DATA ANALYSIS
Year 2019–2020

# #1 Statistical Methods

Master M1–Marine Physics
Université de Bretagne Occidentale

Jonathan GULA
[gula@univ-brest.fr]

# Statistical Analysis

Jonathan GULA
gula@univ-brest.fr

## *Objectif*

Basic knowledge of statistical data analysis methods and application to geophysics data

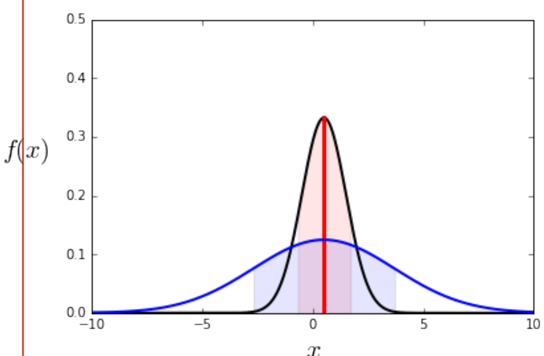
## *Plan*

- Basic notions: random variables, PDF, CDF, distributions, moments, estimators, the central limit theorem.
- Resampling methods (bootstrap, jackknife, Monte Carlo method), construction of confidence intervals, hypothesis testing, Bayesian statistics
- Extreme value theory , generalized Pareto distributions

# Statistical Analysis

Jonathan GULA
gula@univ-brest.fr

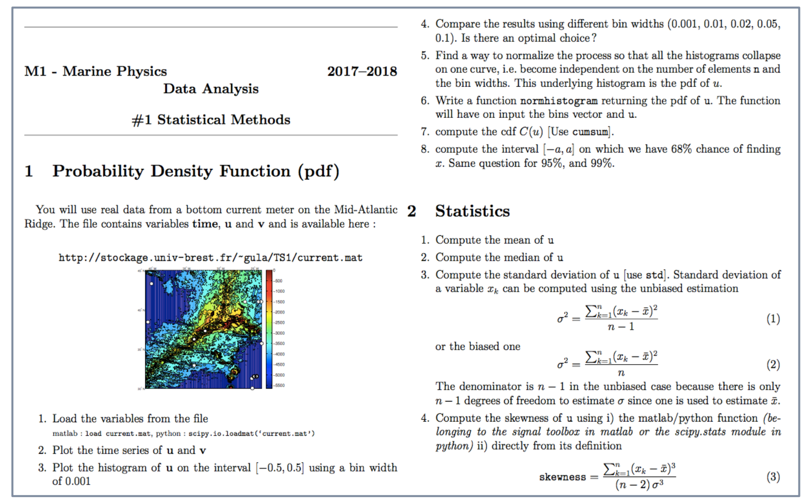
## Lessons (basic theory)




$$\sigma^2 = \int (x-\mu)^2 f(x)\,dx$$

## Activities

- Applications to geophysics data (mooring, argo, model, etc.) using **Python**



**M1 - Marine Physics** **2017–2018**

**Data Analysis**

**#1 Statistical Methods**

### 1 Probability Density Function (pdf)

You will use real data from a bottom current meter on the Mid-Atlantic Ridge. The file contains variables **time**, **u** and **v** and is available here :

`http://stockage.univ-brest.fr/~gula/TS1/current.mat`

1. Load the variables from the file
   matlab : `load current.mat`, python : `scipy.io.loadmat('current.mat')`
2. Plot the time series of **u** and **v**
3. Plot the histogram of **u** on the interval $[-0.5, 0.5]$ using a bin width of 0.001
4. Compare the results using different bin widths (0.001, 0.01, 0.02, 0.05, 0.1). Is there an optimal choice ?
5. Find a way to normalize the process so that all the histograms collapse on one curve, i.e. become independent on the number of elements `n` and the bin widths. This underlying histogram is the pdf of $u$.
6. Write a function `normhistogram` returning the pdf of `u`. The function will have on input the bins vector and `u`.
7. compute the cdf $C(u)$ [Use `cumsum`].
8. compute the interval $[-a, a]$ on which we have 68% chance of finding $x$. Same question for 95%, and 99%.

