Classical statistical inference

Part 1

Associated notebook:

03-Basic_statistics_and_proba_concepts/Basic-statistics_01.ipynb

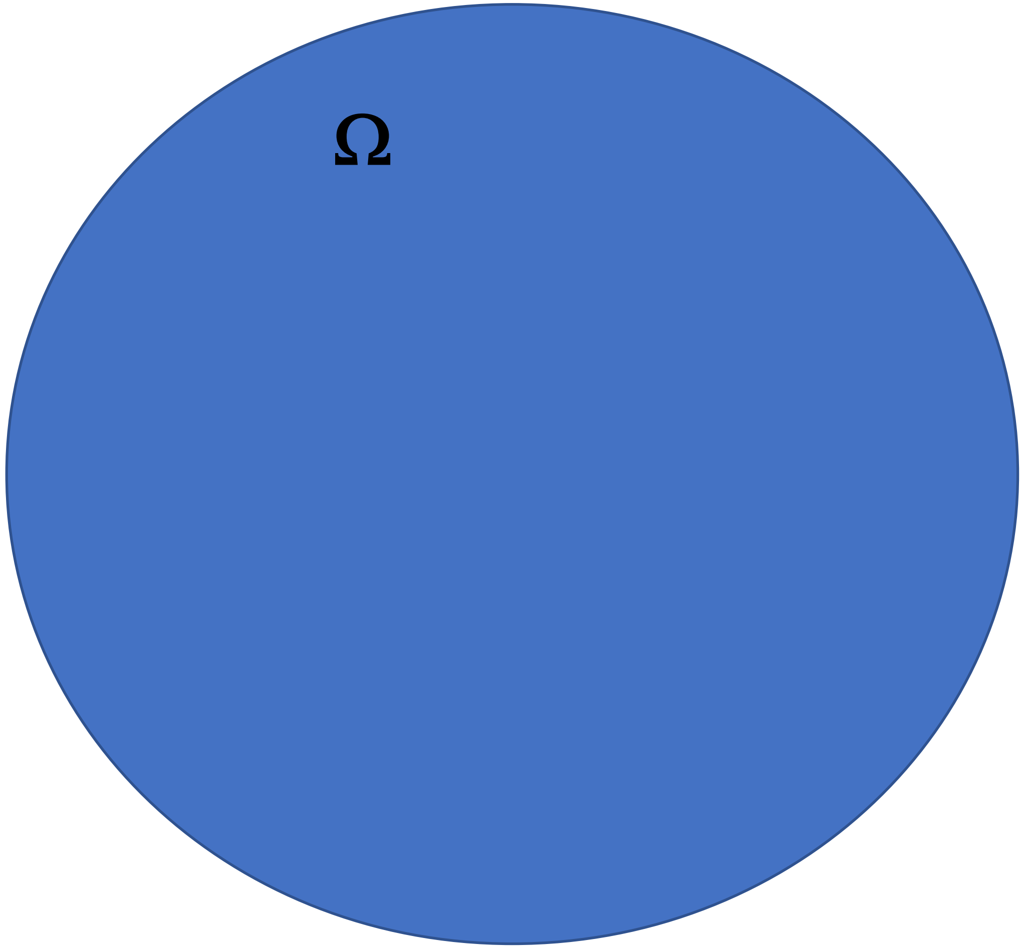
Why some statistics ?

- Python for data (observation / numerical simulations) manipulation
- Data most often contain a stochastic component: observational device, numerical noise, simulation of stochastic process, …

$\Rightarrow$Data ≈ Random variable (RV)

- Statistics is the tool needed to manipulate RV
- *Goals for 2nd part of the lecture:*
  - Uncertainty calculation (no, this is not black magic)
  - Make prediction based on data modelling (first step towards machine learning)

