INTERNATIONAL SCHOOL FOR YOUNG ASTRONOMERS

Lecture 2: Introduction to Astrostatistics

Rafael Martínez-Galarza
Harvard-Smithsonian Center for Astrophysics

Past lecture

- Intro to X-ray astronomy
- How to search for Chandra data, how to download it and visualize it.
- Photons are counted individually! This is very different from other regions of the electromagnetic spectrum.

This lecture:

- What does intuition tell us about the nature of probability?

- Why is probability relevant in astronomy - the Initial Mass Function.

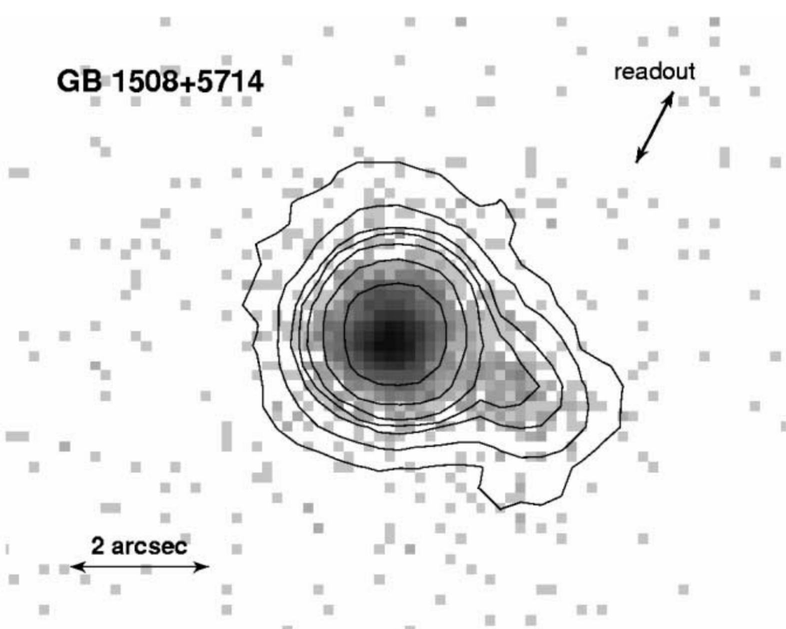
- Random variables

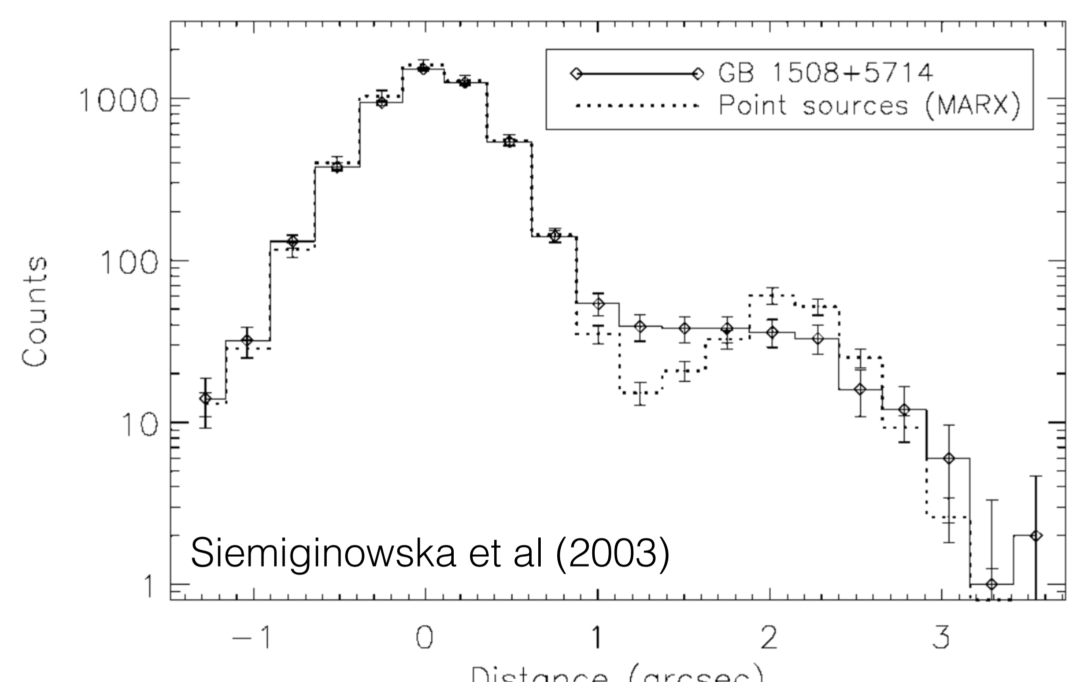
- What is a probability distribution?

- The concept of uncertainty - Possion noise in X-ray astronomy

Why do we need statistics?

- To describe the properties of data. Data collection, classification, reduction. Statistics provides a mean to do so.
- To make decisions. Is this hypothesis correct? why not? Are these data consistent with a given model, or with a different set of data? Do we need more data to answer the question?
- Examples: Is this source a jet, or two point sources?





It is (most likely) a jet. X-ray emission due to inverse Compton scattering of CMB photons, off the relativistic particles.

Intuition on probability: a coin flip

Head

Tail

Head

Tail

Can you predict, for a single toss, whether it will land heads or tails?
Can you predict instead the rate of tails for, say, 10, or 100 tosses?

Why?

Our intuition of probability corresponds to a model of nature

We speak of the probability P(heads) that the coin will land heads.
P(heads) is the fraction of times that the toss results in heads
If the coin is fair, then P(heads) = 0.5

What about a fair dice?
What is P(2)?

What if we have a non-fair coin or dice?

4 coins

What if we toss a coin, say, 4 times?
What is the probability of getting 4 heads (P(4H))?
Or 3 heads and 1 tail P(3H)?

P(4H) = 1/16 = 0.025
P(3H) = 4/16 = 0.25

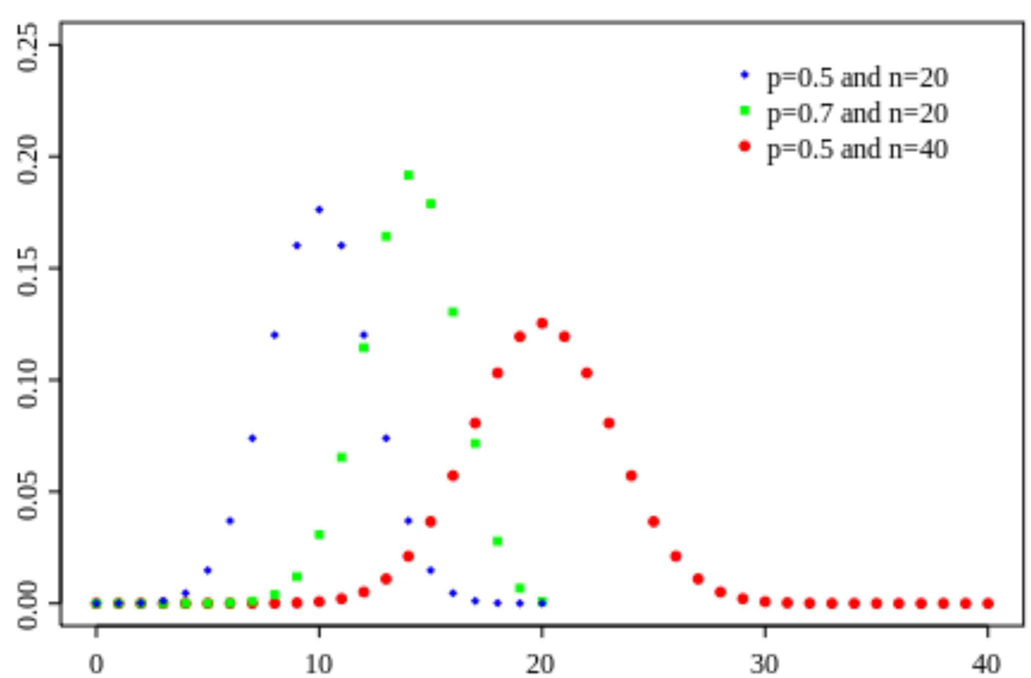
Mathematically, we write this as:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

