Modelling, Uncertainty and Data for Engineers (MUDE)

Week 1.7 : Univariate continuous distributions

Patricia Mares Nasarre

TU Delft

Welcome to…

MUDE

Modelling, Uncertainty, and Data for Engineers

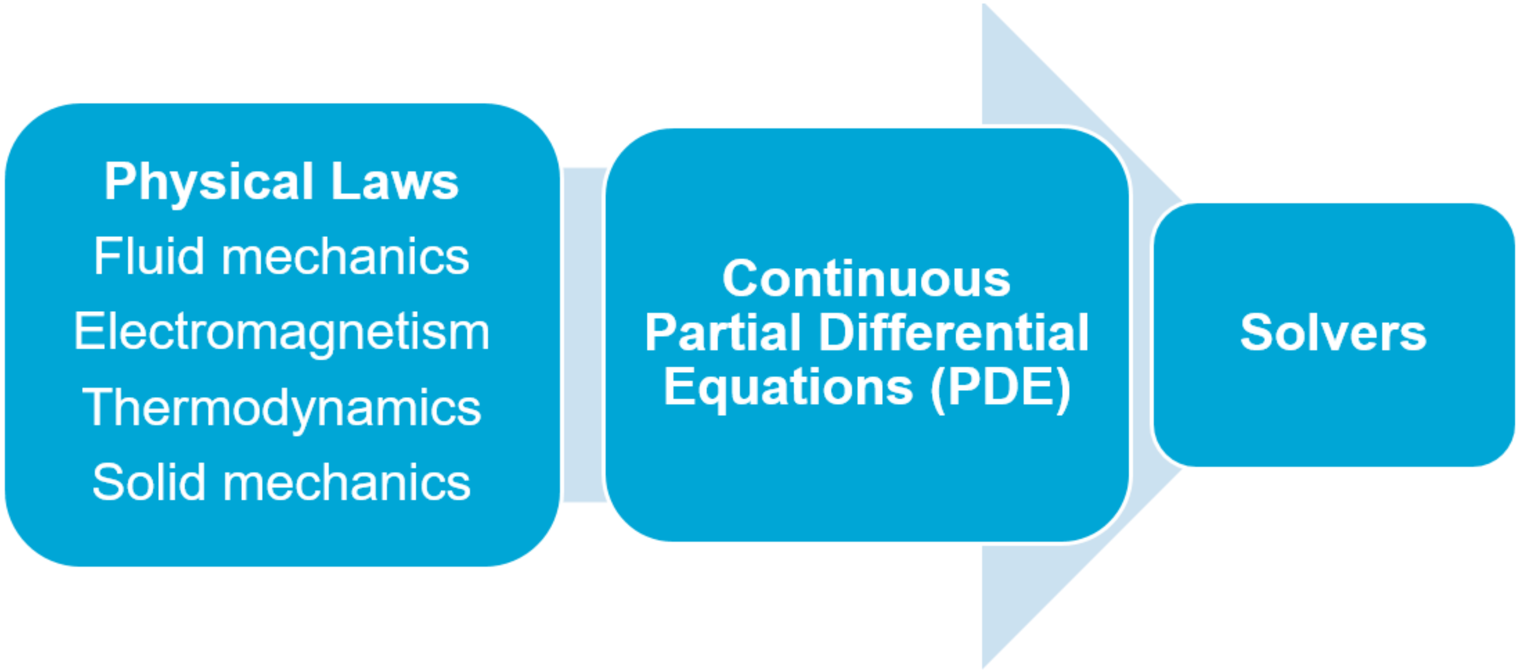
WEEK 7!

TUDelft

Recap of Weeks 5 and 6

Deterministic models over time and space

- If input is ‘a’, output will always be ’b’
- Numerical integration
- FDM



Recap of Week 1

Deterministic vs Stochastic

Deterministic models are those which for some given inputs, always provide the same output. For instance, a equation which gives the average concentration of CO_2 in a city as function of the traffic. For a certain value of traffic, the model will always provide the same concentration of CO_2. Therefore, these models that there is no uncertainty. On the contrary, stochastic models are those which embrace the uncertainty. This is stochastic models will produce different outputs for a given input. In fact, the inputs and outputs of stochastic models are probabilistic distributions (you will learn more about this later!), which relate the values of the variable with the probability of observing it.

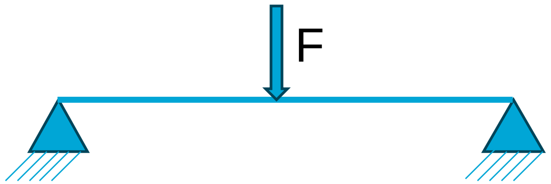
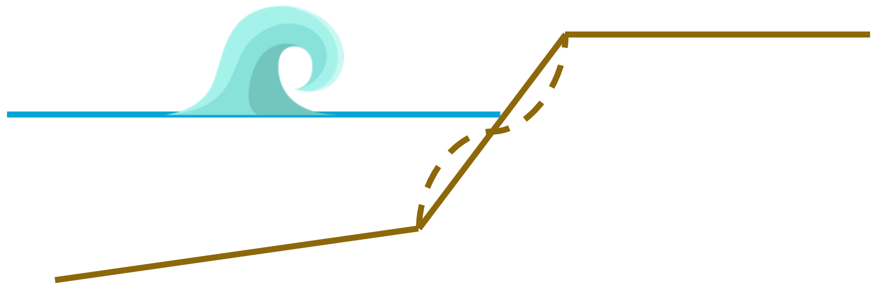
And how do I choose between a deterministic and stochastic model?

All systems, in reality, are stochastic to our eyes, since we never truly know the actual properties and inputs. However, under certain circumstances, this *stochasticity* can be neglected. Let us take a look to some examples of deterministic and stochastic systems:

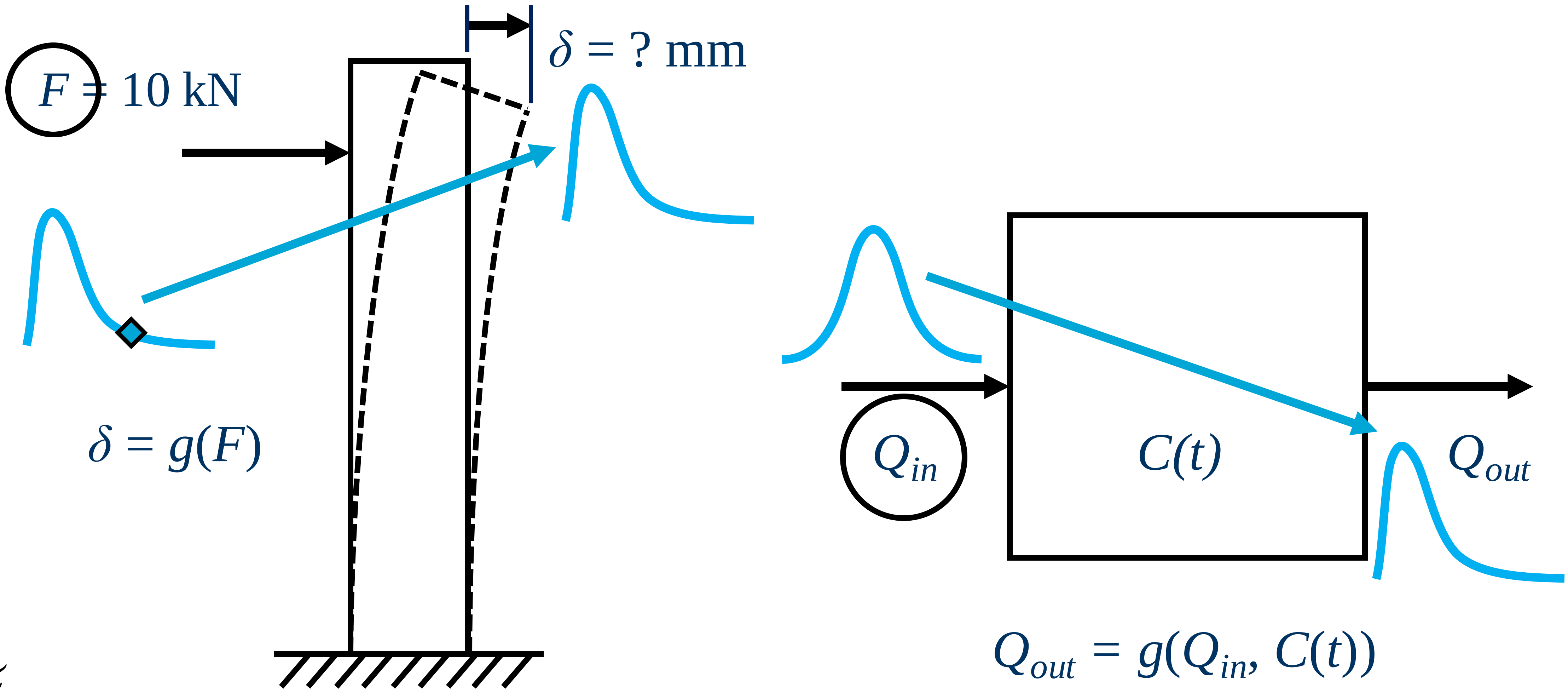
Deterministic → If input is ‘a’, output will always be ’b’

Stochastic → If input is ‘a’, what is the probability of ‘b’

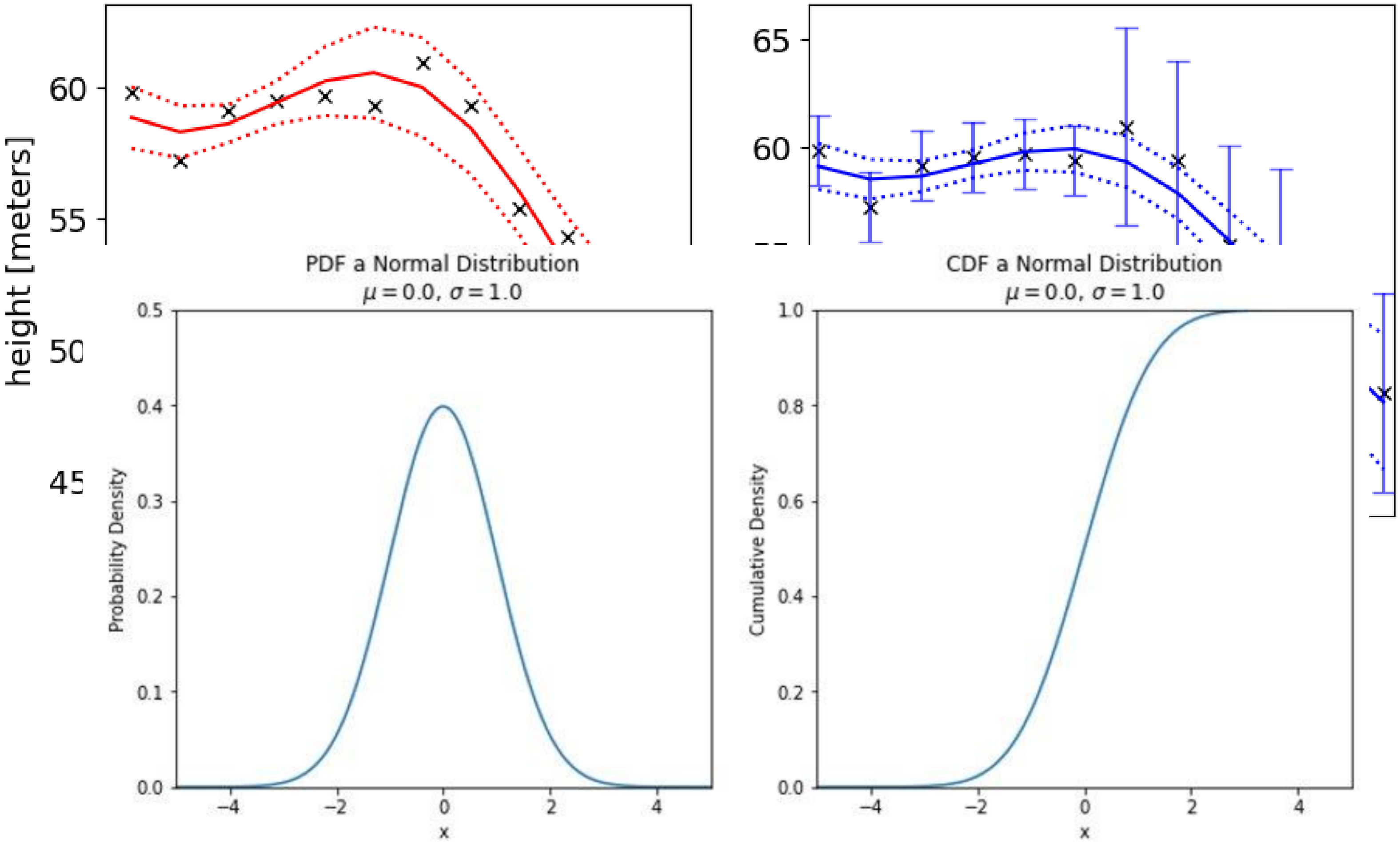
Recap of Week 1

Deterministic – expected deformation	Stochastic – beach profile after storm
	
- Lab experiment - Material properties known (thoroughly tested) - Loading applied by a calibrated machine - Measurements taken from calibrated gauges	- Grain size? - Wave statistics? - Wave trains? - Initial profile? Previous wave storms?

