Statistics and Probability I

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 LEcture notes.

Statistics

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

In general

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

- ¿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?

Probability (some definitions)

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"

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"

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. Next Class.

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 ,
$$F(x_k) \equiv P(\leq x_k) = \sum_{xi \leq x_k} f(x_i) \qquad \text{CDF}$$
- 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. $f(x)$

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

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

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

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

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

Centra Limit Theorem

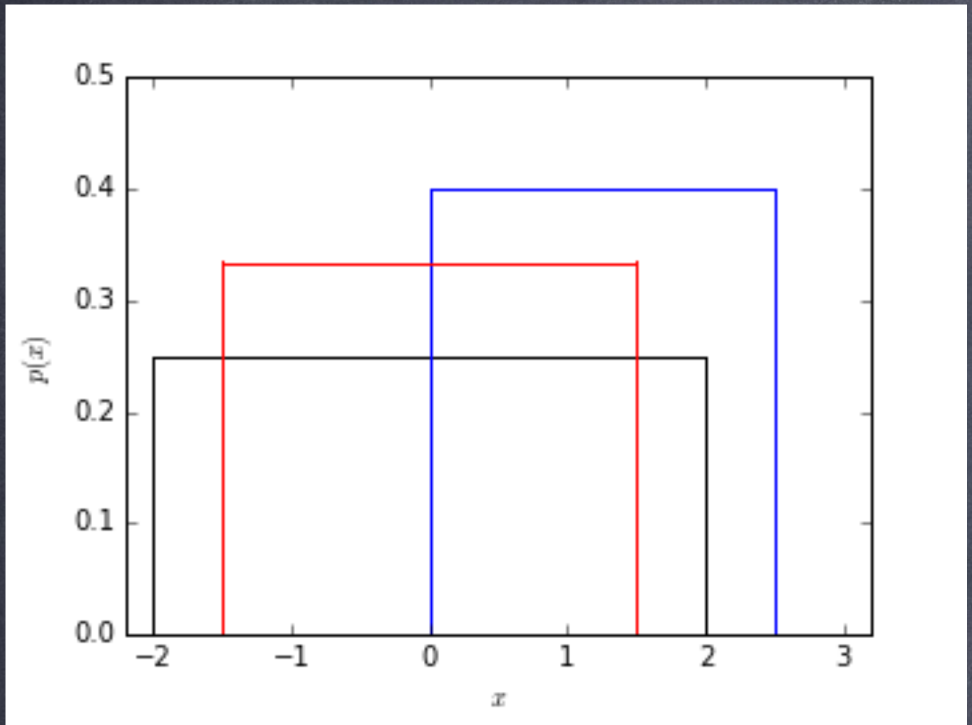
- The mean and variance of a linear combination of random variables is given by:

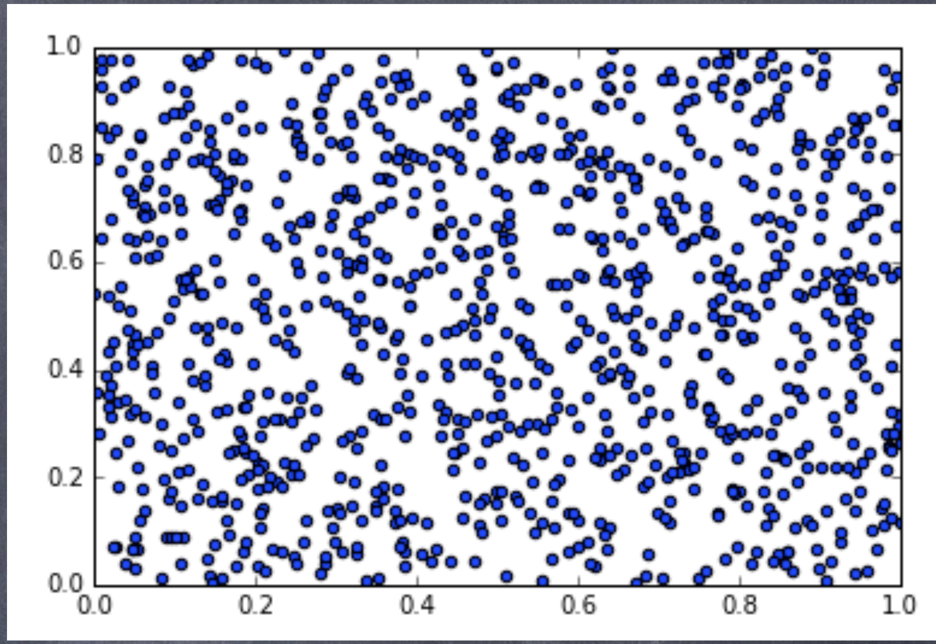
$$Y = \sum_{i=1}^{n} c_i X_i \qquad \sigma_Y^2 = \sum_{i=1}^{n} c_i^2 \sigma_i^2$$