Definitions and notations recap



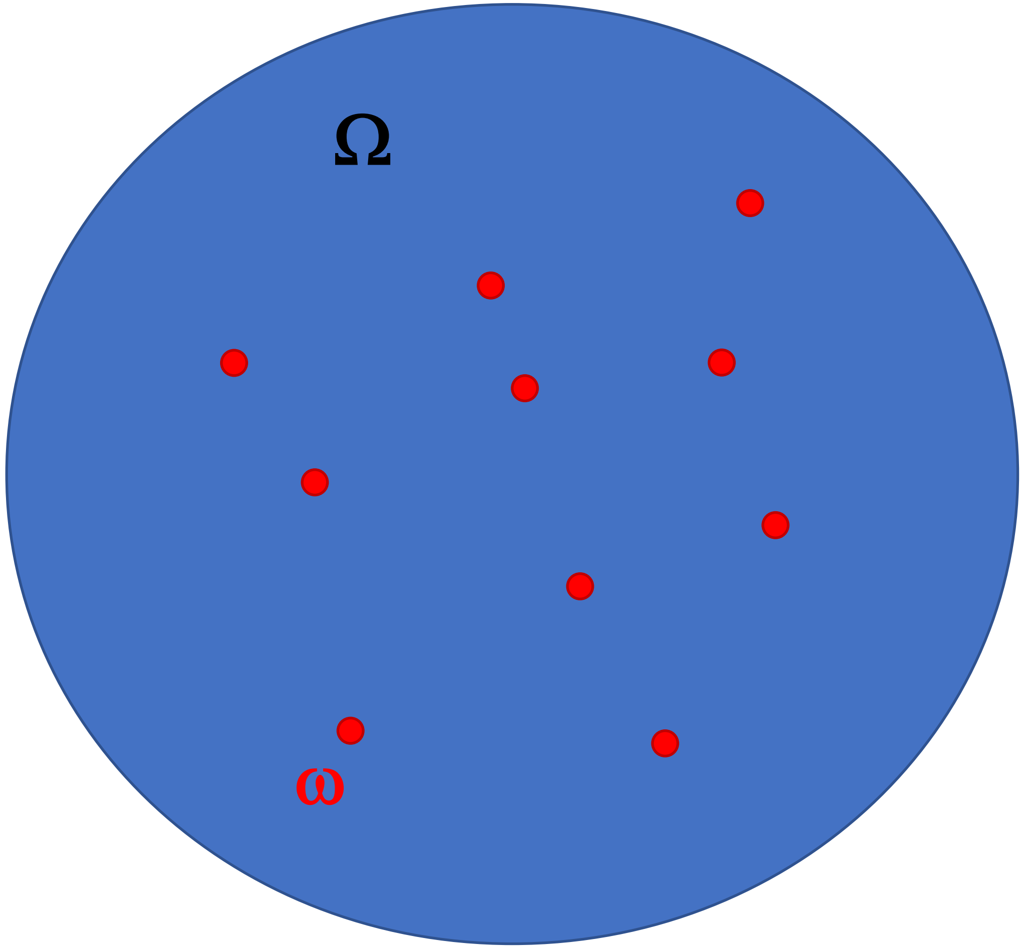
Ω: Sample space ≡ all possible outcome of an experiment

e.g. of experiment

- I measure the magnitude of a star (in a binary system, for a transit, …)
- I count galaxies for different L at a given z
- I obtain the spectrum of a candidate SN
- I measure a GW signal
- …