### 2 Statistics

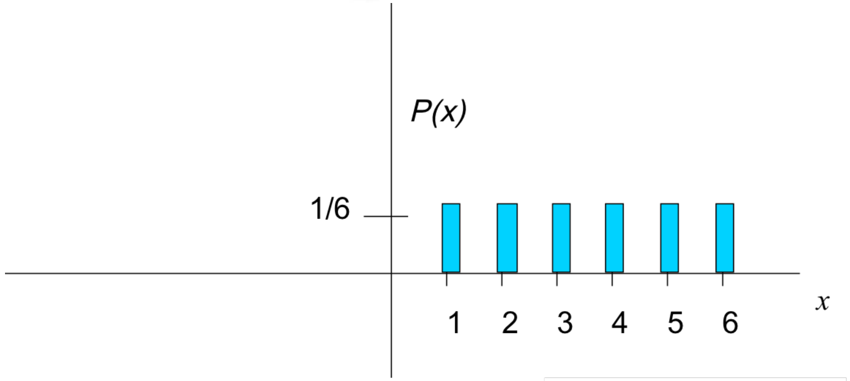
1. Compute the mean of `u`
2. Compute the median of `u`
3. Compute the standard deviation of `u` [use `std`]. Standard deviation of a variable $x_k$ can be computed using the unbiased estimation
$$\sigma^2 = \frac{\sum_{k=1}^{n}(x_k-\bar{x})^2}{n-1} \quad (1)$$
or the biased one
$$\sigma^2 = \frac{\sum_{k=1}^{n}(x_k-\bar{x})^2}{n} \quad (2)$$
The denominator is $n-1$ in the unbiased case because there is only $n-1$ degrees of freedom to estimate $\sigma$ since one is used to estimate $\bar{x}$.
4. Compute the skewness of `u` using i) the matlab/python function *(belonging to the signal toolbox in matlab or the scipy.stats module in python)* ii) directly from its definition
$$\texttt{skewness} = \frac{\sum_{k=1}^{n}(x_k-\bar{x})^3}{(n-2)\,\sigma^3} \quad (3)$$

## Grade

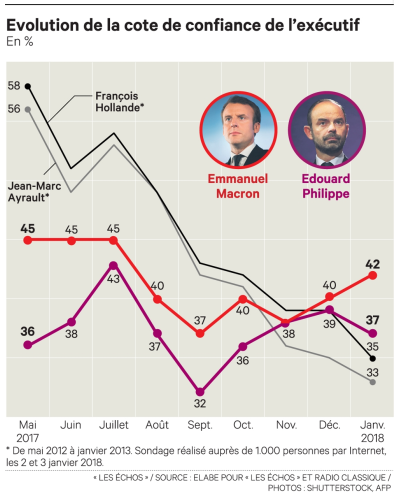
1 homework assignment + 1 computer exam

# Random variable

- A **random variable**, aleatory variable or stochastic variable is a variable whose value is subject to variations due to chance.
- A random variable can take on a **set of possible different values**, each **with an associated probability**.




- For example, if you roll a dice, the outcome is random (not fixed) and there are 6 possible outcomes, each of which occur with probability one-sixth.




- For example, if you poll people about their voting preferences, the percentage of the sample that responds “Yes” is a also a random variable (the percentage will be slightly different every time you poll).

# Random variable

- A random variable’s possible values might represent the possible outcomes of a yet-to-be-performed experiment, or the possible outcomes of a past experiment whose already-existing value is uncertain (for example, due to imprecise measurements or quantum uncertainty).
- They may also conceptually represent either the results of an ”objectively” random process (such as rolling a die) or the ”subjective” randomness that results from incomplete knowledge of a quantity (i.e. quantity of precipitation per day at a given location). The mathematics works the same regardless of the particular interpretation in use.

# Random variable

The mathematical function describing the possible values of a random variable and their associated probabilities is known as a **probability distribution.**

Random variables can be :

- **discrete**, that is, taking any of a specified finite or countable list of values, endowed with a probability mass function, characteristic of a probability distribution ;
- **continuous**, taking any numerical value in an interval or collection of intervals, via a probability density function that is characteristic of a probability distribution.

# Random variable

**Discrete** random variables have a countable number of outcomes.

Examples :

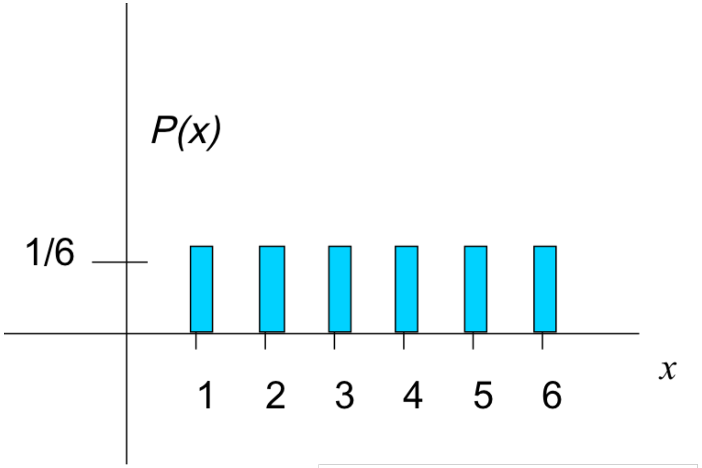
- dice
- yes/no
- Dead/alive
- treatment/placebo
- counts
- etc.

