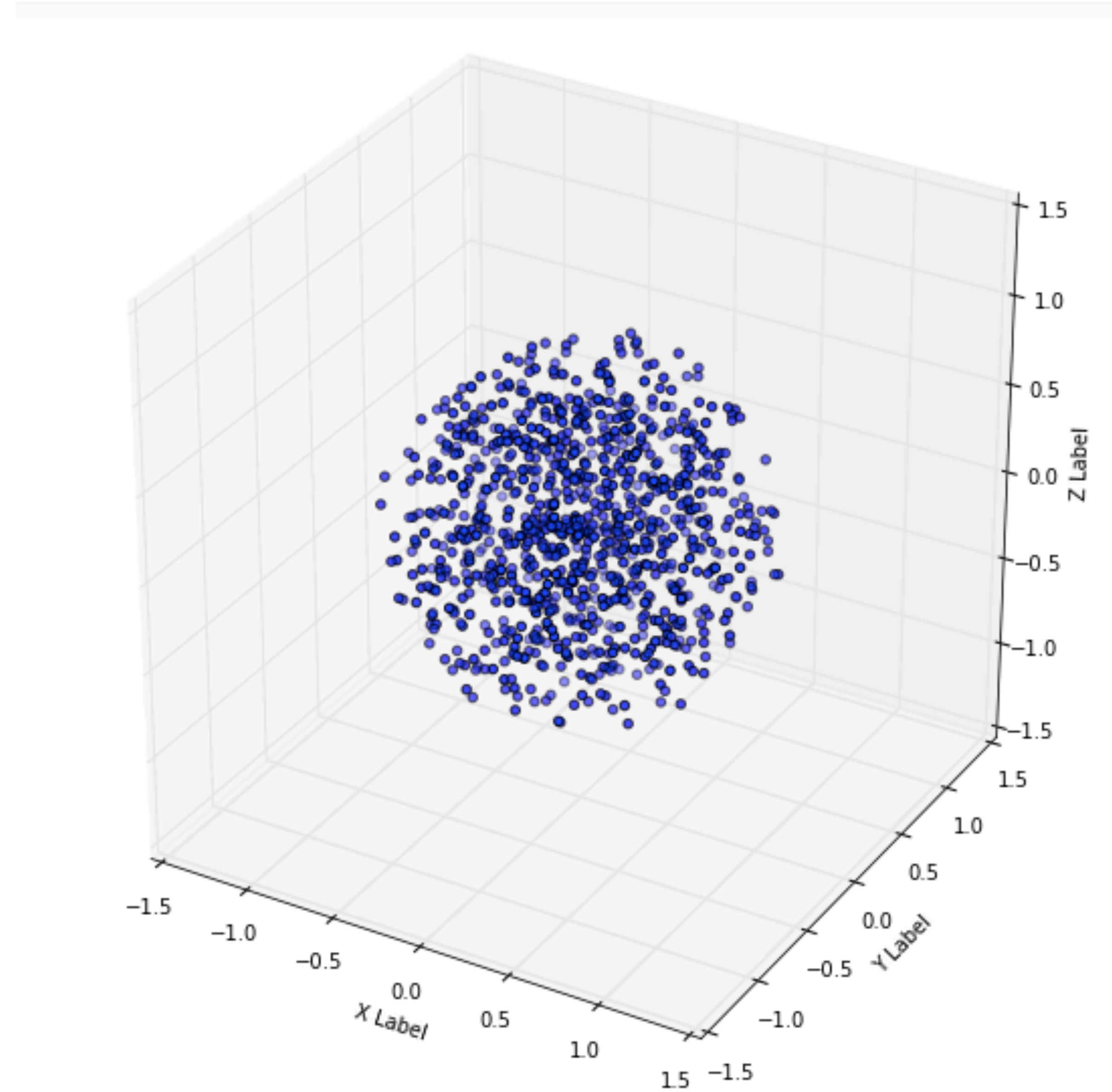


Uniform for r<1

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. Transformation of variables.





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.

