DATA ANALYSIS
Year 2019–2020

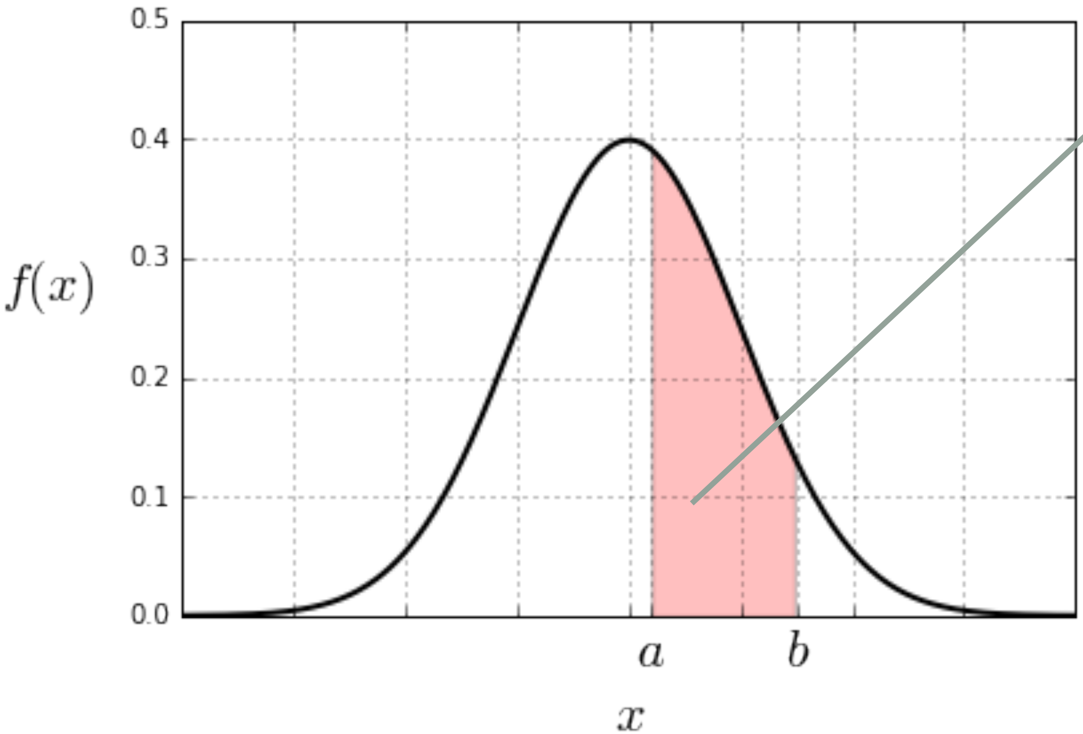
#2 Statistical Methods

Master M1–Marine Physics
Université de Bretagne Occidentale

Jonathan GULA
[gula@univ-brest.fr]

Probability density function (PDF)

Function of a continuous random variable, whose integral across an interval gives the probability that the value of the variable lies within the same interval: $f(x)$



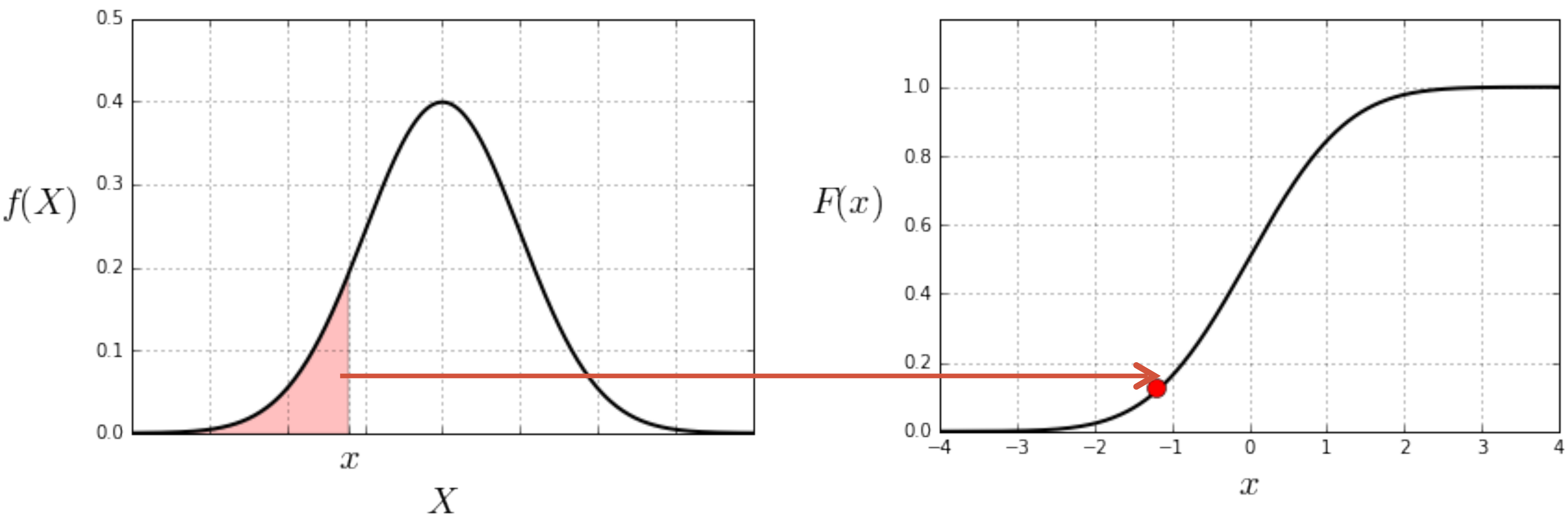
$$\mathrm{P}(a \leq x \leq b) = \int_a^b f(x)dx$$

