MexiCOPAS

Mexican Cosmology Particles and Strings Schools

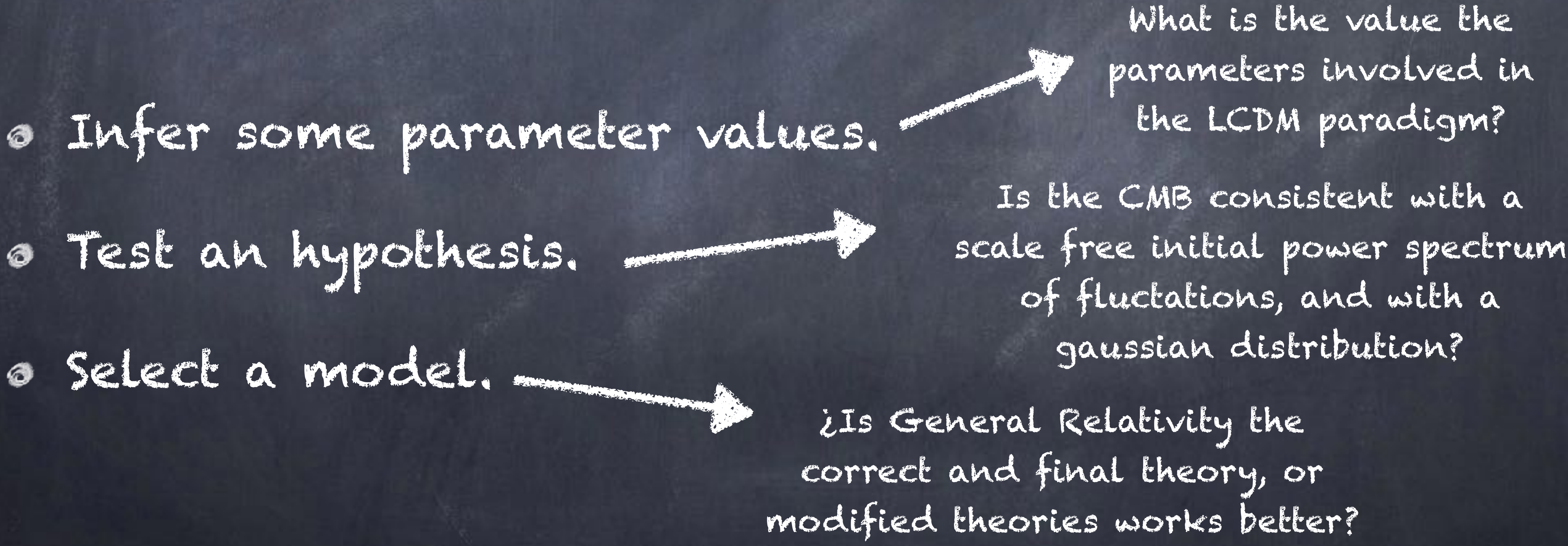
Statistics and Probability I (Introduction to MACSS)

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.

- ¿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.
- Test an hypothesis.
- Select a model.

