Classical statistical inference

Part 1

Associated notebook:

03-Basic_statistics_and_proba_concepts/Basic-statistics_01.ipynb

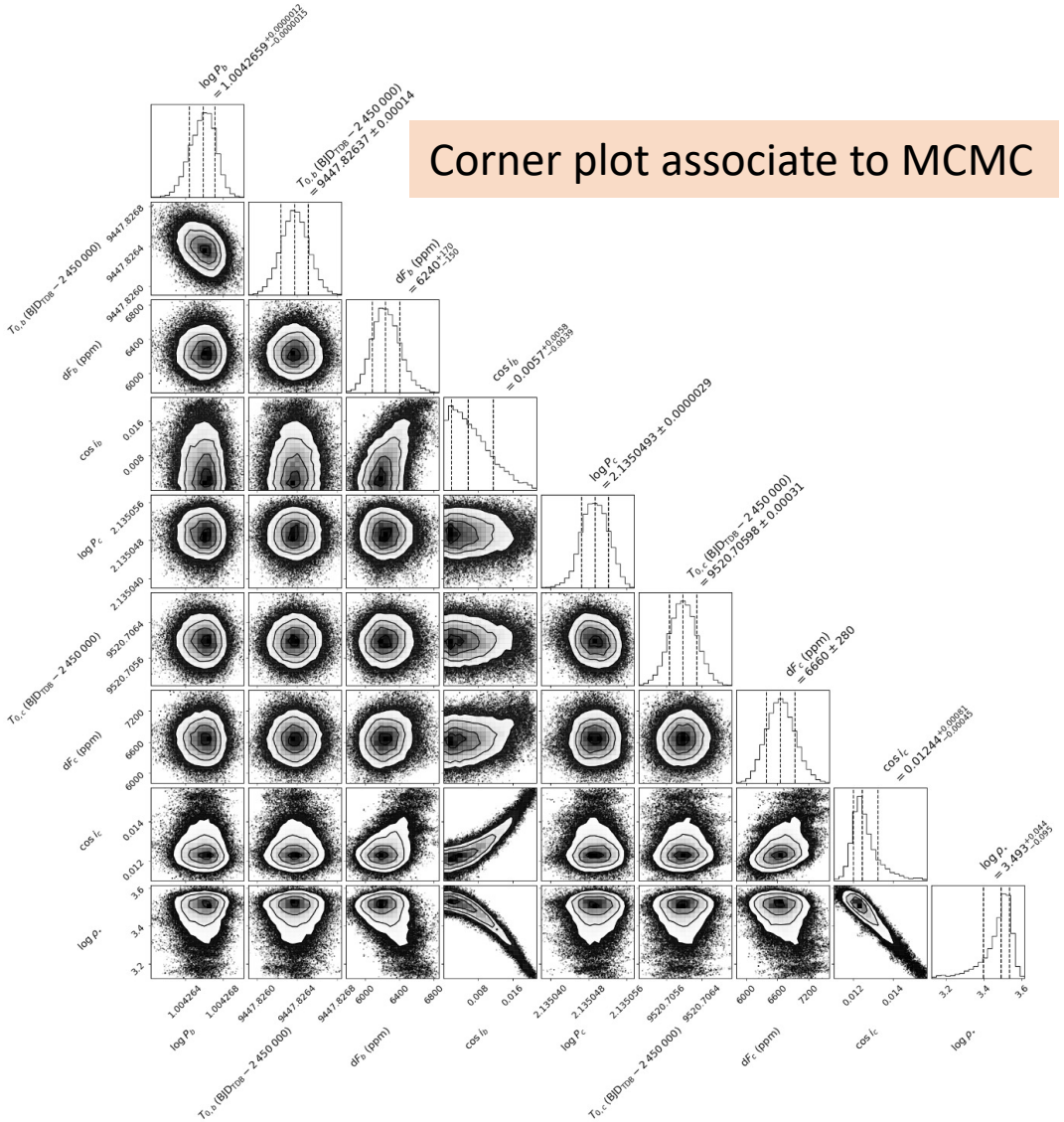
Why some statistics? (Concrete Motivation)