With properties $f(x) \geq 0$ and $\int_{-\infty}^{+\infty} f(x)dx = 1$

Cumulative distribution function (CDF)

It is the primitive of $f(x)$, i.e. the probability that the value of a variable is less than x

$$F(x) = \int_{-\infty}^{x} f(X)\, dX = \mathrm{P}(-\infty \leq X \leq x)$$



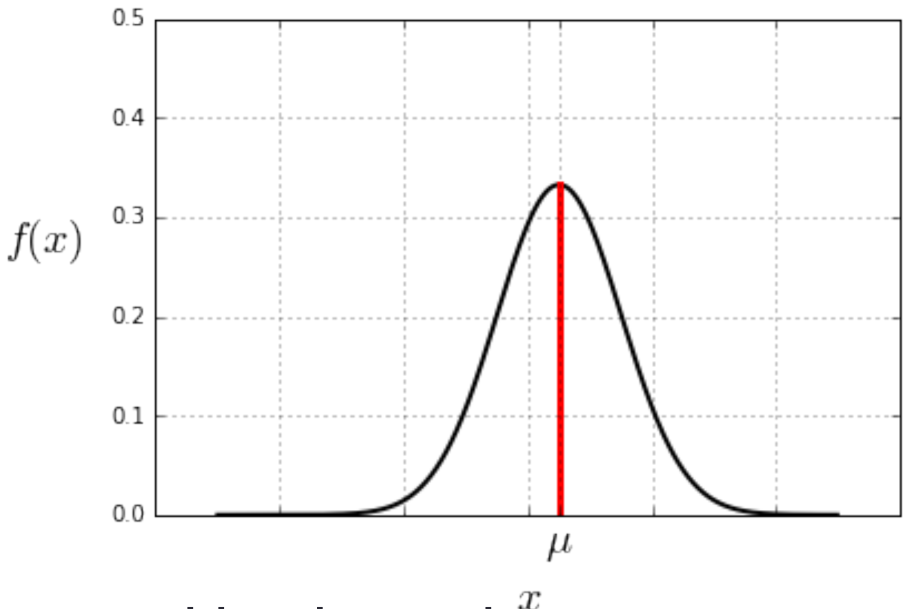
Moments

How to characterize the structure of the observations?

We compute parameters called **moments.**

The first one is the **mean:**

$$\mu = \int x f(x) dx$$



In practice it is estimated by the arithmetic sum:

$$\hat{\mu} = \frac{1}{N} \sum_k x_k$$

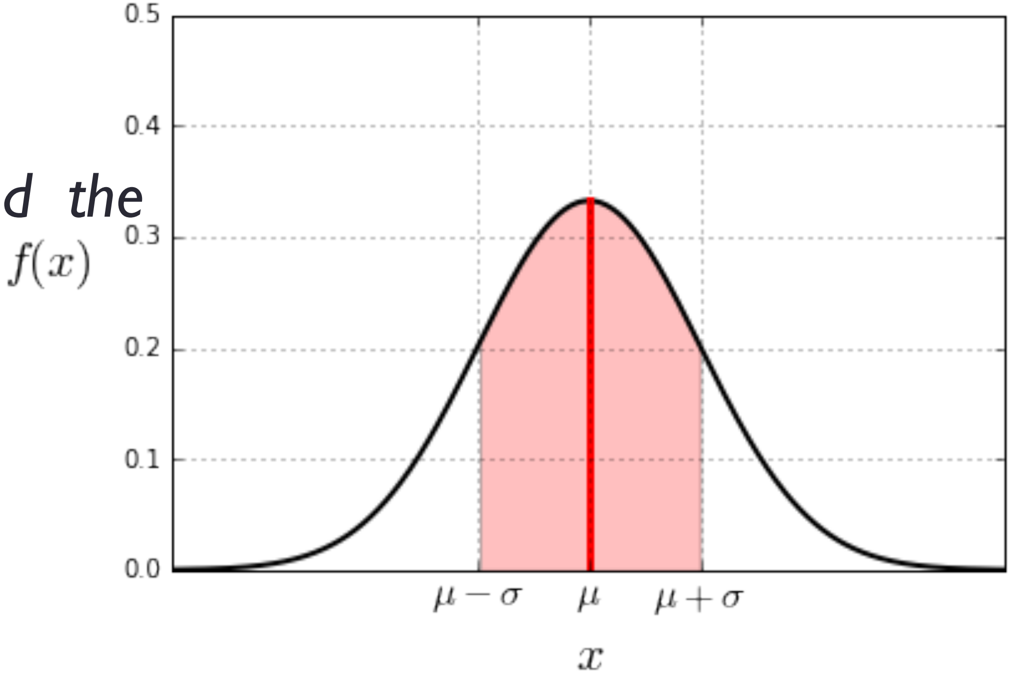
Moments

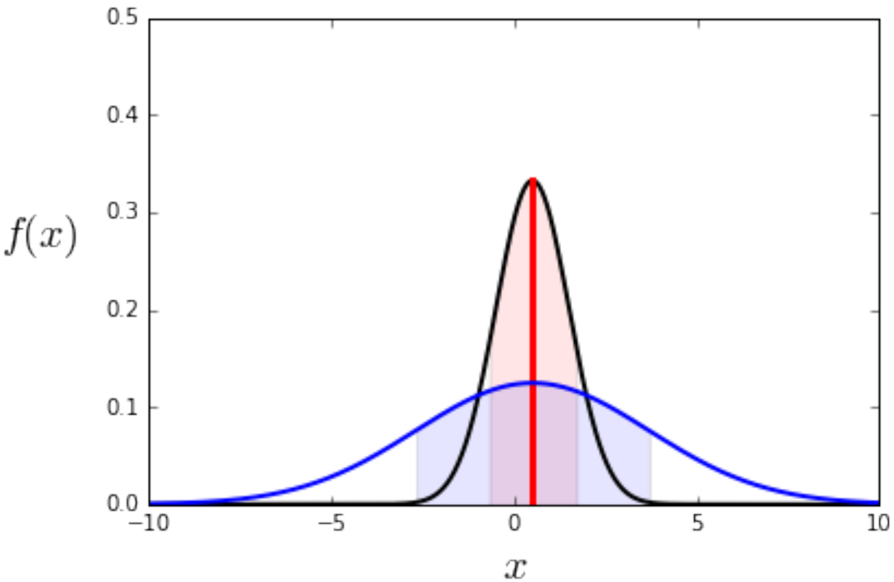
The second one is the **variance:**

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

Where σ is the **standard deviation** (also called rms).