Deterministic design (pre-MUDE) – parameters given



Recap of Weeks 2 and 3



Aleatoric

- intrinsic phenomenon; typically associated with variations that occur in nature

Epistemic

- lack of knowledge; often called model uncertainty

Error

- deficiency in any stage of modelling/simulation not due to lack of knowledge

Variables are Gaussian-distributed!

This week

What would be an example of aleatoric uncertainty in your field?

Aleatoric

- intrinsic phenomenon; typically associated with variations that occur in nature

Epistemic

- lack of knowledge; often called model uncertainty

Error

- deficiency in any stage of modelling/simulation not due to lack of knowledge

Join the Vevox session

Go to **vevox.app**

Enter the session ID: **175-839-818**

Or scan the QR code

TU Delft

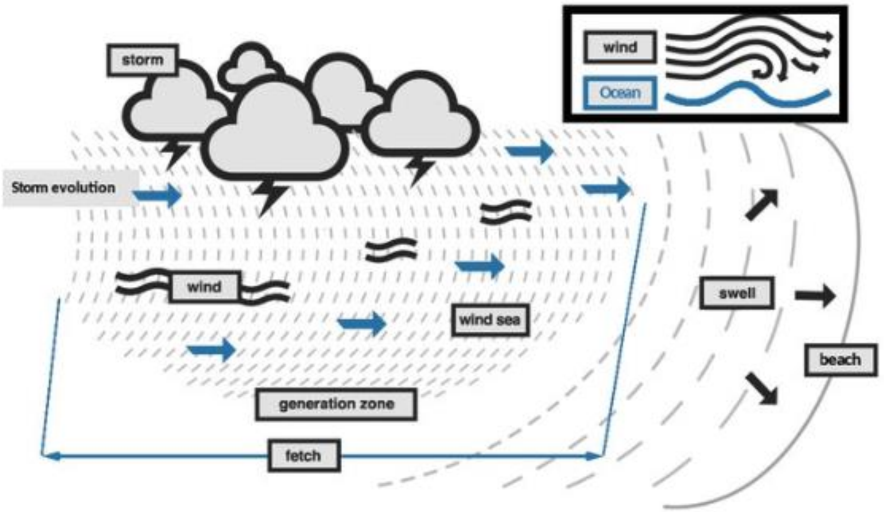
Give an example of aleatoric uncertainty

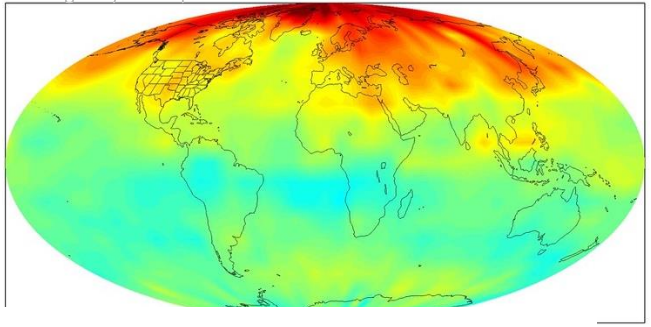
Give an example of aleatoric uncertainty

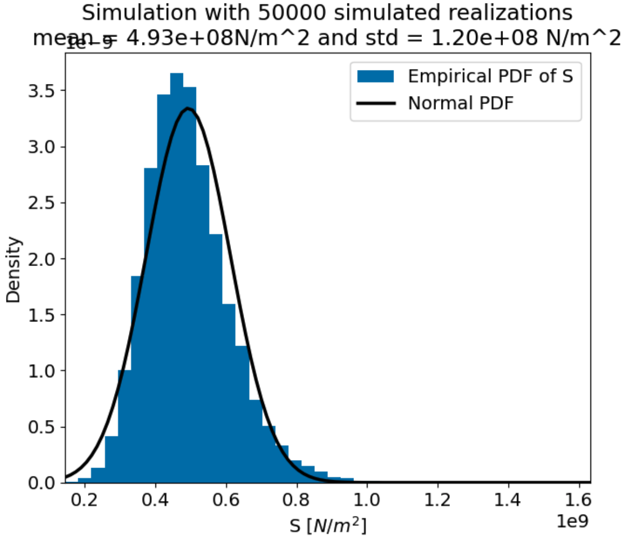
climate change
wind speed
stop and go waves
earthquake
enso
wave spectra
dice roll
sediment transport
avalanche
fire
water parameters
waves
wind
wave height
duration
climate
material defect
wave speed
rainfall
flipping a coin
dice
earthquakes
deep waves
sediment
seismic events
outside temperatur
temperature changes
rain
wood
genetics
wave heigth on a day
trafic jam

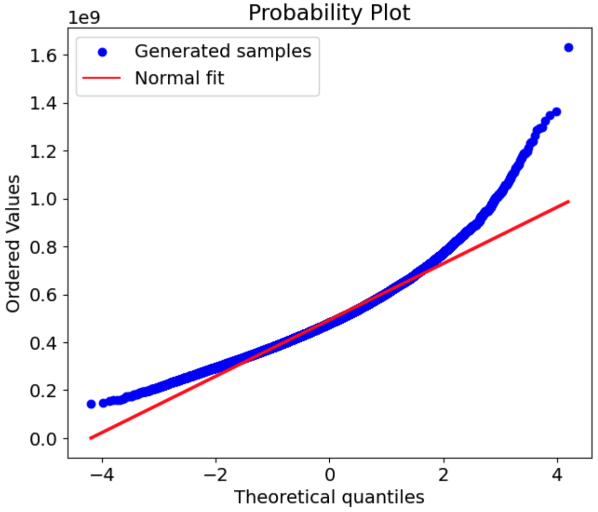
This week

From C.E. Stringari (2020)









Aleatoric

- intrinsic phenomenon; typically associated with variations that occur in nature

Epistemic

- lack of knowledge; often called model uncertainty

Error

- deficiency in any stage of modelling/simulation not due to lack of knowledge

Variables are NOT necessarily Gaussian-distributed!

How do we model this type of uncertainty?

Probability distribution functions

Continuous distribution functions – why?

- Continuous random variables

Continuous distribution functions – concept

- Continuous random variables

- Mathematical model which relates the values of a random variable and their probability

Value/quantile ⟷ probability

Continuous distribution functions – PDF

- Continuous random variables

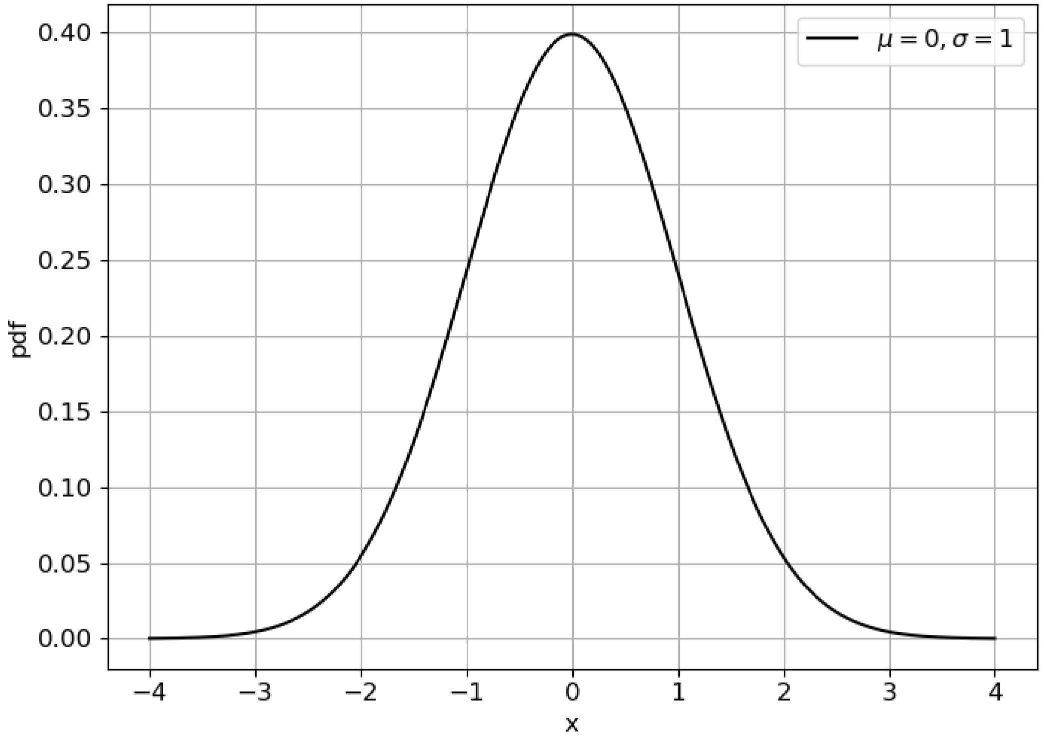
- Mathematical model which relates the values of a random variable and their probability

- Probability density function (PDF) $f_X(x)$

$$f_X(x)dx = P(x < X \leq x + dx)$$

$$f_X(x) \geq 0$$

$$\int_{-\infty}^{+\infty} f_X(x)dx = 1$$



PDF of the Gaussian distribution

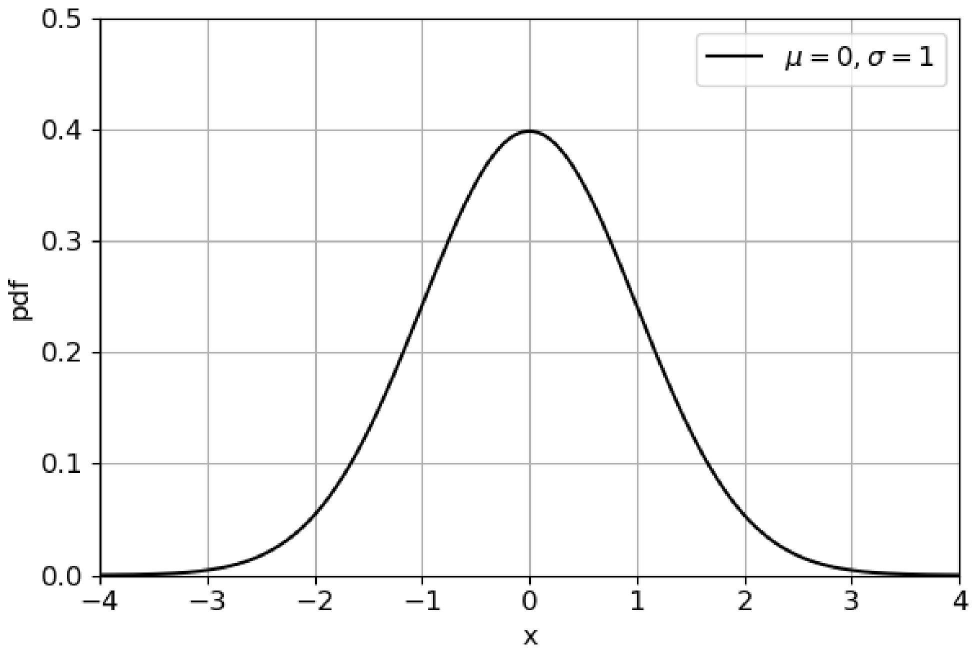
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