Two temperate super-Earths transiting a nearby late-type M dwarf⋆

L. Delrez[1,2], C.A. Murray[3,4], F.J. Pozuelos[1,2,5], N. Narita[6,7,8], E. Ducrot[9,**], M. Timmermans[1], N. Watanabe[10], A.J. Burgasser[11], T. Hirano[7,12], B.V. Rackham[13,14,***], K.G. Stassun[15], V. Van Grootel[2], C. Aganze[11], M. Cointepas[16,17], S. Howell[18], L. Kaltenegger[19], P. Niraula[13], D. Sebastian[20], J.M. Almenara[16], K. Barkaoui[1,13,8], T.A. Baycroft[20], X. Bonfils[16], F. Bouchy[17], A. Burdanov[13], D.A. Caldwell[18,21], D. Charbonneau[22], D.R. Ciardi[23], K.A. Collins[22], T. Daylan[14,24,25],

et al. (2014) relation between these observables and [Fe/H]. We used a Monte Carlo approach to calculate the uncertainty in our estimate, drawing 10^6 samples from normal distributions given by the means and standard deviations of the measurements. We calculated the mean and standard deviation of the resulting values and, adding in quadrature the systematic uncertainty of the relation (0.07), we arrived at our final metallicity estimate of

We performed a fit using NextGen stellar atmosphere models, with the free parameters being the effective temperature ($T_{\rm eff}$) and metallicity ([Fe/H]). The remaining free parameter is the extinction A_V, which we fixed at zero due to the star's proximity. The resulting fit (Fig. 6) has a reduced χ^2 of 1.7, with best fit $T_{\rm eff} = 2850 \pm 75$ K and [Fe/H] $= 0.0 \pm 0.5$. Integrating the model SED gives the bolometric flux at Earth, $F_{\rm bol} = 4.41 \pm 0.15 \times 10^{-11}$ erg s^{-1} cm^{-2}. Taking the $F_{\rm bol}$ and $T_{\rm eff}$, together with the *Gaia* EDR3 parallax, gives the stellar radius, $R_\star = 0.1556 \pm 0.0086\ R_\odot$.



Corner plot associate to MCMC

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, …

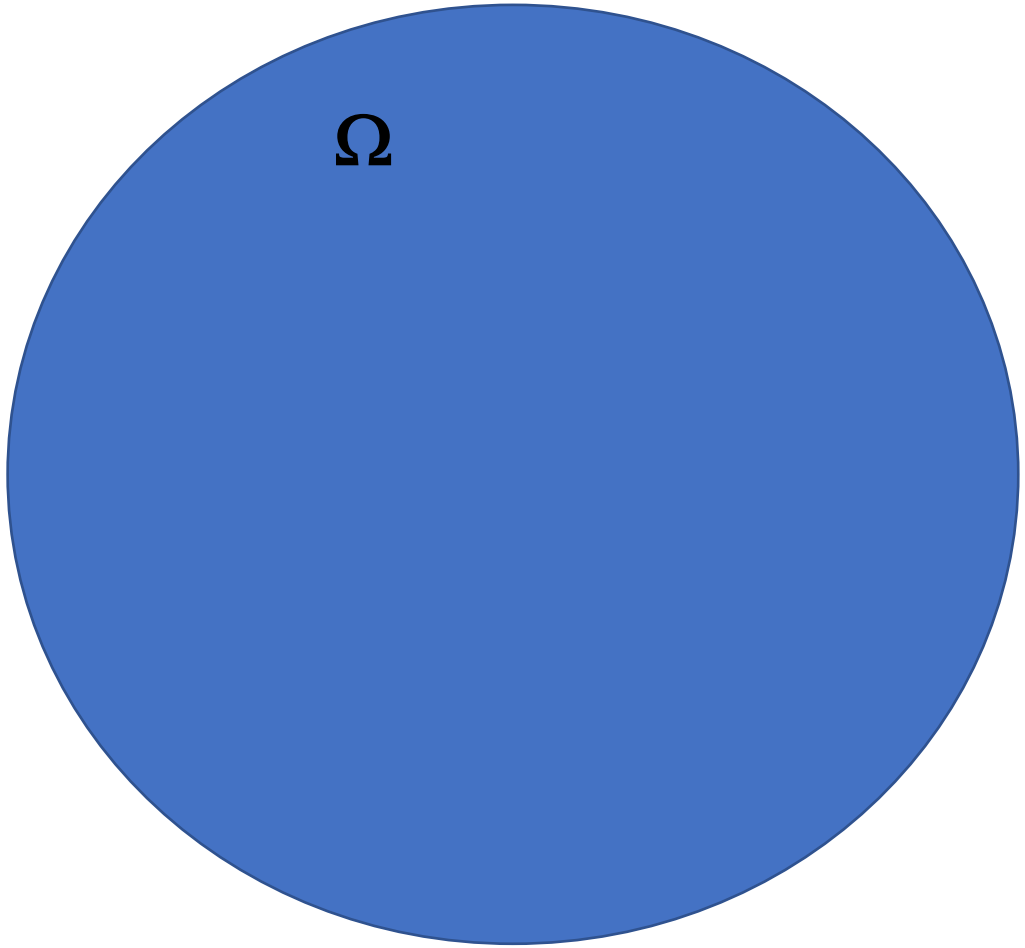
  $\Rightarrow$Data $\approxeq$ Random variable (RV)