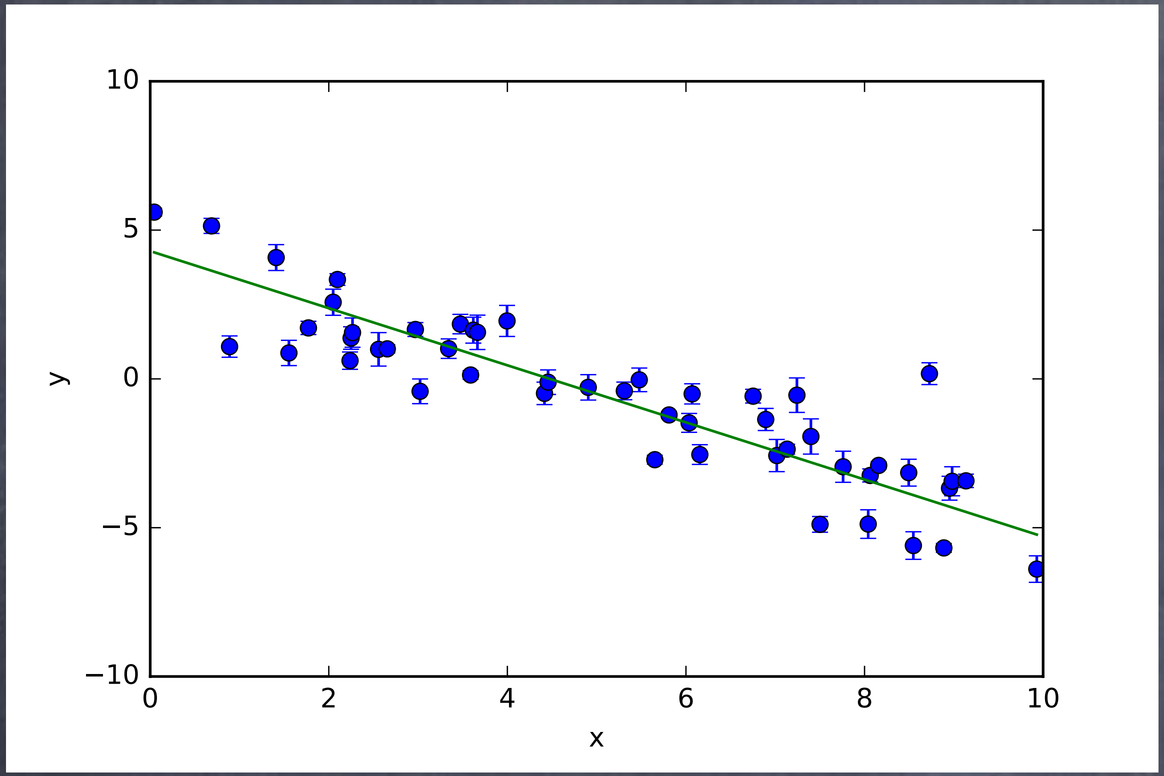
What is the value the parameters involved in the LCDM paradigm?