It describes the *spread of the pdf* *around the mean.*





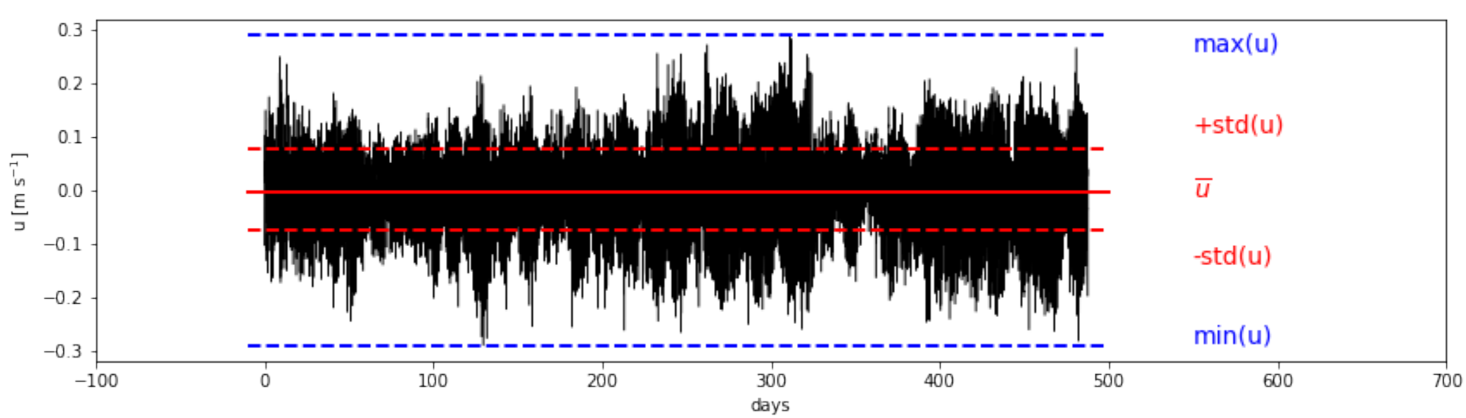
Moments

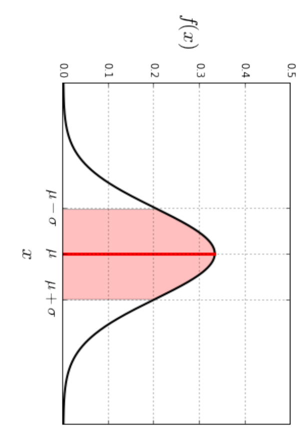
The second one is the **variance:**

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

Where σ is the **standard deviation** (also called rms).