If n=1: Bernoulli

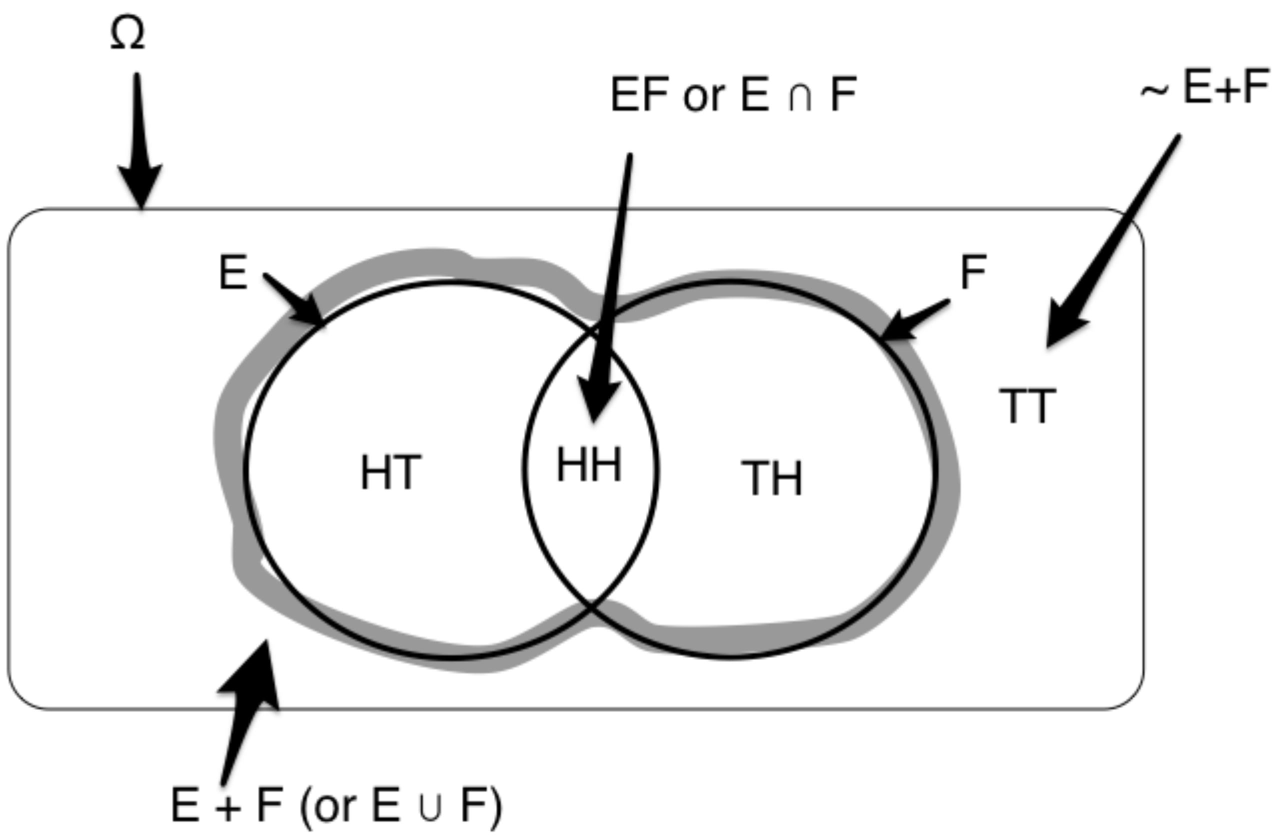
In general (Binomial distribution):

$$f(k,n,p) = \Pr(k;n,p) = \Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$



Number of Heads	Results of Flips	Number of Ways
0	0 0 0 0	1
1	0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0	4
2	0 0 1 1 0 1 0 1 1 0 0 1 0 1 1 0 1 0 1 0 1 1 0 0	6
3	0 1 1 1 1 0 1 1 1 1 0 1 1 1 1 0	4
4	1 1 1 1	1

The rules of probability



1. $p(X) >= 0$; probability must be non-negative

2. $0 \leq p(X) \leq 1$ X has probability range from 0 to 1.

3. $p(X) + p(X^{-}) = 1$ X must either happen or not happen. These last two aximoms can be thought of as saying that the probabilities if all events put tohether must sum to 1.

4. $p(X+Y) = p(X) + p(Y) - p(X,Y)$ X can happen and Y can happen but we must subtract the cases that are happening together so we do not over-count.

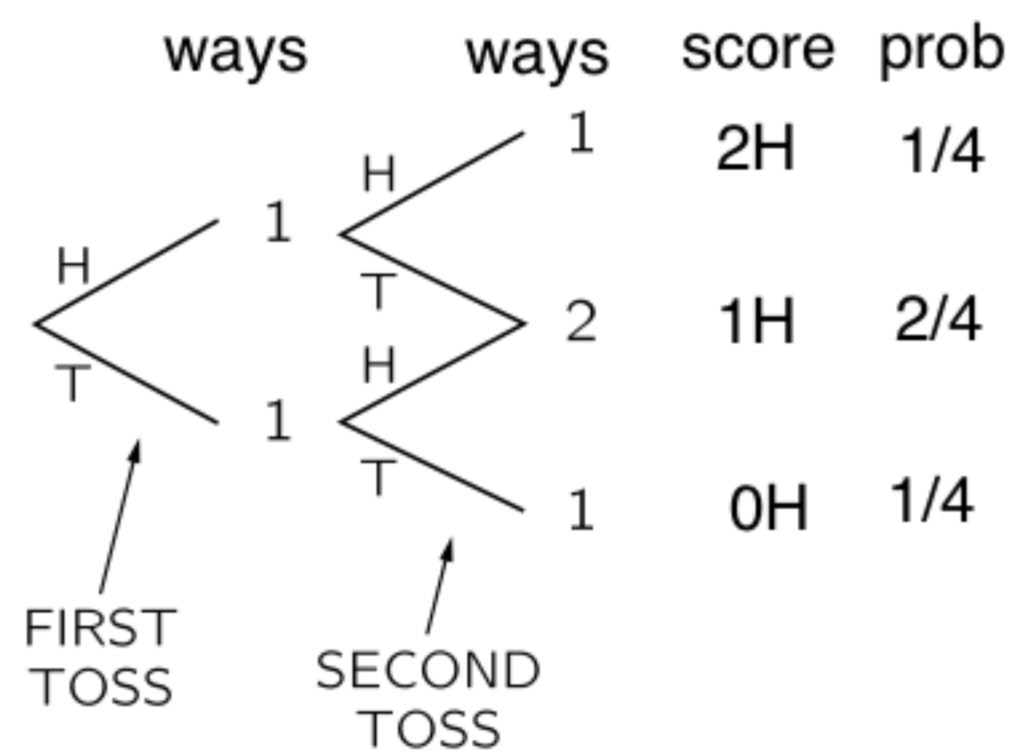
For two coin tosses:

What is the probability of getting two heads?

$$P(EF) = P(E)P(F).$$

What are the odds of at least one heads?

$$P(E+F) = P(E) + P(F) - P(EF).$$



The frequentist view of probability

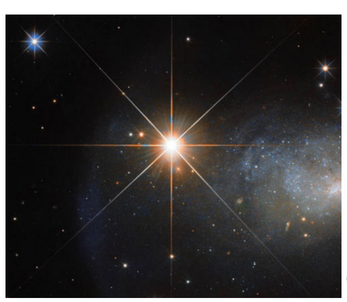
For frequentists probabilities are fundamentally related to frequencies of events.