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

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

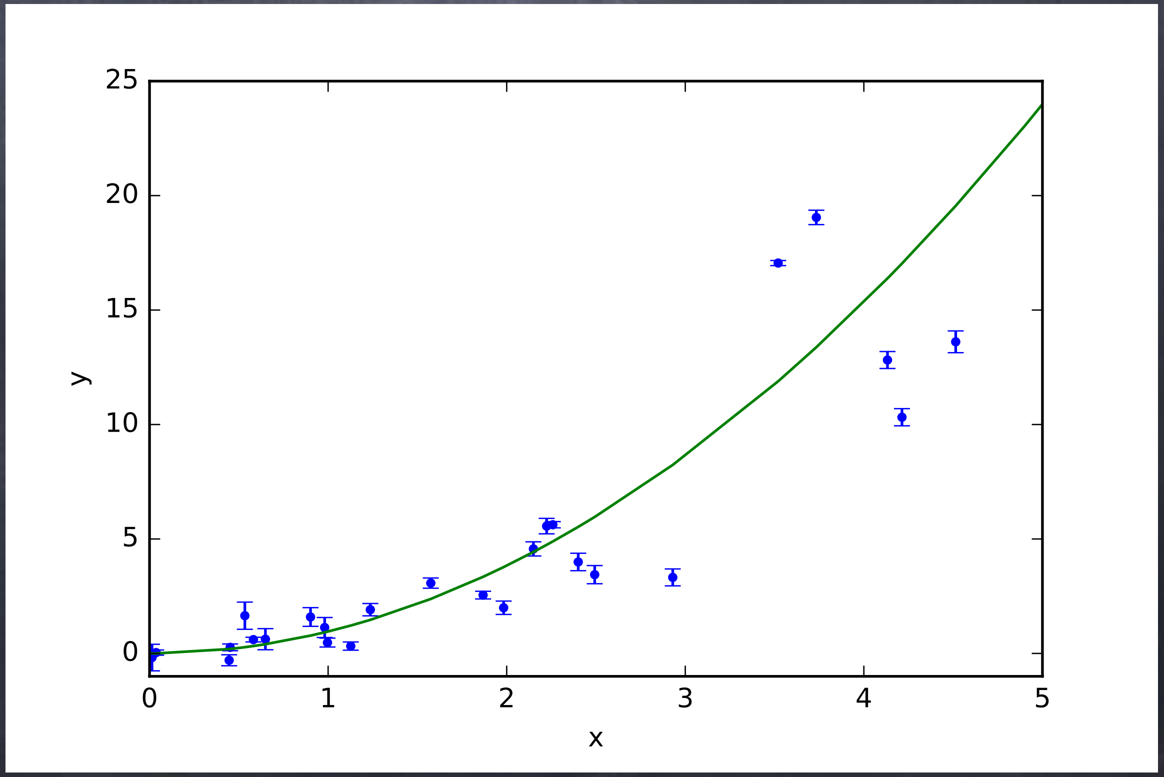
Parameter Estimation

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



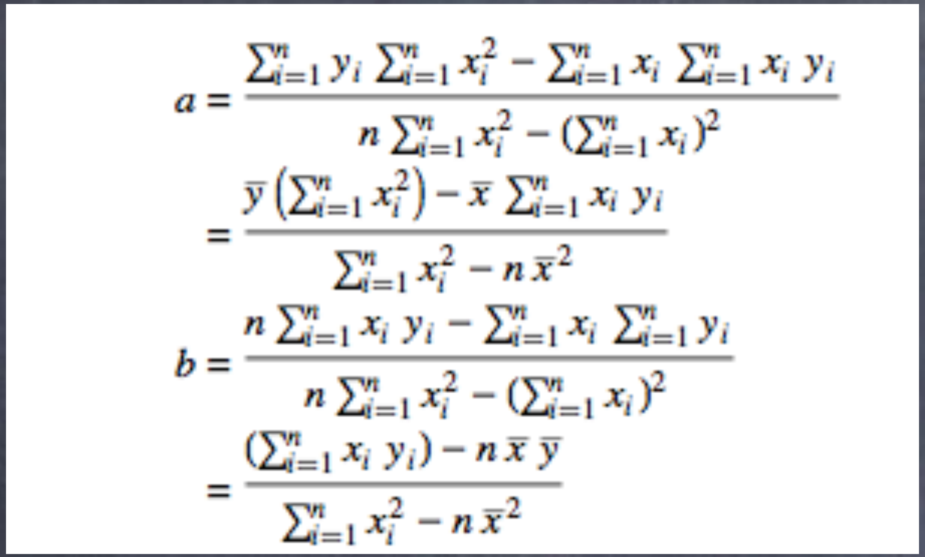
- Linear least square method

- 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*



Parameter Estimation. Level 0

- Least square method.



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

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. 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 ony_i

We will see how least square, and minimum Chi^2 methods are just special cases of Bayesian Inference.