# Probability mass function

The probability mass function $f(x)$ or $p(x)$ for a discrete random variable is the function giving the probability to draw the value $x$

Example of a dice :

| *x* | *p(x)* |
|---|---|
| 1 | *p(x=1)*=1/6 |
| 2 | *p(x=2)*=1/6 |
| 3 | *p(x=3)*=1/6 |
| 4 | *p(x=4)*=1/6 |
| 5 | *p(x=5)*=1/6 |
| 6 | *p(x=6)*=1/6 |




We have necessarily: $$\sum f(x) = 1$$

# Random variable

**Continuous** random variables have an infinite continuum of possible values.

Examples :

- distribution of height among human beings
- distribution of incomes in a society
- distribution of wealth among human beings

- amount of precipitation per day at a given location
- percentage of clouds in the sky at a given location
- intensity of the quakes in a given region
- point-wise velocity in a turbulent flow
- instantaneous dissipation of energy in a turbulent flow
- wave height during a storm
- etc.

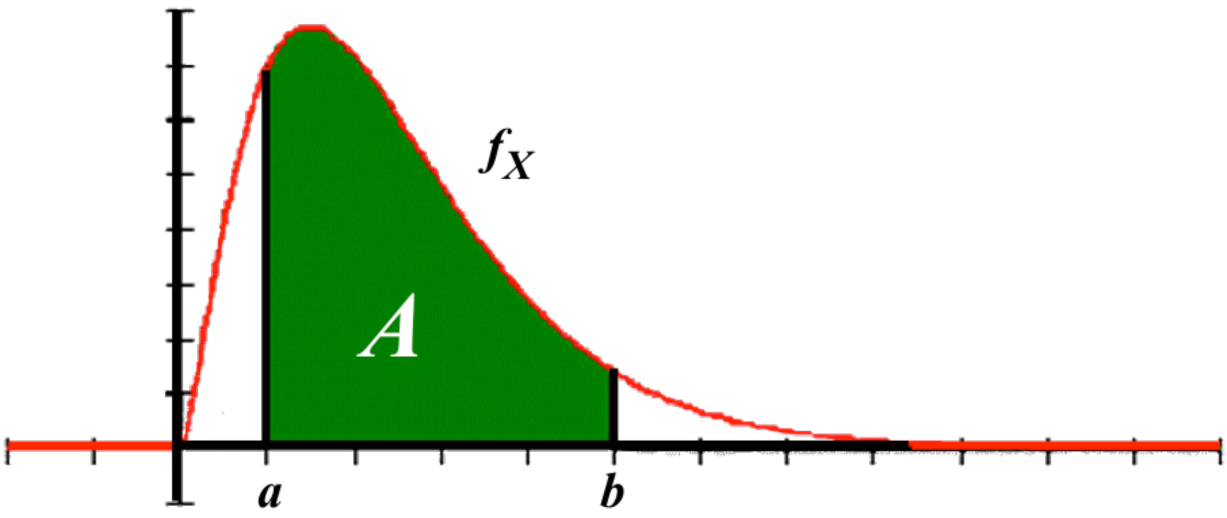
# Probability density function

The pdf $f(x)$ is a probability ***density***.

The probability to draw a value in the interval $[x, x + dx]$ is $f(x)dx$

The probability to get a value in the interval $[a, b]$ is $\Pr(a \leq x \leq b)$

the area under the graph of $f(x)$ over the interval $[a, b]$



# Probability density function

Example:

*Survival times after lung transplant may roughly follow an exponential function:*

$$p(x) = e^{-x}$$

*What is the probability that a patient will die in the second year after surgery (between years 1 and 2) ? :*

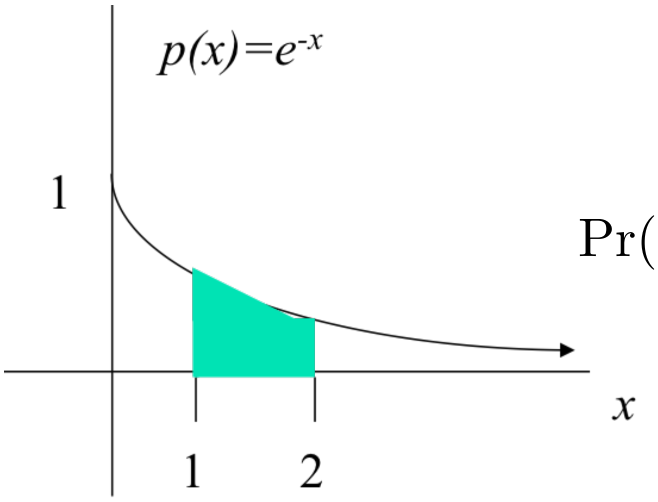
# Probability density function

Example:

*Survival times after lung transplant may roughly follow an exponential function:*

$$p(x) = e^{-x}$$

*What is the probability that a patient will die in the second year after surgery (between years 1 and 2) ? :*




$$\begin{aligned} \Pr(1 \leq x \leq 2) &= \int_1^2 f(x)dx = \int_1^2 e^{-x}dx \\ &= -e^{-2} - +e^{-1} = -0.135 + 0.368 = 0.23 \end{aligned}$$

# Probability density function

The probability that x is any exact particular value is 0.

$$\Pr(x = a) = 0$$

we can only assign probabilities to possible ranges of x.