This is an abstract space. For the mag of a star, $\Omega \equiv \mathbb{R}$

Definitions and notations recap

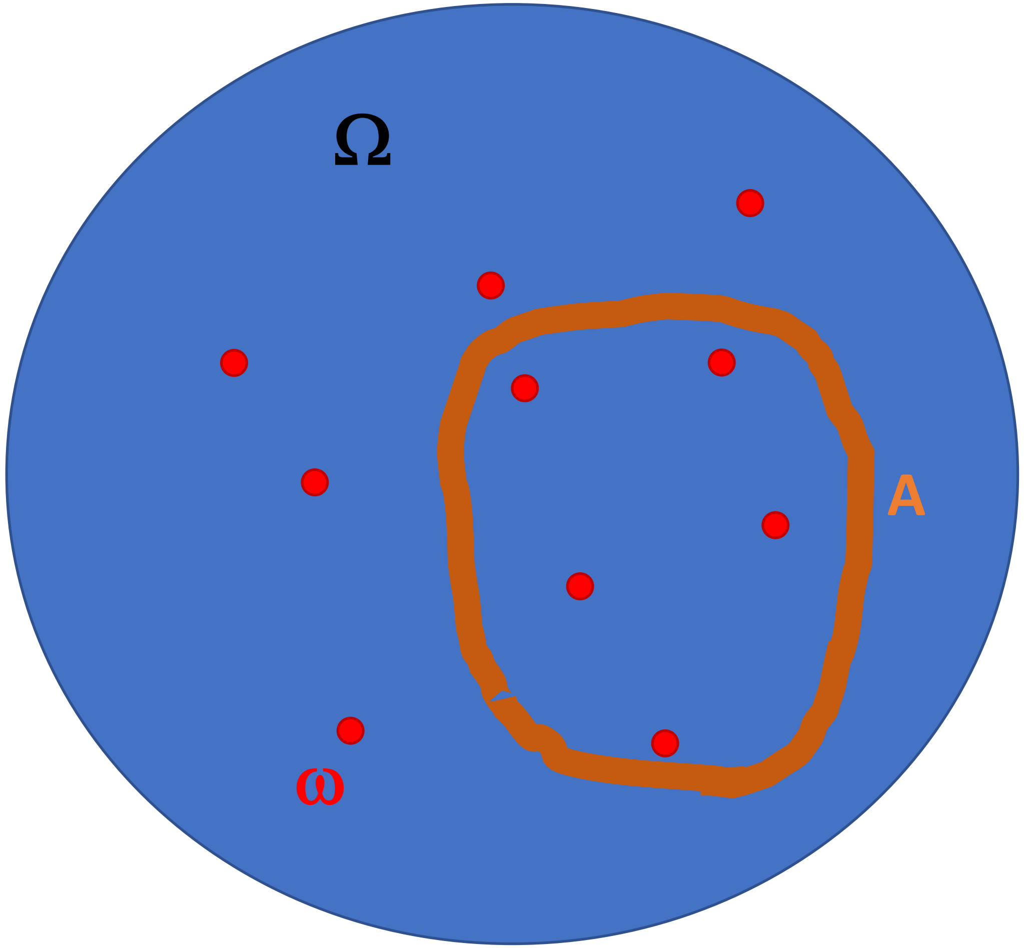


Ω: Sample space ≡ all possible outcome of an experiment

ω: Realisations of the experiment

E.g. There have been 10 measurements of the magnitude of a star.
Each measurement is a different realisation