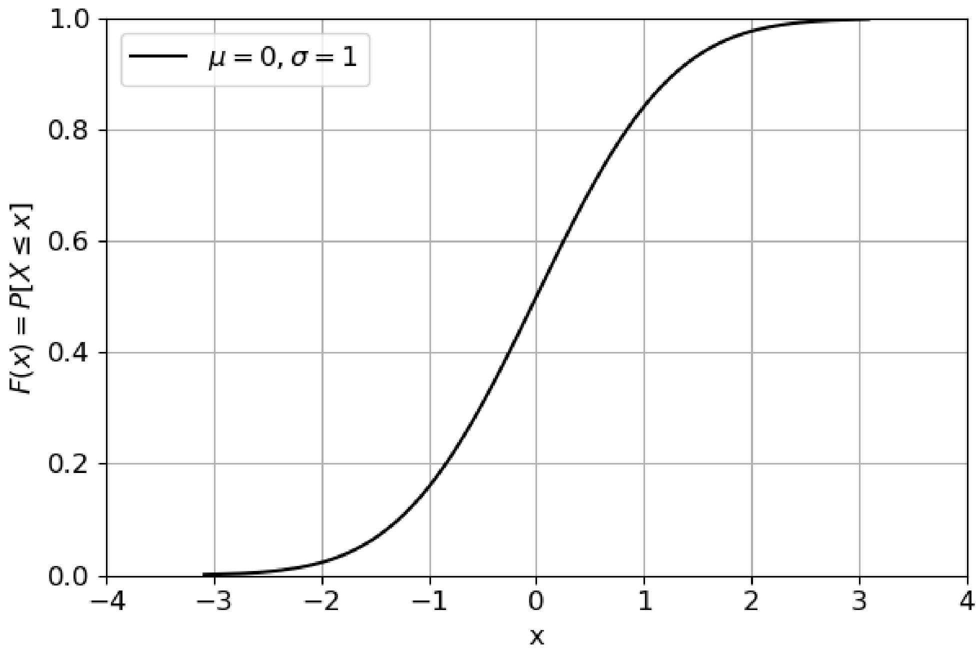
From PDF to CDF

- Probability density function (PDF) $f_X(x)$
- Cumulative distribution function (CDF) $F(x) = \int_{-\infty}^{x} f(x)dx$

CDF of the Gaussian distribution

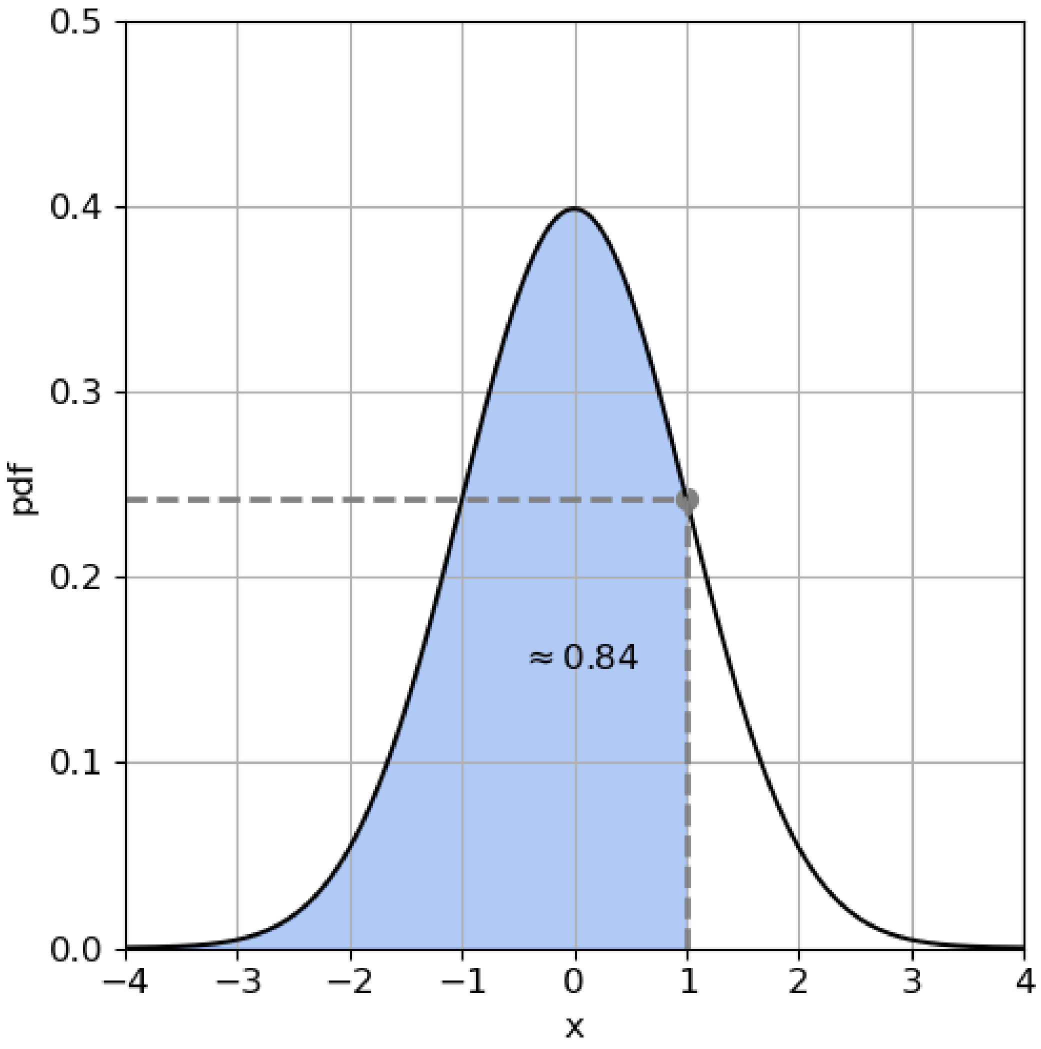
$$F(x) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$$