It describes the *spread of the pdf* *around the mean.*





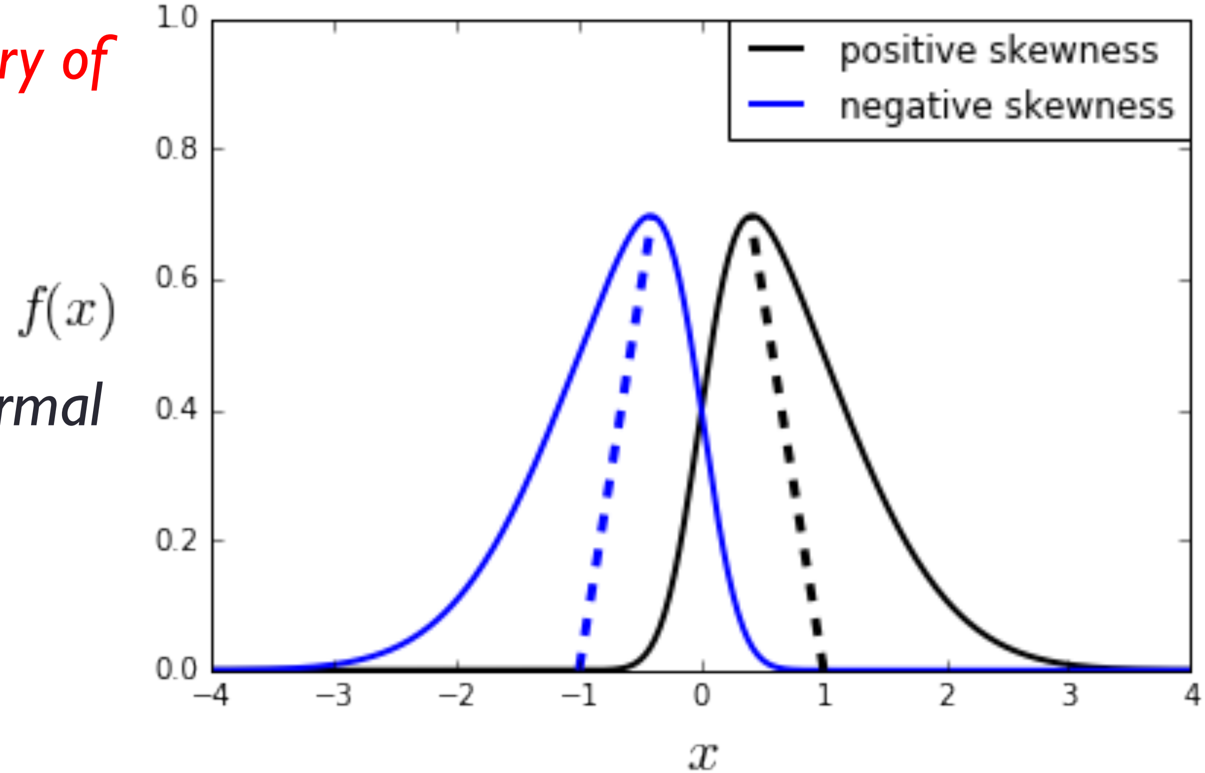
Moments

The third one is the **skewness:**

$$\mu_3 = \frac{1}{\sigma^3} \int (x-\mu)^3 f(x)\, dx$$

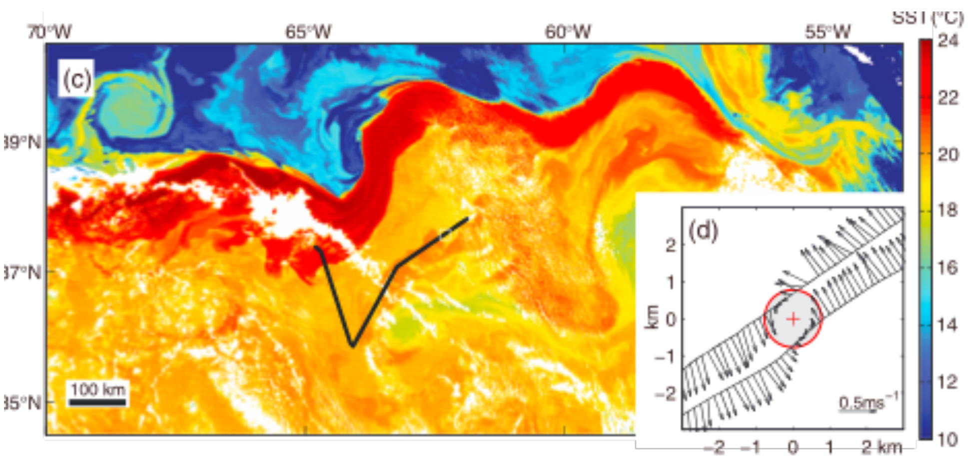
It is a measure of the *asymmetry of the pdf* *about its mean.*

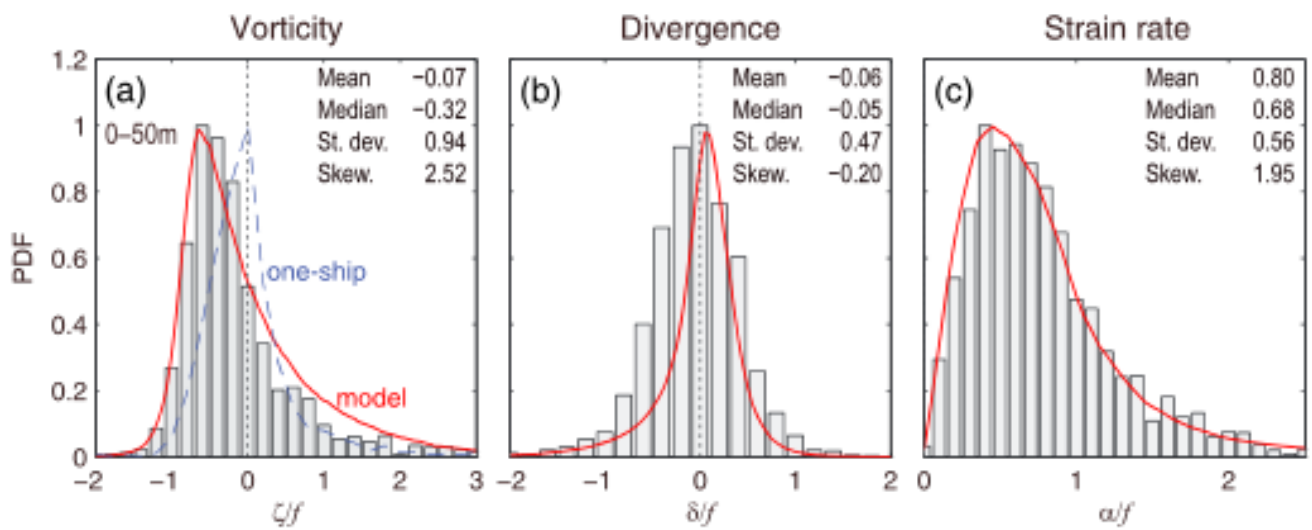
If the pdf is symmetric (e.g. normal distrib.), the skewness is 0



Moments

The third one is the **skewness:** $$\mu_3 = \frac{1}{\sigma^3}\int (x-\mu)^3 f(x)\,dx$$