Definitions and notations recap



Ω: Sample space ≡ all possible outcome of an experiment

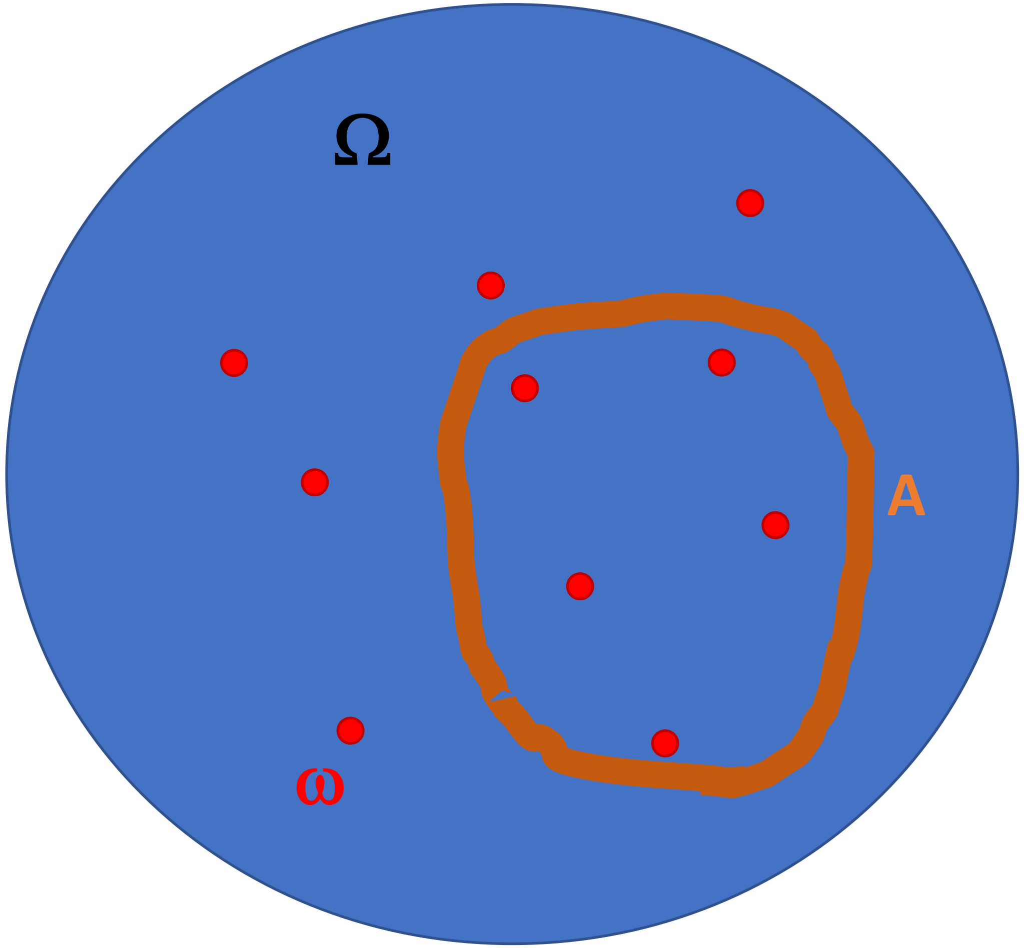
ω: Realisations of the experiment

A: Event ≡ a subsample of ω ≅ Your data set

E.g. You have obtained and are working on 5 measurements of the magnitude of the star.

But an event can be a bit more convoluted quantity, e.g. all measurements you've done that have m < 15 mag

Definitions and notations recap



Ω: Sample space ≡ all possible outcome of an experiment

ω: Realisations of the experiment

A: Event ≡ a subsample of ω ≅ Your data set

p(A): Probability of an event / value to be in [x-dx, x+dx]

e.g. probability that m < 15 mag

What means p(A) in frequentist/classical inference ?

Relative frequency of an event
if experiment is repeated an infinite number of times

Random variable

A random variable is a variable whose value results from the measurement of a quantity that is subject to random variations

In Python: np.random

- **np.random.choice(array):** choice at random in an array
- **np.random.seed(value):** sets the seed of the rnd generator
- **np.random.rand(shape):** random floats drawn from uniform distribution
- **np.random.randint(low, high, shape):** rnd integers btw low and high

Go to: Sect. I.2. of the notebook

Conditional probability p(A | B)

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)}$$ = fraction of times that A occurs given than B occurred

Reads "Probability of A *given* B"

- The calculation of ***p(A | B)*** follows **Bayes** theorem

$$\boxed{p(A \mid B) = \frac{p(B \mid A)\, p(A)}{p(B)}}$$

- The probability to have a flu given that you have fever is different from the probability to have fever given that you have a flu

$$\boxed{p(A|B) \neq p(B|A)}$$

Bayes theorem

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{p(B)}$$

Question:
A: **rare** disease that affects 0.1 % of the population.
B: test that is efficient at 99 % (i.e. 1 % False positive rate).
If you have a positive test (B), what is the probability for you to be affected by this disease (A) ?

Bayes theorem

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{p(B)}$$

Question
A: **rare disease** that affects 0.1 % of the population.
B: test that is efficient at 99 % (i.e. 1 % False positive rate).
If you have a positive test (B), what is the probability for you to be affected by this disease (A) ?