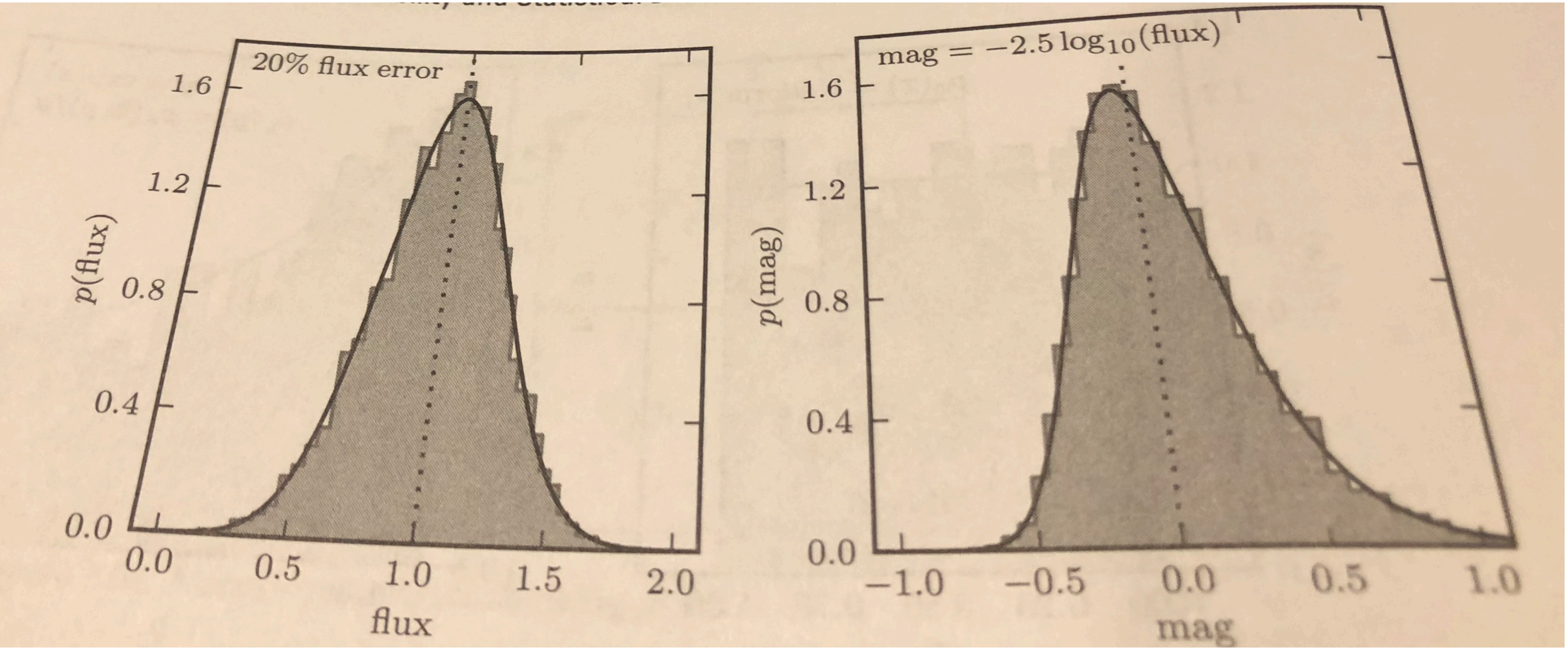
Transformation of variables.

- Any function of a random variable is a random variable itself.