Moments

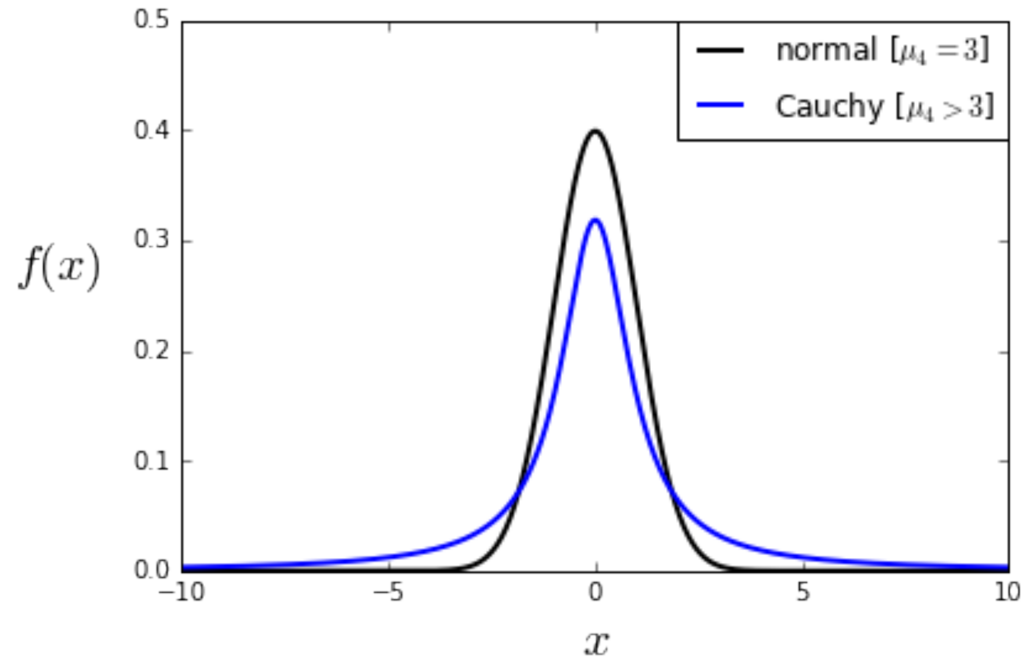
The fourth one is the **kurtosis:**

$$\mu_4 = \frac{1}{\sigma^4} \int (x-\mu)^4 f(x)\, dx$$

The kurtosis measures how fat are the tails of the pdf.

For the normal law the skewness is zero and the kurtosis is 3.

Values larger than 3 indicate more likely extreme values than the normal law.

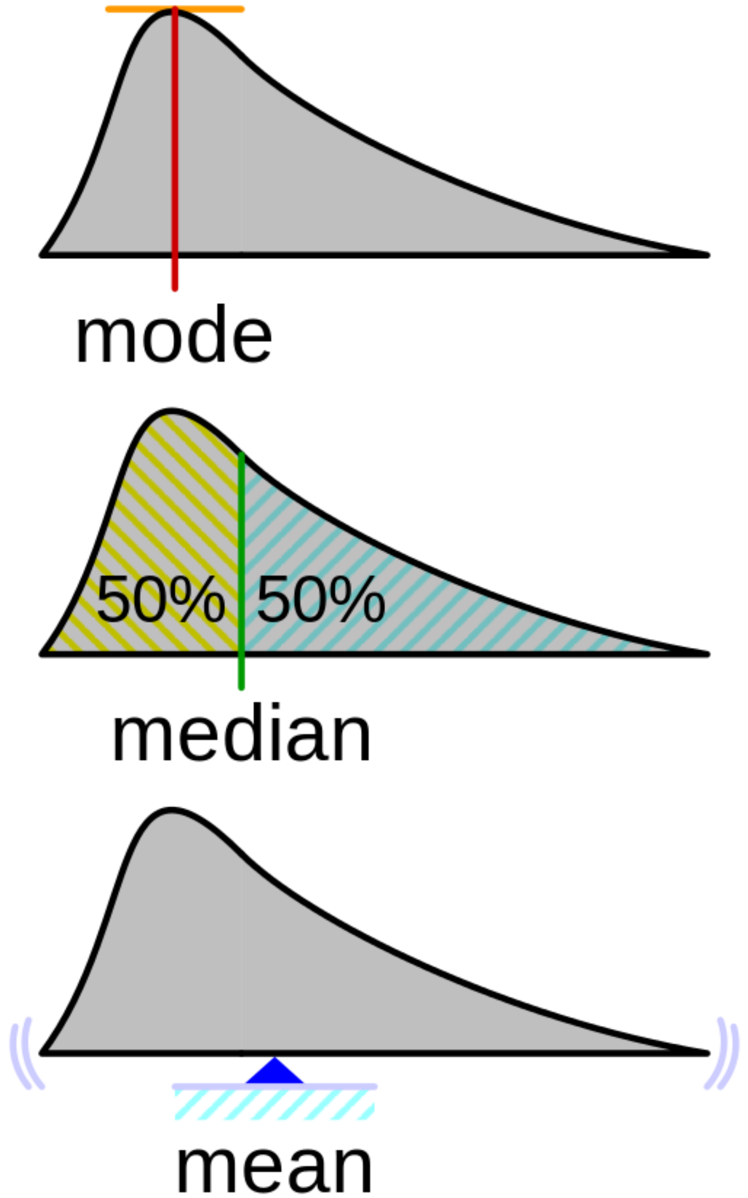


Moments

There is also the **mode** = the most likely result, i.e. the x such that f (x) is maximum.

Another way to characterize a pdf is its **median**, i.e. the x such that the cdf F(x) = 0.5.