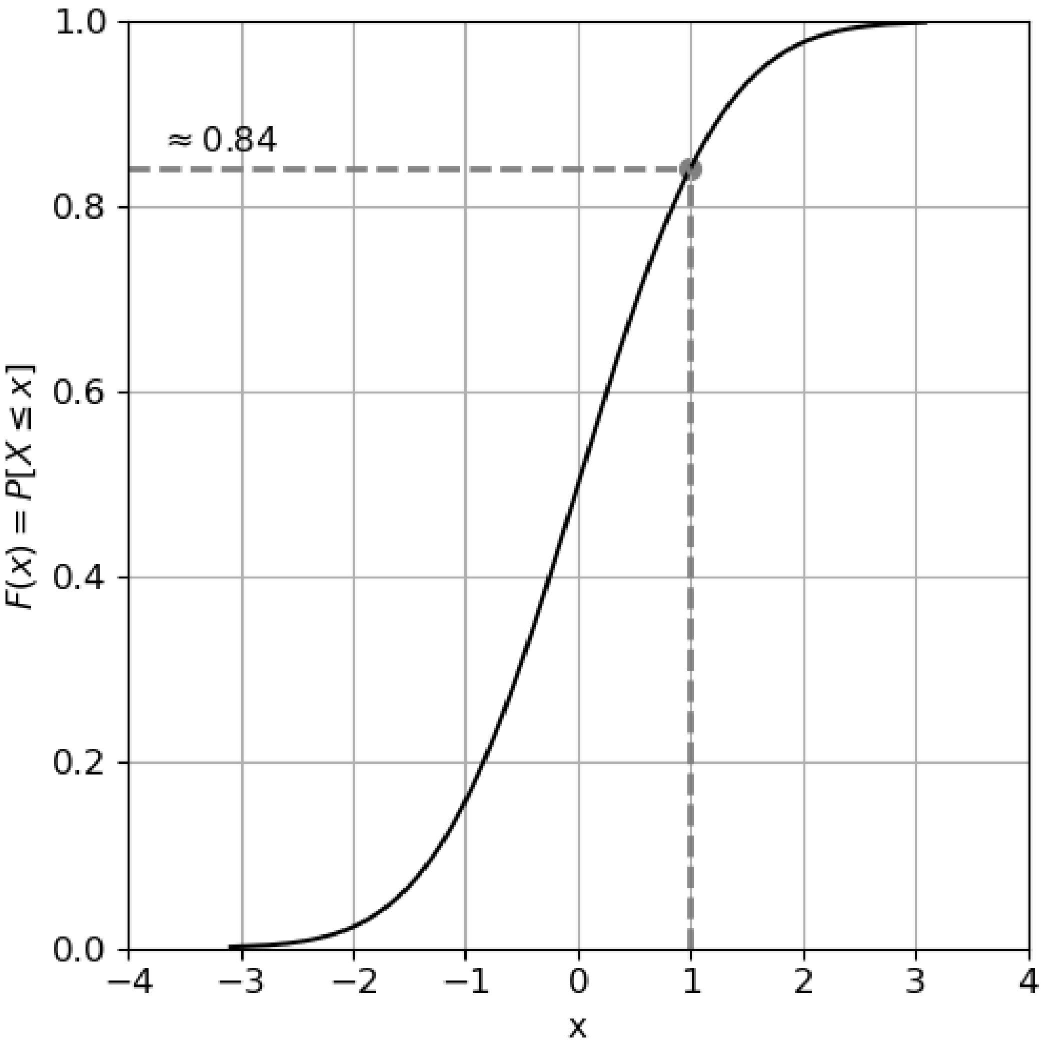




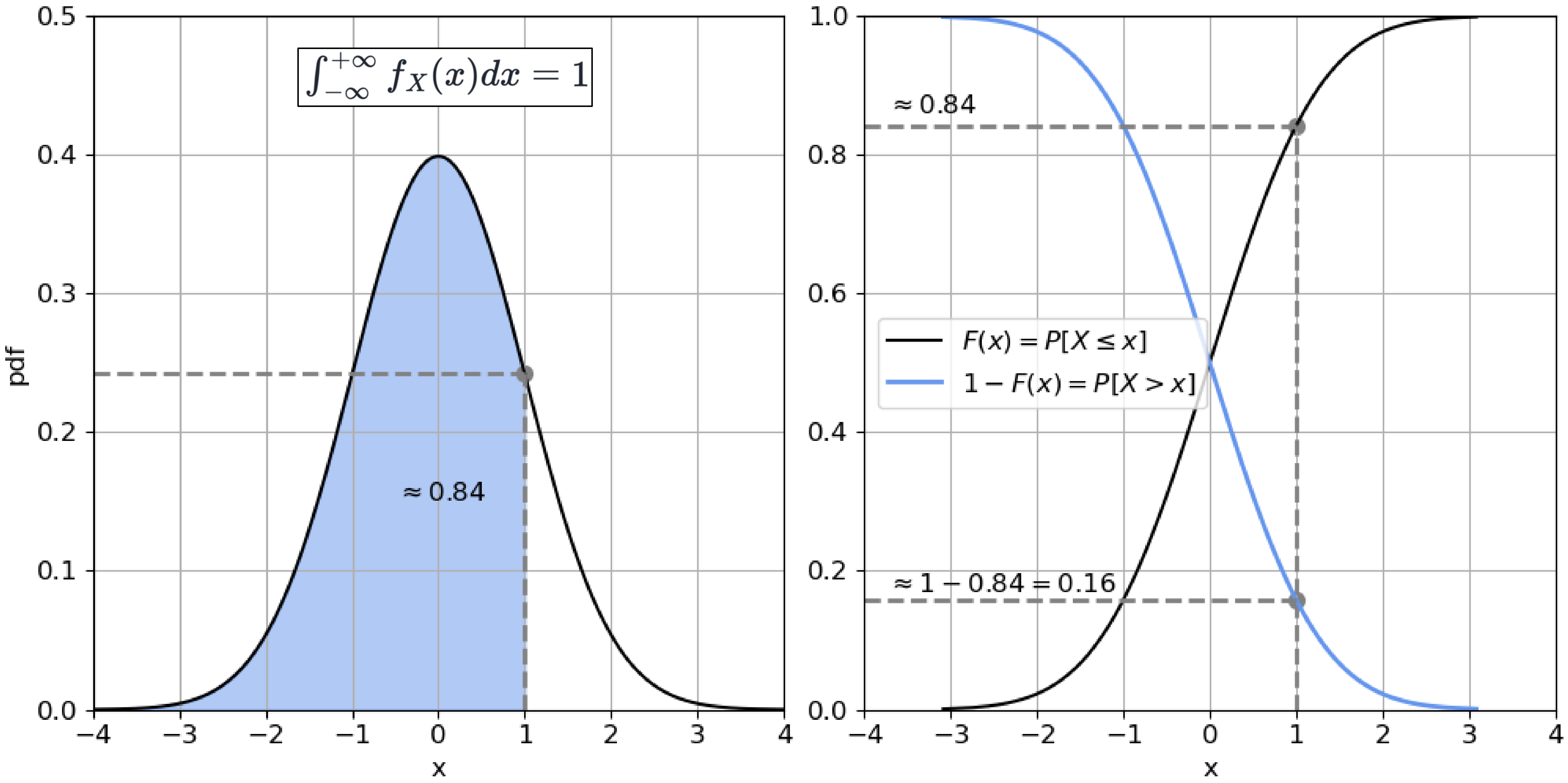
TU Delft

From PDF to CDF





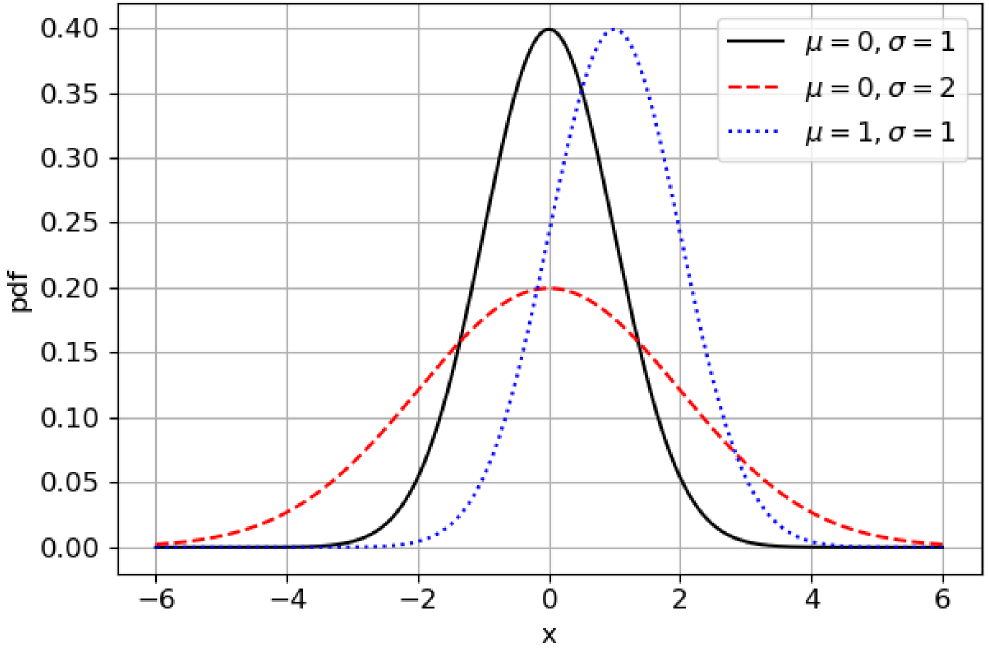
From PDF to CDF – exceedance

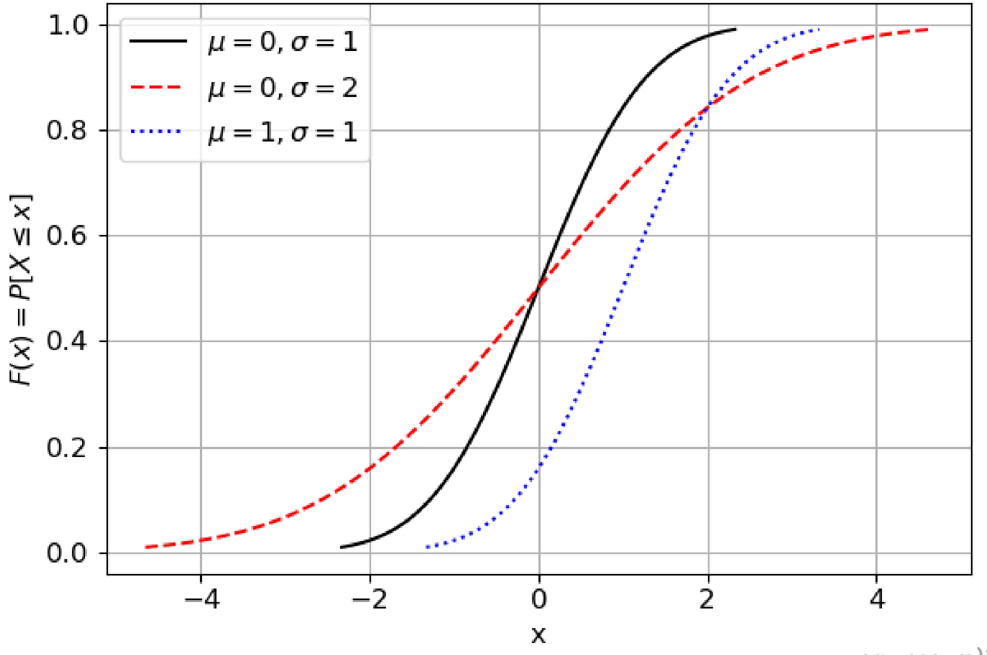


Parameters in PDF and CDF – Gaussian distribution

- Probability density function (PDF) $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- Cumulative distribution function (CDF) $$F(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$$





Empirical distribution functions

Continuous distribution functions

Mathematical model which relates the values of a random variable and their probability

But what do I want to model?

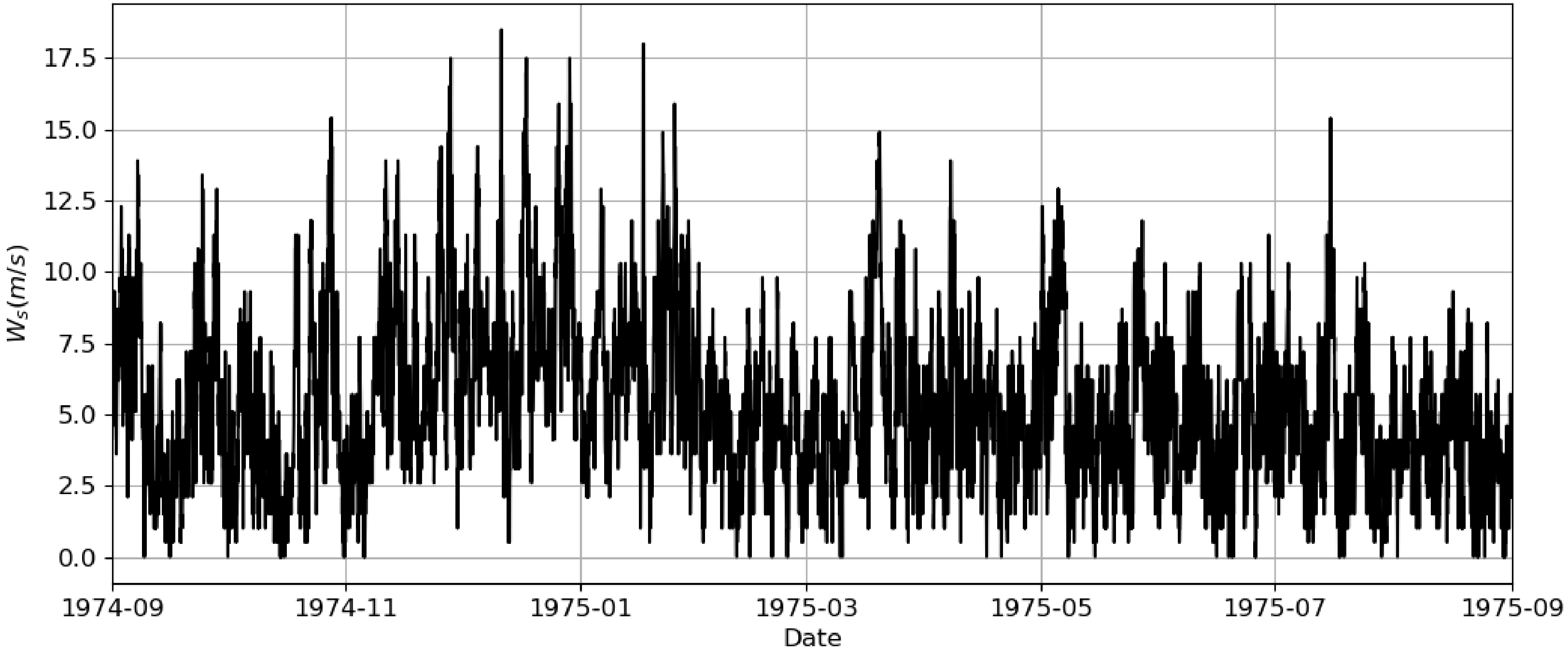
Observations → Empirical distribution function

I want a model which is able to reproduce the probabilistic behavior in the observations

Empirical distribution functions

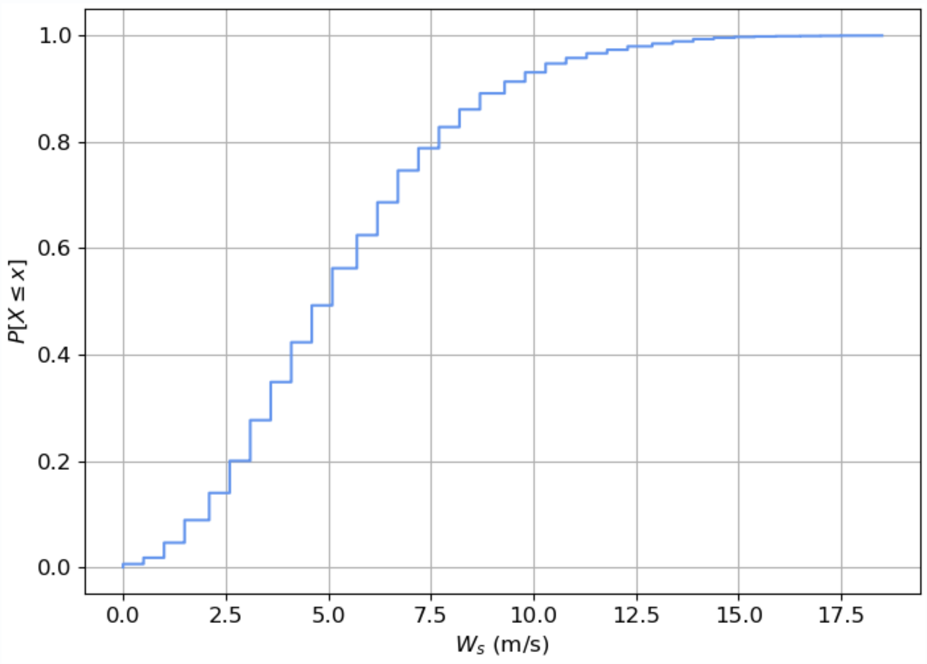
We can define from our observations an empirical PDF and empirical CDF

Let's see it with an example!



Empirical CDF

We need to assign a non-exceedance probability to each observation.




```
>> read observations

>> x = sort observations in ascending order

>> length = the number of observations

>> probability of not exceeding = (range of integer values from 1 \ to length) / length + 1

>> Plot x versus probability of not exceeding
```

Empirical CDF

Let’s do it slowly!