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

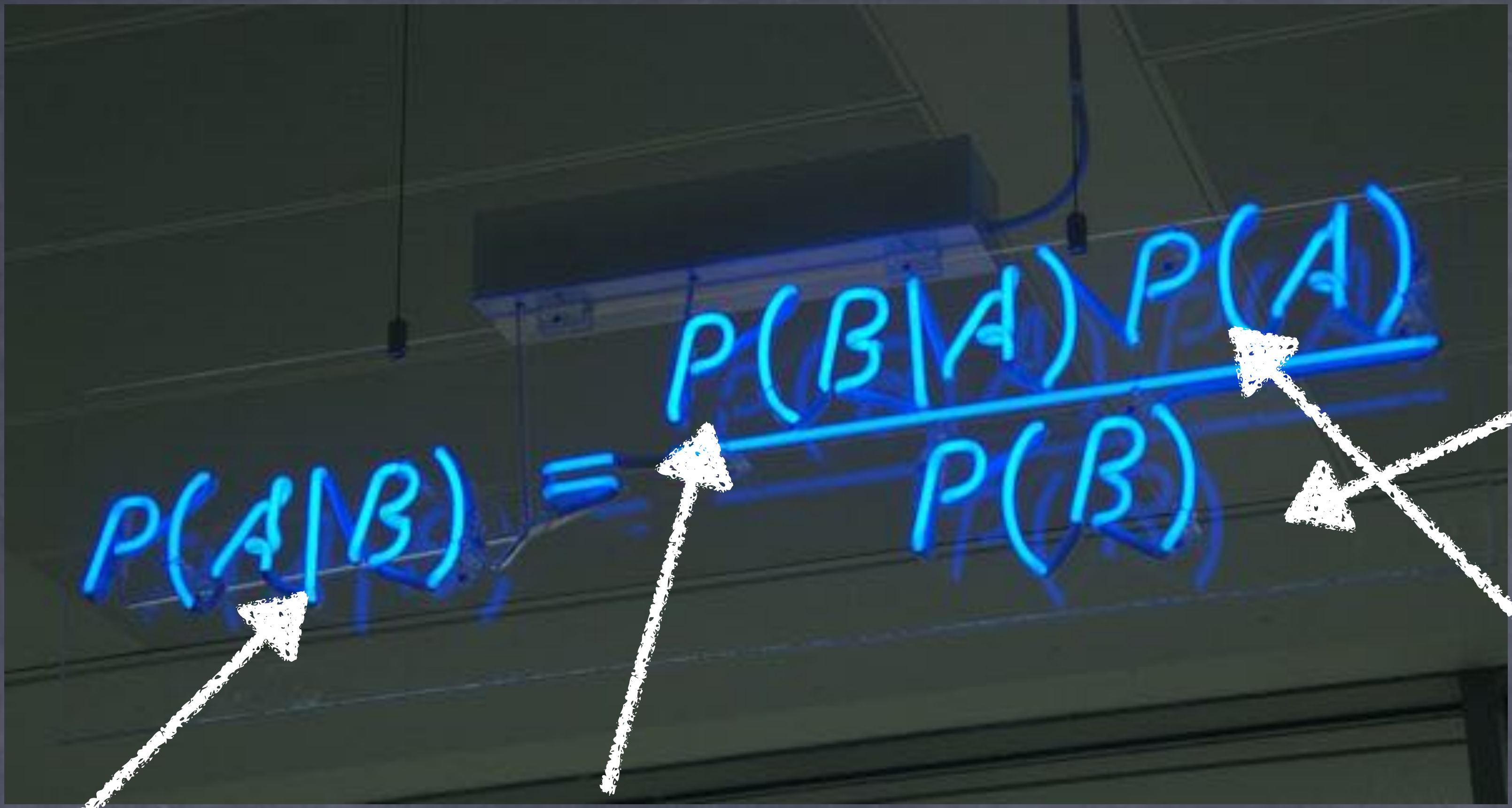
$$p(x) = \sum_i p(x, y_i)$$

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.



Evidence: The probability of de data, in all possibilities

Posterior

Degree of believe that some truth proposition event B implies that A is also truth.

Likelihood

Degree of believe to which some truth proposition A, event, implies that proposition B, event, is also true.

Prior

The degree of our a priori believe of proposition A, event, based on previous knowledge.

A question in cosmology would be: Given the observed CMB data with current experiments (A), what is the probability that the matter density of the Universe is between 0.9 and 1.1 (B).

$P(0.9 < \Omega_m < 1.1|Data)$?

Bayes Theorem (just a taste)

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Derive Bayes theorem from p(x,y)=p(y,x)

Example

Suppose you have the following results for an allergy test

1) It gives a positive result with probability 0.8, when patients have the allergy
2) It gives a false positive with probability 0.1.

If you make yourself the test, and result positive, what is the probability that you actually have the allergy?

P(A): probability of having the allergy

P(T): probability of test being true

What is the question in terms of probability?

Example

Suppose you have the following results for an allergy test

1) It gives a positive result with probability 0.8, when patients have the allergy
2) It gives a false positive with probability 0.1
3) The probability of having thee allergy in the population is 0.01

If you make yourself the test, and result positive, what is the probability that you actually have the allergy?