- Sometimes we measure a variable x, but the interesting final result is y(x). If we know the PDF p(x), what is the PDF p(y)?, where y=Phi(x).

$$p(y) = p[\Phi^{-1}(y)] \left| \frac{d\Phi^{-1}(y)}{dy} \right|$$

Eg. Flux Vs Mag

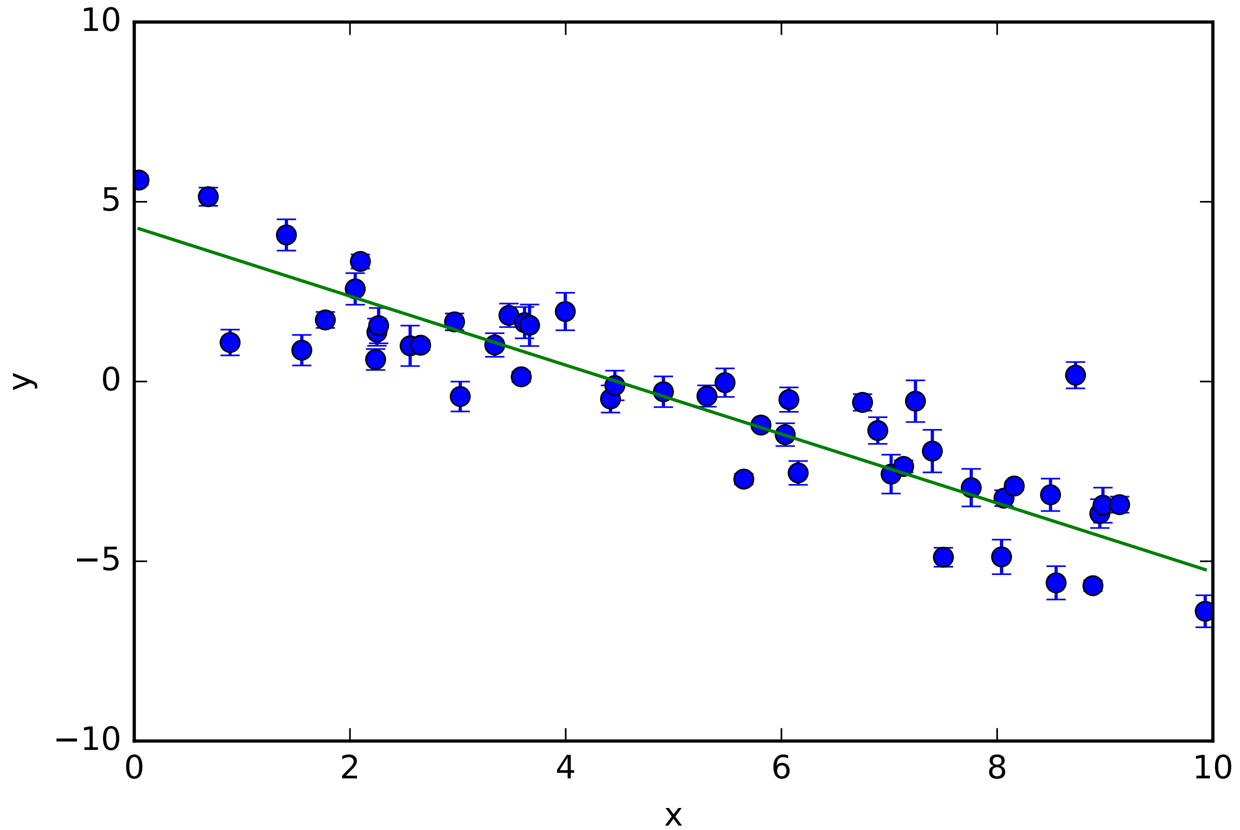


Exercise: If $y = \Phi(x) = \exp(x)$ and p(x)=1 for $0 \leq x \leq 1$ (a uniform distribution). What is the resultant distribution for y.

