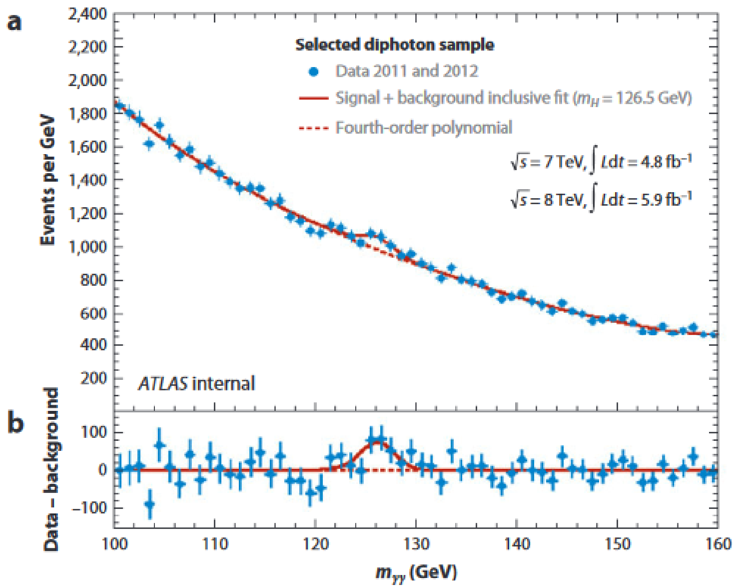
Data analysis

What is data analysis? (for us)



1.- Scientific method: modelling, quantify falsifiability of Hypothesis

Is there a peak?

2.- Learn from the parameters of the model

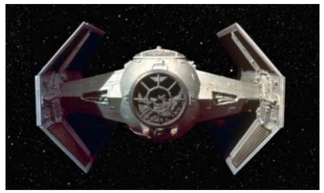
What is the mass?

Data analysis

What is data analysis? (for us)

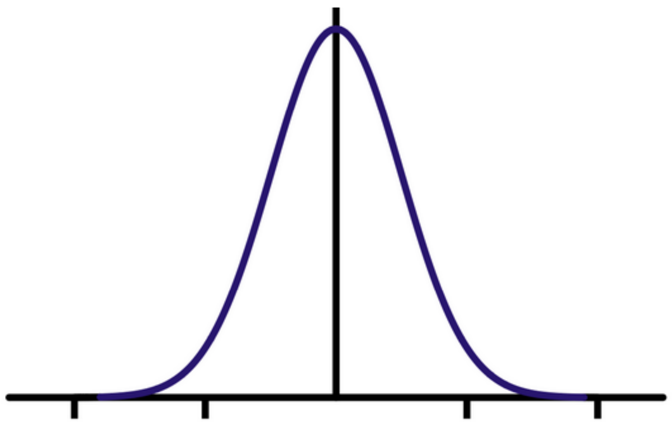
1.- Classical data analysis: STATISTICS

Numbers



2.- Non-classical data analysis: PROBABILITY (Bayes theorem)

Probability distribution functions



Statistics: basic definitions

Population (N members): ALL members under study: *all students from UPC*
Sample (n members): a subset of the whole population: *the people sitting in this room*

Random variable: X coming from a random phenomenon
it can take the values $x_i = x_1, x_2, x_3, \ldots$

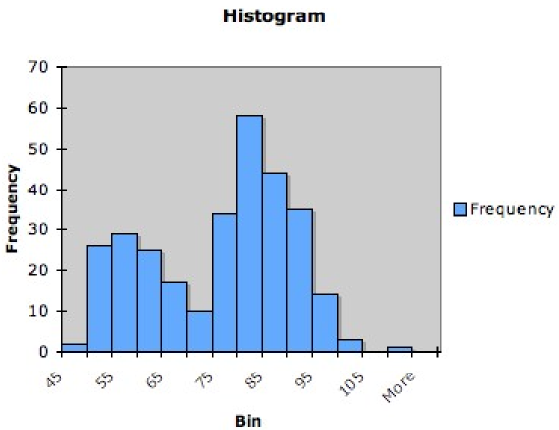
Probability distribution function (pdf):

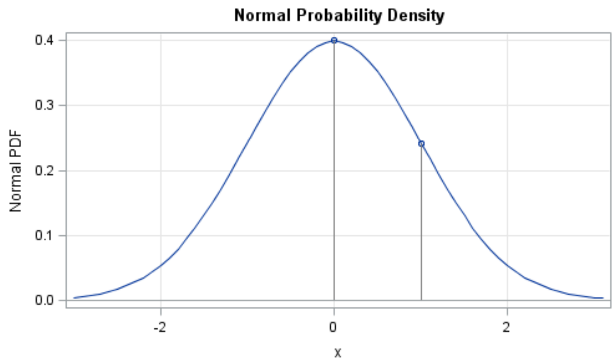
discrete case

$$P(X = x_i) = f(x_i)$$

continuous case

$$P(X = x) = f(x)$$



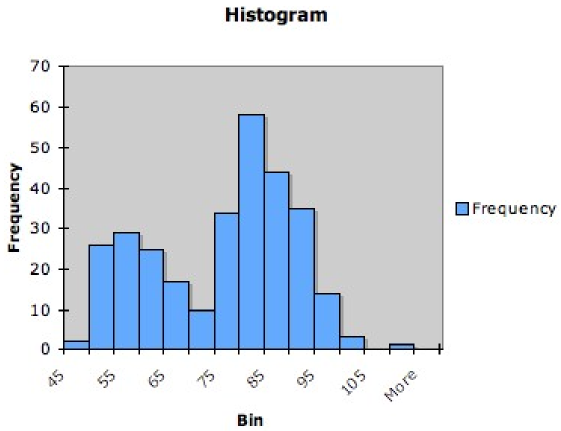


Statistics: definitions

Expectation value:

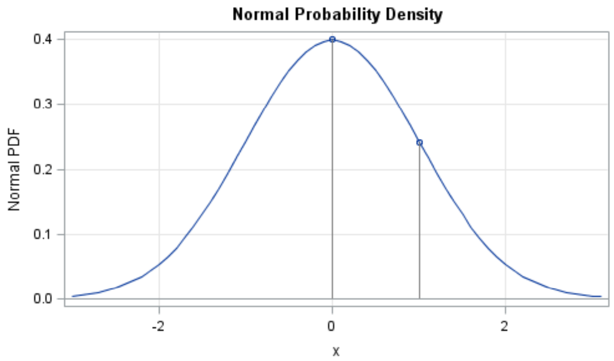
discrete case

$$P(X = x_i) = f(x_i)$$



continuous case

$$P(X = x) = f(x)$$



For random variable X

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$E(X) = \int x \cdot f(x) dx$$

... and for any function of g(X) like X^2

$$E(X) = \sum_{i=1}^{n} g(x_i) \cdot P(X = x_i)$$

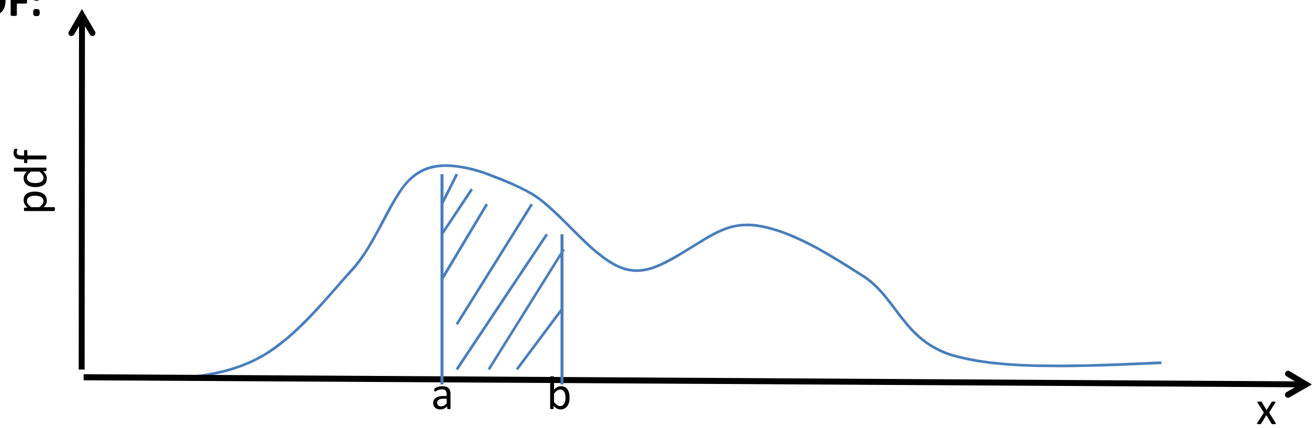
$$E(X) = \int g(x) \cdot f(x) dx$$

Statistics: definitions

	“Location” of data	“Spread” of data
sample	**Average or mean** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	**Mean square deviation or sample variance** $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Population	**Average or mean** $\bar{X} = E(X) = \frac{1}{n}\sum_{i=1}^{N} x_i$	**Variance** $\sigma_x^2 = Var(X) = E\left[(X - \bar{X})^2\right] = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$ $\sigma_x^2 = S^2$ for N large (typically N>30) **Standard deviation** $\sigma_x = \sqrt{\sigma_x^2}$

Some defintions about PDF's

Let's consider a PDF:



Relation with probabilities

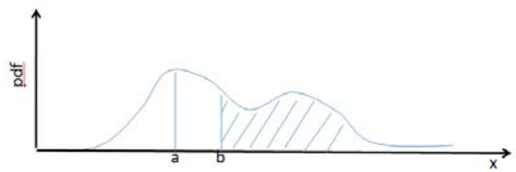
$$p(a < x < b) = \int_a^b f(x)dx$$

Therefore

$$\int_{-\infty}^{\infty} f(x) = 1$$

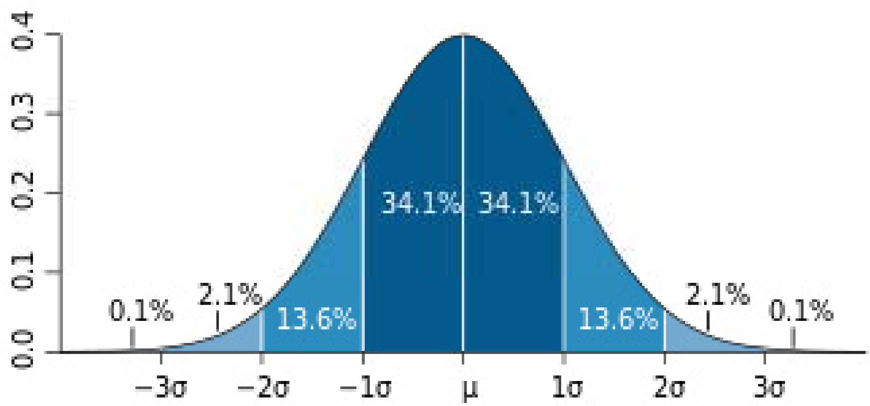
P-value for b (one side)

$$p(x > b) = \int_b^{\infty} f(x)dx$$



Some typical PDF's

Normal distribution



$$f(x) = \frac{1}{\sqrt{2\pi\sigma_x}} \exp\left(-\frac{(x-\mu)^2}{2\sigma_x^2} \right)$$