P(A): probability of having the allergy

P(T): probability of test being true

What is the question in terms of probability?

We want p(A|T)

Example (Solution)

We want p(A|T)

We know p(T|A)=0.8, p(T|~A)=0.1   , P(A)=0.01

Bayes Theorem:

$$p(A|T) = \frac{p(T|A)p(A)}{p(T)}$$

$$p(A|T) = \frac{p(T|A)p(A)}{p(T,A) + p(T, \sim A)}$$

$$p(A|T) = \frac{p(T|A)p(A)}{p(T|A)p(A) + p(T, \sim A)p(\sim A)}$$

$$p(A|T) = \frac{0.8 * 0.01}{0.8 * 0.01 + 0.1 * 0.99} = 0.075$$

Ejemplo: Las catafixias de Chabelo (adaptado de The Monty Hall problem)

Hay 3 puertas. Detrás de una hay un premio. Supongamos que eliges la puerta 2. Chabelo abre la puerta 1 y allí no hay un premio. ¿Te quedas con tu elección original o cambias a la puerta 3?

Eventos:

b: Eliges la puerta 2.

A : Chabelo abre la puerta 1 y no está allí el premio

Pregunta: ¿Cuál es la probabilidad de que el premio esté en la puerta 2, dado que no está en 1?

$P(a)$: Prob de que el premio esté en la puerta 1

$P(b)$: Prob de que el premio esté en la puerta 2

$P(c)$: Prob de que el premio esté en la puerta 3

$P(A,b)$: Prob de que Chabelo abra la puerta 1 dado que el concursante eligió la puerta 2.

Pregunta: ¿Cuál es la probabilidad de que el premio esté en 2, dado que no está en la puerta 1 abierta por Chabelo?

La pregunta en términos de probabilidades sería:

¿$P(b|A,b)$?

P(a): Prob de que el premio esté en la puerta 1 1/3

P(b) : Prob de que el premio esté en la puerta 2

P(c) : Prob de que el premio esté en la puerta 3

P(A) : Prob de que Chabelo abra la puerta 1.

Teorema de Bayes: $P(b|A) = \frac{p(A|b)P(b)}{p(A)}$

Solución

Calculamos p(b|A)

$$P(b|A) = \frac{p(A|b)P(b)}{p(A)}$$

$$P(b|A) = \frac{P(A|b)P(b)}{P(A|a)P(a) + P(A|b)P(b) + P(A|c)P(c)}$$

Prob de que abra la puerta 1, dado que el premio está en la puerta 1.

Prob de que abra la puerta 1, dado que el premio está en la puerta 2

Prob. de que abra la puerta 1, dado que el premio está en la puerta 3.

$$P(b|A) = \frac{(1/2)(1/3)}{(0)(1/3) + (1/2)(1/3) + (1)(1/3)} = 1/3$$

P(~b|A)=2/3

i.e. Si cambias a la puerta 3 tienes 2/3 de probabilidad de ganar!

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.

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

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

- The Variance and standard deviation.

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

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

- Example: Binomial distribution $X \sim B_{n,p}$

$$f(x|B_{n,p}) = \frac{n!}{(n-x)!x!} p^x (1-p)^{(n-x)},$$

$$n = 1, 2, 3..., n$$

$$0 \leq p \leq 1$$

$$x = 0, 1, ..., n$$

$$\mu = np$$

$$\sigma = \sqrt{(np(1-p))}$$

Probability density function (continus variable)