Solution: (See Sect. I.3. of the notebook)

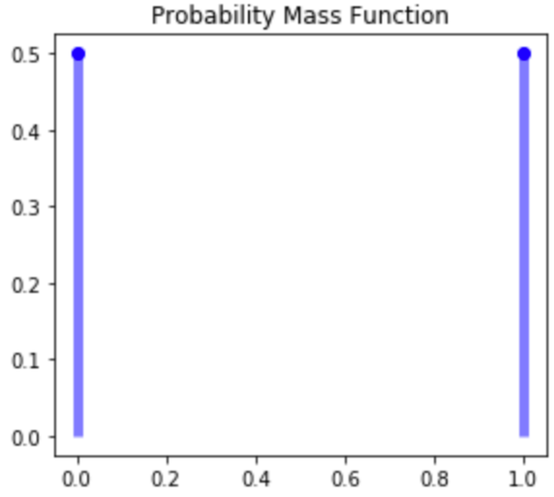
Among 1000 persons, 1 has the disease (it touches 0.1 % of the population = p(A)).
The test has 99% efficiency (=p(B | A)). Which means that 1% of the people will be tested positive while not being sick.
=> 10 people will be positive while healthy. You should also have ≈ 1 being positive while being effectively sick. p(B)=0.01 + 0.001 =0.011
=> p(disease} | +) ≈ 1/11 = 9 %

BEWARE

RARE events common in astronomy

Conditional probabilities are often implicit

Probability density / mass function



Coin Toss (Bernouilli PMF): The PDF is the normalised histogram we had obtained

$$\boxed{Ber(k \mid p) \;=\; p^k\,(1-p)^k}$$

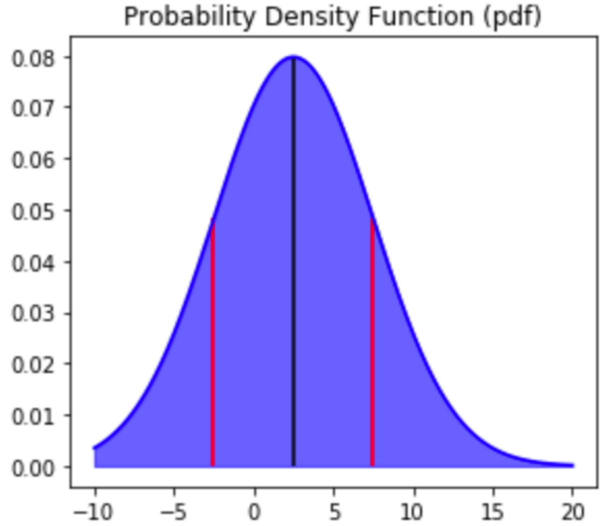
k in {0, 1} ≡ {failure, success}

parameter (success rate)

$$\int p(x)\,\mathrm{d}x = 1$$

Uniform PDF:

$$\boxed{\begin{aligned} h(x) &= \tfrac{1}{b-a} \text{ if } a \leq x \leq b \\ h(x) &= 0 \;\text{ otherwise} \end{aligned}}$$



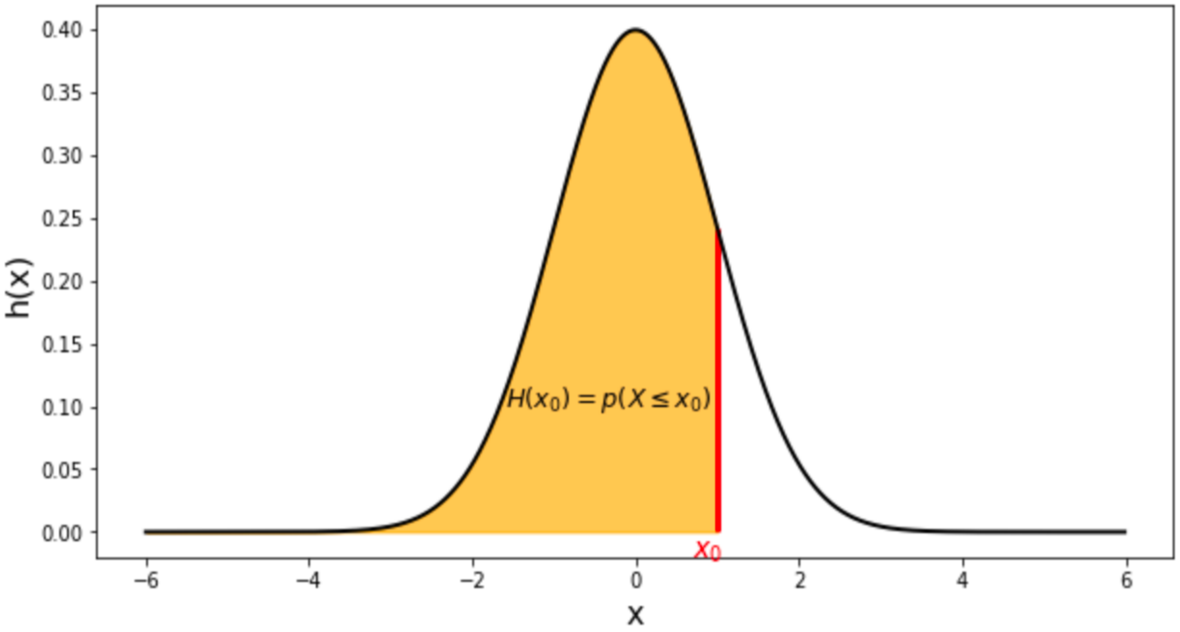
Gaussian PDF:

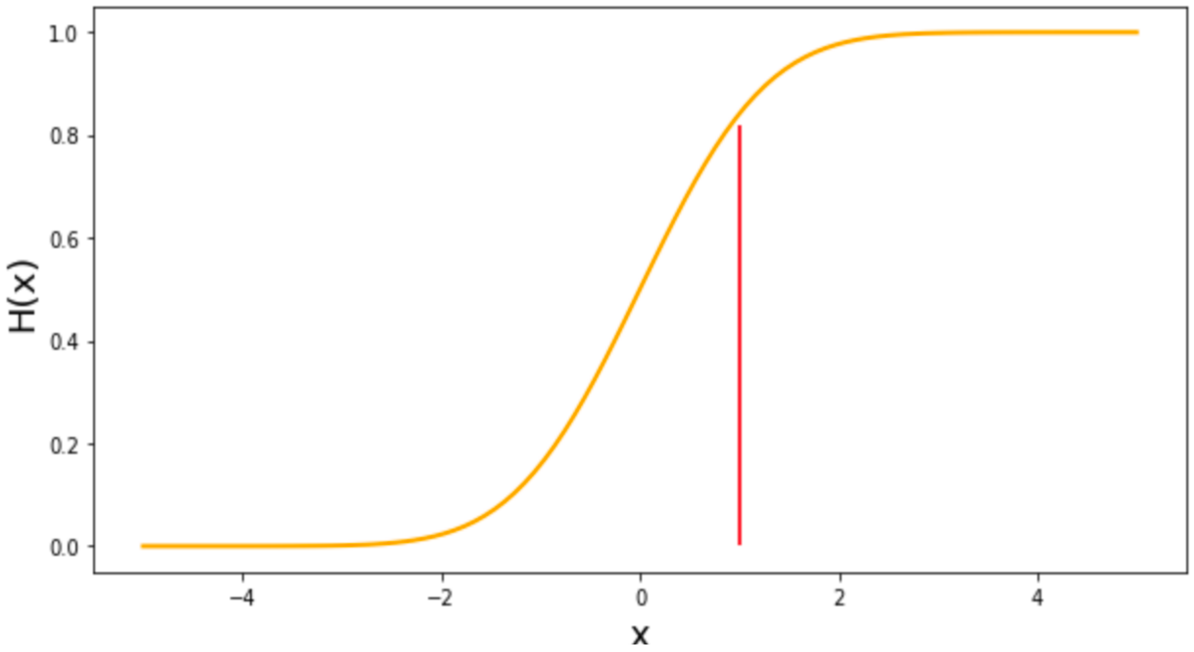
$$\boxed{h(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)}$$

PDF In Python: go to Sect. I.4 of the notebook

Cumulative density function

This is the integral of the PDF: $$p(X \leq x) = H(x) = \int_{-\infty}^{x} h(x')\,\mathrm{d}x'$$





$$H(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\frac{(x'-\mu)^2}{\sigma^2}\right)\mathrm{d}x'.$$

CDF In Python: go to Sect. I.5 of the notebook

Probability enclosed between 1-2-3 σ for N(μ, σ)

See the last exercise of Sect. I.5 of the notebook