*Why : the area of the pdf for a single value is zero because the width of the interval is zero ! This DOES NOT imply that x cannot take the value a, it simply means that the probability that it takes this value exactly is infinitely small.*

**Example:** *Suppose a species of bacteria typically lives 4 to 6 hours. What is the probability that a bacterium lives exactly 5 hours? The answer is actually 0%. A lot of bacteria live for approximately 5 hours, but there is a negligible chance that any given bacterium dies at exactly 5.0000000000000000000000000000000000000000... hours.*

# Probability density function

The probability density is always positive: $f(x) \geq 0$

And by definition: $$\int f(x) dx = 1$$

summed over the full interval on which it is defined.

The probability $f(x)dx$ is dimensionless, such that $f(x)$ has the dimension of the inverse of $x$

# Cumulated density function

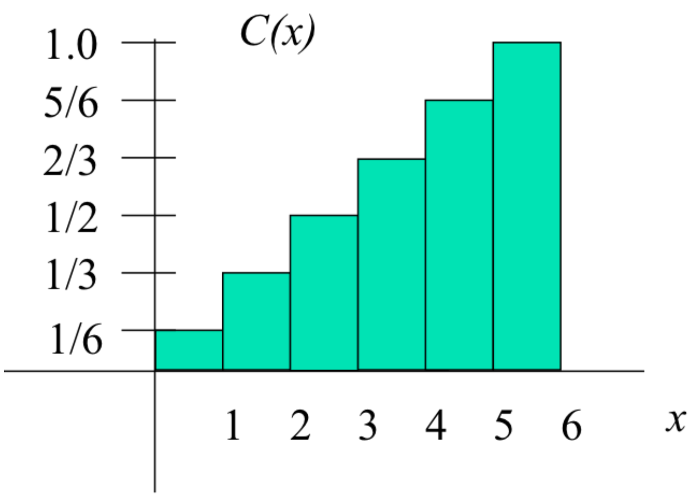
It is the primitive of $f(x)$

$$C(x) = \int_{-\infty}^{x} f(z)\, dz$$

By construction it is an increasing function from 0 to 1.

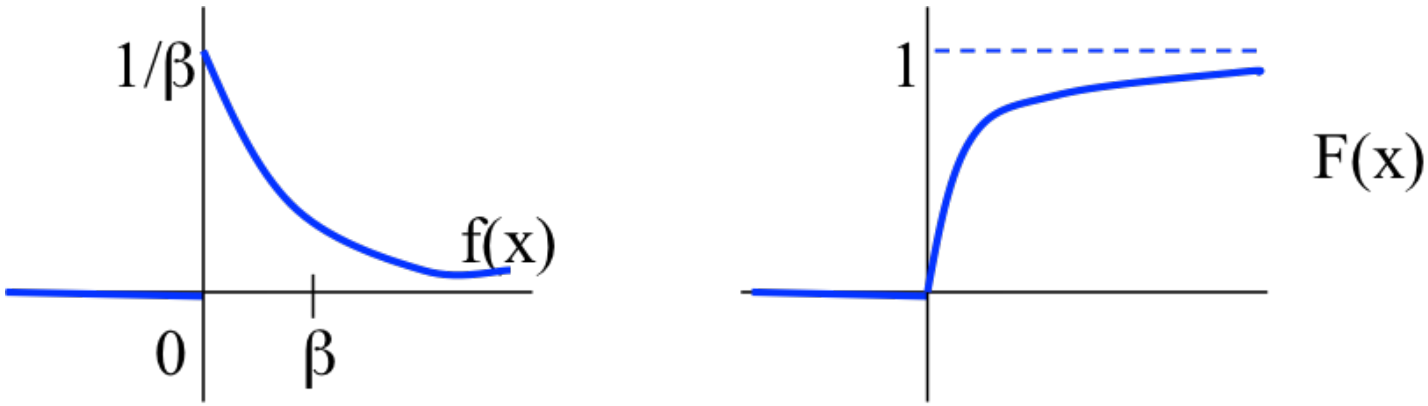
Example of a dice :

| *x* | *P(x≤A)* |
|---|---|
| 1 | *P(x≤1)*=1/6 |
| 2 | *P(x≤2)*=2/6 |
| 3 | *P(x≤3)*=3/6 |
| 4 | *P(x≤4)*=4/6 |
| 5 | *P(x≤5)*=5/6 |
| 6 | *P(x≤6)*=6/6 |