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

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

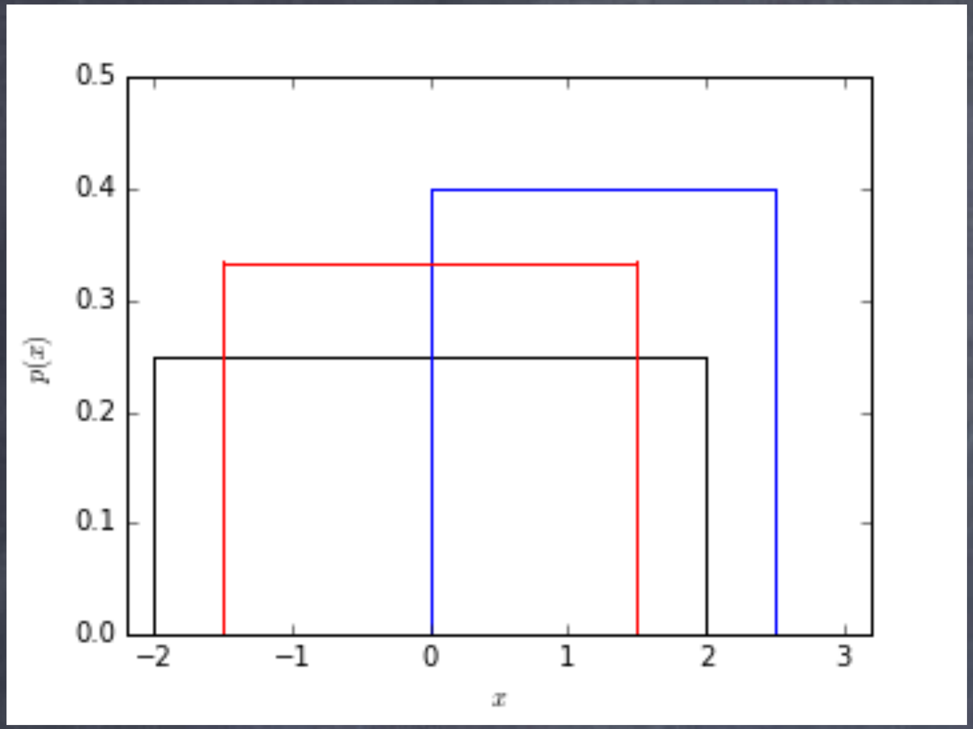
$$F(x) = P(X \leq x) = \int_{-\infty}^x f(x')dx'$$ CDF

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

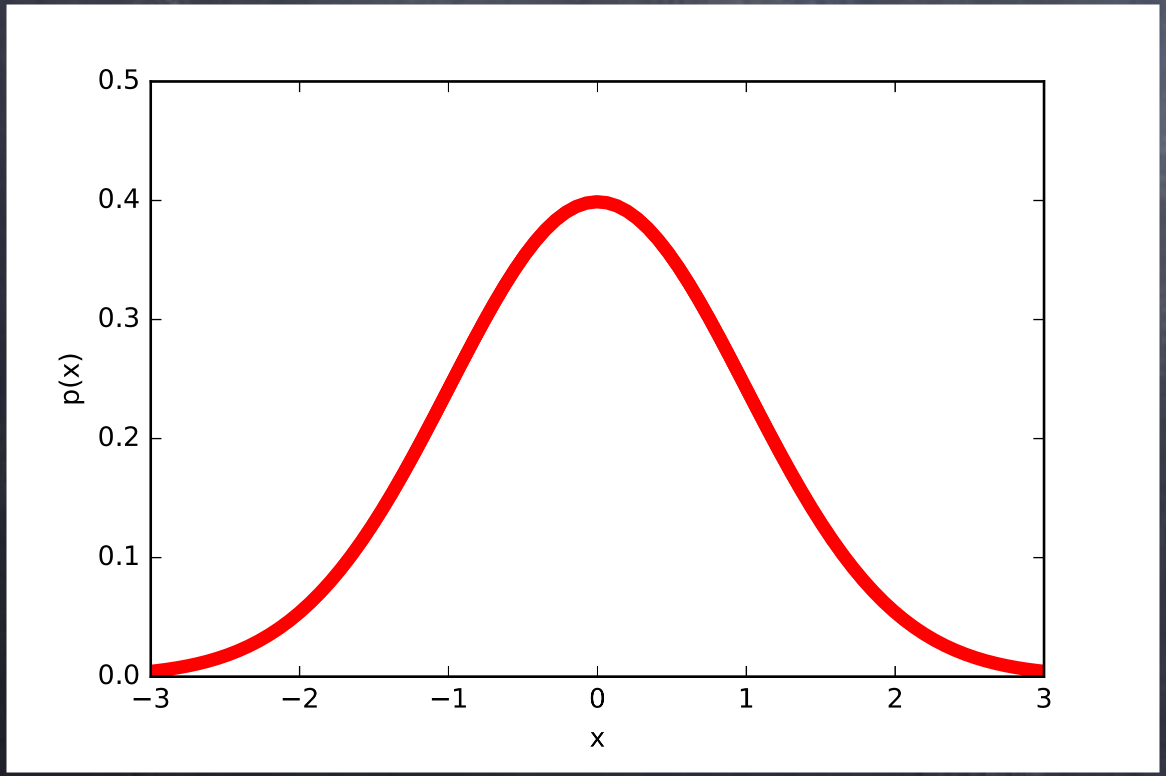
$$E(g(X)) = \int_{\infty}^{\infty} g(x)f(x)dx$$