Length = 5

x
3.2
4.5
3.8
7.5
2

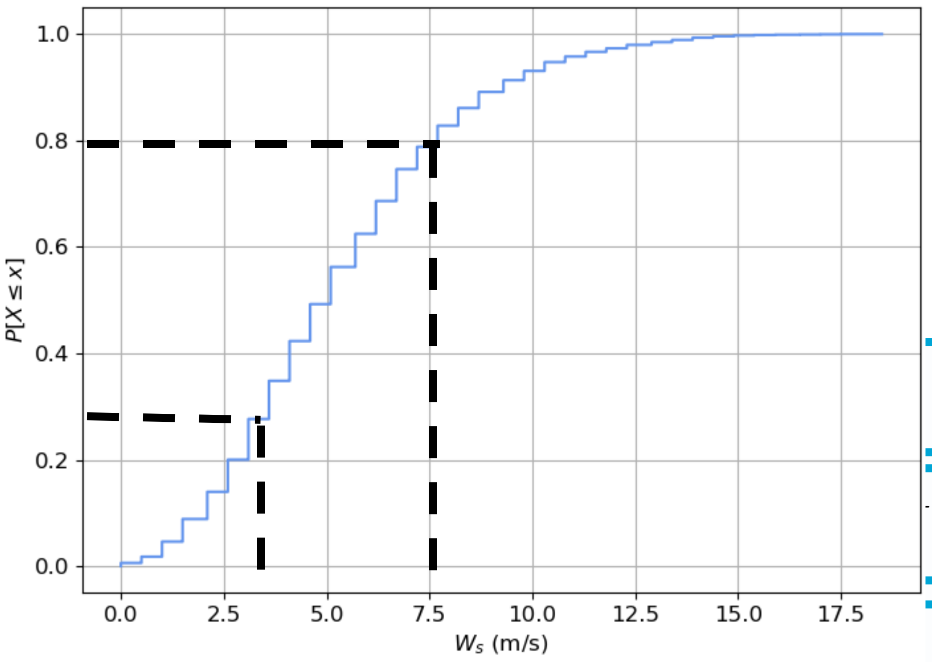
```
>> re
```

```
>> x                                        ng
order
```

```
>> le
```

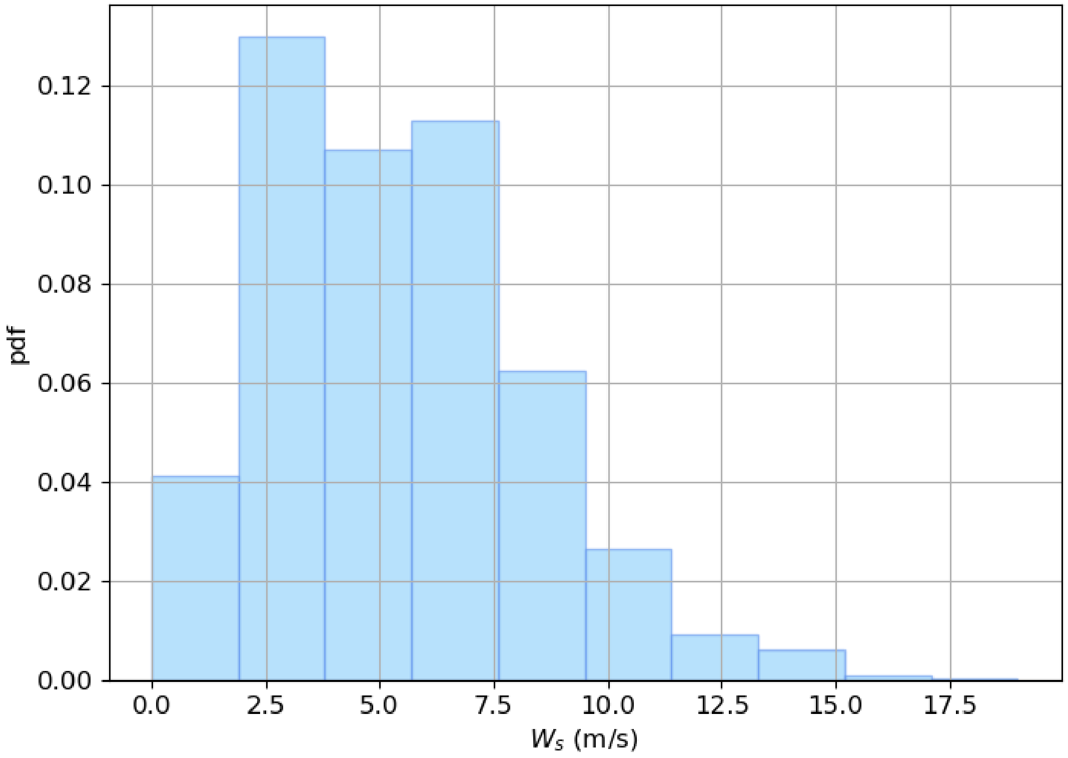
```
>> probability of not exceeding = (range
of integer values from 1 to length) /
length + 1
```

```
>> Plot x versus probability of not
exceeding
```



Empirical PDF

$$f(x) = F'(x) = \lim_{\Delta x \to 0} \frac{F(x+\Delta x)-F(x)}{\Delta x}$$

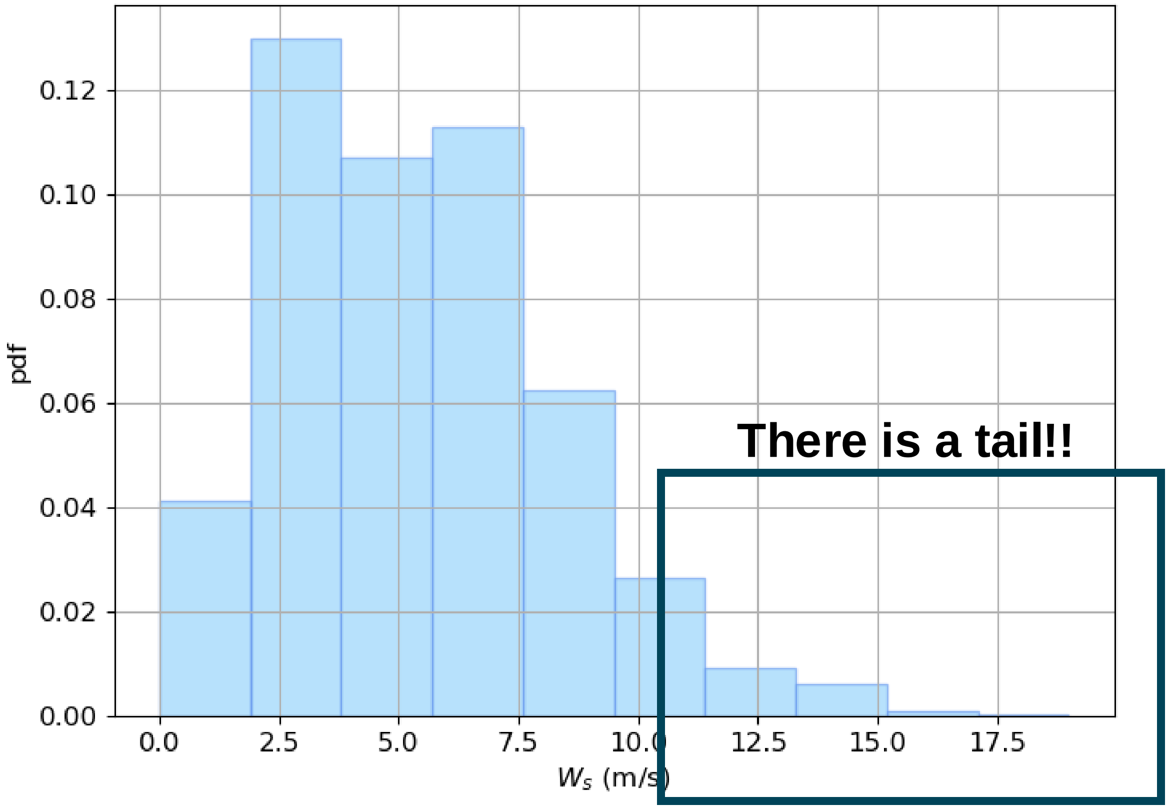


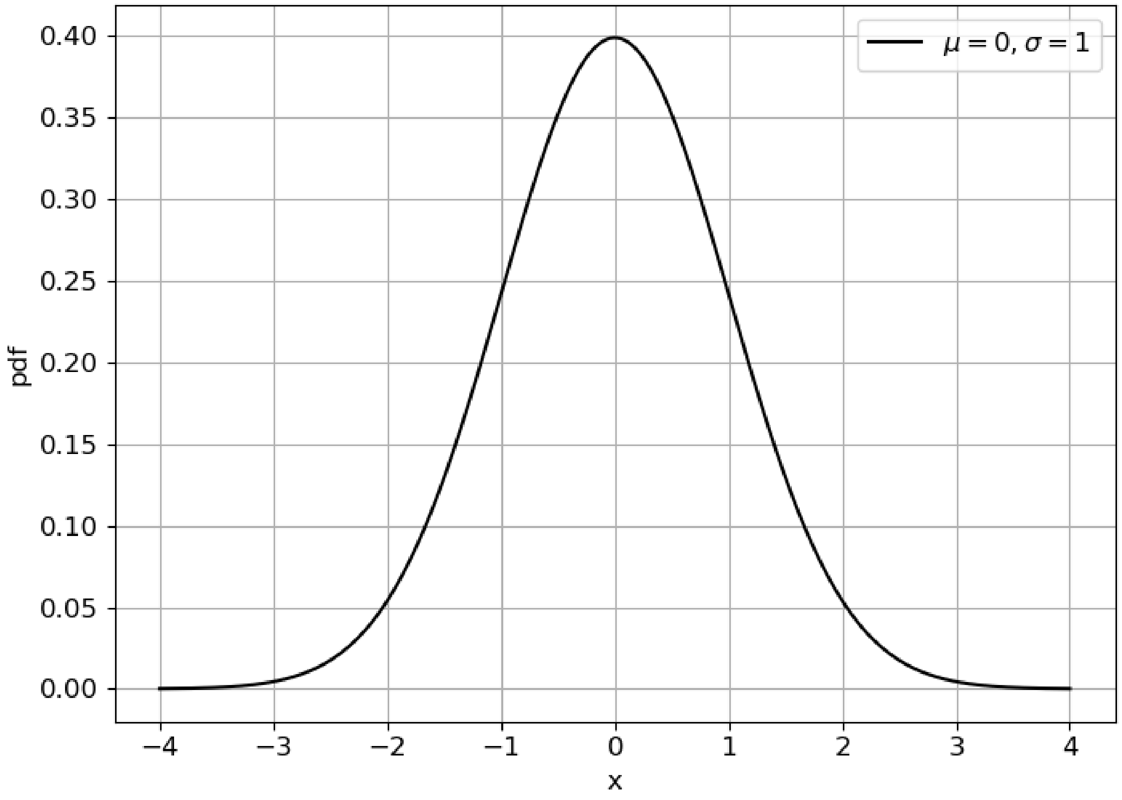

```
>> read observations

>> bin_size = 2   #delta x

>> min_value = minimum value of observations
   max_value = maximum value observations
   n_bins = (max_value - min_value)/bin_size
   bin_edges = range of n_bins + 1 values
   between the truncated value of min_value
   and the ceiling value of max_value

>> bin count = empty list
   for each bin:
         append the number of observations  between
the bin_edges to count

>> freq = count / number of observations

>> densities = freq / bin_size

>> Plot barplot densities
```

Why non-Gaussian?