$$f(x) \equiv N(\mu, \sigma)$$

Relation with probabilities

$$p(-\sigma_x < x < \sigma_x) = \int_{-\sigma}^{\sigma} f(x)dx = 68.2\%$$

P-value for normal distribution for a value a: is related with the variance

$$p(-\sigma_x < x < \sigma_x) = \int_{-\sigma}^{\sigma} f(x)dx = 68.2\%$$

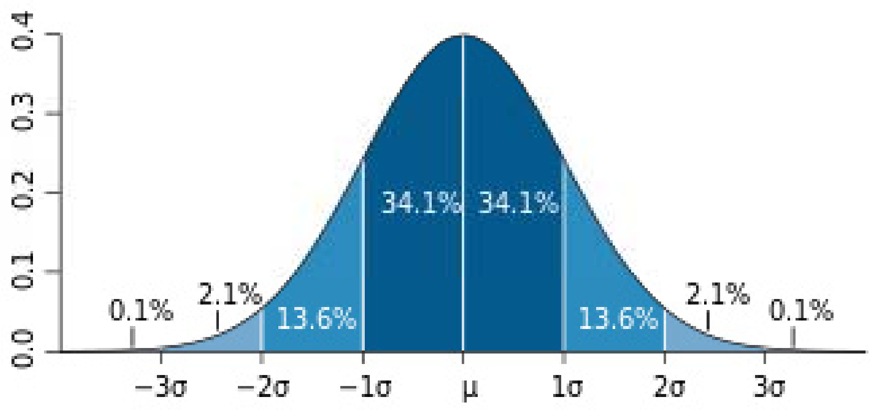
So we have a relationship between "error" and probability!

There is a 68,2% of chances that your real value is inside your "error "

(assuming that x has a normal distribution;-)

Standard Normal distribution

Standard Normal distribution (μ=0, σ=1)



$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$

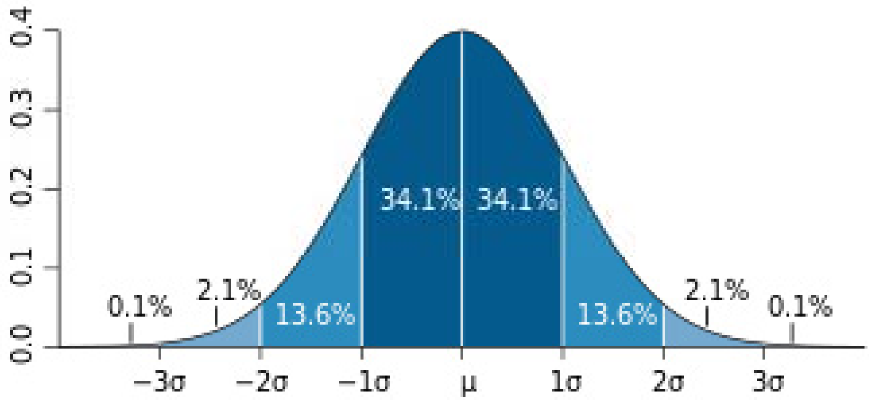
Therefore, defining a "z-value" as $z = \frac{x-\mu}{\sigma_x}$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_x}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma_x}\right)^2\right) \Rightarrow f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$$

Z follows a standard normal distribution function...

a flavour of fitting

How to obtain the value of μ from data?

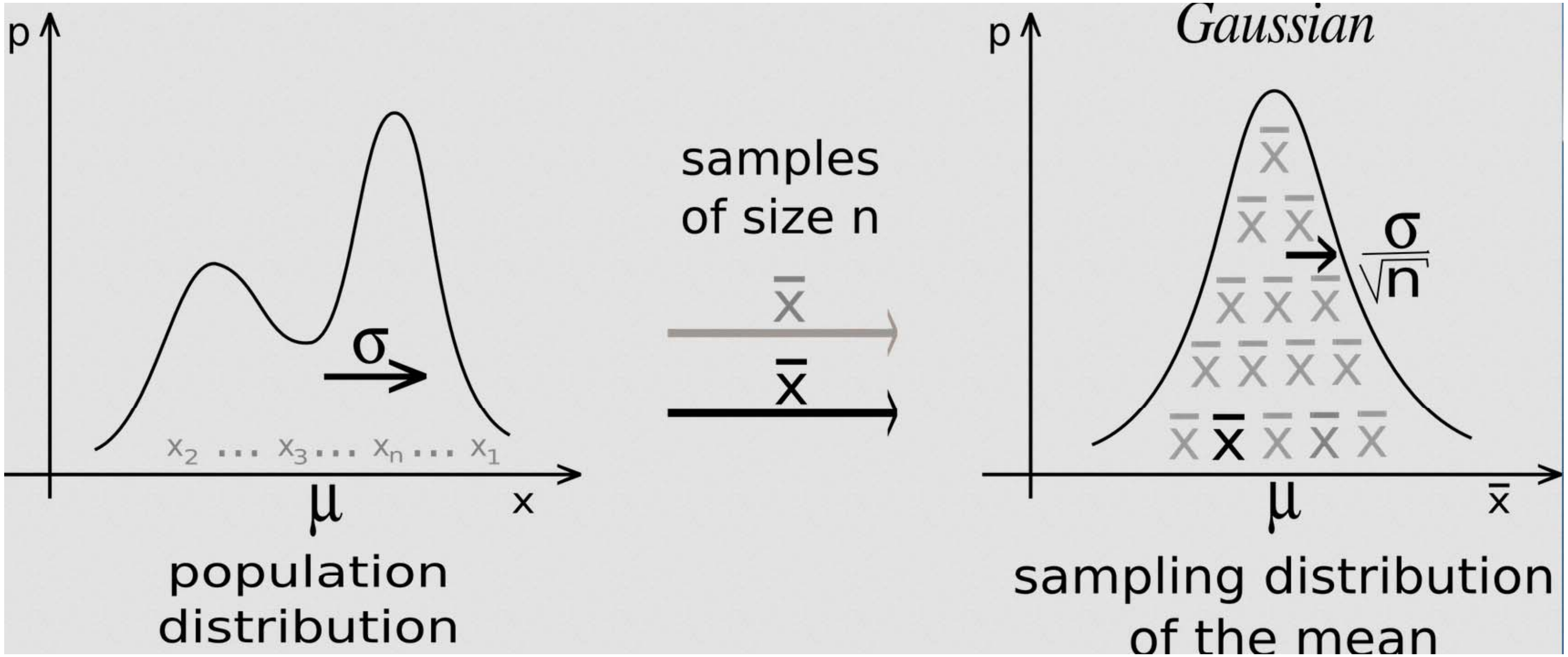


In orther to obtain the value of μ , when x is normal distributed, we have to minimize the quantity:

$$z^2 = \left(\frac{x - \mu}{\sigma_x} \right)^2$$

Least squartes fitting strategy

Central limit theorem

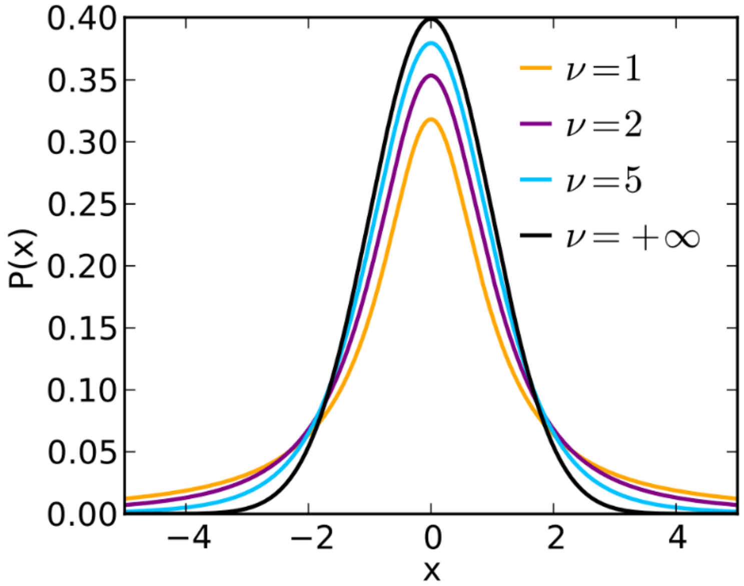


The distribution of the mean tends to be gaussian when increasing n
no matters the pdf that originated the mean!!!!

This is the good and the bad thing from statistics is based on the normal PDF!!!!

t-Student distribution

Imagine we do not know σ, i.e. the error of the data... shall we panic? NO, take a Guiness!



$$f(t) \propto \left(1+\frac{t^2}{k}\right)^{-\frac{k+1}{2}}$$

$$\mu=0; \mathrm{var}=\frac{k}{k-2} \to 1; \max=0$$

And it depends only on the degrees of freedom (number of points for the moment).
And of course tends to the Normal distributin for k big N(0,√k)

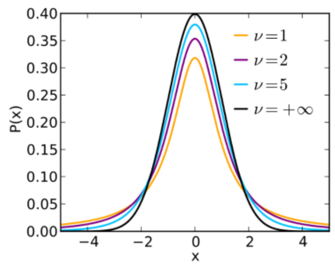
t-Student distribution

Usefull to compare two mean values (if they follow a Normal distribution!!!)
Lets assume that the degrees of freedom of X and Y are m and n , then

$$S_y = \frac{\sum (y_i - \overline{Y})^2}{m} \qquad S_x = \frac{\sum (x_i - \overline{X})^2}{n} \qquad \nu = m + n - 2$$

$$s^2 = \frac{nS_x + mS_y}{\nu}$$

$$t = \frac{\overline{X} - \overline{Y}}{s\sqrt{m^{-1} + n^{-1}}}$$



$$f(t) \propto \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}$$

And it depends only on the degrees of freedom (number of points for the moment)

generalized t-Student distribution

Usefull (agaion)to compare two mean values (if they follow a Normal distribution!!!)

but now imagine that a variable X that can be expressed as

$X = \mu + \hat{\sigma} T$ where T follows a T-student distribution (might be positive or negative!!)
where $\hat{\sigma}$ is a scale factor

$$T = \frac{X - \mu}{\hat{\sigma}}$$ follows therefore a T-student distribution with n degrees of freedom

... it can be proved that $\hat{\sigma}^2 = \mathrm{var}(x)$ for n large enough,
and thus is sometimes assumed as "the error" in μ (-sic-)

and for large n will be the same as for the Normal distribution with $\sigma = \sqrt{n}$

We can therefore compare X (with an "error") with a fixed value μ

Chi-Squared distribution

Do you remember the normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_x}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma_x}\right)^2 \right) \Rightarrow f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2 \right)$$

We minimize Z^2 in order to find μ... the following question arises:

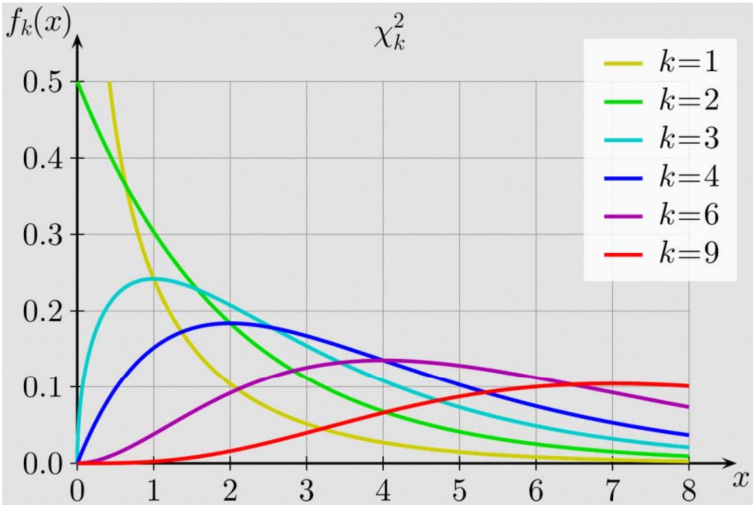
What is the PDF for Z^2 itself?