Inference

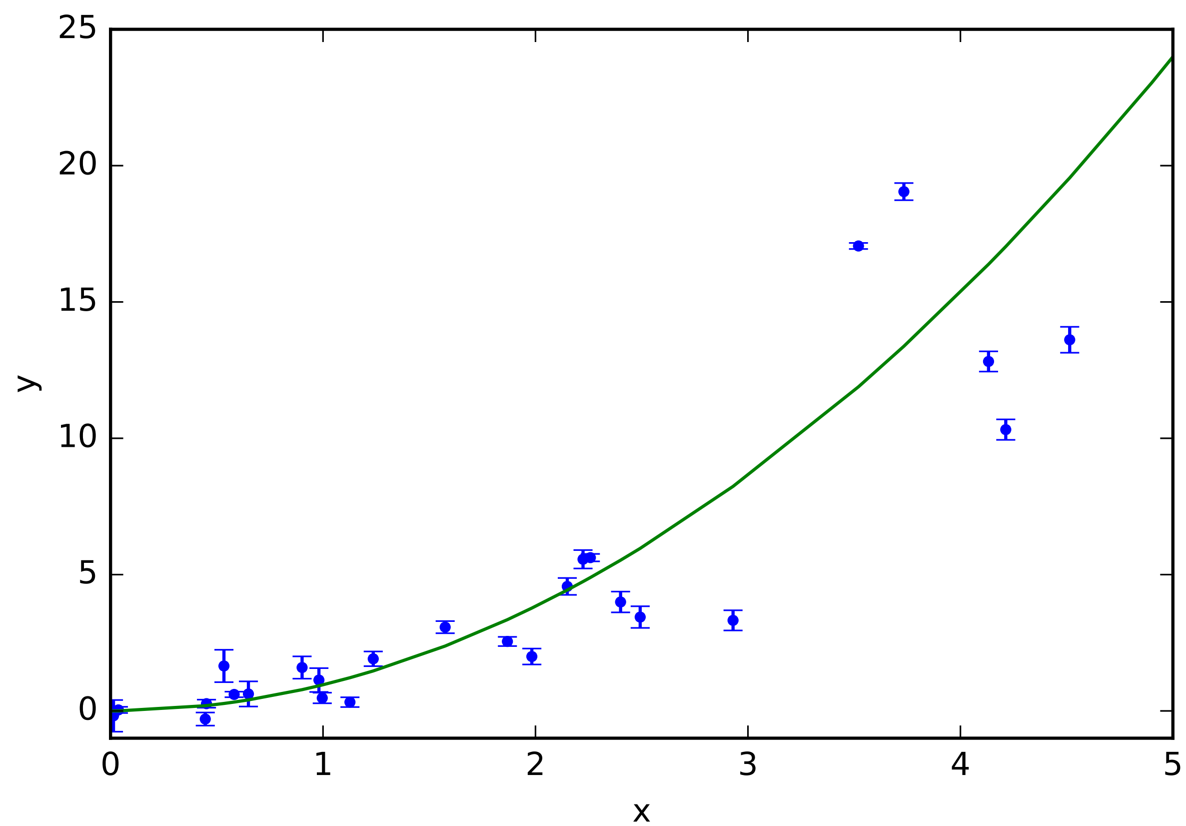
Parameter Estimation. Level 0

- 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 don't believe too much on the error bars?



Parameter Estimation. Level 0

Least square method.

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

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

Where this comes from?

Minimize the residual

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

assuming:

- A linear function f=ax+b.

- Errors are Gaussian and uncorrelated.

Minimization implies:

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

Parameter Estimation, Level 1

χ^2 Minimization

$$\frac{\partial \chi^2}{\partial \theta} = 0$$

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

θ : free parameters

σ_{y_i} : variance on y_i

Homework: show explicitly that the linear least square method is derived from the minimization of the chi-square when the model is a straight line.

Parameter Estimation and optimization

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

θ : free parameters

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

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.

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

- 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

For a Gaussian Likelihood , the minimization of the Chisq is equivalent to the maximization of the Likelihood. i.e. the best fit and the most likely model coincide.

Homework: Write down a Poisson likelihood, and identify which parameter(s) correspond to the data and which to the model.

Likelihood

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

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

What if the likelihood is not Gaussian...?

How do we estimate errors on our inferred parameters?

Exercise:

Infer the parameters for a linear model that describes the data (available in "data" directory) using the three methods.

1.- Liner least square

2.- Minimum Chisq.

3.- Sampling from the maximum likelihood.

Tip: Use libraries like: numpy.linalg.lstsq, scipy.optimize.minimize

Define generic functions that could be used for different data sets, and different models.

What are the values of the parameters in each case, and the errors?