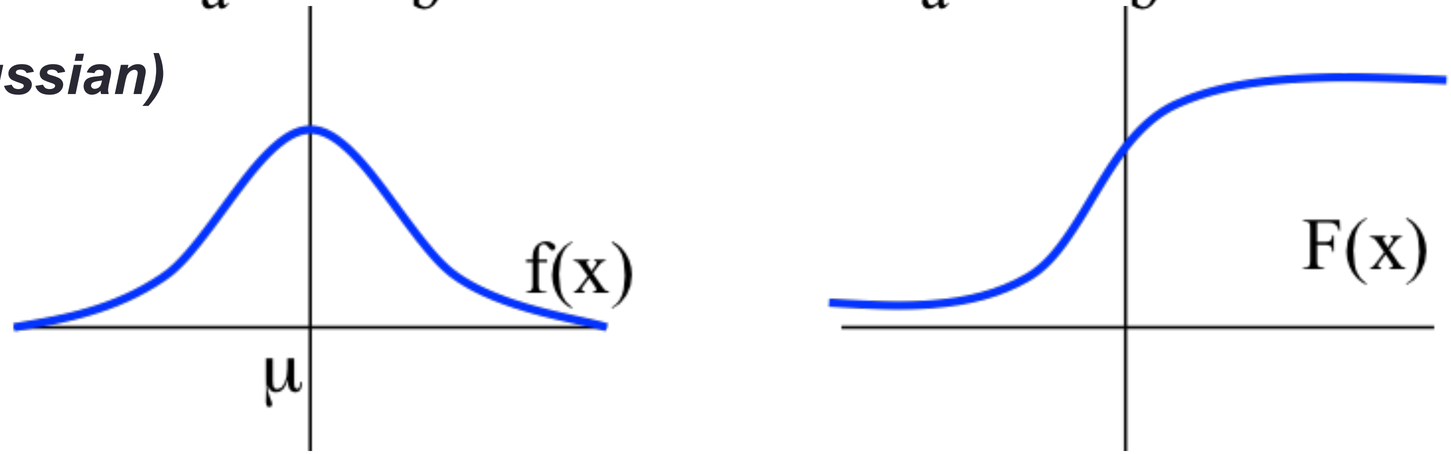
# Classical examples of Probability density function

- ***Exponential law***




- ***Uniform distribution***


- ***Normal law (aka Gaussian)***




***poisson law, power-law, log-normal law, chi-squared distribution, Pareto distribution, Cauchy distribution, etc.***

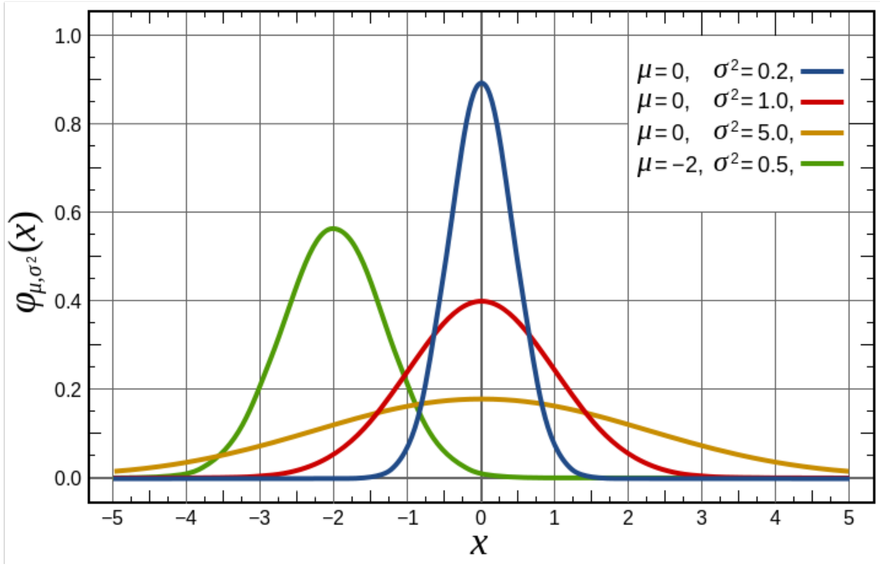
# Classical examples of Probability density function

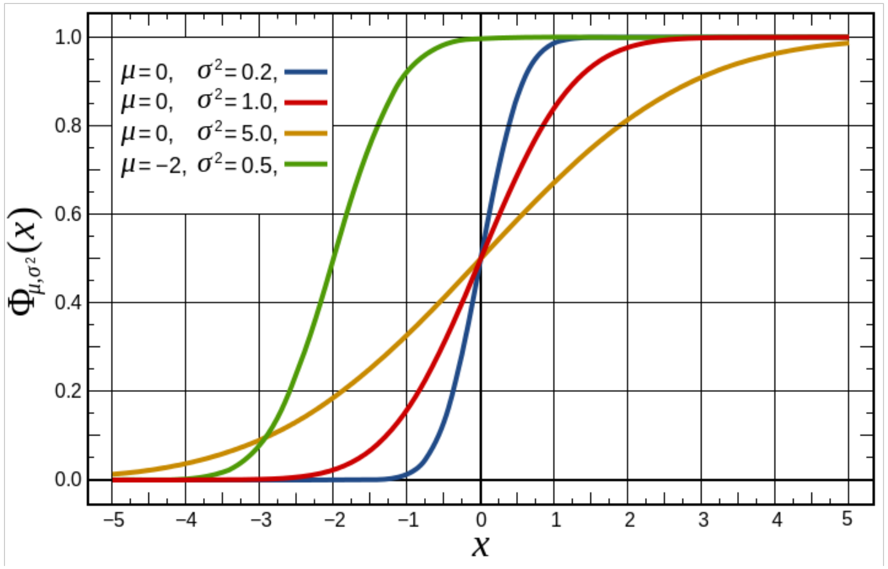
- The normalized centered normal distribution reads

$$\mathcal{N}(0,1) \sim f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

- And for a given mean and standard deviation $(\mu, \sigma)$

$$\mathcal{N}(\mu,\sigma) \sim \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$