- 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



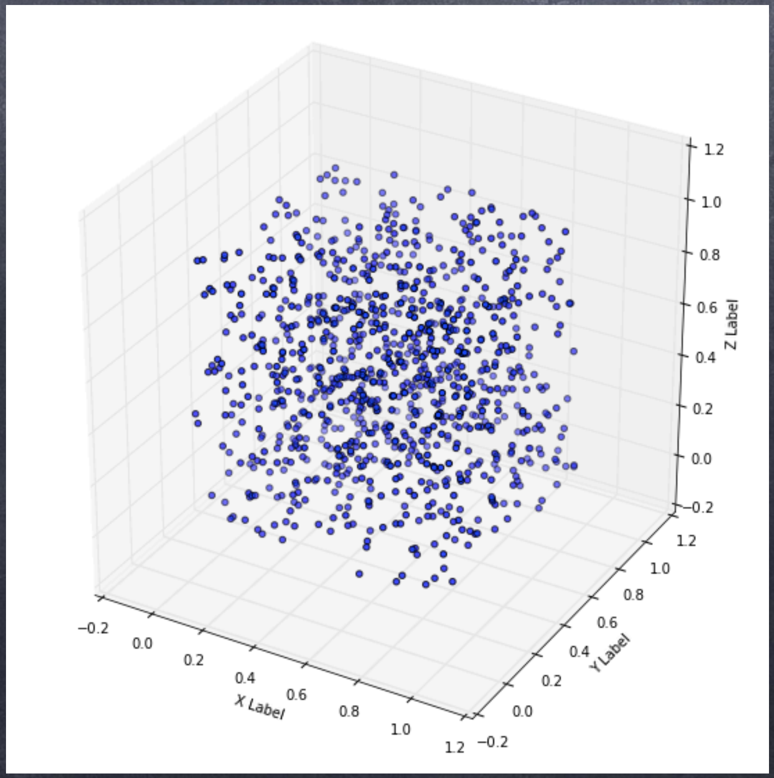


2D

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

$$W = b - a$$

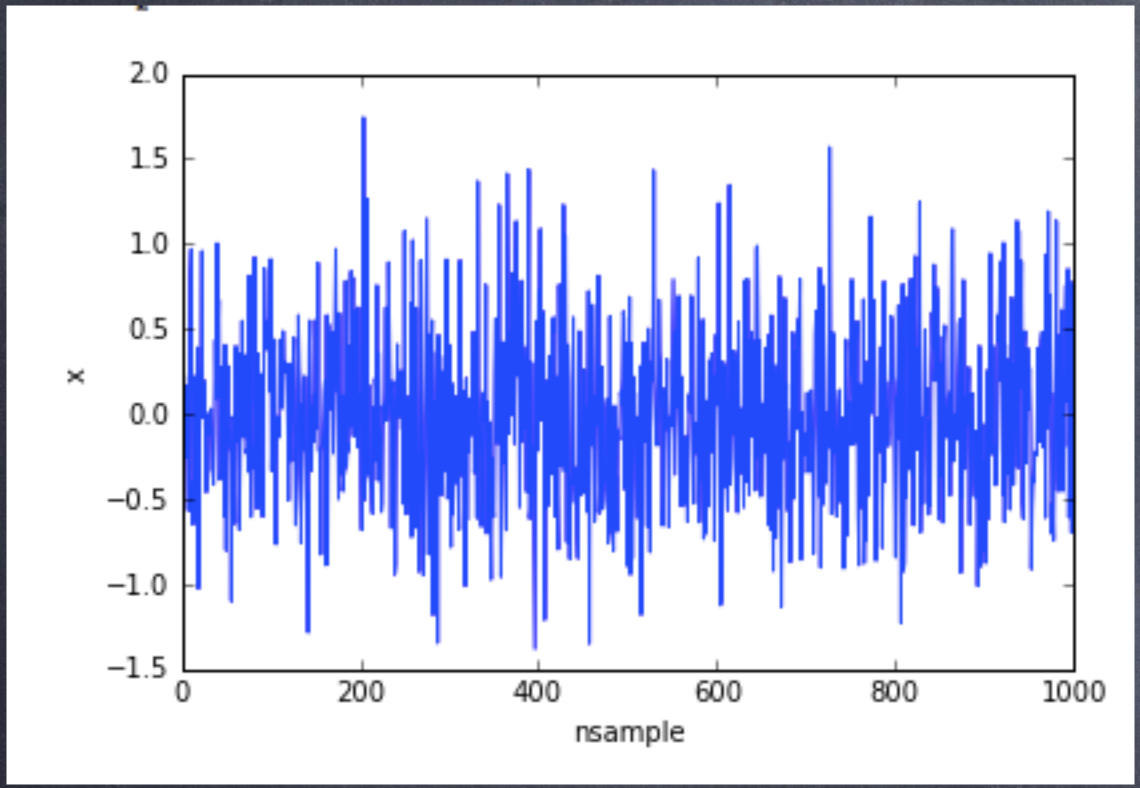
1D

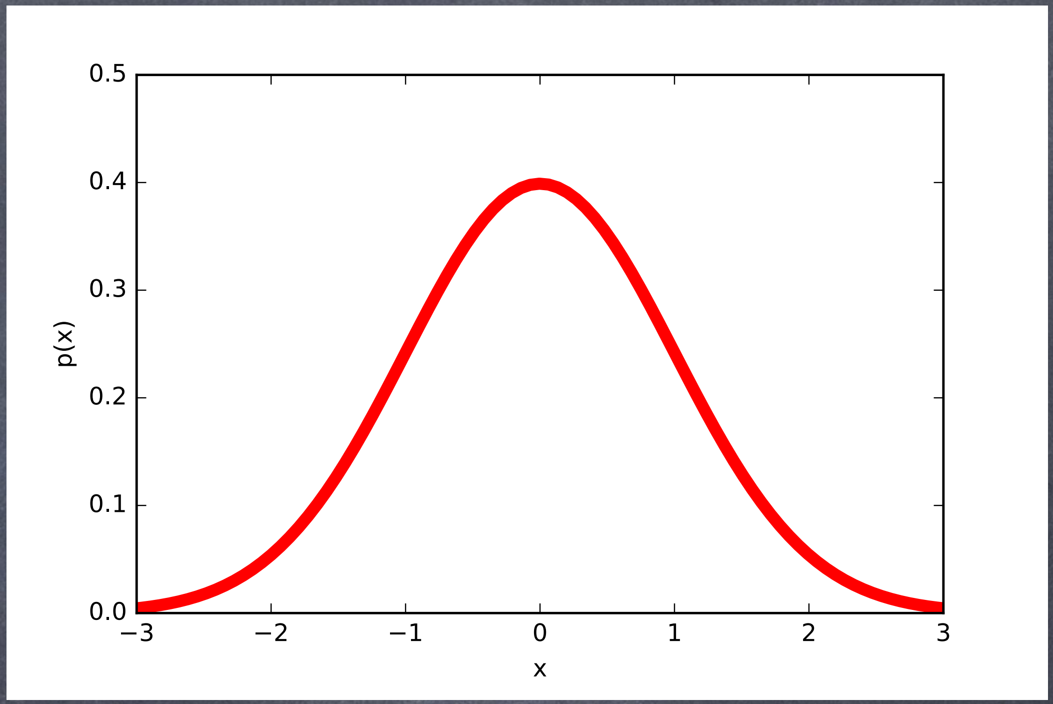


3D

Probability distribution functions:

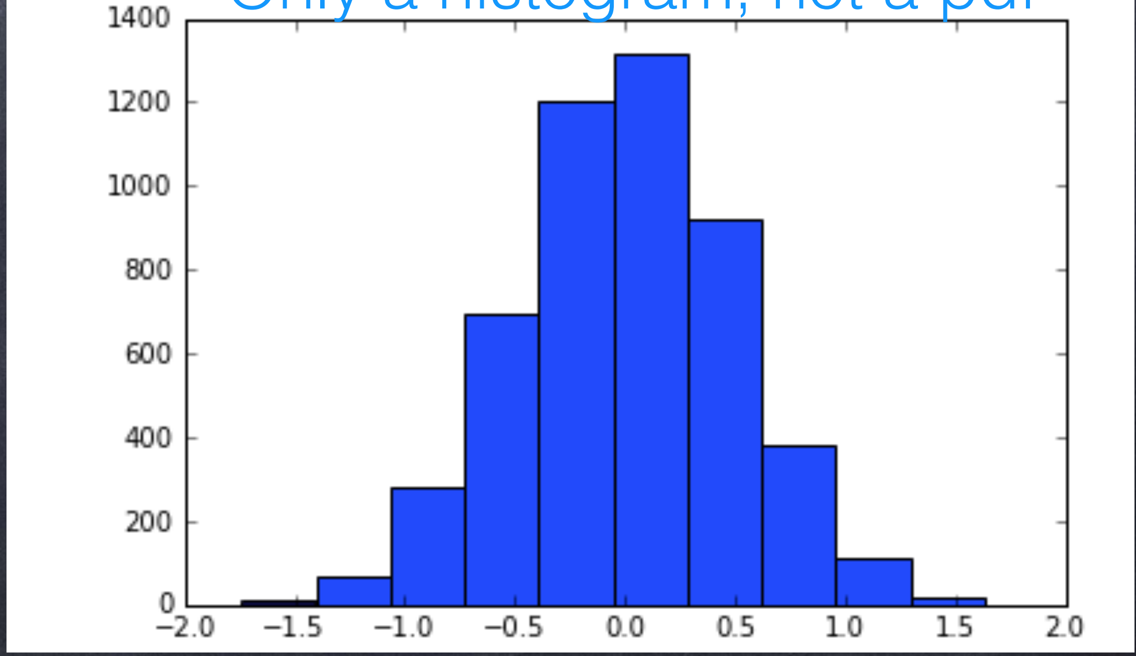
The next basic one: Gaussian/Normal





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

Only a histogram, not a pdf



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_0 and sigma_0 defines the distribution of some quantity we want to measure, and b and sigma_e defines de error distribution.