Concept of tail

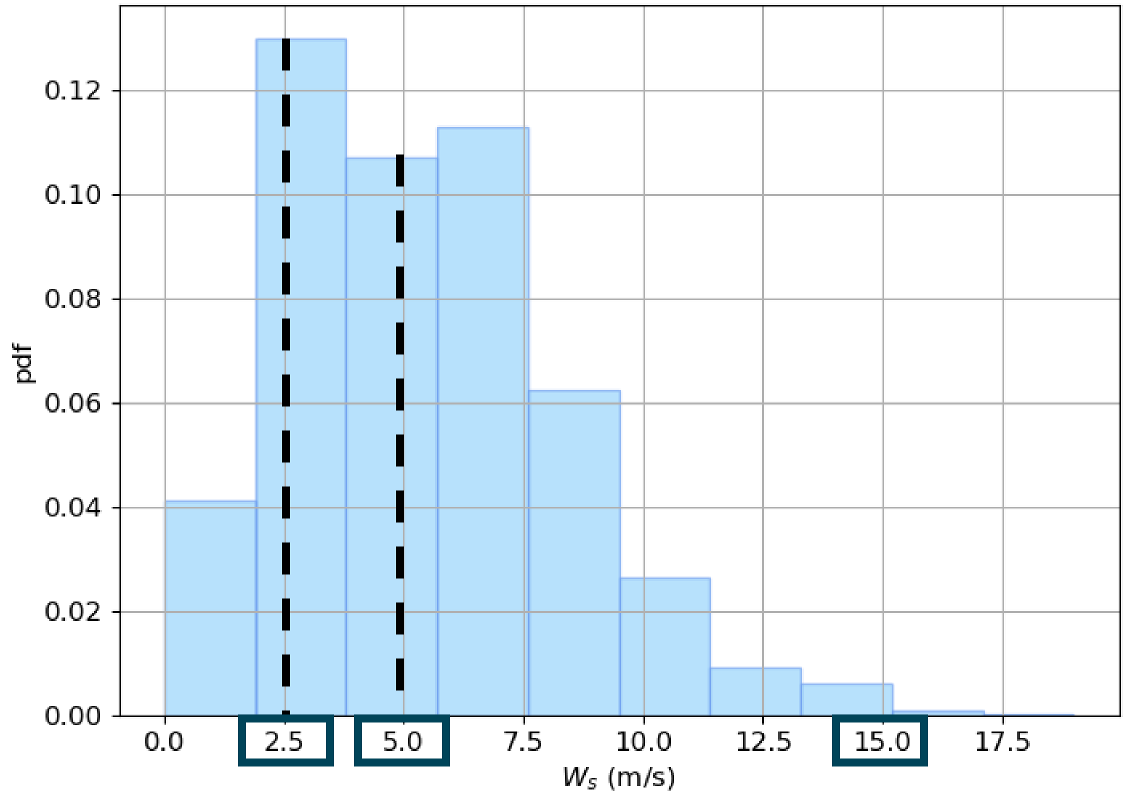
Does this look Gaussian?





TU Delft

Why is the tail important?



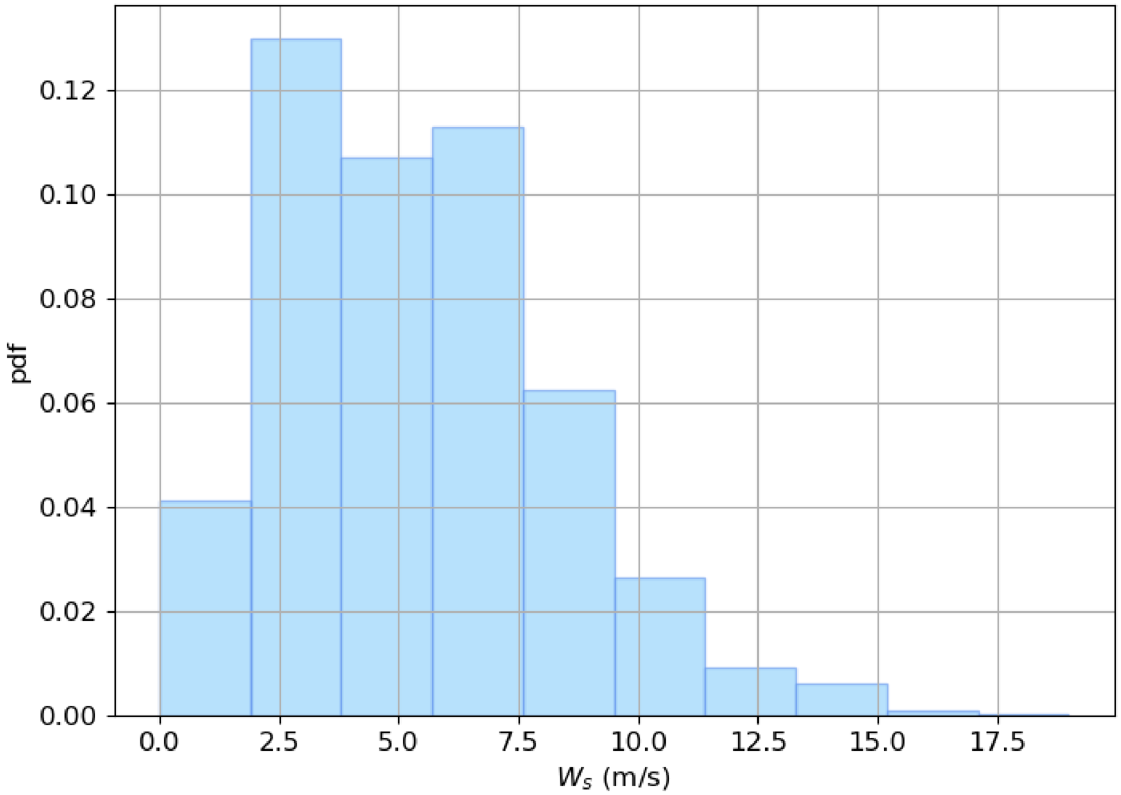
- You are designing a building against wind loading
- Which value would you use for design?
- You vote!

Which design value would you choose?

2.5 m/s (mode of the ecdf)	0%
5.0 m/s (mean)	0%
15.0 m/s (approx. max observation)	0%

Which design value would you choose?

We typically design to withstand extreme values



- We want the building to perform in ordinary conditions (around central moments)
- We also want the building to withstand the storms
- Tails can also be negative!
  - E.g.: nutrients concentration to ensure the survivability of species

Brief intro to a selection of parametric distributions

TU Delft

Parametric distributions in the book

5. Continuous Distributions
5.1. PDF and CDF
5.2. Empirical Distributions
5.3. Parametric distributions
Revisiting Gaussian distribution
Non-Gaussian distributions
Uniform distribution
Exponential distribution
Gumbel distribution
Lognormal distribution
Summary of parametric distributions
5.4. Fitting parametric distributions
5.5. Parameterization of continuous distributions

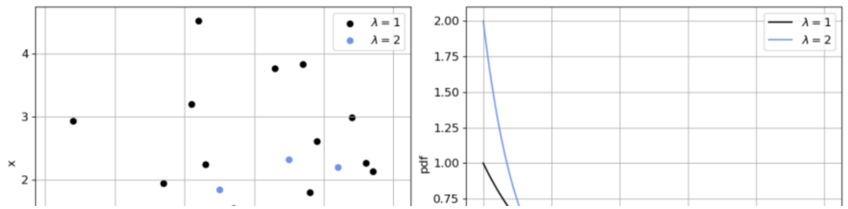
Exponential distribution

Another widely used distribution function is the Exponential distribution. For instance, it is applied to model the waiting time between succesive events of a Poisson process. The PDF of the Exponential distribution is given by

$$f(x) = \lambda e^{-\lambda x} \qquad for\ x \geq 0, \lambda > 0$$

$$f(x) = 0 \qquad otherwise$$

where λ is the parameter of the distribution, which is often called *rate*. In the right pannel of the figure below, an example of two Exponential distributions with $\lambda = 1$ and $\lambda = 2$ is shown. As you can see, the maximum density in the PDF of en Exponential distribution is located at zero and it is followed by an Exponential decay. The higher the parameter λ, the higher the value of the density in $x = 0$ and the faster the decay. In other words, the higher the parameter λ, the more concentrated the values of the random variable which are likely to occur and, thus, the lower the standard deviation. This can be seen on the left pannel of the figure, where random samples of the distribution are plotted. There you can see how higher values of the random variable x appear when $\lambda = 1$, presenting then a higher dispersion.



1. **Gaussian**
2. Uniform
3. **Exponential**
4. **Gumbel** (left- and **right-tailed**)
5. Lognormal

There are a lot more in the literature!!

- Read about the rest in the book.
- What do I need to know? how the distribution looks (PDF/CDF), how it responds to changes in the parameters and some basic properties (symmetry or bounds).

TU Delft

Exponential distribution

- PDF

$$f(x) = \lambda e^{-\lambda x} \qquad for\ x \geq 0, \lambda > 0$$

$$f(x) = 0 \qquad otherwise$$

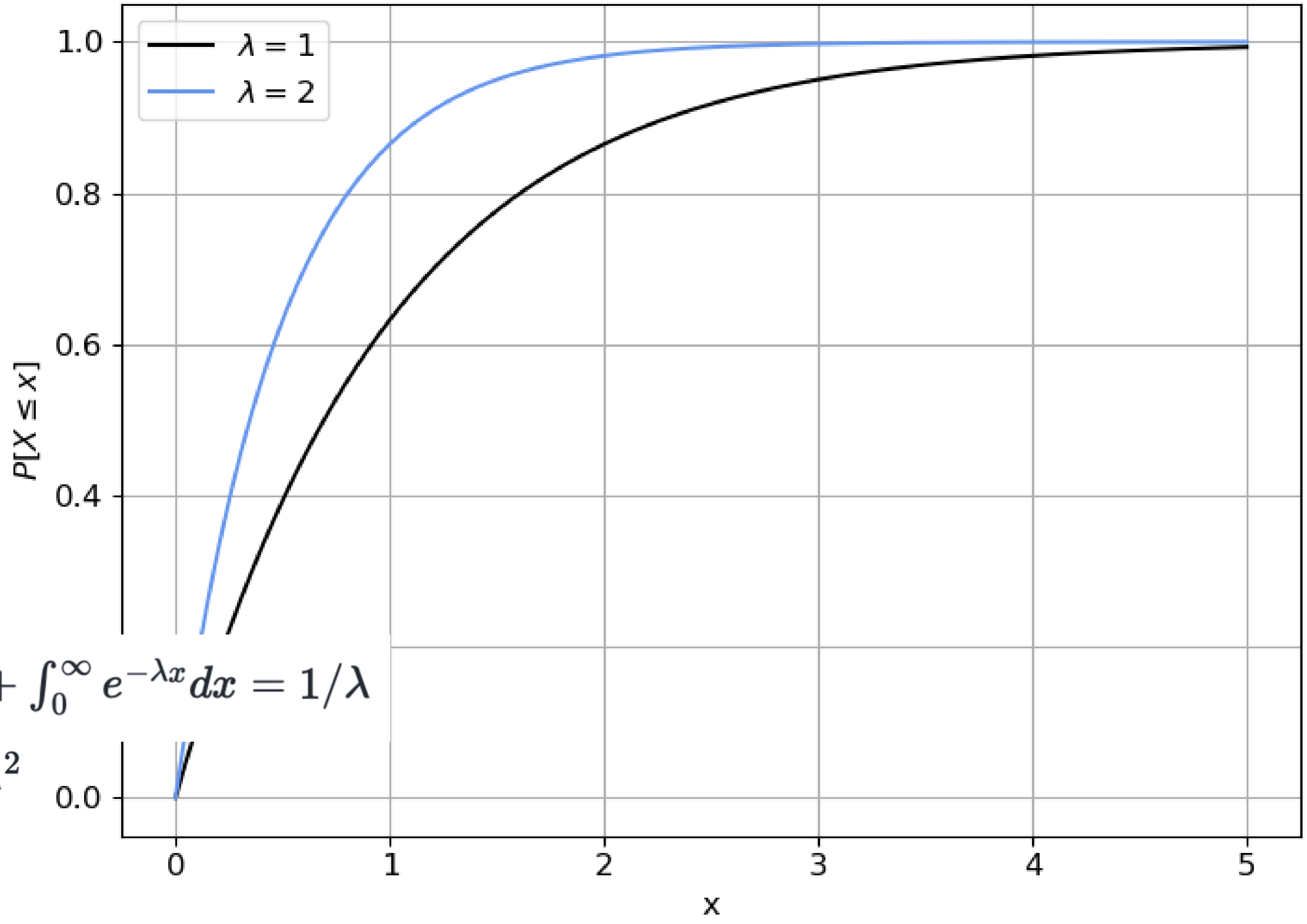
- CDF

$$F(x) = 1 - e^{-\lambda x}$$

- Some properties

$$E[X] = \int_o^{\infty} x\lambda e^{-\lambda x} dx = [-xe^{-\lambda x}]_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx = 1/\lambda$$

$$Var[X] = E[X^2] - (E[X])^2 = 1/\lambda^2$$



TU Delft

Gumbel distribution (right-tailed)

- PDF

$$f(x) = \frac{1}{\beta} e^{-\left(\frac{x-\mu}{\beta} + e^{-\left(\frac{x-\mu}{\beta}\right)}\right)}$$