$F_1 = 10.1 \pm 0.1$ mJy

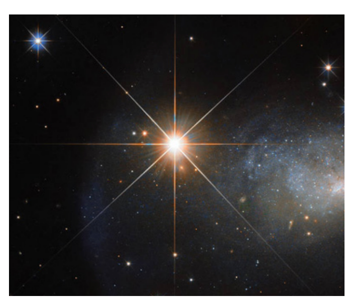

$F_2 = 10.2 \pm 0.2$ mJy


$F_3 = 10.1 \pm 0.1$ mJy

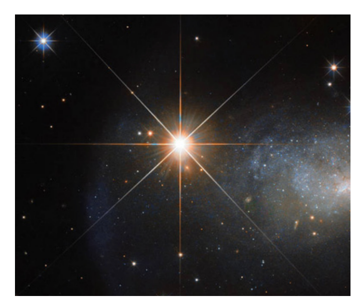

$F_4 = 10.3 \pm 0.2$ mJy




$F_5 = 10.0 \pm 0.2$ mJy

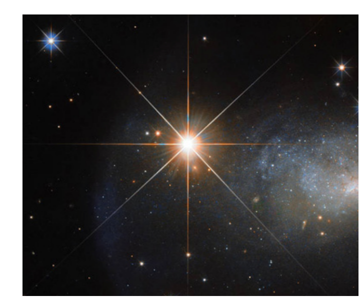



$F_6 = 10.1 \pm 0.1$ mJy




$F_7 = 10.2 \pm 0.3$ mJy

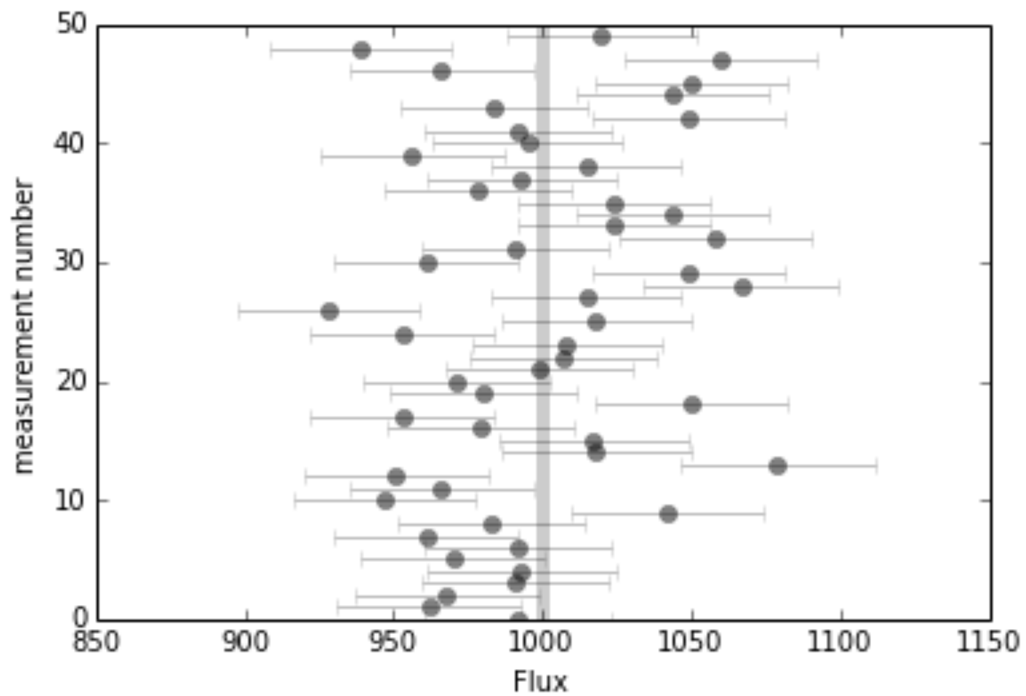
. . .




$F_N = 9.9 \pm 0.2$ mJy

The frequentist view of probability

The question is, given this set of measurements $D=\{F_i,e_i\}$, what is our best estimate of the true flux F_{true}? We assume Gaussian errors (we'll see why later).



We assume that a true flux exists, and we attempt to infer its value from our measurements.

The question we are asking is: what is the probability of getting this dataset given that the true flux is, say, 1000 units?

We can compute the probability of measurement *i* given the true flux:

$$P(D_i \mid F_{\mathrm{true}})$$

If we assume that each measurement is independent of each other, then the *joint* likelihood function is:

$$\mathcal{L}(D \mid F_{\mathrm{true}}) = \prod_{i=1}^{N} P(D_i \mid F_{\mathrm{true}})$$

where $D=\{D_i\}$ represents the entire set of measurements. This reads as "The likelihood of obtaining the set of measurements D, given the true flux F_{true}."

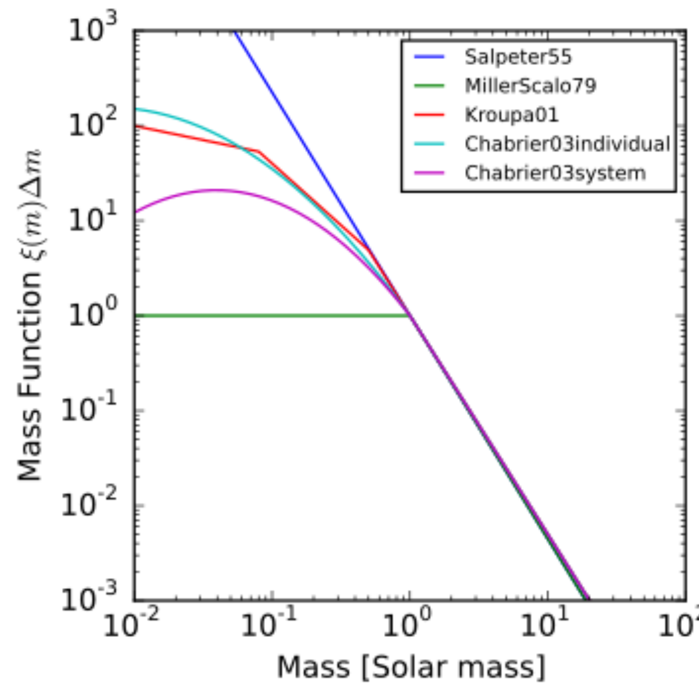
Q: What does $L(D|F_{true}) = 0.85$ mean?

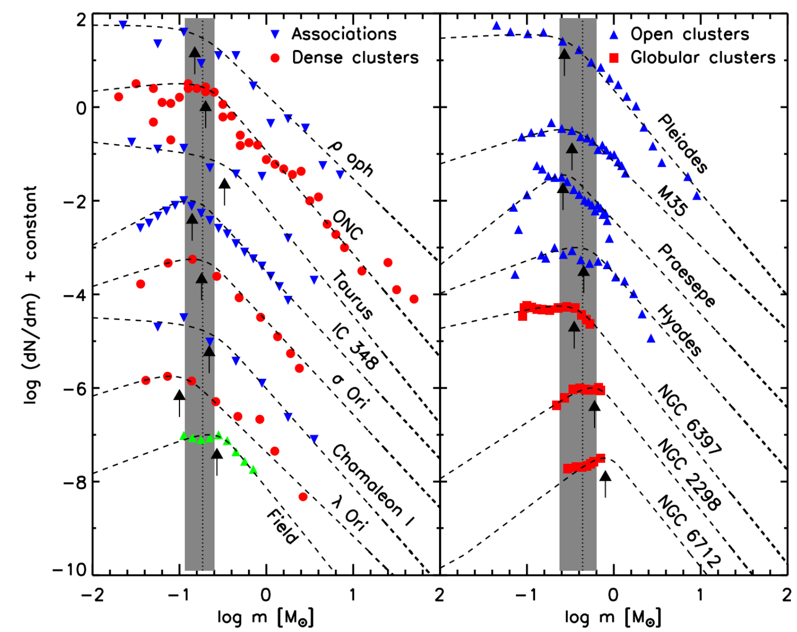
A motivation for astrostats: The IMF

Stars form in clusters:

R136 in 30 Doradus, LMC

How do stellar mass distribute?
the Initial Mass Function (IMF)





The exact shape of the IMF (and whether it is Universal or not) has deep implications on the physics of the interstellar medium. Measuring it is hard. But once you know it, you can understand the processes that drive star formation

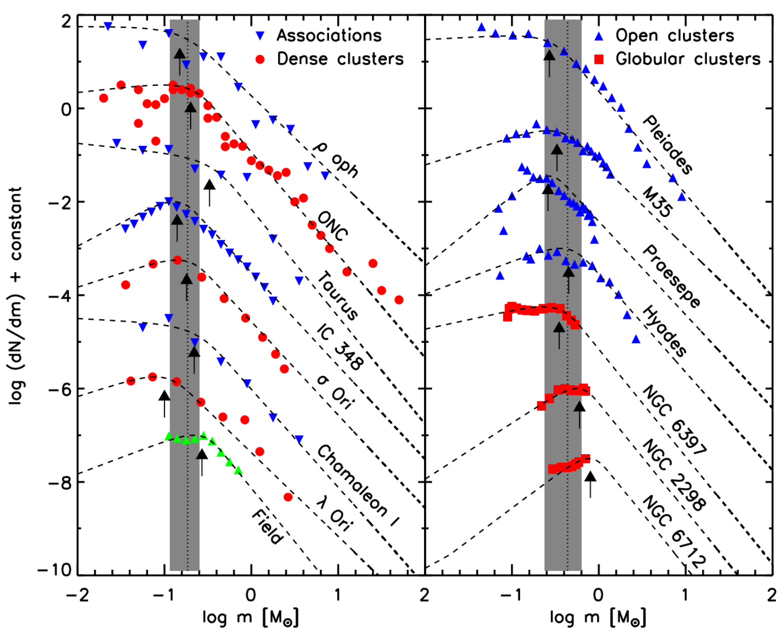
The IMF is an power law probability distribution that tells you how likely it is to find a star of a given mass in a globular cluster, or a galaxy.

In other words, it tells you about the frequency of, say solar masses stars in a given stellar system. Hence the *frequentist* interpretation of probability

Common PDFs in astronomy

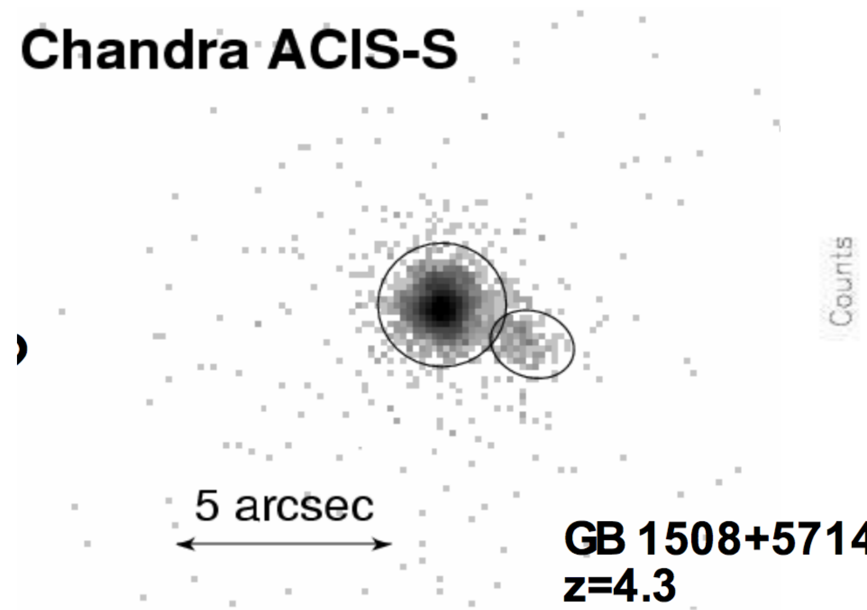
Power law: IMF

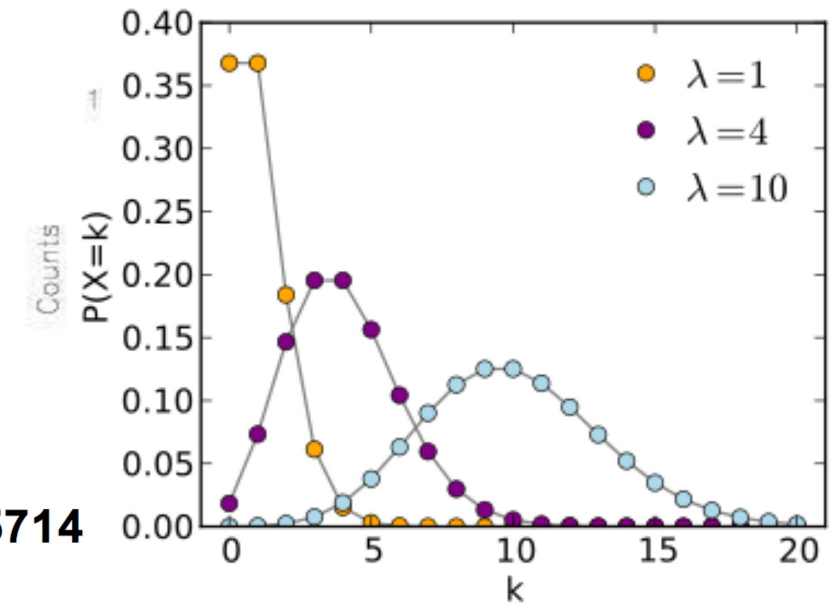
$$\overline{F}(x) = \Pr(X > x) = \begin{cases} \left(\frac{x_\mathrm{m}}{x}\right)^\alpha & x \geq x_\mathrm{m}, \\ 1 & x < x_\mathrm{m}, \end{cases}$$



Poisson: X-ray photons

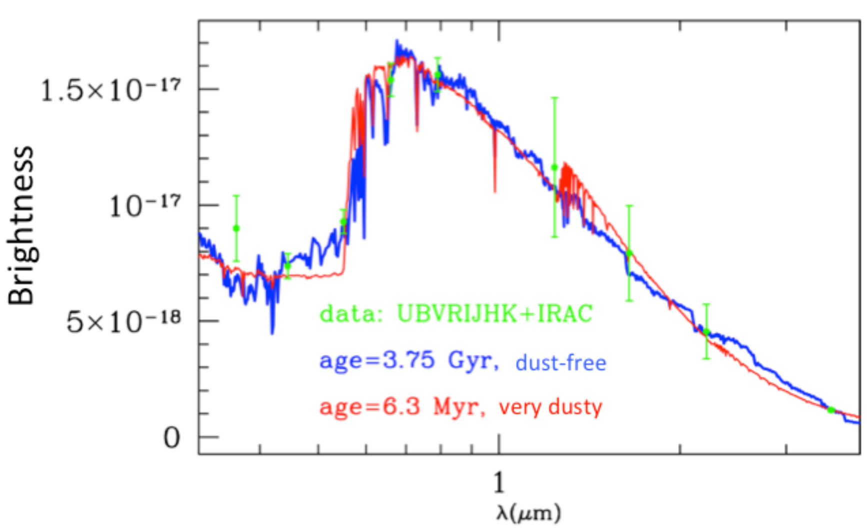
$$P(k \text{ events in interval}) = e^{-\lambda}\frac{\lambda^k}{k!}$$