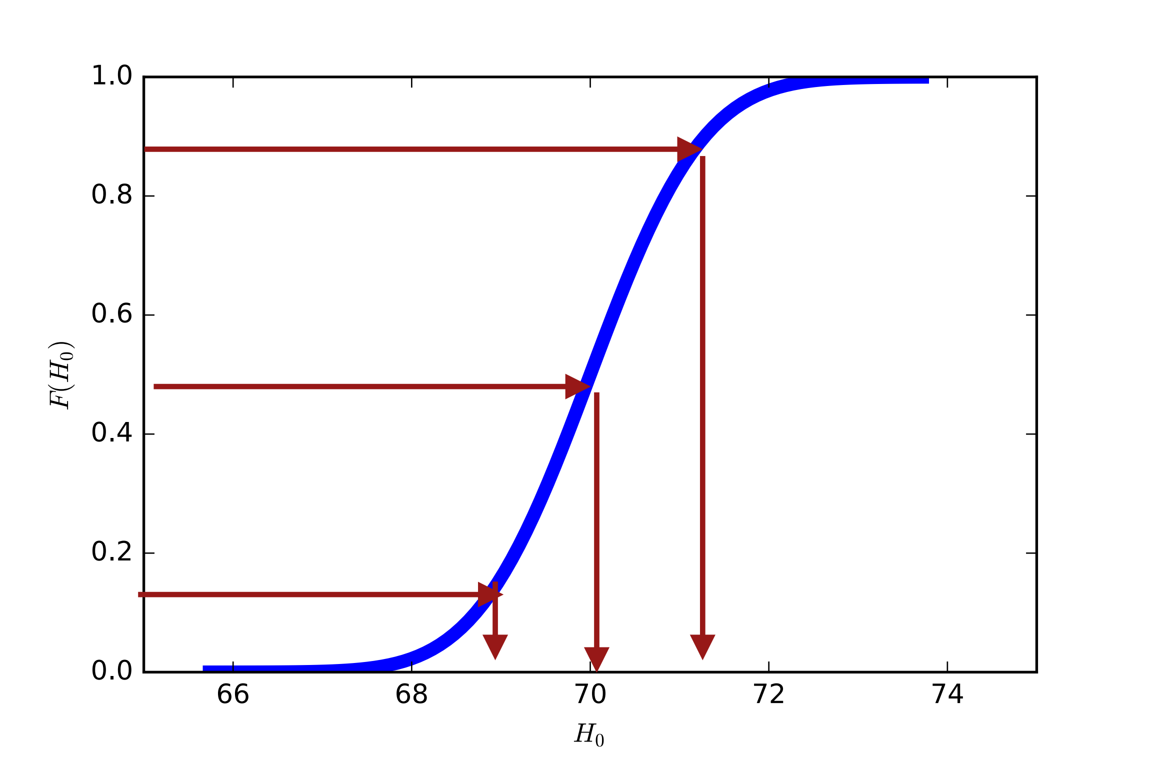
Convolution

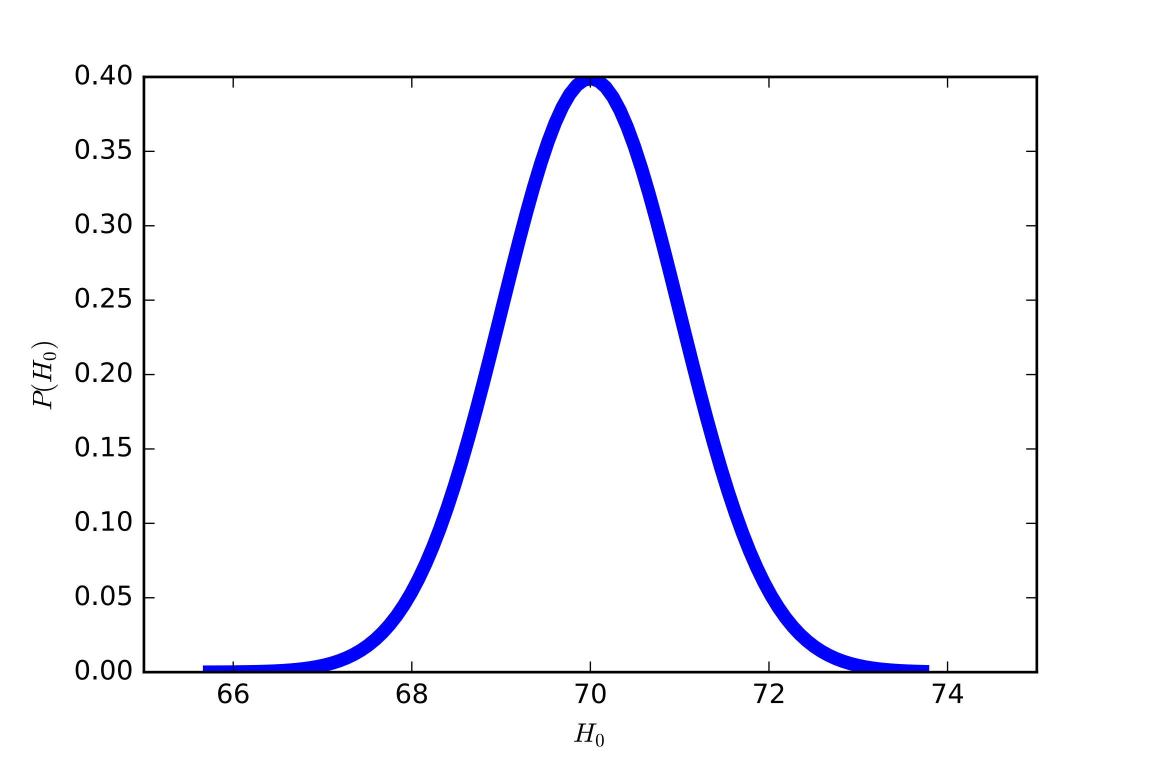
$$(f \star g)(x) = \int_{-\infty}^{\infty} f(x')g(x-x')dx' = \int_{-\infty}^{\infty} f(x-x')g(x')dx'$$

- 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.