- Statistics is the tool needed to manipulate RV. Python allows one to have a more *practical understanding* of statistics.

- *Main Goals:*
  - Uncertainty calculation (no, this is not black magic)
  - Make prediction based on data modelling (first step towards machine learning)

http:// www.wooclap.com/RSLBIT

Definitions and notations recap

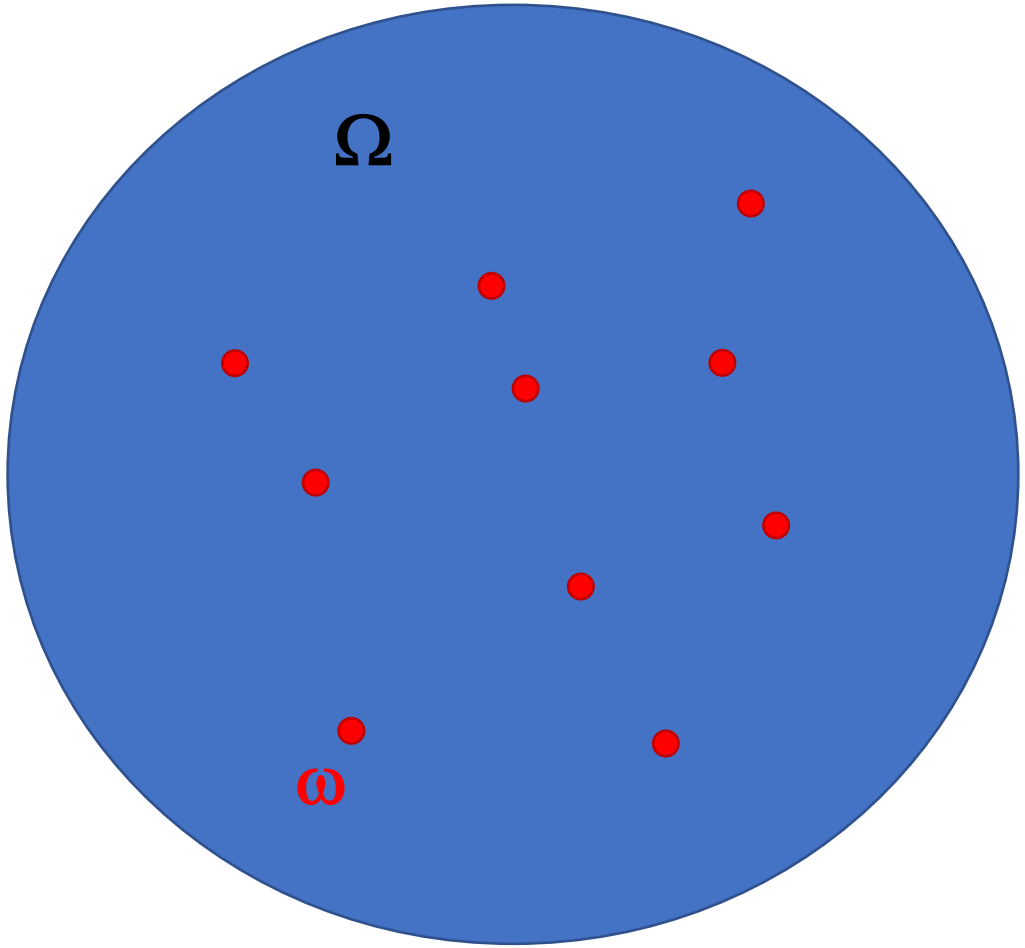


Ω: Sample space ≡ all possible outcome of an experiment

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

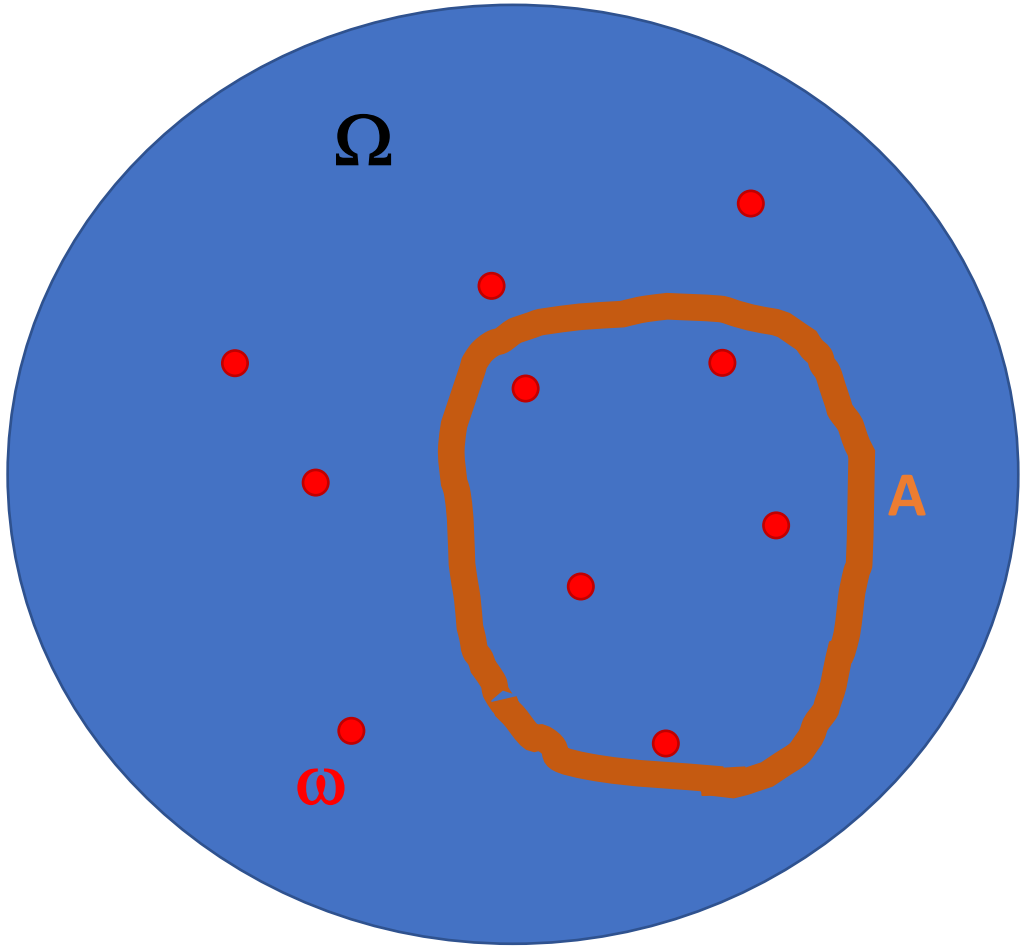


Ω: Sample space ≡ all possible outcome of an experiment

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

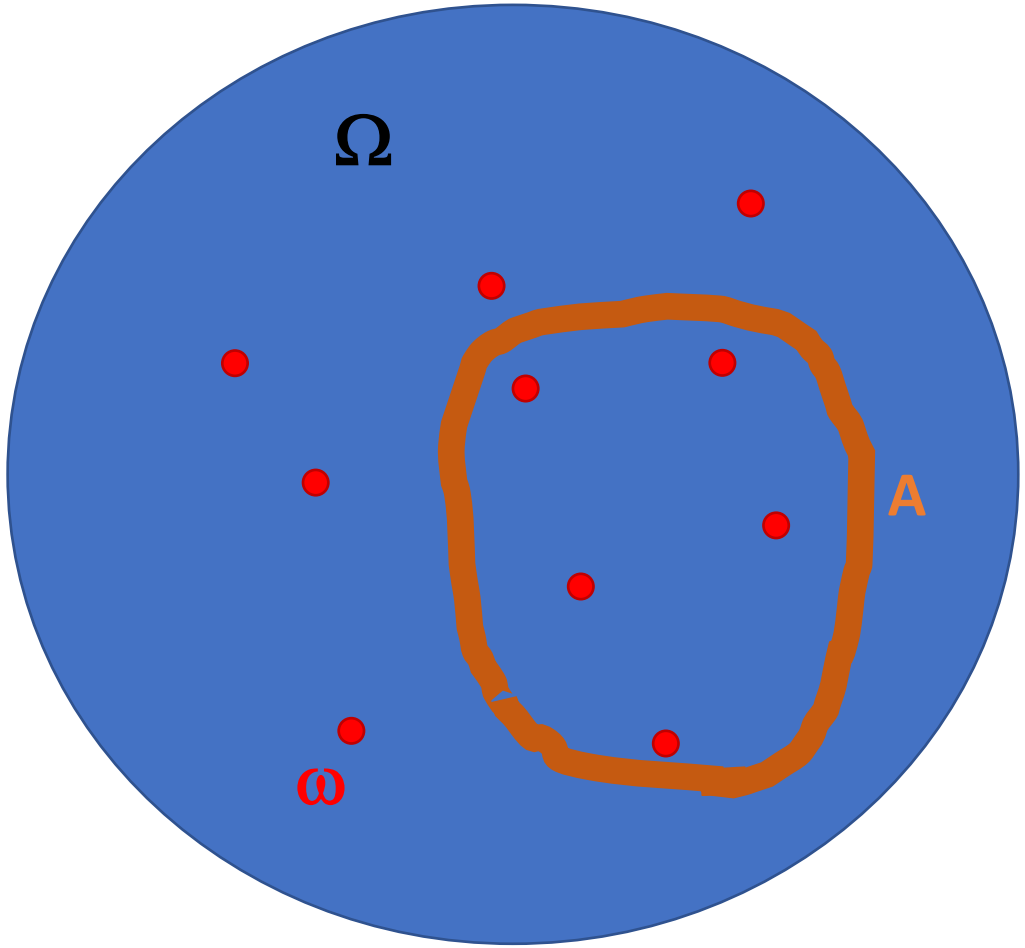
ω: Realisations of the experiment

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 that B occurred