Chi-Squared distribution

If X_i follow a standard normal distribution then $\chi^2 = X_1^2 + X_2^2 + X_3^2 + \ldots + X_k^2$

follow a so called χ^2 distribution with n degrees of freedom defined as:

$$f(x) = A \cdot x^{k/2-1} e^{-x/2} \qquad A = \frac{1}{2^{k/2}\Gamma\left(k/2\right)}$$



$$\mu = k; \text{var} = 2k; \max = k - 2$$

If k➔∞ then $\frac{\chi^2 - k}{\sqrt{2k}}$ follows a standard normal distribution , i.e. χ^2 follows N(k,√2k)

F distribution

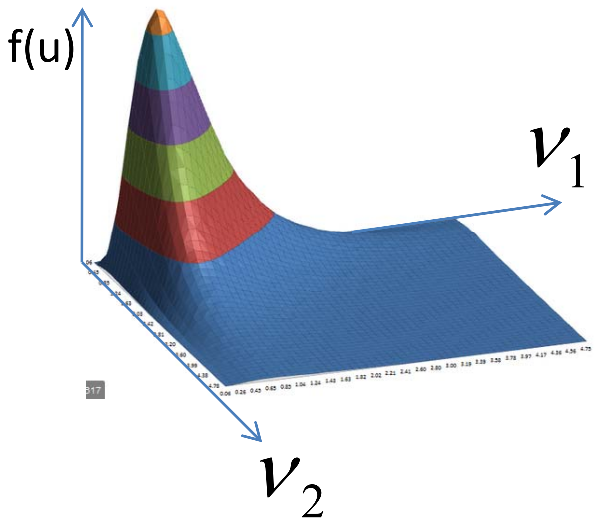
Usefull to compare two variances (again if they come from a Normal distribution!!!)

Lets assume that Z_1^2 and Z_2^2 follow a chi-square distribution... then:

$$u = \frac{Z_1^2 / \nu_1}{Z_2^2 / \nu_2}$$

follows an F distribution defined as

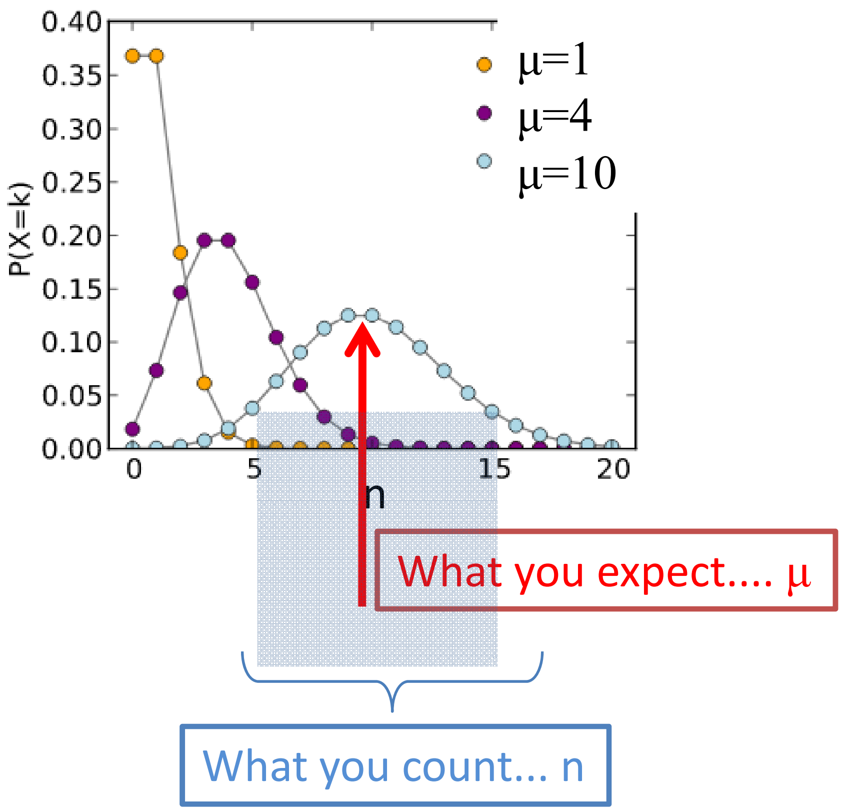
$$f(u) = A \cdot u^{\nu_1/2 - 1} (\nu_2 + \nu_1 u)^{-\frac{(\nu_1 + \nu_2)}{2}}$$



Poisson- distribution

Let's imagine we perform a counting experiment, we might ask:

What is the probability that n neutrons hit the detector in one hour?



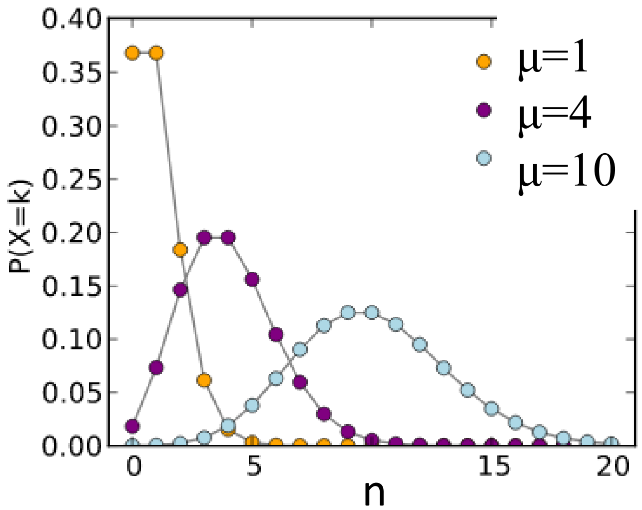
$$f(k) = \frac{\mu^n e^{-\mu}}{n!}$$

$$\max = \mu$$

Poisson- distribution

Let's imagine we perform a counting experiment, we might ask:

What is the probability that n neutrons hit the detector in one hour?



$$f(k) = \frac{\mu^n e^{-\mu}}{n!}$$

max= μ

Which is not a Gaussian distribution... but as always, if k is large enough

$$f(k) = \frac{\mu^n e^{-\mu}}{n!} \Longrightarrow N\left(\mu, \sqrt{k}\right)$$

The dumb rule σ=√n comes from the poisson distribution!!