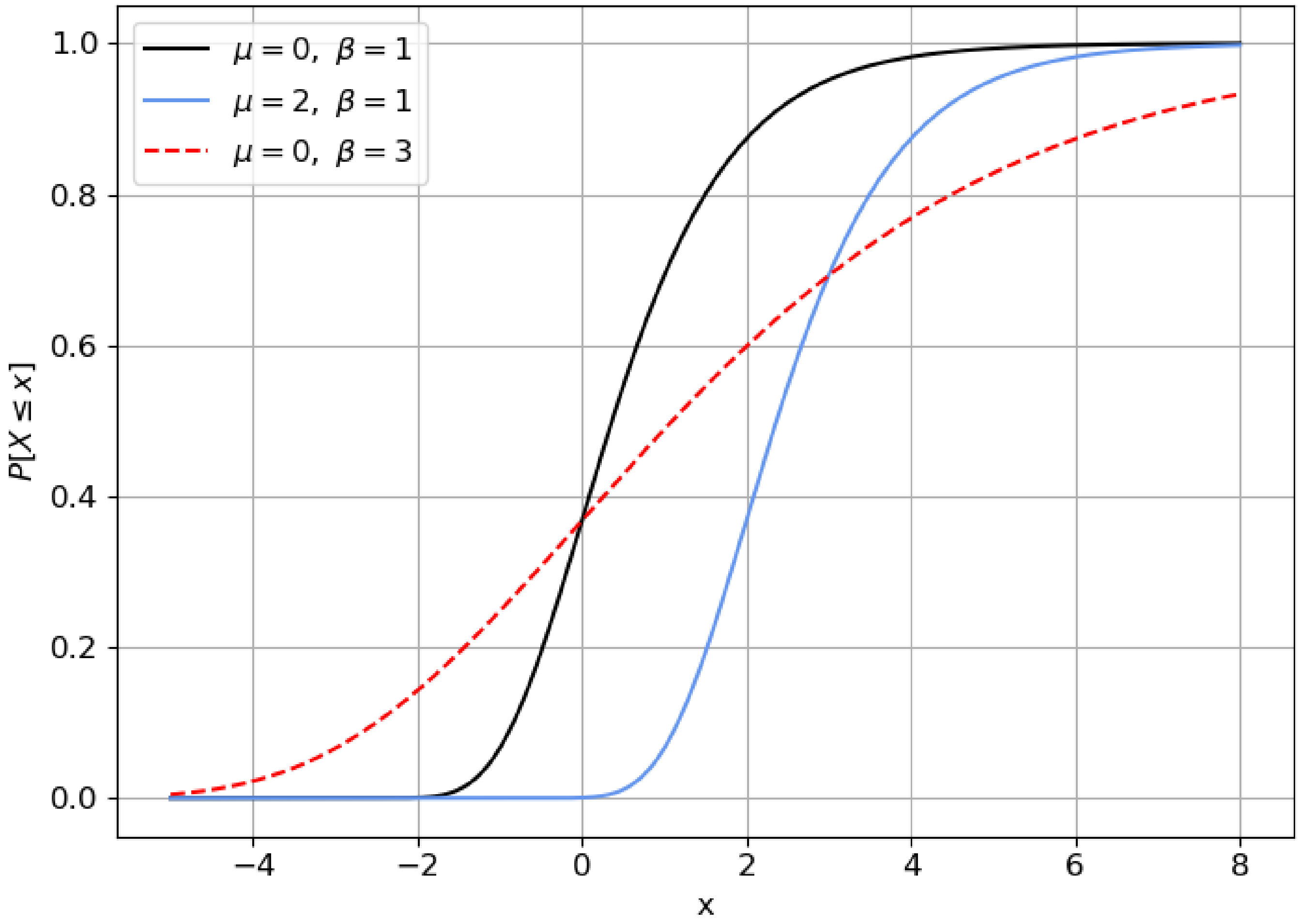
- CDF

$$F(x) = e^{-e^{-\frac{x-\mu}{\beta}}}$$

- Some properties

$$E[X] = \mu + \gamma\beta \qquad \gamma \approx 0.577$$

$$Var[X] = \frac{\pi^2}{6}\beta^2$$



Fitting distribution functions

Fitting distributions

- Given:
  - An empirical distribution function
  - A parametric distribution function (e.g.: Gumbel)

$$f(x) = \frac{1}{\beta} e^{-\left(\frac{x-\mu}{\beta} + e^{-\left(\frac{x-\mu}{\beta}\right)}\right)}$$

- Which is the value of the parameters of the distribution that best fits our data?

How to choose the parametric distribution function, next part of the lecture!

- Different methods: **moments** and MLE here.

Fitting distributions by moments

- Equate the moments of the observations to those of the distribution function

- Moments for the Gumbel distribution

$E[X] = \mu + \gamma\beta$ $\gamma \approx 0.577$ → Mean of the observations

$Var[X] = \frac{\pi^2}{6}\beta^2$ → Variance of the observations

Fitting distributions by moments - Example

- The intensity of earthquakes in Rome (Italy) is a random process.
- Using ‘Catalogo dei terremoti italiani dall’anno 1000 al 1980’ (the Catalog of Italian earthquakes from year 1000 to 1980) edited by D. Postpischl in 1985, we want to fit a Gumbel distribution to the observations using the method of moments.

- Mean intensity = 3.02
- Variance of intensity = 0.99

Gumbel distribution:

$$E[X] = \mu + \gamma\beta \qquad \gamma \approx 0.577$$

$$Var[X] = \frac{\pi^2}{6}\beta^2$$

Equating them to the observations:

$$3.02 = \mu + 0.577\beta$$

$$0.99 = \frac{\pi^2}{6}\beta^2$$

Thus, $\mu \approx 2.57$ and $\beta \approx 0.77$.

TU Delft

Assessing the goodness of fit

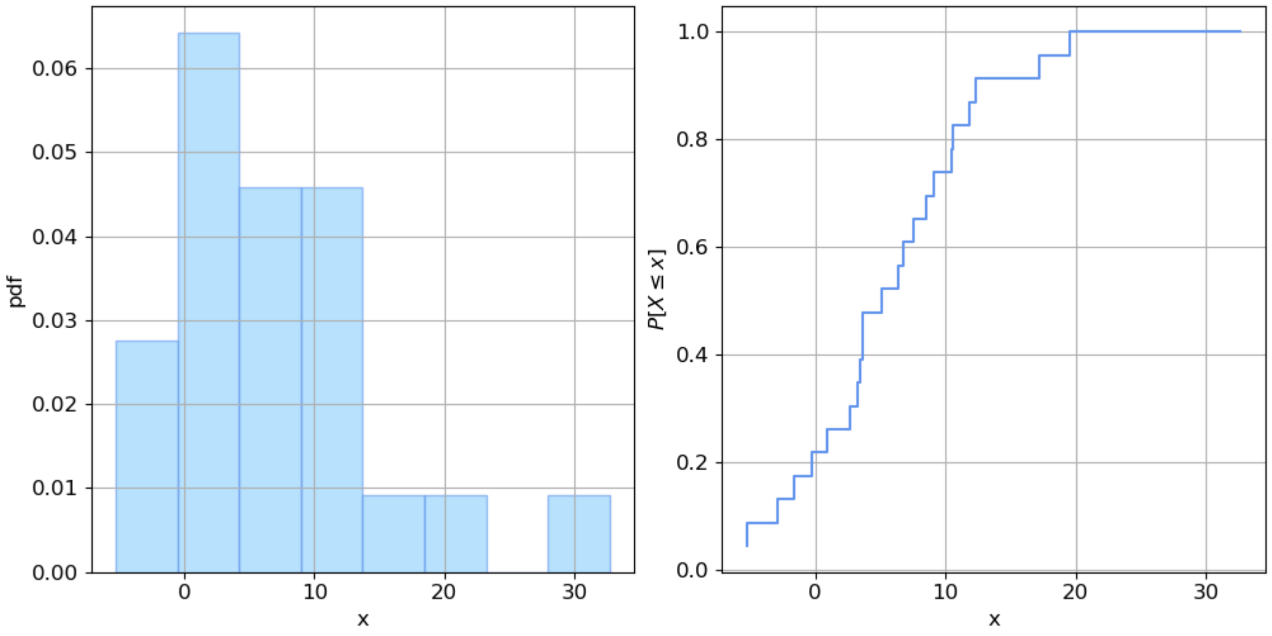
How do I choose a distribution?

- Physical constrains
  - E.g. Do negative values have physical meaning?
- Statistics of the observations

- Goodness of fit techniques
  - Not a ground truth
  - Objective way to compare models
  - You may obtain contradictory results!

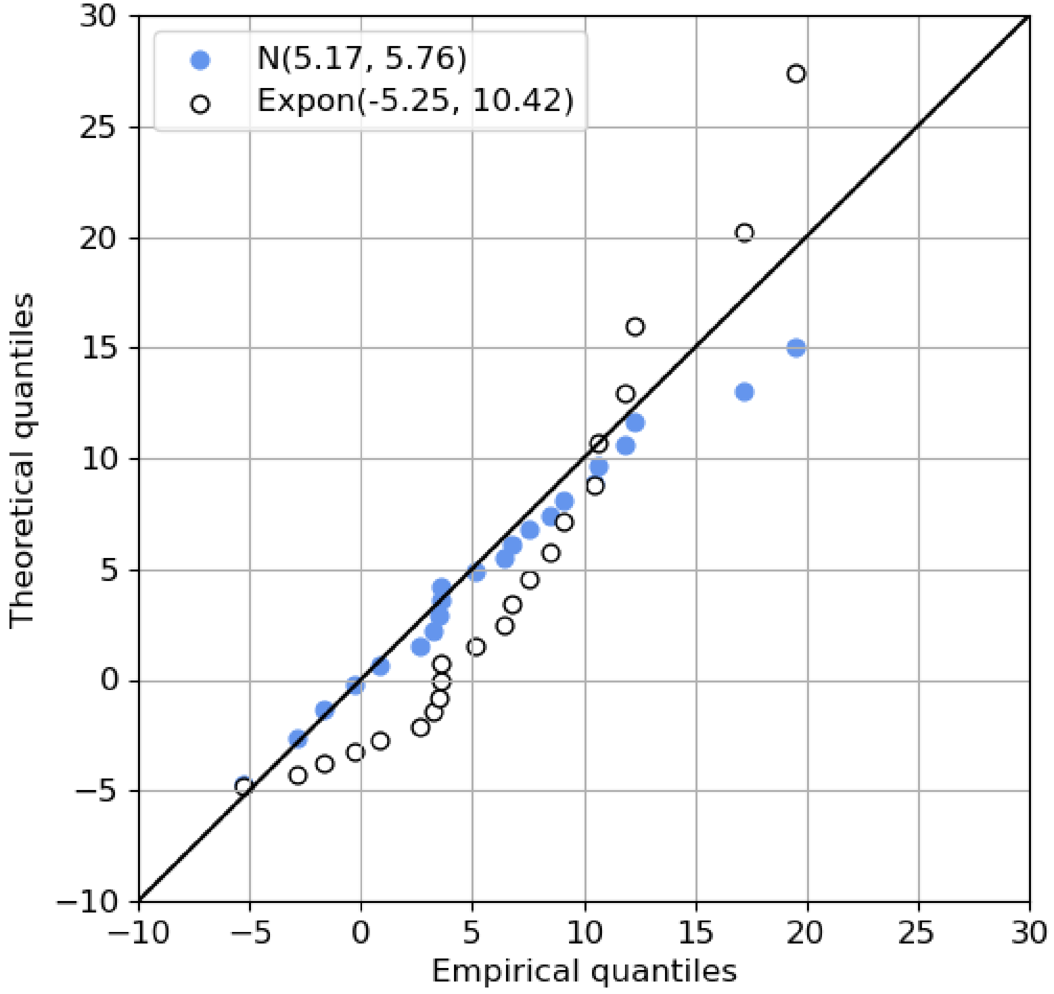
- **As professionals, the choice is yours!**

EXAMPLE:
- Toy dataset
- Exponential or Gaussian?