Reads "Probability of A *given* B"

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

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

- 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

$$p(A|B) \neq p(B|A)$$

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) ?

NB: Efficiency is NOT sensitivity (sensitivity generally means fraction of true positive).

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

Bayes theorem: disease problem

Contingency Table

E \ H	Disease = H	No Disease = !H	Σ
Test + = E	P(E | H) p(H)	P(E | !H) p(!H)	P(E)
Test - = !E	P(!E | H) p(H)	P(!E | !H) p(!H)	P(!E)
Σ	P(H)	P(!H)	1

E \ H	Disease = H	No Disease = !H	Σ
Test + = E			
Test - = !E			
Σ	100	999000	100000

Fill this row first →

0.1 % of the people have a disease => 100 out of 100000

Bayes theorem: disease problem

Contingency Table

E \ H	Disease = H	No Disease = !H	Σ
Test + = E	P(E | H) p(H)	P(E | !H) p(!H)	P(E)
Test - = !E	P(!E | H) p(H)	P(!E | !H) p(!H)	P(!E)
Σ	P(H)	P(!H)	1

E \ H	Disease = H	No Disease = !H	Σ
Test + = E	0.99 * 100 = 99		
Test - = !E	(1-0.99) * 100 = 1		
Σ	100	999000	100000

The test is efficient at 99%: P(E | H) = 0.99

Bayes theorem: disease problem

Contingency Table

E \ H	Disease = H	No Disease = !H	Σ
Test + = E	P(E | H) p(H)	P(E | !H) p(!H)	P(E)
Test - = !E	P(!E | H) p(H)	P(!E | !H) p(!H)	P(!E)
Σ	P(H)	P(!H)	1

E \ H	Disease = H	No Disease = !H	Σ
Test + = E	0.99 * 100 = 99	(1 - 0.99) * 999000 = 999	1098
Test - = !E	(1-0.99) * 100 = 1		
Σ	100	999000	100000