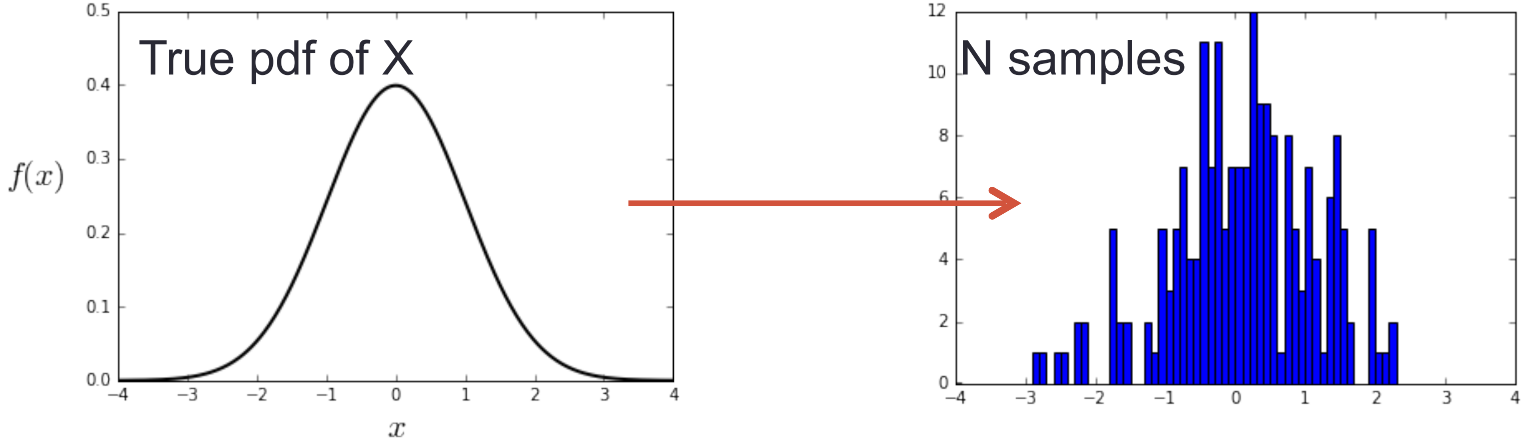
For the normal distribution the median, mode and mean are the same. This is not always the case.



Estimators

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

$$x_k \text{ for } k = 1..N$$



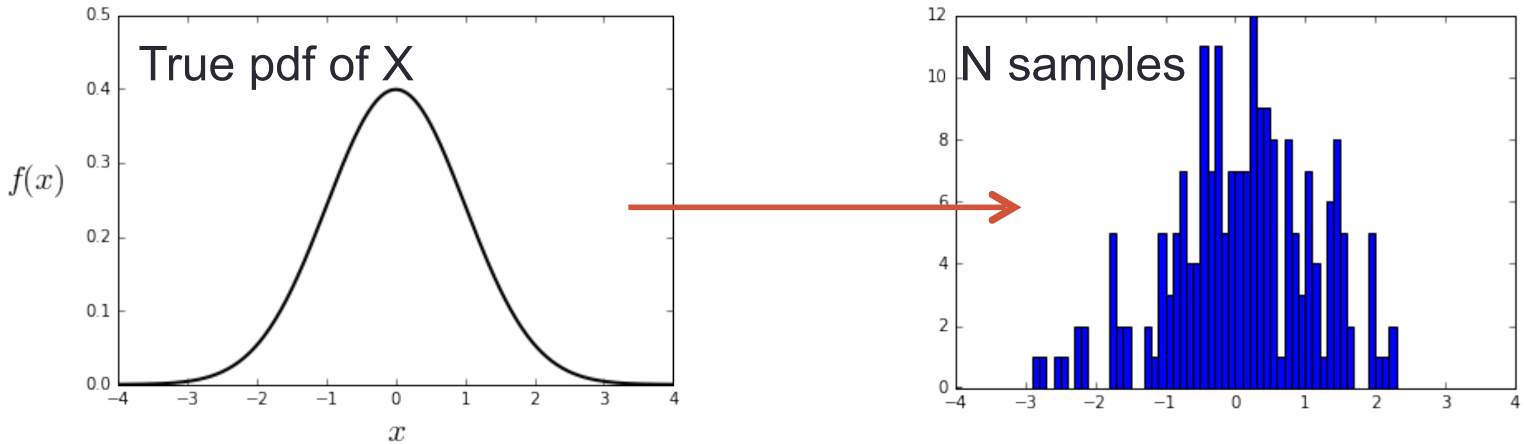
In practice we never know the true pdf but we can **estimate** it using the N samples.

We have access to the properties of X only via the empirical N samples.

Estimators

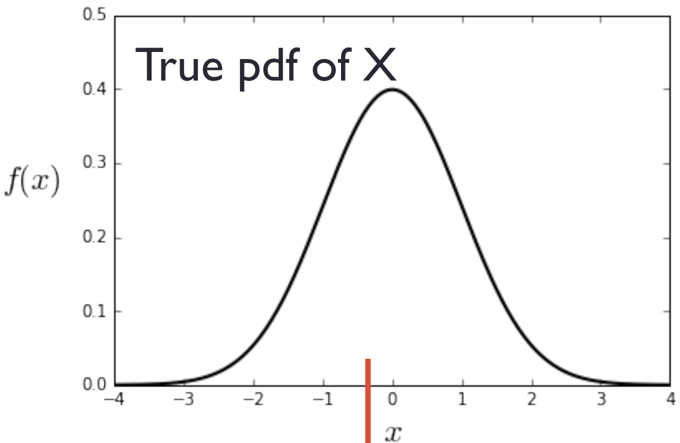
In practice, if X is a random variable, we will deal with a finite number N of empirical realizations of the random variables :

$$x_k \text{ for } k = 1..N$$



Estimator: if θ is a property of X (e.g. mean, variance, skewness, etc.), we denote $\hat{\theta}$ the estimate of θ. It is the rule used to estimate θ out of a given set of samples.