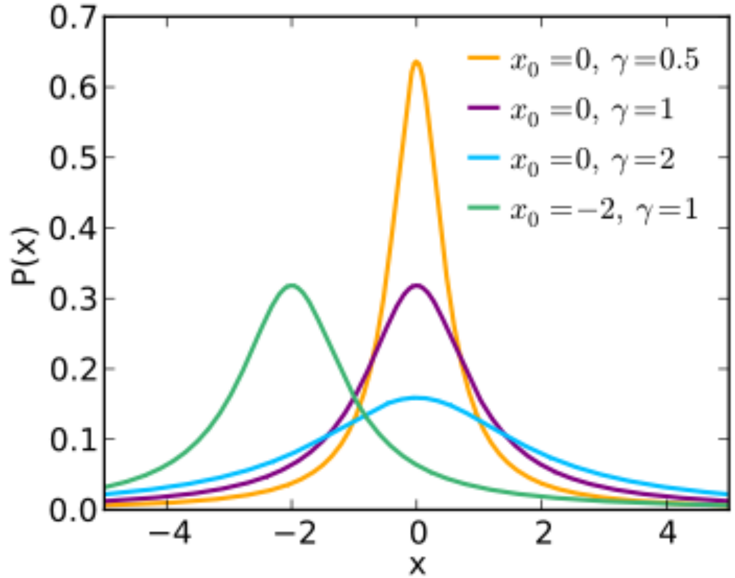



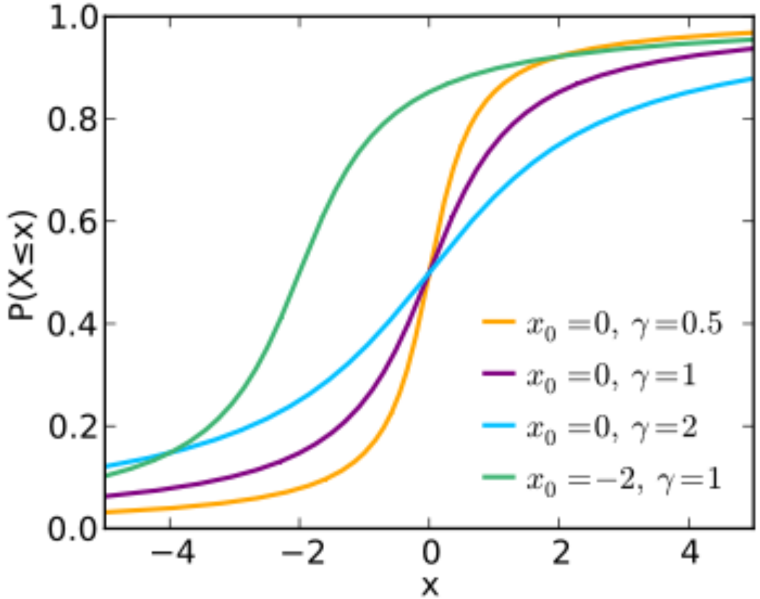
# Classical examples of Probability density function

- The Cauchy distribution

$$f(x) = \frac{1}{\pi(1+x^2)}$$

$$C(x) = \frac{1}{\pi} arctan(x) + \frac{1}{2}$$



# Estimating the PDF

- In practice, if X is a random variable, we will deal with a finite number N of empirical realizations of the random variables :