Summarizing

Normal (Z): $N(\mu, \sigma)$

Central limit theorem

Usefull to compare two values A and B when error σ=variance is known

χ^2-distribution: $N(\nu, \sqrt{2\nu})$

Distribution for Z^2 or for a sumation of ν (degrees of freedom)

Usefull to perform fitting

t-distribution: $N(\mu, \sqrt{n})$

"central limit theorem" for small n

Usefull to compare two values A and B when error σ=variance is unknown

F-distribution:

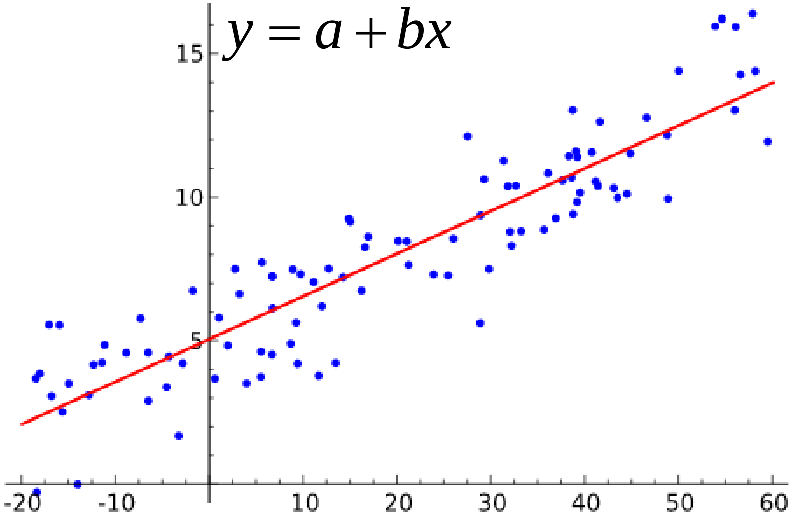
Usefull to compare variables that follow a χ^2 distribution

Usefull to compare the "errors" of $A \pm \sigma_A$ and $B \pm \sigma_B$

Poisson-dist: $N(\mu, \sqrt{n})$

PDF for a counting experiment (assuming n counts)

Fitting data: least squares



Remember the goals:

1.- Parameter estimation
2.- Modelling (Hypothesis testing)

Goal: to minimize the "total distance" of n exp. points $d_1+d_2+\ldots\ d_n$ to the model

or more seriously.... to minimize the following variabel that follows a χ^2-distribution and therefore D **is normally distributed** around H

$$\chi^2 = \sum_{i=1}^{n} \frac{(D_i - H_i)^2}{H_i}$$

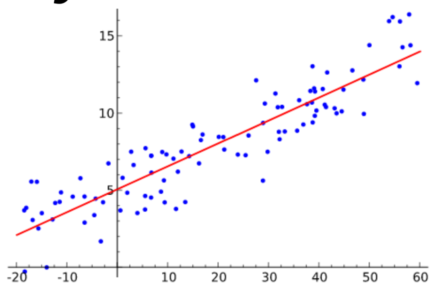
Where D_i are the data pints, and H_i are the model expected values (Hypothesis)

... since for n large a χ^2-distribution for point i tends to a gaussian with σ^2=n=H_i

$$\chi^2 = \sum_{i=1}^{n} \frac{(D_i - H_i)^2}{H_i} \rightarrow \sum_{i=1}^{n} \frac{(D_i - H_i)^2}{\sigma_i^2}$$

Fitting data: parameter estimation

$$y = a + bx$$



After minimization of χ^2 and defining

$$\Delta = n\sum x^2 - \left(\sum x\right)^2$$

	a	b
VALUE and therefore μ	$a = \dfrac{\sum x^2 \sum y - \sum x \sum xy}{\Delta}$	$b = \dfrac{n\sum xy - \sum x \sum y}{\Delta}$
ERROR and therefore σ	$\sigma_b = \sigma_y \sqrt{\dfrac{n}{\Delta}}$	$\sigma_b = \sigma_y \sqrt{\dfrac{\sum x^2}{\Delta}}$

REMEMBER: **EVERYTHING MUST BE NORMALLY DISTRIBUTED**!!!!!

Data around H, and therefore also the obtained parameters around **a** and **b**

Fitting data: is the model right?

Let’s do the question again (more precisse):

•Is the χ^2 arising after minimization
•When assuming that the data are normally distributerd around the model (hypothesis)
•Following a χ^2-distribution of (as it should?)

Rule of thumb (or the joys of the χ^2-distribution)

The χ^2 should follow a χ^2-distribution with n-m degrees of freedom arising for the data (with n points) and the model (with m=2 parameters)

For n data and 2 parameters μ=n-m and therefore the calculated χ^2 :

$$\chi^2 \approx (n-m)$$

For this reason it seems reasonable to define a reduced χ^2 that should be about one

$$\chi^2_{red} = \frac{\chi^2}{n-m} \approx 1$$

... for a good fit

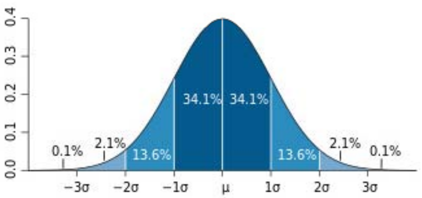
Fitting data: is the modelling right?