Estimators



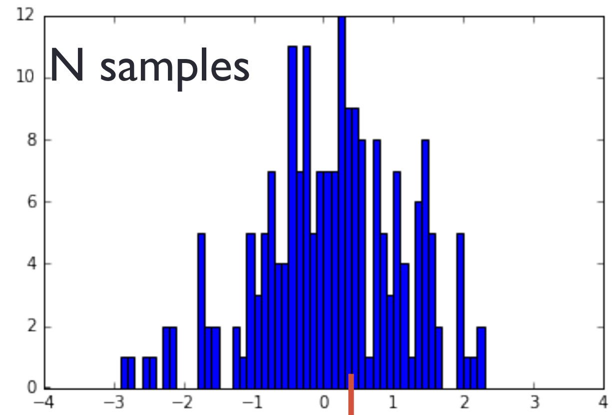
True population mean:

$$\mu = \int x f(x) dx$$

True population variance:

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

$$x_k \text{ for } k = 1..N$$



Mean estimator (= sample mean)

$$\hat{\mu}(x) = \frac{1}{N} \sum_k x_k$$

Variance estimator:

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

Normality tests

- **Normality test** = Check if your data sample deviates from a Gaussian distribution

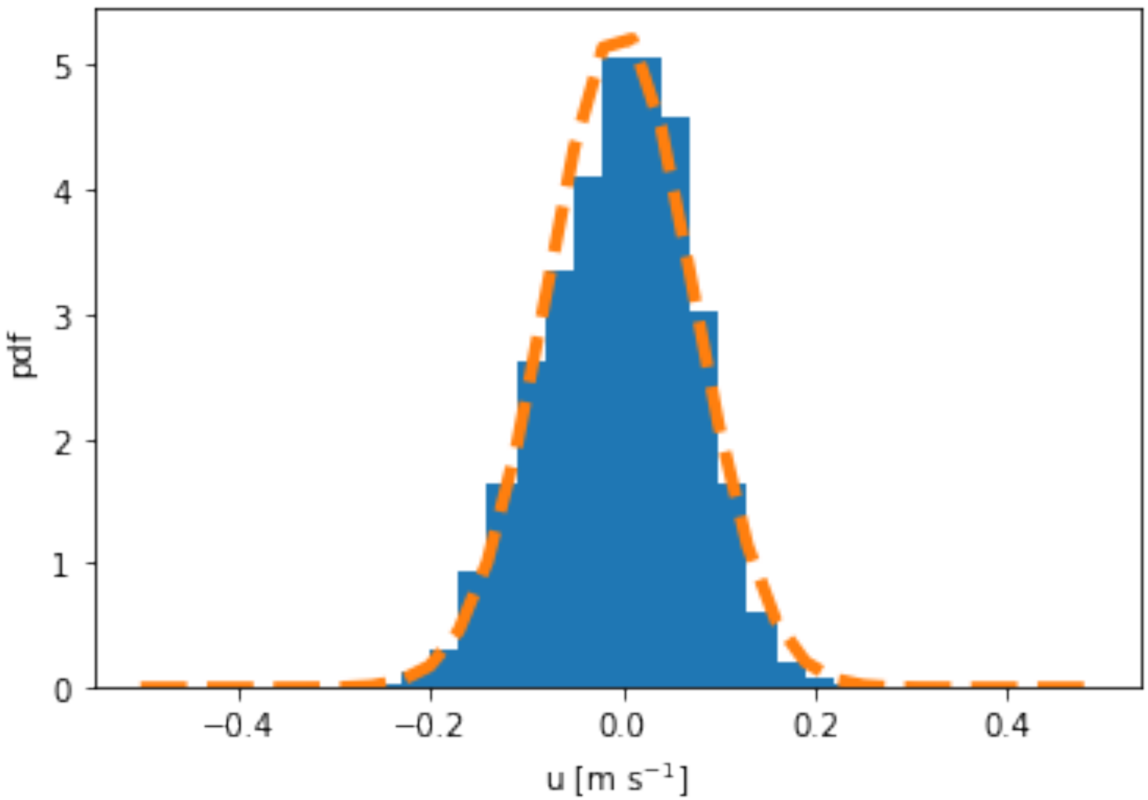
  - 1. **Graphical Methods**. These are methods for plotting the data and qualitatively evaluating whether the data looks Gaussian.

  - 2. **Statistical Tests**. These are methods that calculate statistics on the data and quantify how likely it is that the data was drawn from a Gaussian distribution