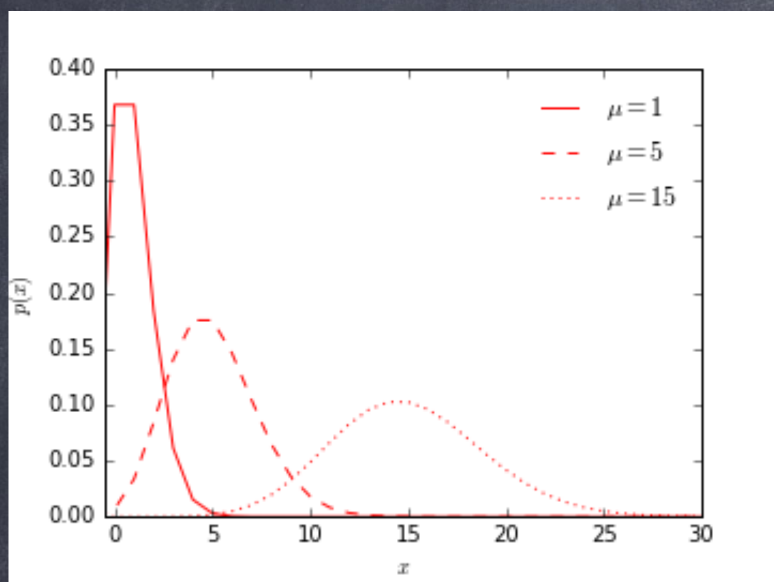


$$F(x) = \int_{-\infty}^{x} f(x')dx'$$

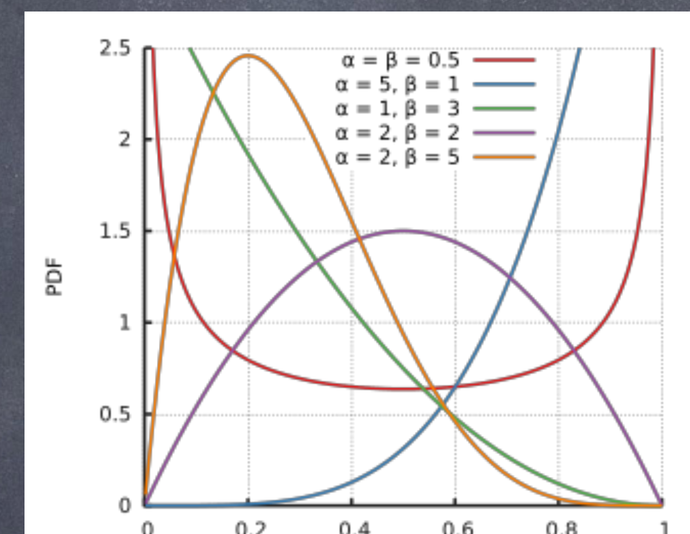
Exercise: Find the cumulative distribution function for the Gaussian Distribution, and reproduce the plots. Choose a random number between 0 and 1, and use the CDF to assign the corresponding value of H0. Generate as many as you want, and make the histogram of H0 to verify you did it right. Use a mean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

Other Probability distribution

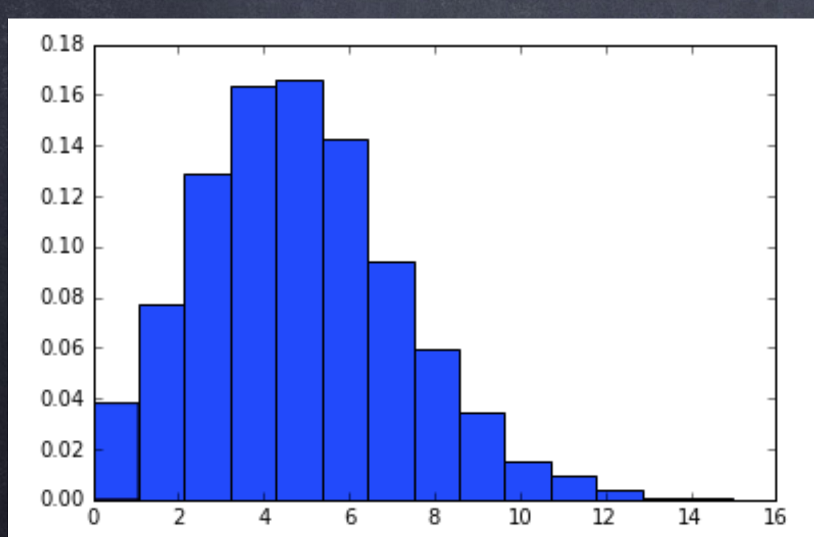
- Binomial, Poisson and χ^2 distribution can be approximated, for large numbers, by a Gaussian distribution. Other distributions