Let's do the question again (more precisse):

•Is the χ^2 arising after minimization
•When assuming that the data are normally distributerd around the model (hypothesis)
•Following a χ^2-distribution of (as it should?)

Probability(or the joys of the χ^2-distribution)

The χ^2 should follow a χ^2-distribution with n-m degrees of freedom
that we can calculate from the number of points and number of parameters.



e question is now: what is the probability (p)
to get the calculated value of χ^2 or greater?

We calculated the PDF and get this number... for n big $\chi^2 \approx N(0, \sqrt{2n})$ and we use

Hypothesis testing

We have two competing models
H_0: or null hypothesis
H_1: or alternative hypothesis

What is the probability that H_1 is compatible with H_0?

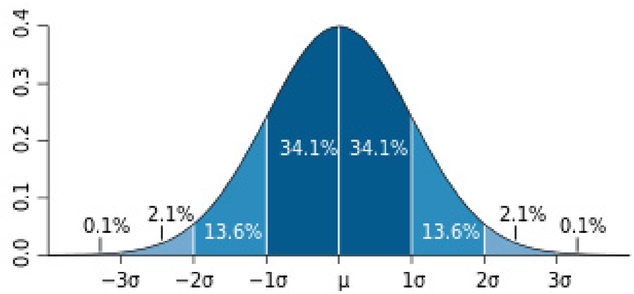
Example: is there a linear correlation?

Or better:

What is the probability that the obtained slope (H1) equals some value b, in our case b=0 (H0))

$$T = \frac{b_{H1} - b_{H0}}{\sigma_b}$$

... it follows a t-student distribution but for n big follows $N(0,\sigma_b)$



... and you can calculate the P value

"Hypothesis testing"

We have two competing models
H_0: or null hypothesis
H_1: or alternative hypothesis

What is the probability that H_1 is compatible with H_0?

Example: is there a linear correlation?

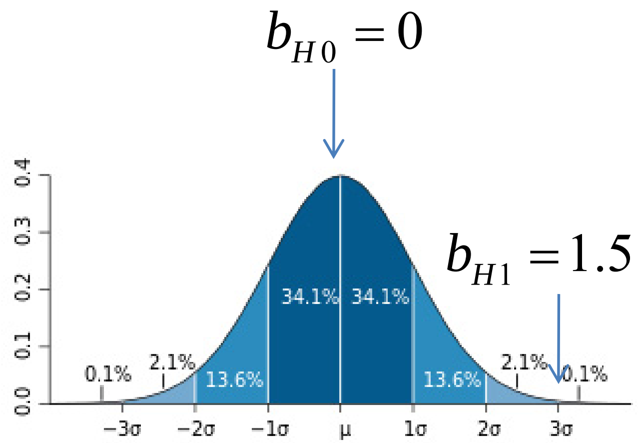
Imagine you get y=a+bx, being a=3.0±0.1 and b=1.5±0.5

What is the probability that the obtained slope (H_1) equals some value b, in our case b=0 (H_0))

$$T = \frac{b_{H1} - b_{H0}}{\sigma_b} = \frac{1.5 - 0}{0.5} = 3$$

Or as a "distance" H_1 is at 3σ from H_0

and this means that P(b≥1.5)=0.001



if we establish a significance level of α=0.05 (is an accepted value) to reject the hypothesis...
Therefore THERE IS a linear correlation

“Hypothesis testing”

We have two competing models
H_0: or null hypothesis
H_1: or alternative hypothesis

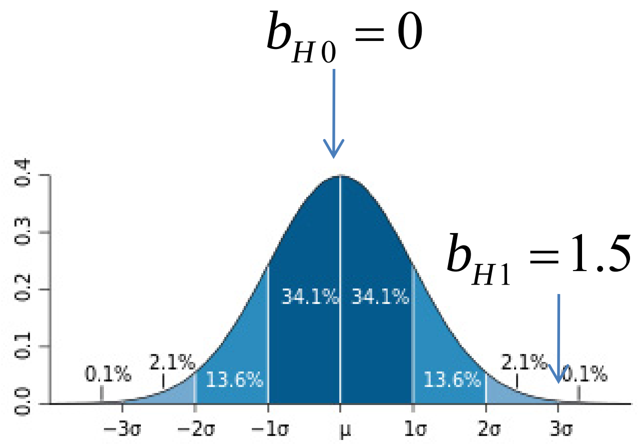
What is the probability that H_1 is compatible with H_0?

Example: is there a linear correlation?

Imagine you get y=a+bx, being a=3.0±0.1 and b=1.5±1.5

What is the probability that the obtained slope (H_1) equals some value b, in our case b=0 (H_0))

$$T = \frac{b_{H1} - b_{H0}}{\sigma_b} = \frac{1.5 - 0}{1.5} = 1$$



Or as a “distance” H_1 is at σ from H_0

and this means that P(b≥1.5)=0.34

if we establish a significance level of α=0.05 (is an accepted value) to reject the hypothesis...
Therefore THERE IS NOT a linear correlation (b_{H0} might come from an error. is “inside the error”)

"Hypothesis testing"

We have two competing models
H_0: or null hypothesis
H_1: or alternative hypothesis

What is the probability that H_1 is compatible with H_0?

Example: is there a linear correlation?

Problems:
1.-We need nested models!!!

i.e. we need to set some parameters $\neq 0$ to perform "model selection"
(a+bx, setting b=0, what if we want to compare y=ax+b with y=A3exp(bx)???)

2.- We are NOT doing "model selection"

only setting the probability that a given parameter is not zero....

“Hypothesis testing”

We have two competing models

H_0: or null hypothesis

H_1: or alternative hypothesis

What is the probability that H_1 is compatible with H_0?

Parameter free: let’s really compare the two models!!!