  - *https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/*

Normality tests

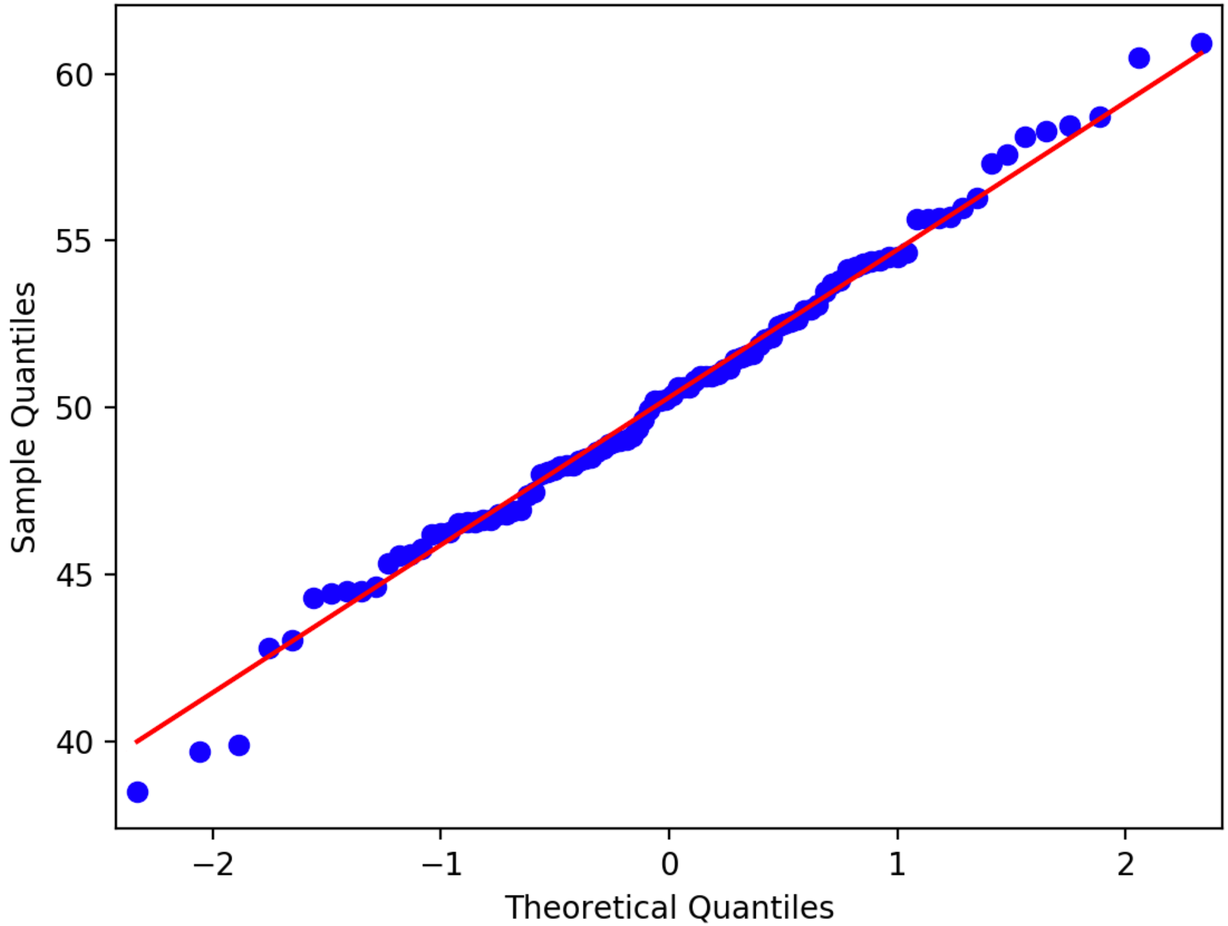
1. **Graphical Methods.** These are methods for plotting the data and qualitatively evaluating whether the data looks Gaussian.



Normality tests

1. **Graphical Methods**. These are methods for plotting the data and qualitatively evaluating whether the data looks Gaussian.

quantile-quantile plot:



Normality tests

2. **Statistical Tests**:

There are many statistical tests that we can use to quantify whether a sample of data looks as though it was drawn from a Gaussian distribution.

Each test makes different assumptions and considers different aspects of the data.

Each test will return at least two things:

- **Statistic**: A quantity calculated by the test that can be interpreted in the context of the test via comparing it to critical values from the distribution of the test statistic.
- **p-value**: Used to interpret the test, in this case whether the sample was drawn from a Gaussian distribution.

Normality tests

2. **Statistical Tests**:

The tests assume that that the sample was drawn from a Gaussian distribution. Technically this is called the null hypothesis, or H0. A threshold level is chosen called alpha, typically 5% (or 0.05), that is used to interpret the p-value.

In the SciPy implementation of these tests, you can interpret the p value as follows.

- **p <= alpha**: reject H0, not normal.
- **p > alpha**: fail to reject H0, normal.

This means that, in general, we are seeking results with a larger p-value to confirm that our sample was likely drawn from a Gaussian distribution.

A result above 5% does not mean that the null hypothesis is true. It means that it is very likely true given available evidence. The p-value is not the probability of the data fitting a Gaussian distribution; it can be thought of as a value that helps us interpret the statistical test.

Normality tests

A. The Shapiro-Wilk test

```
# Shapiro-Wilk Test
from numpy.random import seed
from numpy.random import randn
from scipy.stats import shapiro
# seed the random number generator
seed(1)
# generate univariate observations
data = 5 * randn(100) + 50
# normality test
stat, p = shapiro(data)
print('Statistics=%.3f, p=%.3f' % (stat, p))
# interpret
alpha = 0.05
if p > alpha:
     print('Sample looks Gaussian (fail to reject H0)')
else:
     print('Sample does not look Gaussian (reject H0)')
```

Normality tests

B. D’Agostino’s K^2 Test

It calculates summary statistics from the data, namely kurtosis and skewness, to determine if the data distribution departs from the normal distribution.

```
# D'Agostino and Pearson's Test
from numpy.random import seed
from numpy.random import randn
from scipy.stats import normaltest
# seed the random number generator
seed(1)
# generate univariate observations
data = 5 * randn(100) + 50
# normality test
stat, p = normaltest(data)
print('Statistics=%.3f, p=%.3f' % (stat, p))
# interpret
alpha = 0.05
if p > alpha:
	print('Sample looks Gaussian (fail to reject H0)')
else:
	print('Sample does not look Gaussian (reject H0)')
```

Normality tests

C. Anderson-Darling Test

It is a statistical test that can be used to evaluate whether a data sample comes from one of among many known data samples, it is a modified version of a more sophisticated nonparametric goodness-of-fit statistical test called the Kolmogorov-Smirnov test.

```
# Anderson-Darling Test
from numpy.random import seed
from numpy.random import randn
from scipy.stats import anderson
# seed the random number generator
seed(1)
# generate univariate observations
data = 5 * randn(100) + 50
# normality test
result = anderson(data)
print('Statistic: %.3f' % result.statistic)
p = 0
for i in range(len(result.critical_values)):
      sl, cv = result.significance_level[i], result.critical_values[i]
      if result.statistic < result.critical_values[i]:
            print('%.3f: %.3f, data looks normal (fail to reject H0)' % (sl, cv))
      else:
            print('%.3f: %.3f, data does not look normal (reject H0)' % (sl, cv))
```

Moments and estimators

- See TD1 – Statistics (#2)
- See TD1 – Normal distribution (#3)