P(E | !H) = proba. To be positive without being sick = (1-0.99)

Bayes theorem: disease problem

Contingency Table

E \ H	Disease d = H	No Disease = !H	Σ
Test + = E	P(E | H) p(H)	P(E | !H) p(!H)	P(E)
Test - = !E	P(!E | H) p(H)	P(!E | !H) p(!H)	P(!E)
Σ	P(H)	P(!H)	1

$$P(H \mid E) = \frac{P(E \mid H)\, P(H)}{P(E)}$$

E \ H	Disease d = H	No Disease = !H	Σ
Test + = E	0.99 * 100 = 99	(1 - 0.99) * 999000 = 999	1098
Test - = !E	(1-0.99) * 100 = 1	98901	100000-1098=98902
Σ	100	999000	100000

$$P(d \mid +) = \frac{P(+ \mid d)\, P(d)}{P(+)} = \frac{99}{1098}$$

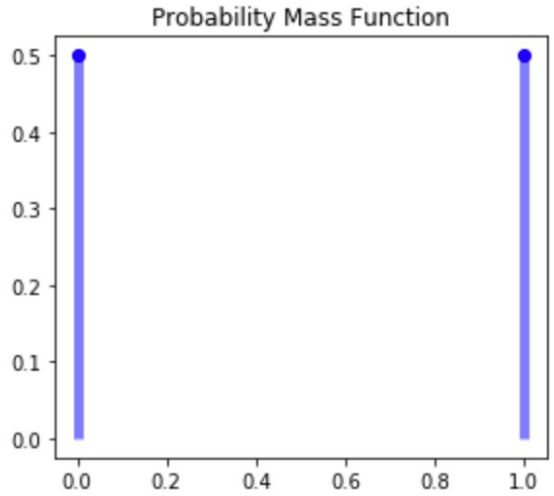
Probability density / mass function

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

$$Ber(k \mid p) = p^k (1-p)^k$$

k in {0, 1} ≡ {failure, success}

parameter (success rate)



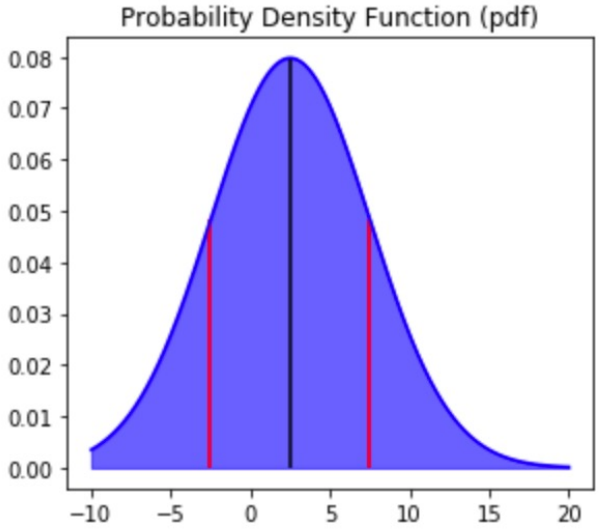
$$\int p(x)\,dx = 1$$

Uniform PDF:

$$h(x) = \frac{1}{b-a} \text{ if } a \le x \le b$$
$$h(x) = 0 \text{ otherwise}$$

Gaussian PDF:

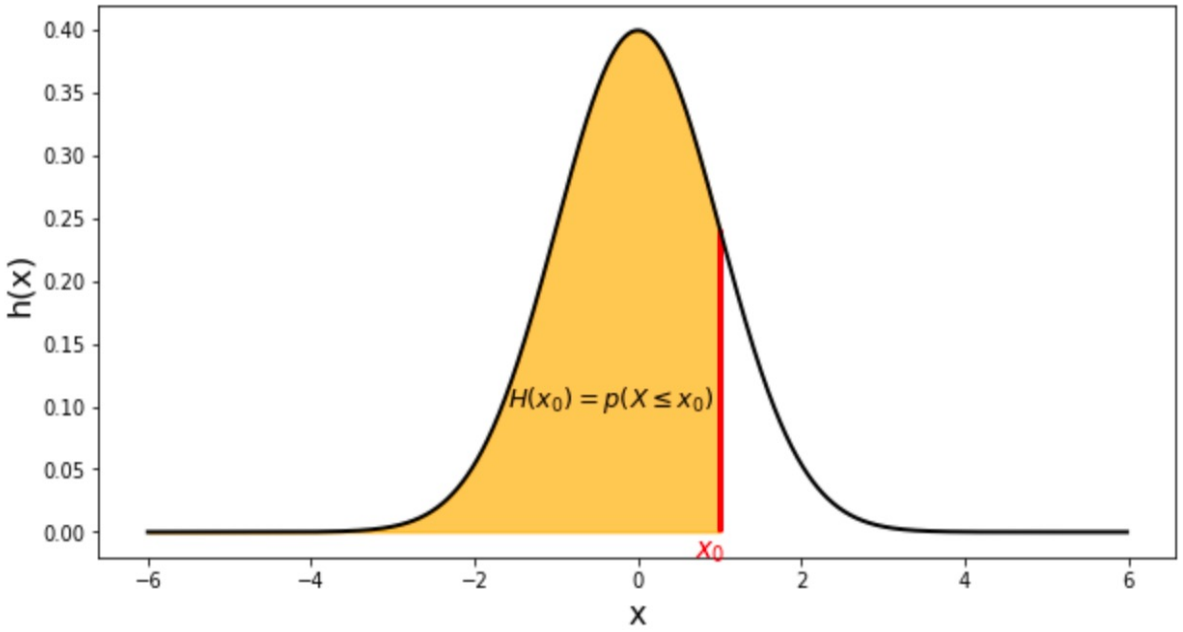
$$h(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

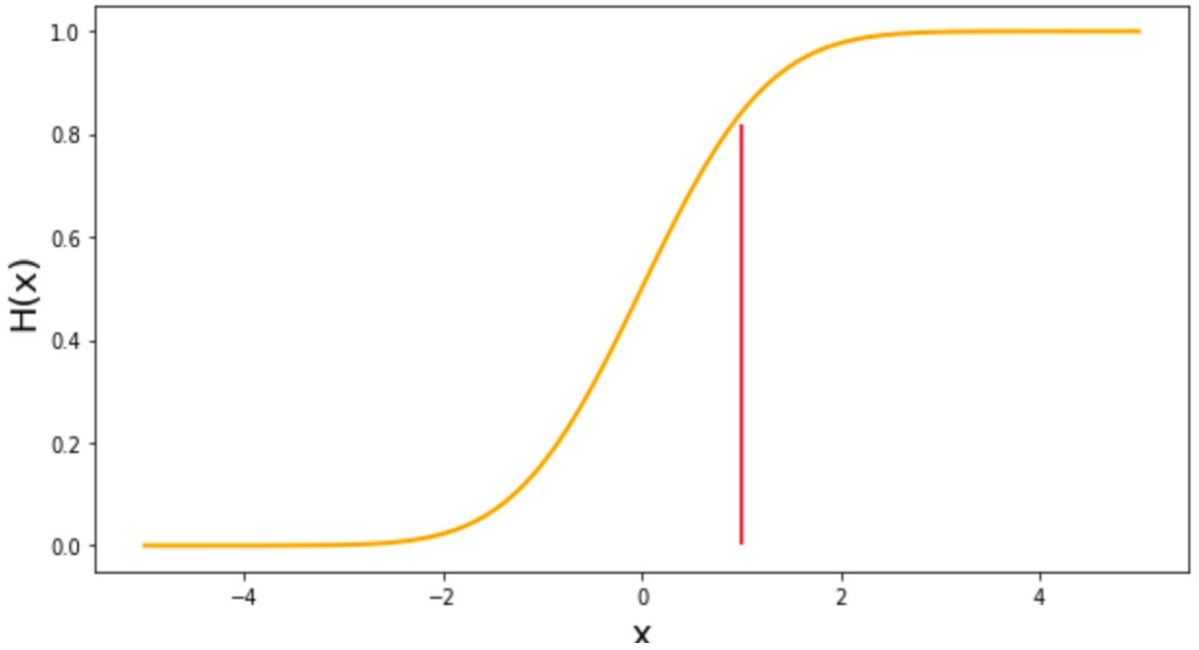


PDF In Python: go to Sect. I.4 of the notebook

Cumulative distribution 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 supplementary exercise of Sect. I.6.2 of the notebook
`Basic_statistics_02.ipynb`