Let’s use the F-distribution

$$u = \frac{\chi^2_{H1} / (n - m_{H1})}{\chi^2_{H1} / (n - m_{H0})}$$

We might now calculate the probability that we get the diference between models

“Hypothesis testing”

Even more cool: We perform a **Kolmogorov-Smirnov** test

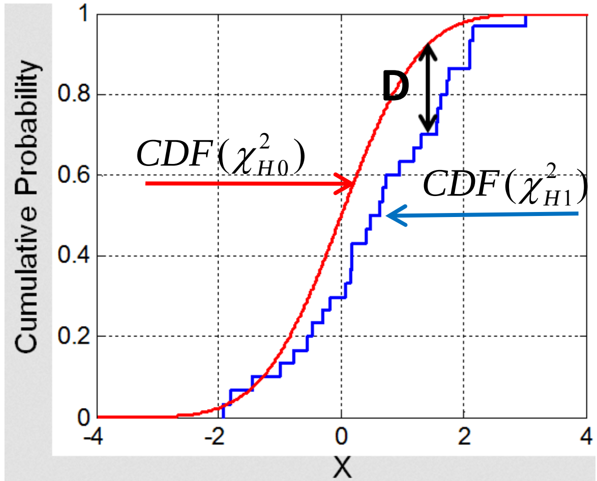
1.- We calculate the cumulative distribution function CDF from the PDF of the two models: H_0 (with n-m_{H0} dof), and H_1 (with n-m_{H1} dof),

$$CDF(\chi^2) = \int_0^{\chi^2} f(\chi^2) d\chi^2$$

2.- We look at the maximum distance of the two of them

3.- We look at the maximum distance of the two of them D

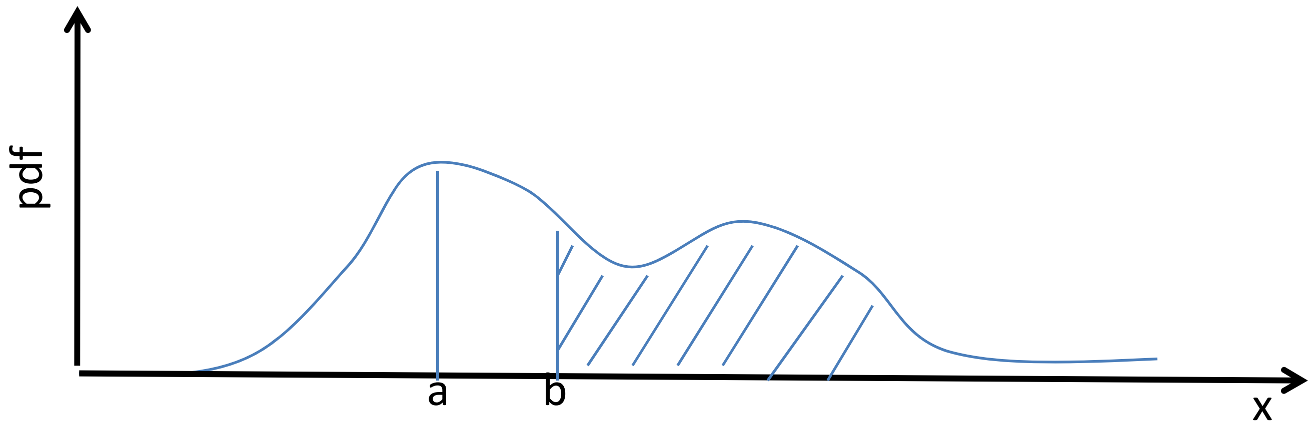
4.- for n and m large, the quantity



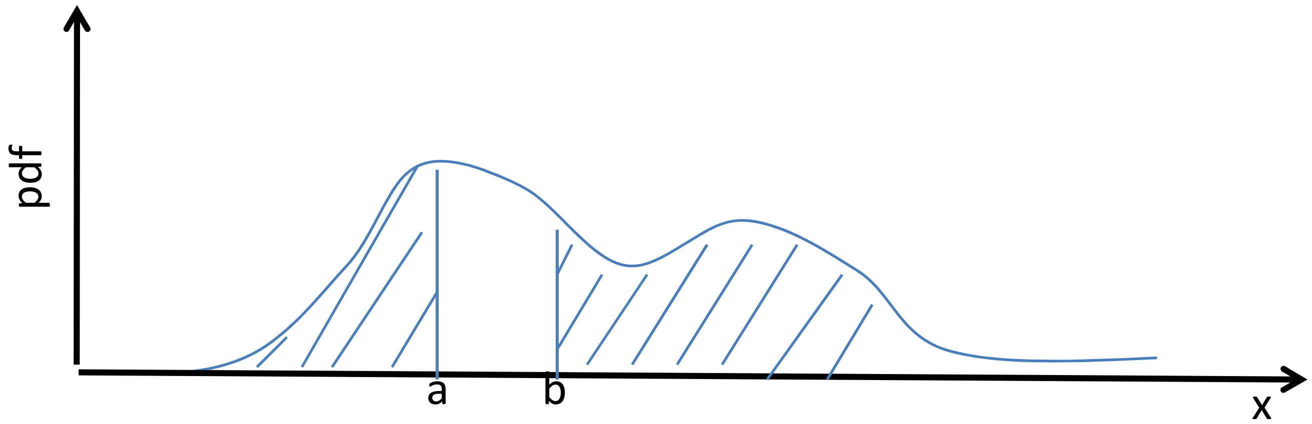
$$4D^2 \frac{dof_{H0} \cdot dof_{H1}}{dof_{H0} + dof_{H1}} = 4D^2 \frac{(n - m_{H0})(n - m_{H1})}{(n - m_{H0}) + (n - m_{H1})}$$

follows a χ^2 distribution with two degrees of freedom

and now we might calculate the P value...



pdf
a
b
x



pdf
a
b
x