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

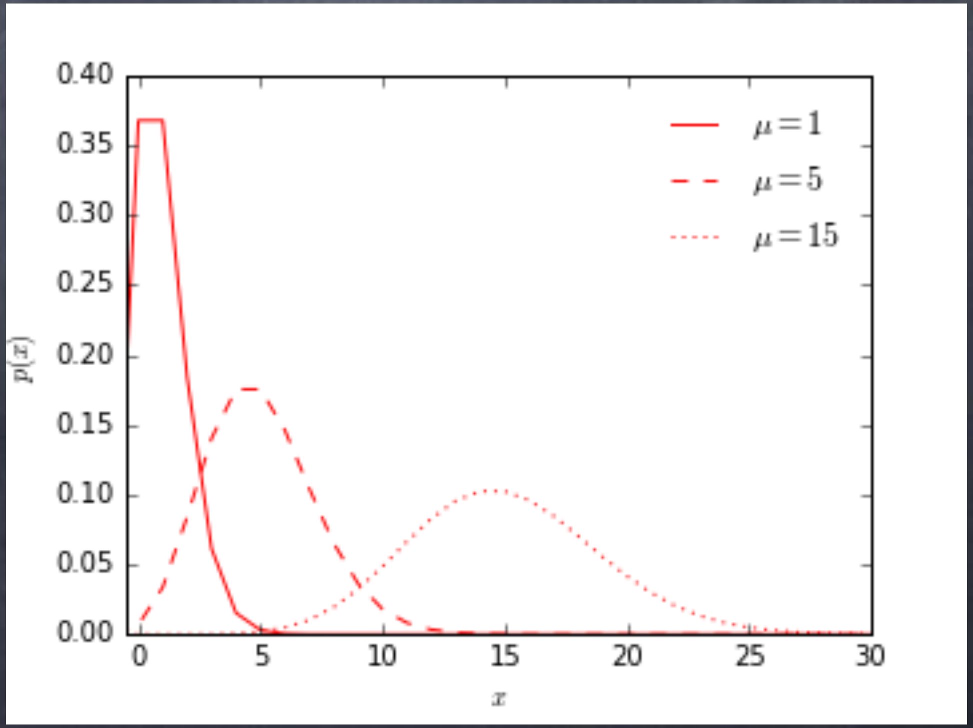
Common Probability Density Functions



- Flat

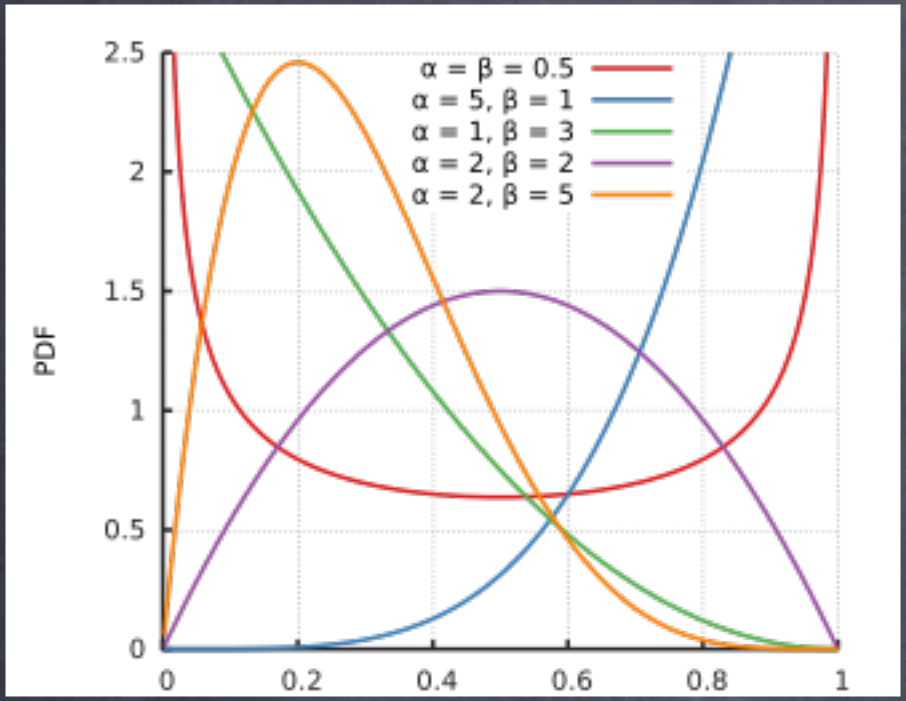


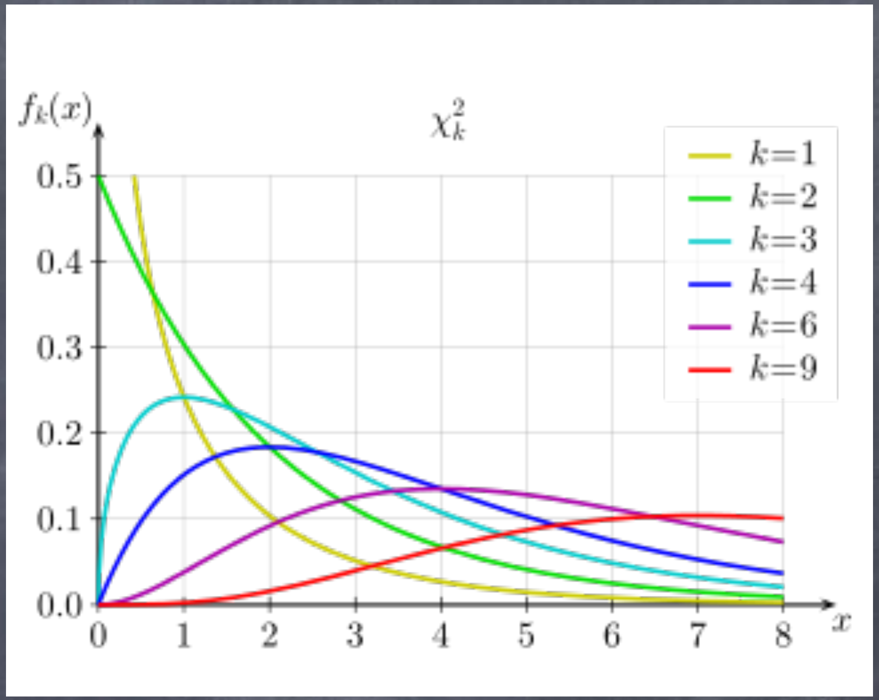
- Gaussian



- Poisson

Also useful





Beta distribution

- χ^2 distribution

Compute what are the mean and variance for each distribution

Centra Limit Theorem

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

$$Y = \sum_{i=1}^{n} c_i X_i$$

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

- 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.

Two applications of CLT

- 1) A sample avergage $\bar{X}_n$ of n independent identical distributed variables ,

$$\bar{X}_n = \sum \frac{1}{n} X_i$$

is normally distributed, since it is a linear combination of n variables X_i with c_i=1/n then:

$$\bar{X}_n \sim N(\mu_{\bar{X}_n}, \sigma_{\bar{X}_n})$$

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