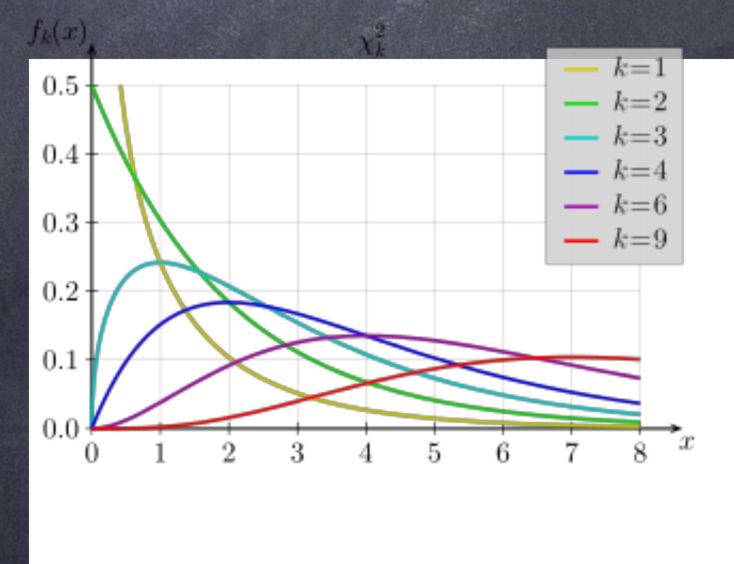
- Poisson



- Beta distribution



- Binomial



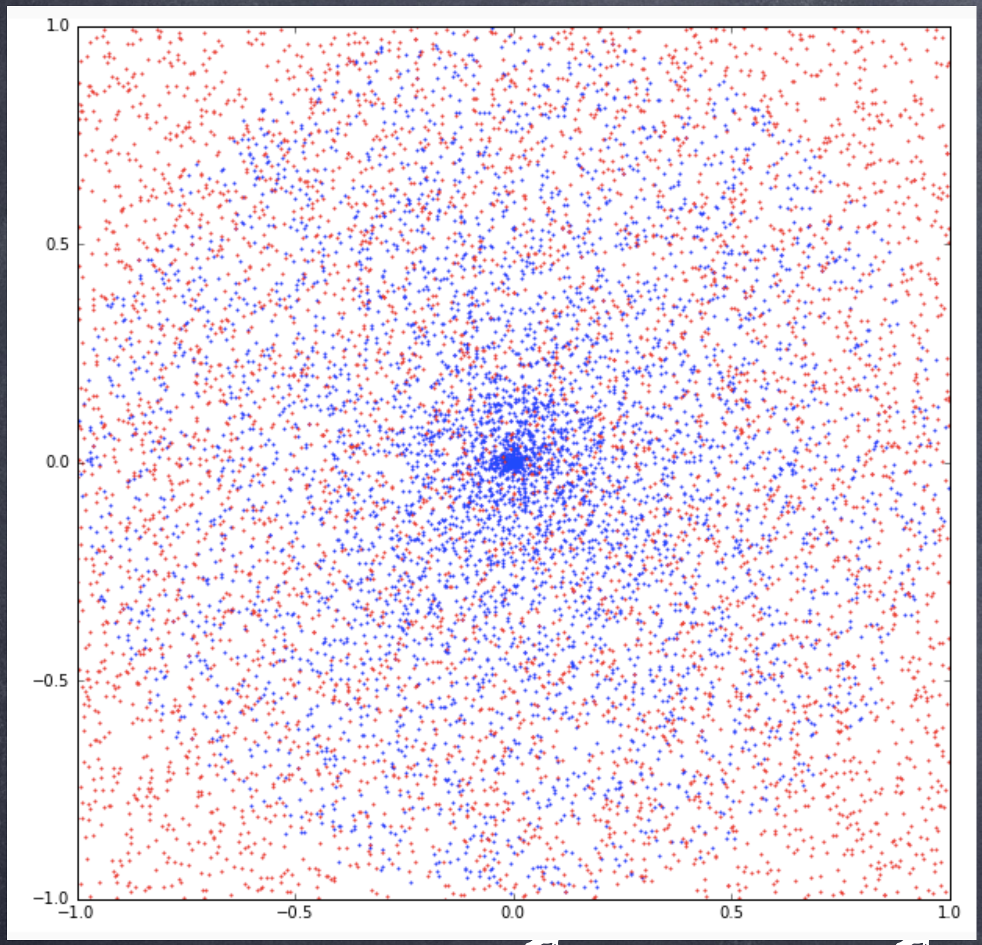
- χ^2 distribution

Exercise

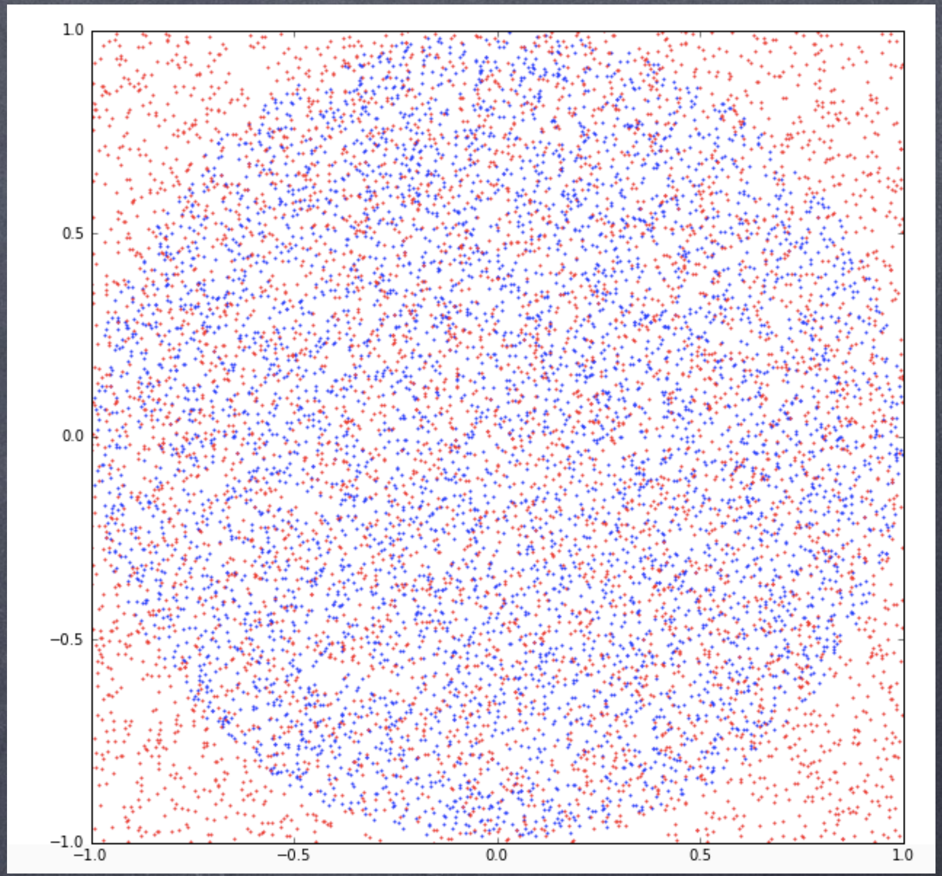
- For the distributions mentioned, find the CDF mean and standard deviation. Plot both the PDF and CDF, for some different values of mean and sigma. . Use a random number generator to compute sample the pdf and compare it with the CDF using the corresponding python function.
- Investigate about other useful distribution functions.

Draw samples from a specific distribution. Transformation of variables.