Normal: measurement error

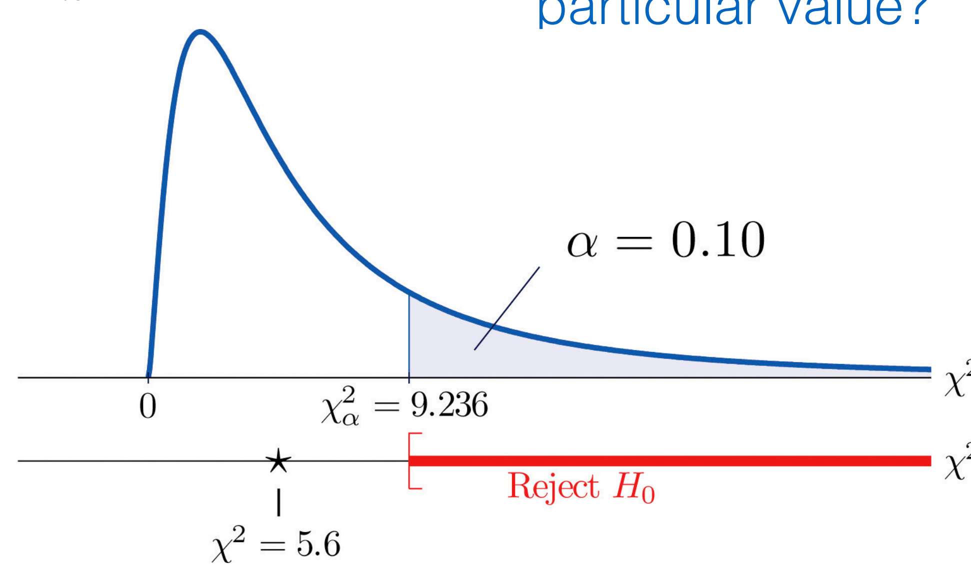
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Chi-squared: goodness of fit

$$\chi^2 = \sum_{n=1}^{N} \frac{(y_n - y_n')^2}{\sigma_n^2}$$

This is the χ^2 statistic
How likely is it to obtain a particular value?

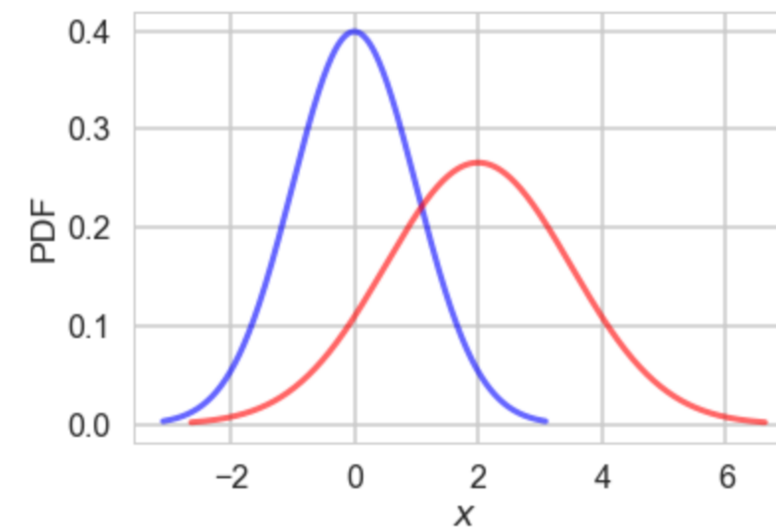


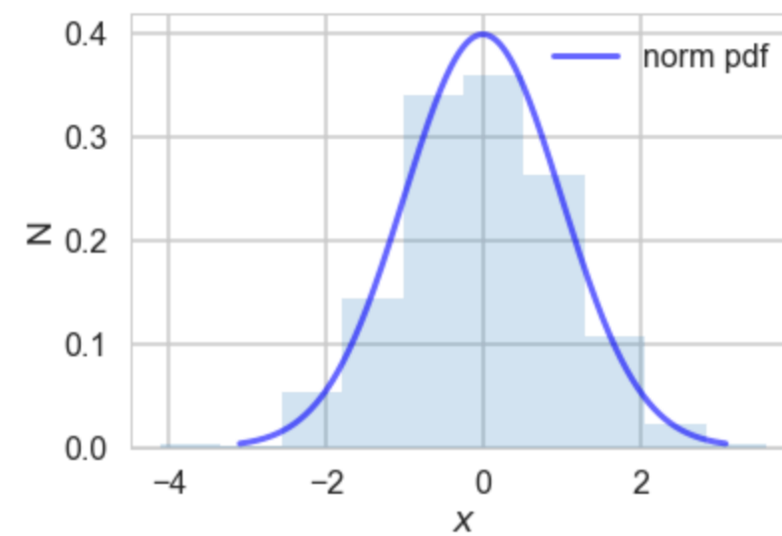
Properties of PDFs

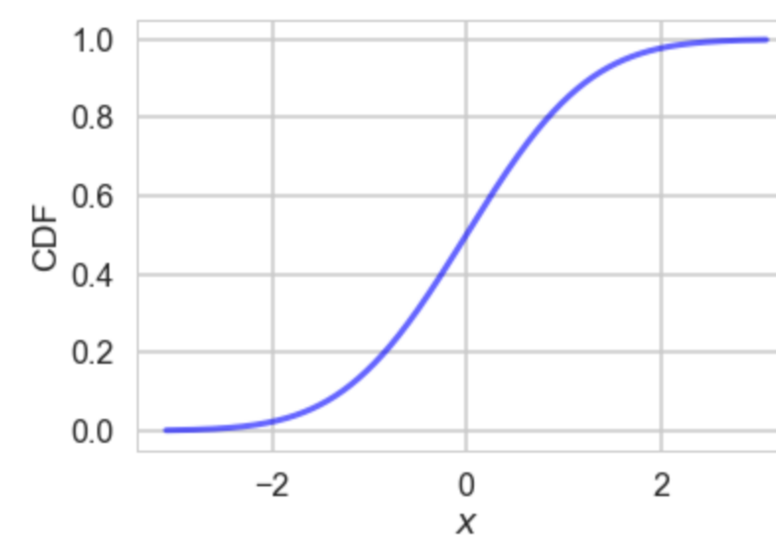
- PDFs tell us about the probability of a random variable adopting a given value.

- They also have a predictive power: you can use them to randomly draw possible values of the random variable, and over time they should resemble the PDF itself. The process of drawing random values is known as 'sampling'.

- We can also get the cumulative probability distribution, that tell us how likely it is to get a value that is larger or smaller than certain value.





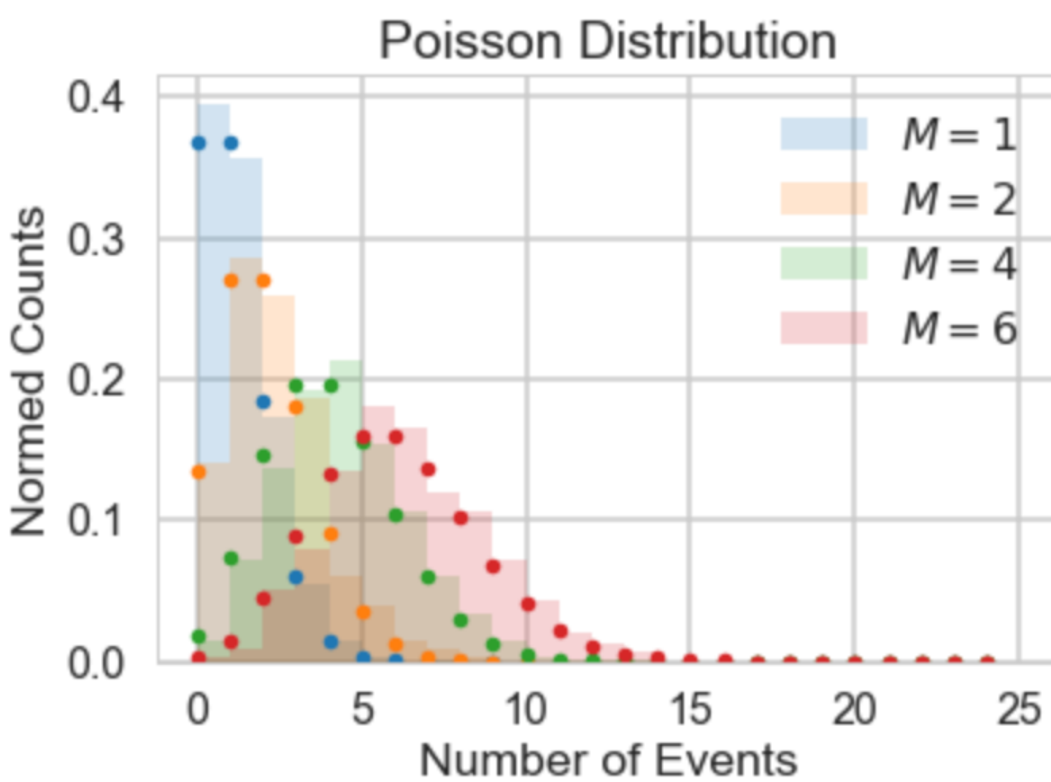


Counting photons hitting an X-ray detector - Poisson distribution

The discrete Poisson distribution:

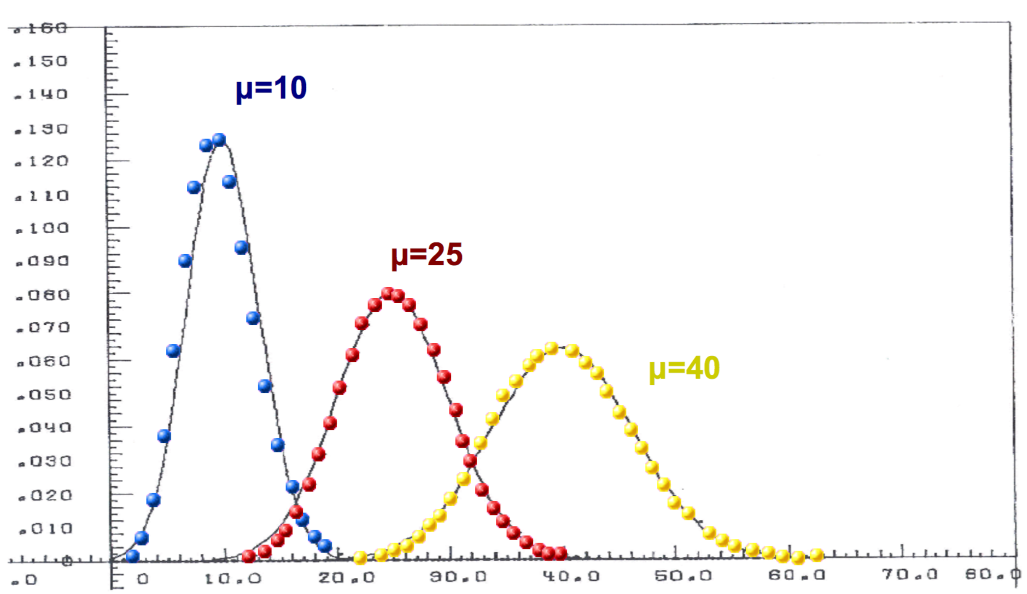
$$p(D_i \mid M_i) = \frac{M_i^{D_i}}{D_i!} e^{-M_i}$$

probability of finding D_i events (*counts*) in bin *i* (*energy rage*) of dataset *D* (*spectrum*) in a given length of time (exposure time), if the events occur independently at a constant rate M_i *(source intensity).*



As M -> ∞ Poisson distribution converges to Gaussian distribution:
$N(\mu = M_i; \sigma_2 = M_i)$

The continuous Gaussian distribution is more suitable for optical astronomy, where photon detection efficiency is much larger,

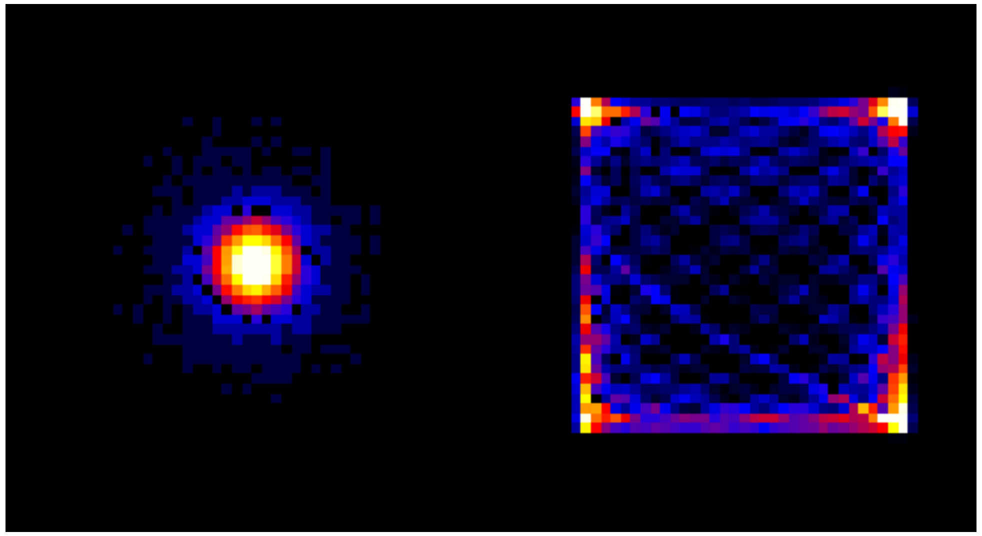


A Chandra dataset: evt2 file

An event file is a list of events (photons hitting the detector that are detected). An event file is *not* an image (but can easily be turned into one).

The calibration is performed by transforming detector coordinates of the events into sky coordinates by applying an aspect solution, that accounts for telescope motion

The time of each event recorded using a correlation between telemetry and an absolute reference



How a point source might look on the detector

Chandra-ACIS
Some extension columns

Column Name	Description
TIME	S/C TT corresponding to mid-exposure
CCD_ID	ccd reporting event
NODE_ID	ccd serial readout amplifier node
EXPNO	exposure number of ccd frame reporting event
CHIPX	X position of center pixel of event
CHIPY	Y position of center pixel of event
TDETX	X position of event in tiled detector coordinates
TDETY	Y position of event in tiled detector coordinates
DETX	X position of event in ACIS detector coordinates
DETY	Y position of event in ACIS detector coordinates
X	X position of event in sky coordinates
Y	Y position of event in sky coordinates
ENERGY	nominal energy of event

Random variables

X-ray photons hitting a detector -> Random process
In order to link random events to probabilities we need to define random variables.

Definition: A random variable is a mapping:

$$X : \Omega \rightarrow \mathbb{R}$$

That assigns a real number X(ω) to each outcome ω.

Said differently, a random variable is a variable whose possible values are the outcomes of a random phenomenon.

For example, given an X-ray flux F_{true}, we assign a probability Poisson(k=14; M=F_{true}) to the event of detecting exactly 14 photons in a given observation.

We can easily code random variables in Python

```
from random import randint
repeat = True
while repeat:
    print("You rolled",randint(1,6))
    print("Do you want to roll again?")
    repeat = ("y" or "yes") in input().lower()
```

```
y
 You rolled 5
 Do you want to roll again?
 You rolled 2
 Do you want to roll again?
 You rolled 1
 Do you want to roll again?
 You rolled 5
 Do you want to roll again?
 You rolled 5
 Do you want to roll again?
 You rolled 6
```

What is the random variable here?
What probability is assigned to each possible outcome?

Example

- Two ways of using the Poisson distribution in X-ray datasets.