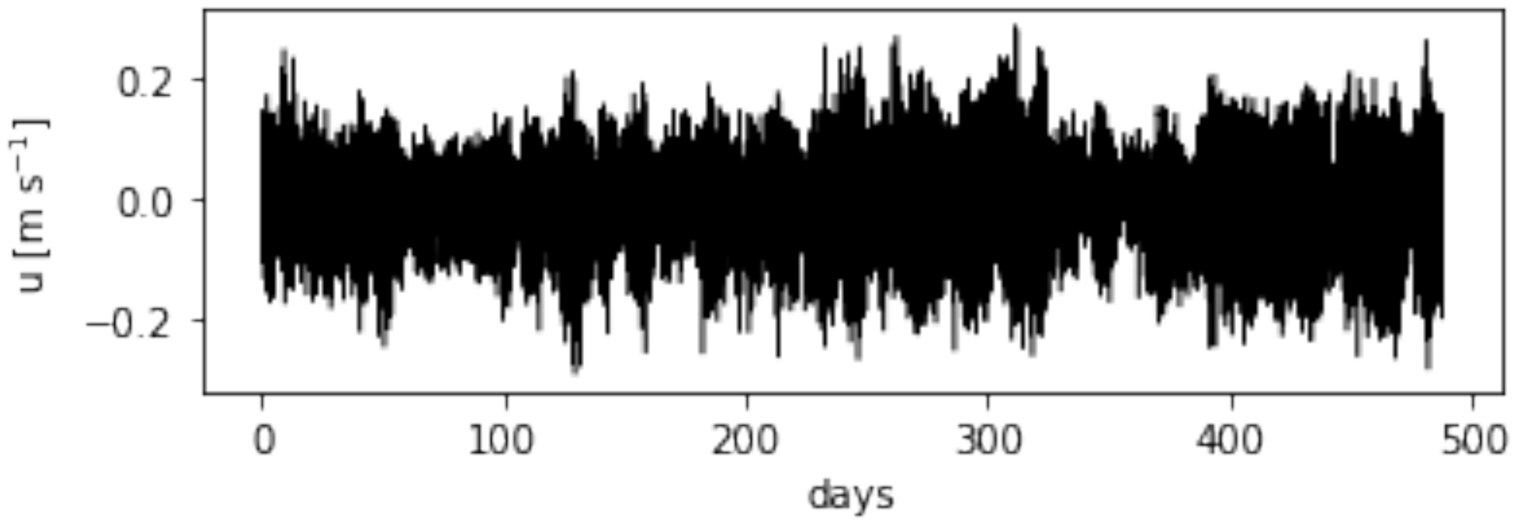
$$x_k,\ k = 1 \ldots N$$

- We have access to the properties of X only via the empirical N samples. Intuitively we see that the larger N the better we know X.

- In practice we never know the pdf but we can estimate it, by computing an histogram of the. $x_k$

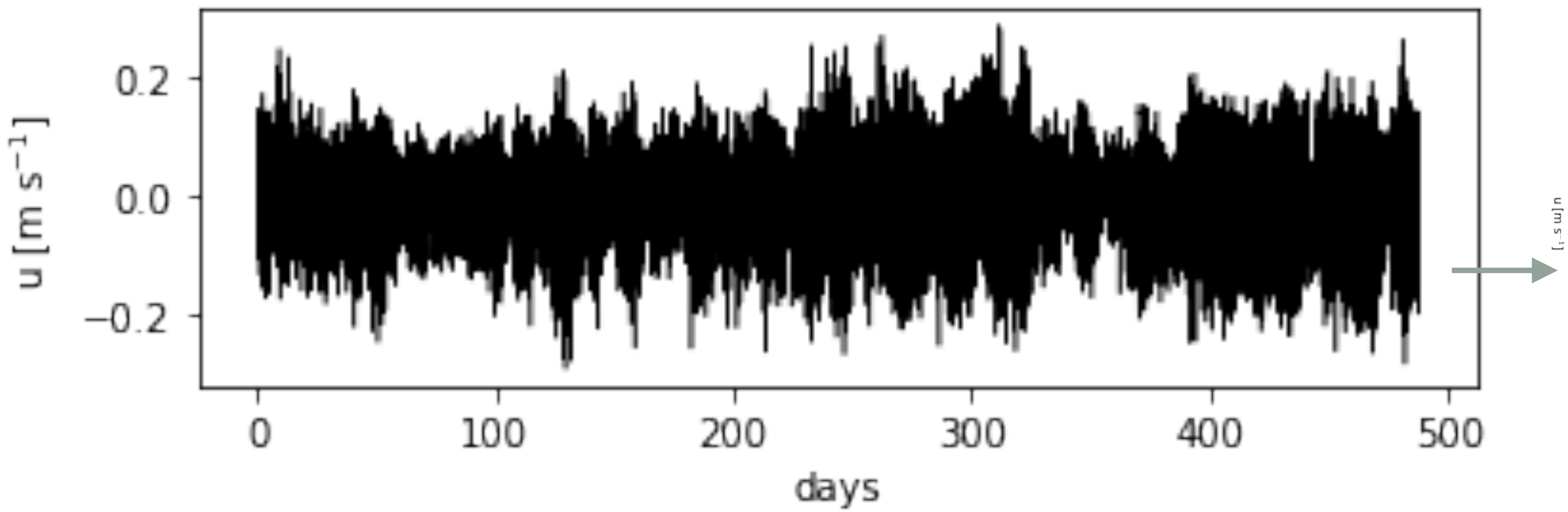
# Sample distributions

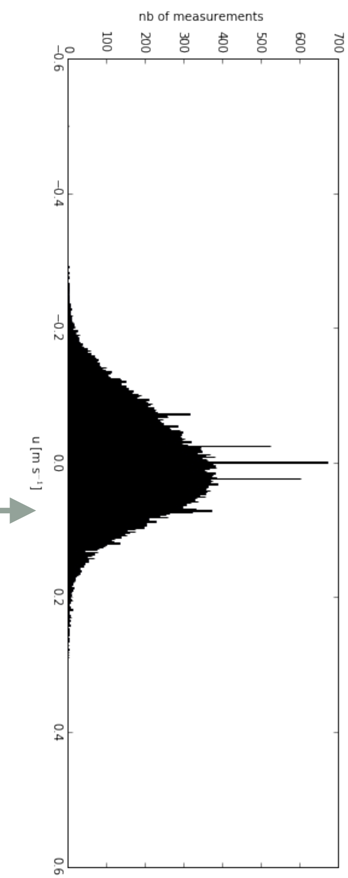
Example:



# Sample distributions

The most basic estimate of a population distribution can be made by using the histogram of the measured data points.



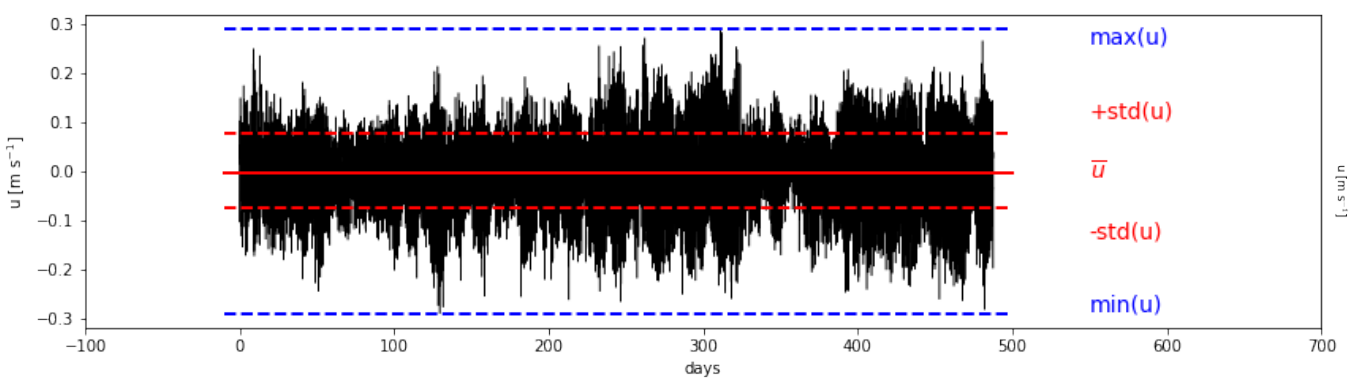
# Sample distributions

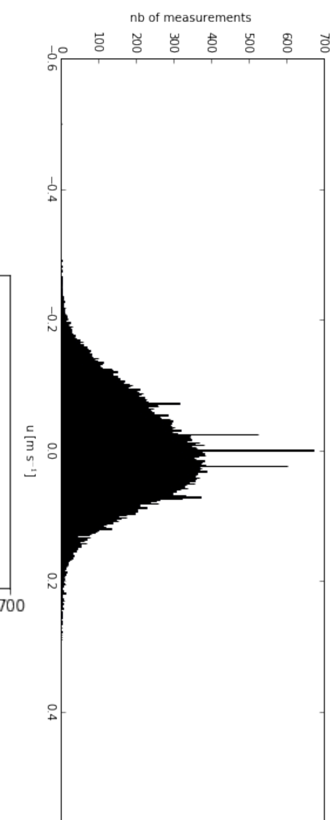
The most basic estimate of a population distribution can be made by using the histogram of the measured data points.

The most basic descriptive parameter is the sample mean: $\hat{\mu} = \frac{1}{N}\sum_k x_k$

To determine how the data are spread about the mean, we can compute the standard deviation:

$$s = \sqrt{\frac{1}{N-1}\sum_k (x_k - \overline{\mu})^2}$$




In papers you'll see: velocity is $u = -0.0025 \pm 0.0756\,\mathrm{m\,s^{-1}}$

# Estimating the PDF

- See TD1 – Statistics (#1)

# Sample distributions

The most basic estimate of a population distribution can be made by using the histogram of the measured data points.

The most basic descriptive parameter is the sample mean:

$$\overline{\mu} = \sqrt{\frac{1}{N}\sum_k x_k}$$

To determine how the data are spread about the mean, we can compute the variance:

$$s^2 = \frac{1}{N-1}\sum_k (x_k - \overline{\mu})^2$$

or standard deviation:

$$s = \sqrt{\frac{1}{N-1}\sum_k (x_k - \overline{\mu})^2}$$

# Sample distributions

The most basic estimate of a population distribution can be made by using the histogram of the measured data points.

The most basic descriptive parameter is the sample mean: $\hat{\mu} = \frac{1}{N}\sum_k x_k$

To determine how the data are spread about the mean, we can compute the variance:

$$s^2 = \frac{1}{N-1}\sum_k (x_k - \hat{\mu})^2$$