(Ej. 2D)



!=



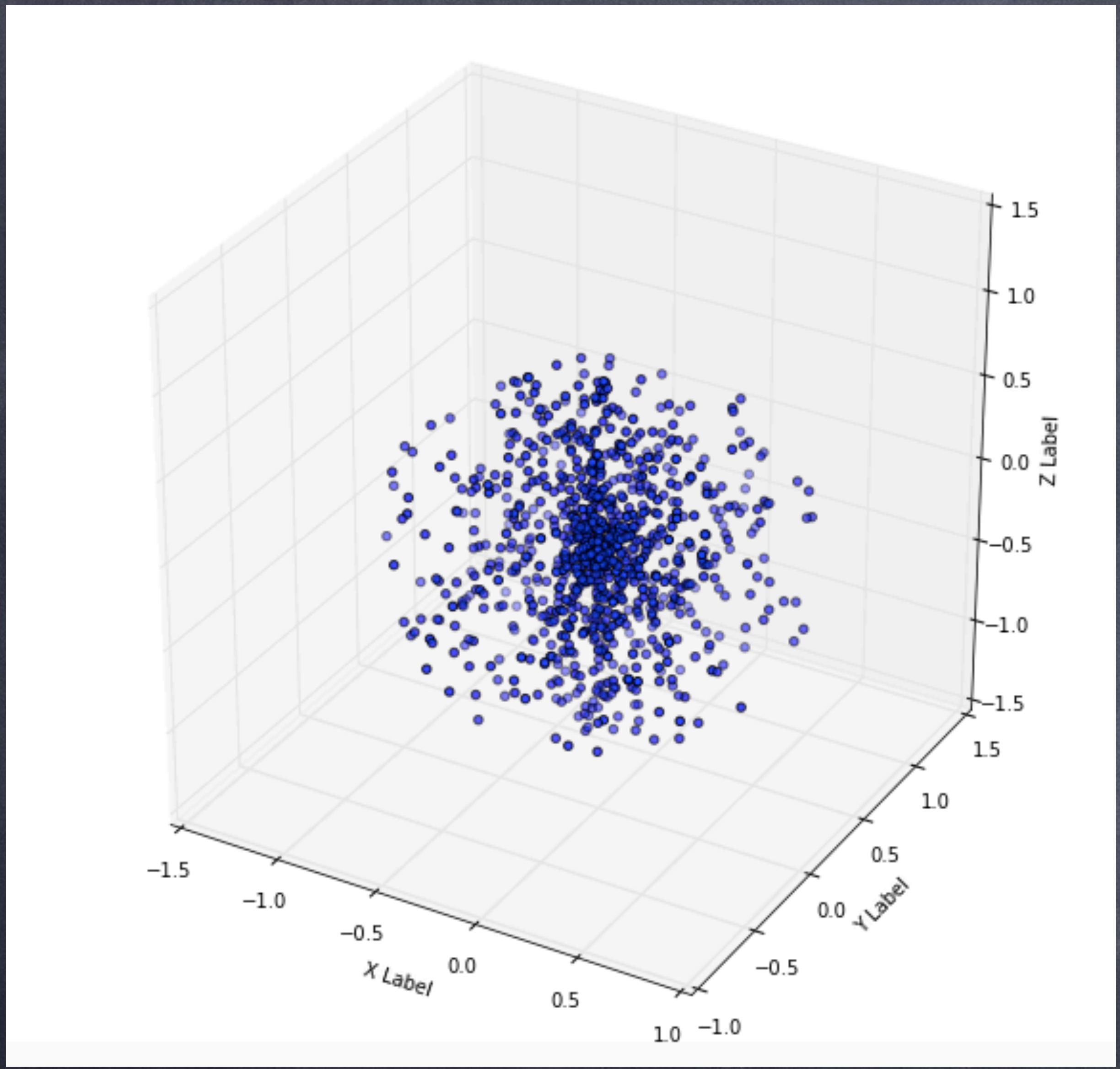
non-Uniform for r<1

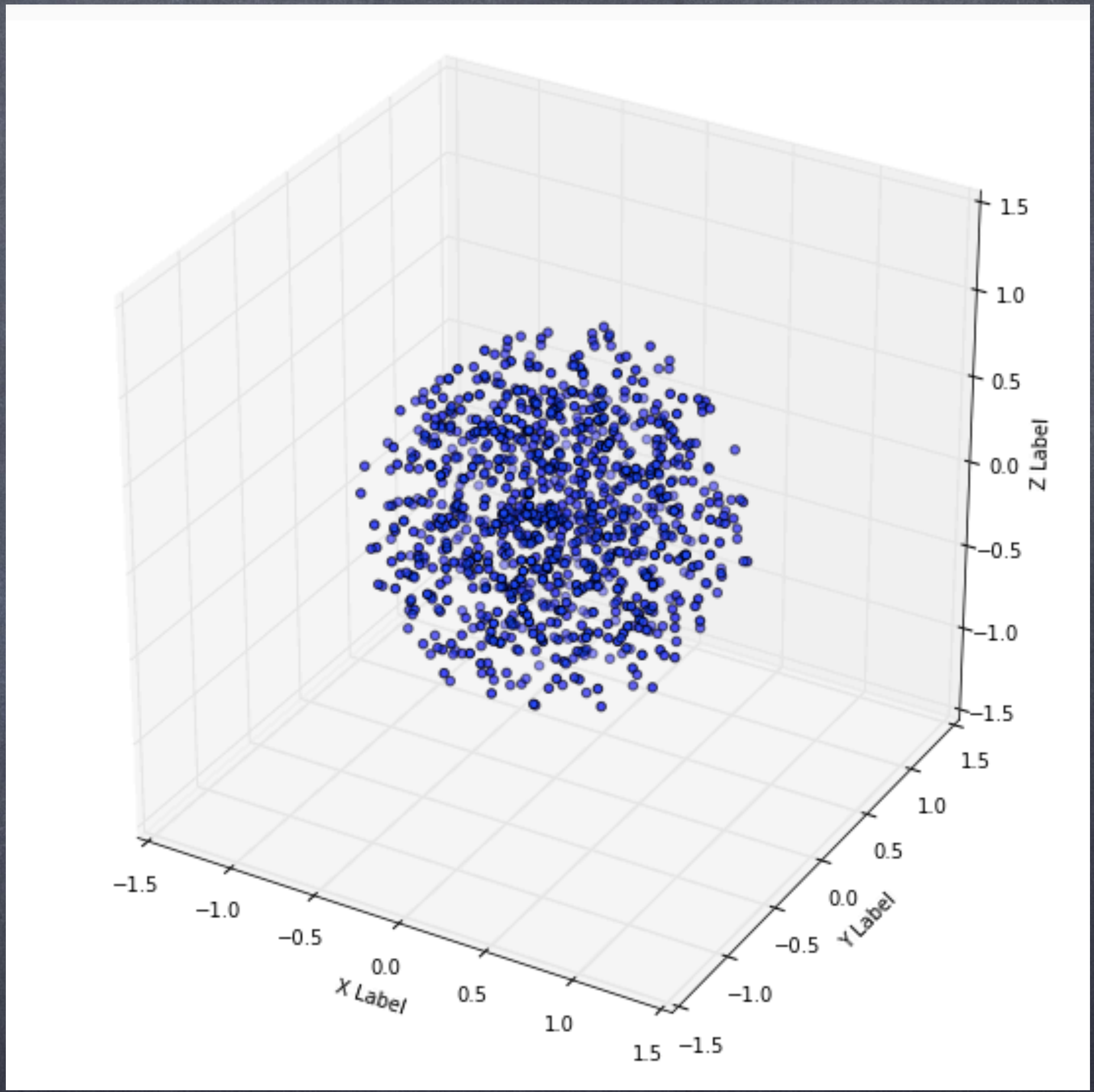
1) generate r and theta, from U.

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.

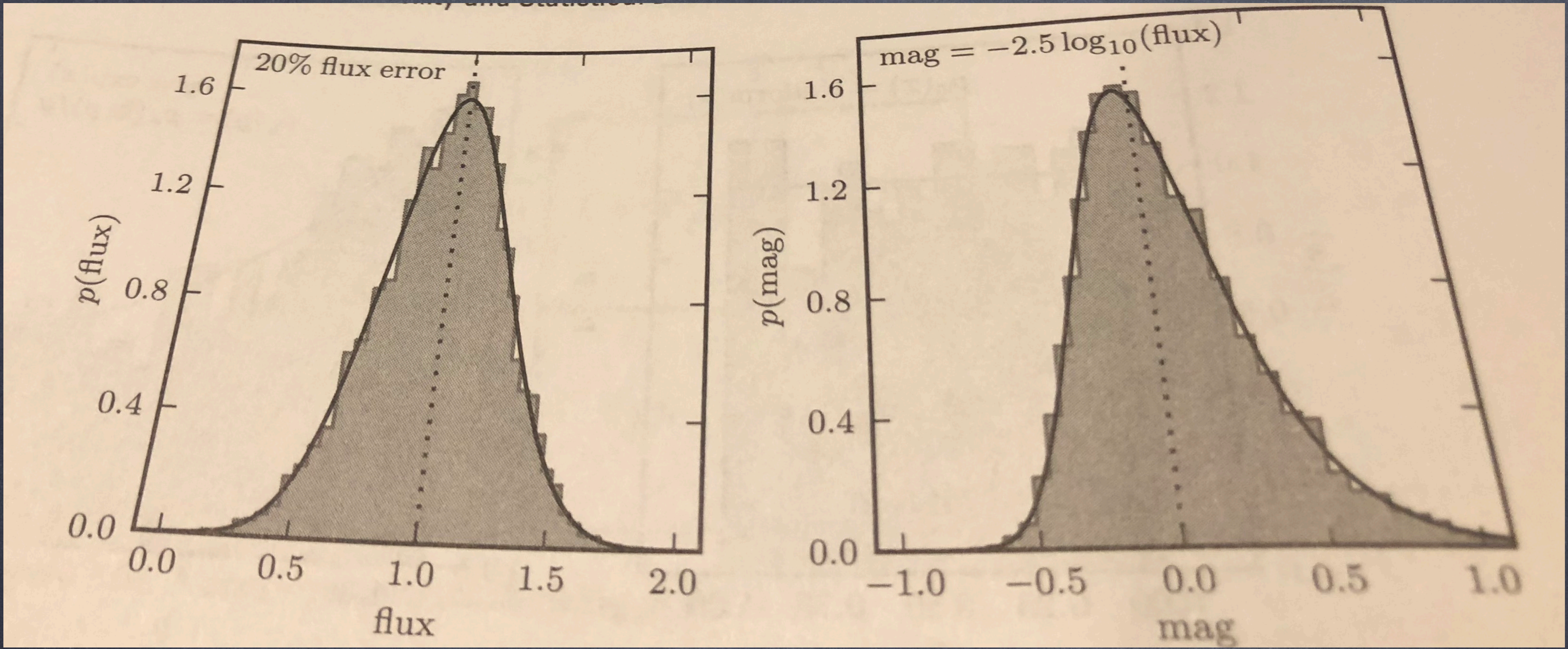
1) generate x,y from U
r=sqry(x^2+y^2+z^2)<=1
2) generate sqrt[r], theta and z=cos(phi) from U.

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). In 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.

Exercise

- Reproduce the plots of the Uniform distribution of points inside a circle, and a sphere.