  - 1) Number of photons hitting the detector at a particular location (dx,dy) per unit time, regardless of their energy

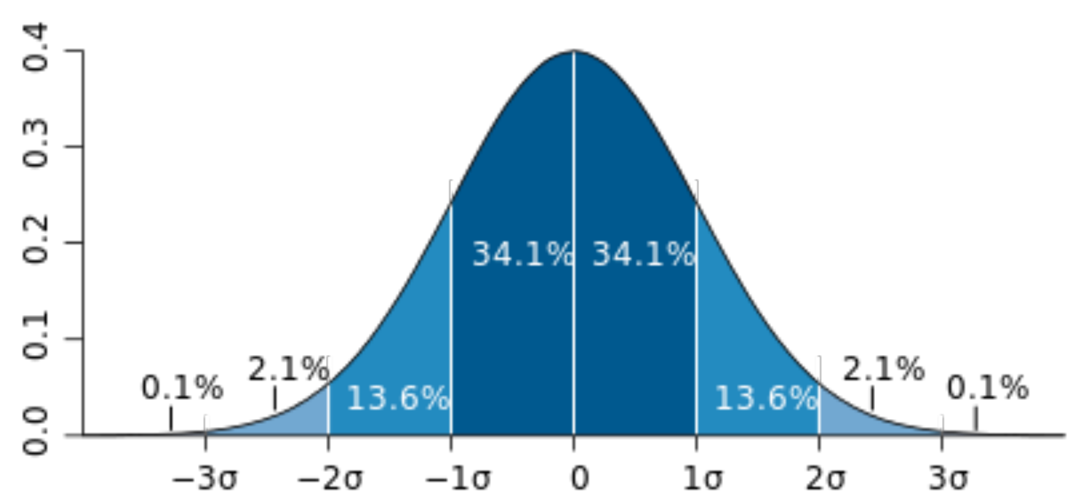
  - 2) Numbers of photons of energy Ei hitting the detector per unit time, regardless of their position.

  - (See Jupyter notebook).

Measuring random variables

- In general, we use two statistics to describe random variables: location and scatter.
  - (a) Location. Where are most of my samples?:
    - Mean: this is just the average: $\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$
    - Median: the value who is exactly in the middle of the arranged set of values
    - Mode: the value that occurs more frequently
  - (b) Scatter. How far are the samples from the are where most of the samples are:
    - Standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}.$$



    - What does F = 1.2 +/- 0.1 mJy mean?

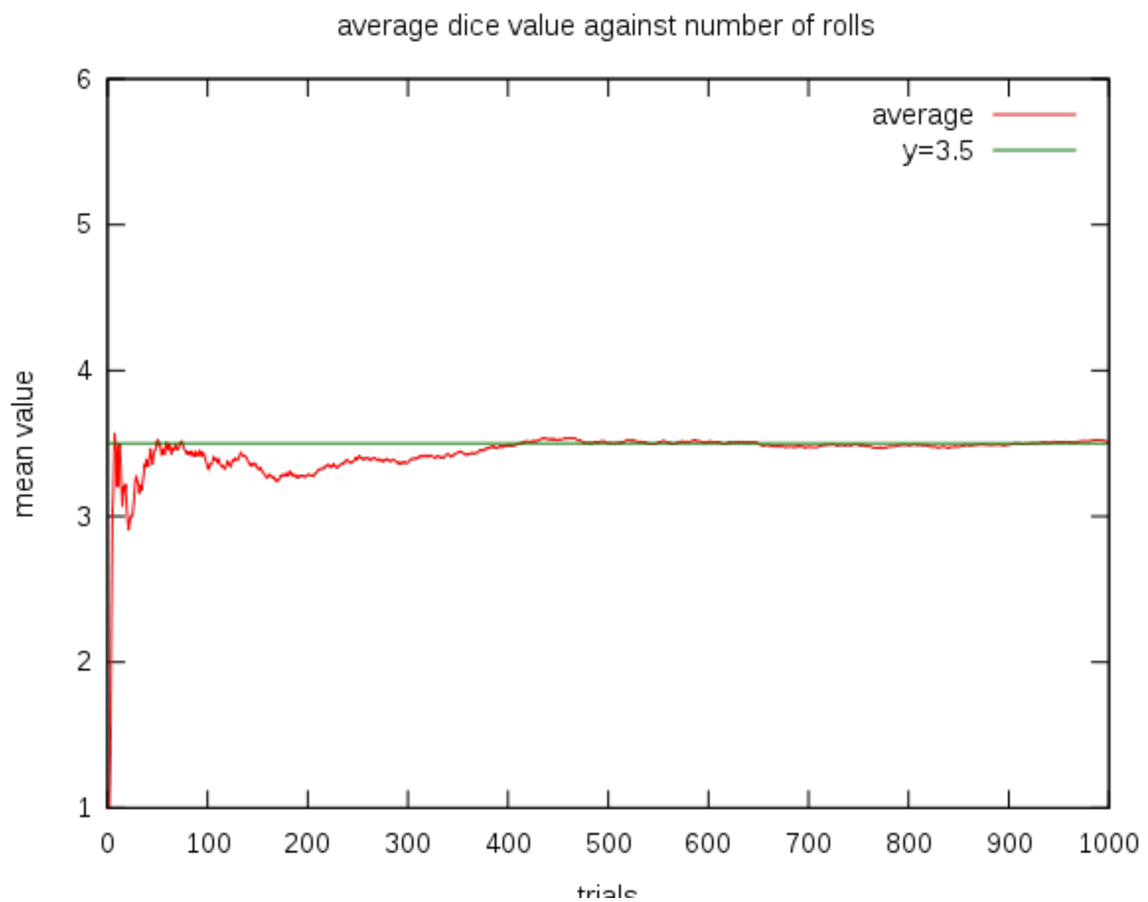
Expected values

We assume that random variables are distributed according to certain distribution, which has certain parameters. Expected values relate the data to these distributions:

What is the expectation, or expected value of a random variable?

Intuition: long-run average value of repetitions of the experiment it represents.

Practice: the probability-weighted average of all possible values.



Formally: $\mathrm{E}[X] = \int_{\mathbb{R}} x f(x)\, dx.$

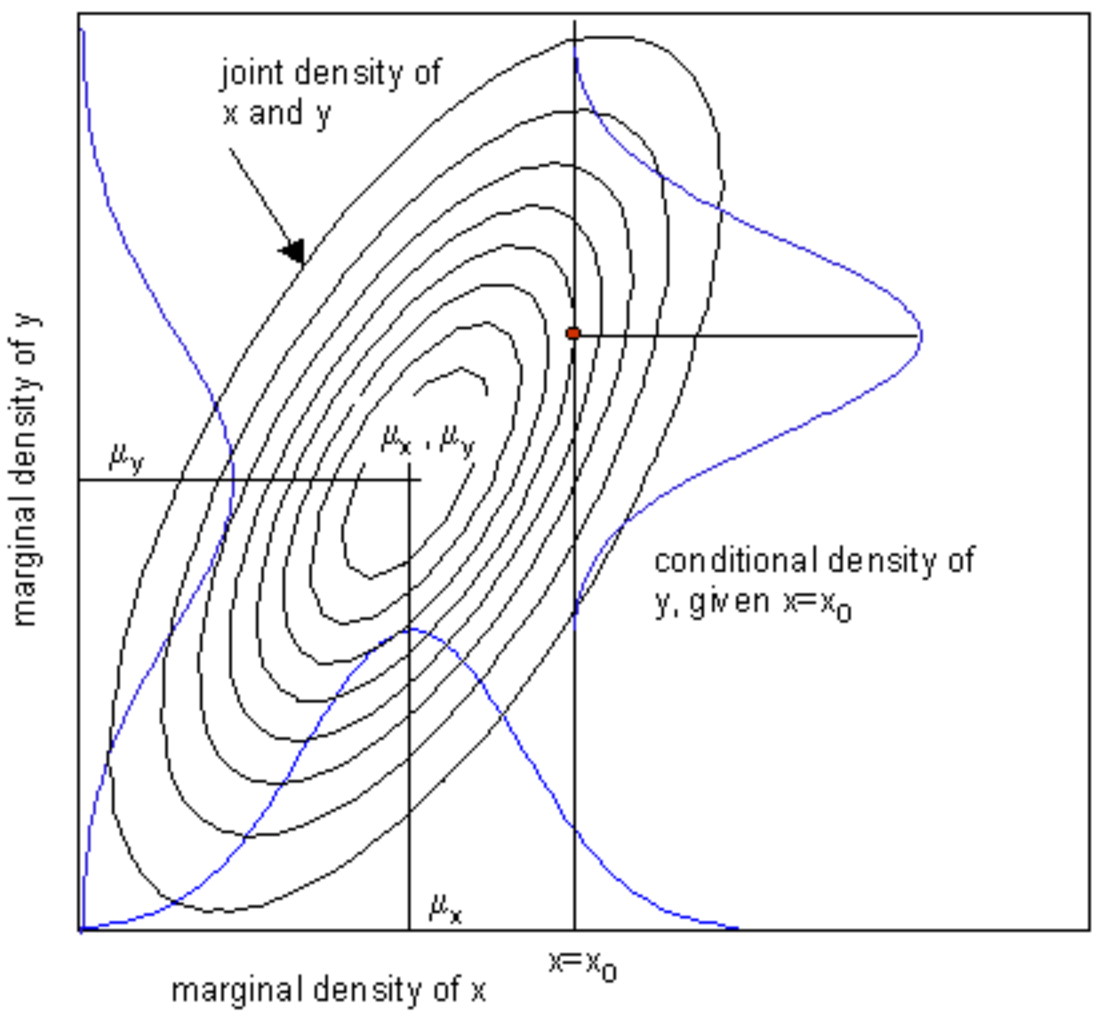
$$\mathrm{E}[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5.$$

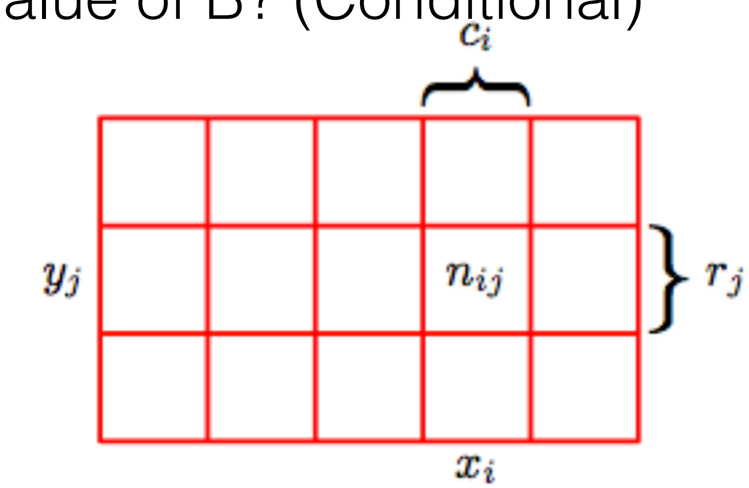
The sum of all possible values of x, weighted by their probability of occurrence. The result of repeating an experiment many times.

Joint, marginal, and conditional probabilities

We need to deal with probabilities when more than one random variables are involved. Three questions:

- What is the probability of A and B happening together? (Joint)
- What is the probability of A happening, when all possible outputs of B are considered? (Marginal)
- What is the probability of A, given a specific value of B? (Conditional)





Joint: $p(X = x_i, Y = y_j) = n_{ij}/N$

Marginal: $p(X = x_i) = \sum_j p(X = x_i, Y = y_j)$

Conditional: $p(Y = y_j \mid X = x_i) = \frac{n_{ij}}{c_i}$

Note that (product rule of probability):

$$p(Y = y_j \mid X = x_i) = \frac{n_{ij}}{c_i} = \frac{n_{ij}}{N} / \frac{c_i}{N}$$

Joint, marginal, and conditional probabilities

Example: rolling two dice a thousand times.

$$P(d_1 = 3, d_2 = 3) = \frac{n_{3,3}}{N}$$

$$P(d_1 = 3) = \frac{\sum_j n_{3,j}}{N}$$

$$P(d_1 = 3 | d_2 = 2) = \frac{n_{3,2}}{\sum_i n_{i,2}}$$

Die 1

Die 2

	1	2	3	4	5	6
1	27	29	31	30	32	28
2	21	26	30	30	33	24
3	27	33	32	20	31	28
4	17	28	35	25	34	18
5	26	24	21	28	21	32
6	26	34	35	23	25	36

Joint:
$P(d_1{=}3, d_2{=}2) = 30/1000 = 0.03$

Marginal:
$P(d_1{=}3) = (31+30+32+35+21+35)/1000 = 0.18$

Conditional:
$P(d_1{=}3|d_2{=}2) = 30/(21+26+30+30+33+24) = 0.18$

Statistical in(dependence)

Two random variables are statistically independent if the occurrence of one of them does not affect the probability of the other one occurring. In the case:

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

Another consequence of independence:

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

We do not learn anything on x by conditioning on a particular value of y.

If the two variables conditionally depend on each other, then:

$$p(x, y) = p(y \mid x) p(x)$$

We can apply this repeatedly for situations involving more than two variables:

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

Example: if you have a deck of 52 cards, and draw a card at random that happens to be an ace, what is the probability of drawing a second card that is also an ace?

$$\mathrm{P(Ace)} = \frac{\text{number of Aces in a deck of cards}}{\text{number of cards in a deck}}$$

$$\mathrm{P(Ace)} = \frac{4}{52} = \frac{1}{13}$$

$$\mathrm{P(Ace)} = \frac{\text{number of Aces remaining in the deck of cards}}{\text{number of cards remaining in a deck}}$$

$$\mathrm{P(Ace)} = \frac{3}{51}$$

Bayes’ Rule

Derivation of the Bayes' Theorem

We start with the product rule of probabilities. We know that:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

And also:

$$P(B|A) = \frac{P(A,B)}{P(A)}$$

Rearranging and combining these two equations, we get:

$$P(A|B)P(B) = P(A|B)P(A)$$

Now we divide both sides by $P(B)$:

$$P(A|B) = \frac{P(A|B)P(A)}{P(B)}$$

This is know as the Bayes' rule

Rev. Thomas Bayes

“An Essay towards solving a Problem in the Doctrine of Chances”, 1763

The Bayes’ theorem is just an implication of the rule of probabilities. But its interpretation has a deep impact in the way we interpret probabilities, as we will see in the next lecture. In astronomy, Bayesian inference has received lots of attention over the last years.

Example: naked-eye supernovae

http://www.astro.ubc.ca/

< 1987, 4 naked-eye supernovae had been recorded in 10 centuries. What, before 1987, was the probability of a bright supernova happening in the 20th century?

There are three possible answers.

(1) Probability is **meaningless** in this context. This is physics, deterministic, and timing can be calculated. They are not random events.

(2) **Frequentist** point of view: best estimate of the probability is 4/10, although it is obviously not very well determined. (Assumes equally likely to be reported throughout ten centuries - some degree of belief about detection efficiency will have to be made explicit in this kind of probability assignment.)

(3) We could try an **a priori** assignment. We might know

- the stellar mass function,
- the fate and lifetime as a function of mass,
- the stellar birth rate, and
- detection efficiency.

From this we could calculate the mean number of supernovae expected in 1987, and we would put some error bars around this number to reflect unknowns......

Exercise

- Using CIAO, download observation 12020 from the Chandra archive, and reprocess it with the latest calibration. This corresponds to the Cassiopeia A supernova remnant.
- Produce an RGB image like last time, and identify regions dominated by photons of different energies.
- Using DS9, load and save a few regions of interest.
- What is the distribution of photon energies in the range 0.5-7 keV for Cas A? What kind of distribution do you think that is?
- Find the mean and the standard deviations of the distribution, and record them
- Now do the same, but now limit yourself to each of the regions of interest.
- In each case, compare the histograms of energy with normal distributions with μ corresponding to the mean and σ corresponding to the standard deviation.
- What is the probability of detecting a photon in each of the three ands s,m, and h energy bands for each of the regions?
- What is the probability of detecting hard or soft X-ray photons coming from a white dwarf?
- What does this tell you about the physical processes taking place in the different regions?