Graphical methods - QQplot

- Measured against predicted values
  - Fitted distribution: estimate the values of the random variable with the observed empirical probabilities
- 45 degree-line is the perfect fit

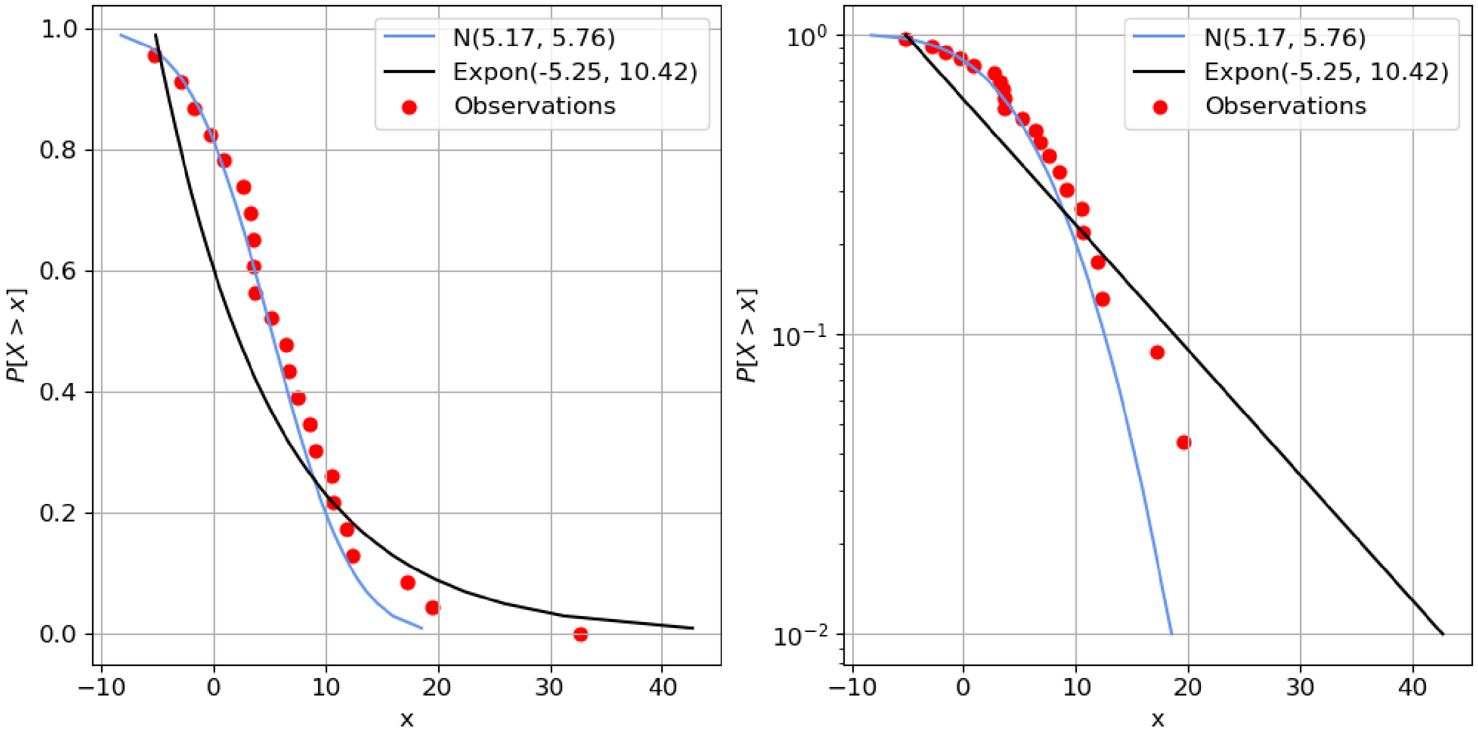


- Simple
- Fast to implement
- Central moments + tail

Graphical methods – log-scale

- Exceedance probability plot (1-F(x))

- How does the fitted distribution fit the observations in log-scale?



- Simple
- Fast to implement
- Focus on the tail: key element!

TU Delft

Formal hypothesis tests – Kolmogorov-Smirnov

- Widely used nonparametric hypothesis test

- Two variants:
  - Two samples: same population?
  - **One sample: GOF to a distribution**

Hypothesis tests:

H_0: null hypothesis
H_1: alternative hypothesis

Statistic ~ distribution → p-value

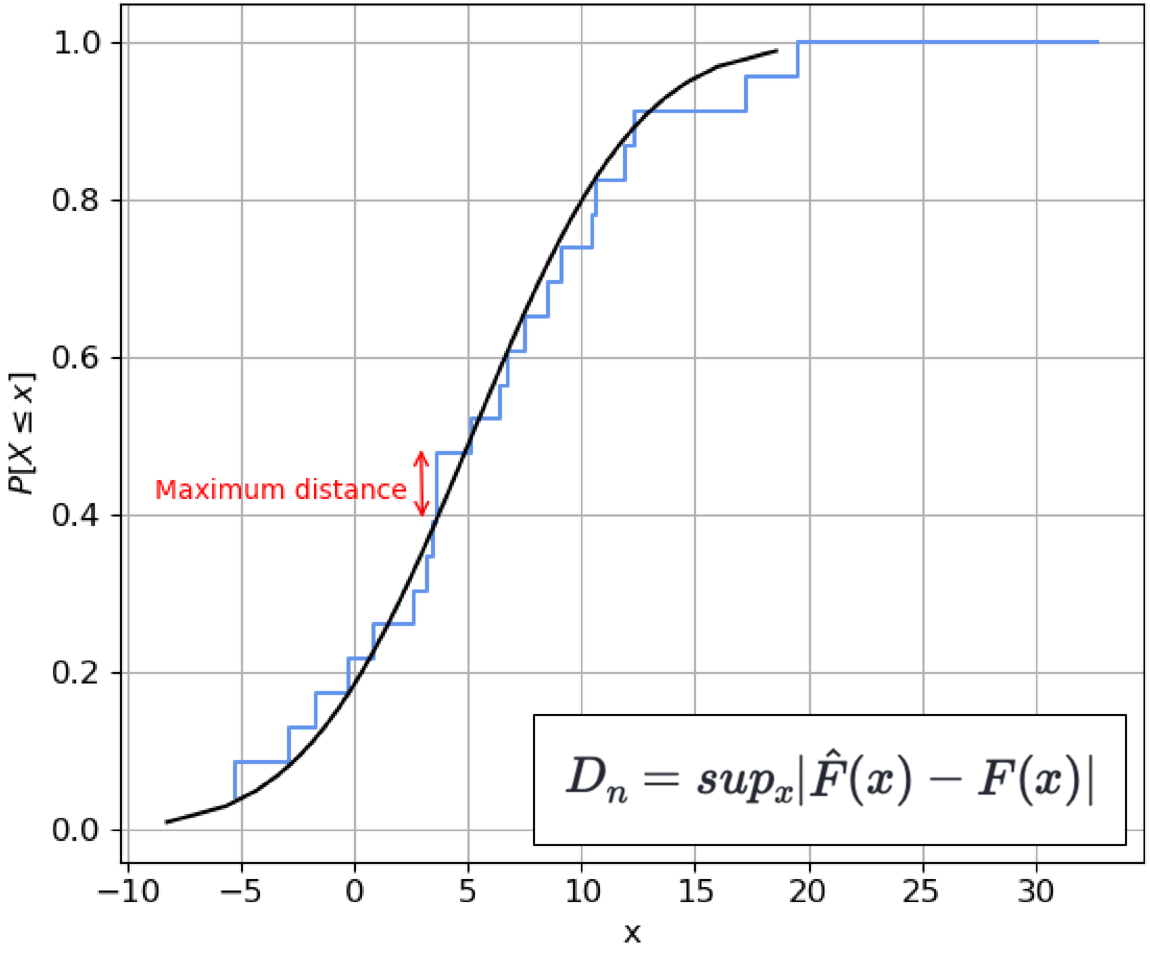
p-value: probability of the null hypothesis being true

Significance (typically $\alpha = 0.05$)

If p-value> α: We accept H0
If p-value< α: We reject H0

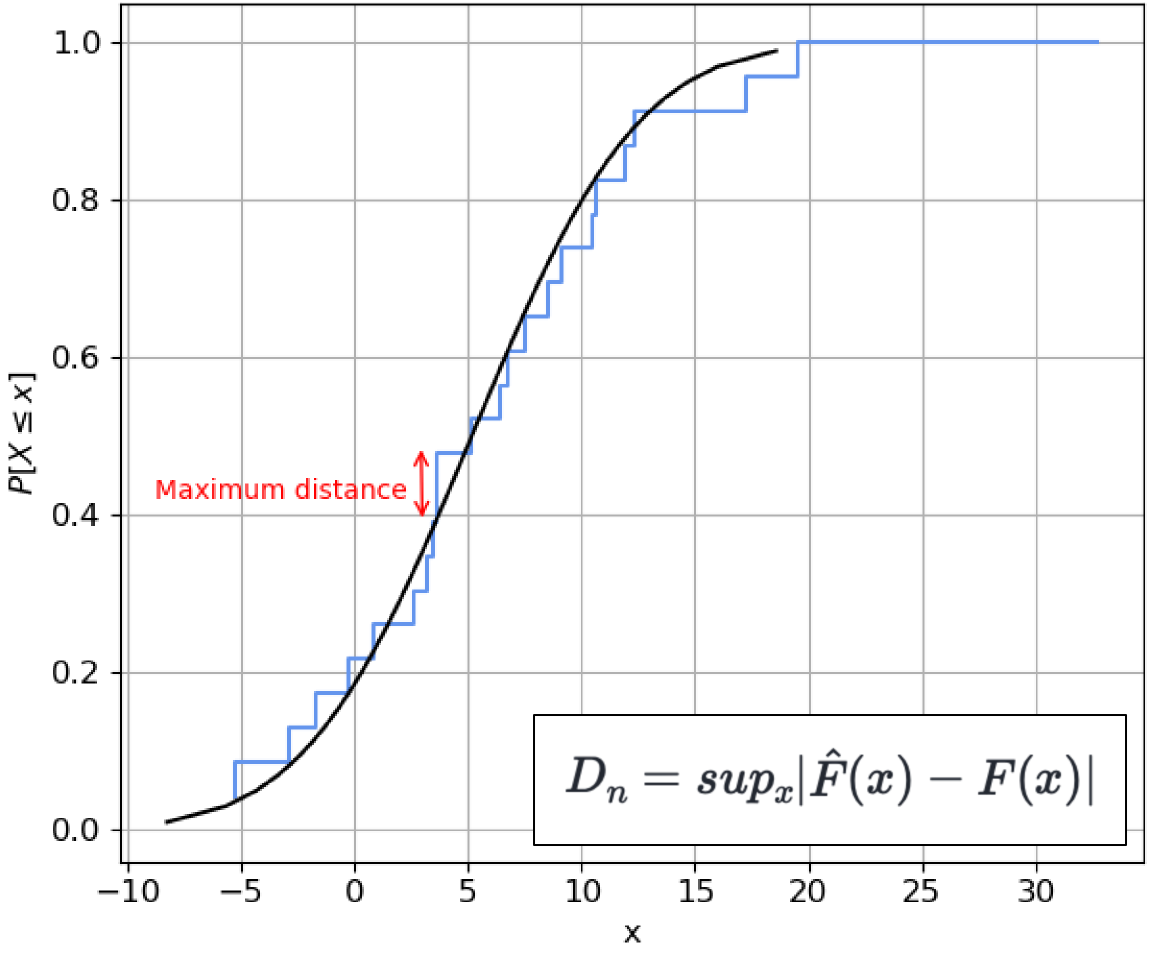
Formal hypothesis tests – Kolmogorov-Smirnov

- **One sample: GOF to a distribution**
- Based on the KS statistic: (roughly) the maximum distance between the ECDF and the fitted CDF
- $H_0 : \hat{F} \sim F$
- P-value > $\alpha = 0.05 \longrightarrow$ I cannot reject that the observations follow the distribution



Formal hypothesis tests – Kolmogorov-Smirnov

- $H_0 : \hat{F} \sim$ Normal distribution
- P-value = 0.93
- P-value =0.93 > $\alpha = 0.05 \longrightarrow$ I cannot reject that the observations follow a Normal distribution



TU Delft

What's next?

- There is more in the textbook!
  - 5.5 Parameterization of continuous distributions
- Wednesday workshop: concrete compressive strength
- Friday project: your choice!
  - Traffic and CO_2 emissions
  - Waves and impacts
  - Velocities, depths and discharges

5.5. Parameterization of continuous distributions

In the previous sections, you have studied different parametric distributions that can be applied to model the univariate uncertainty in our data. Those distributions were characterized by a set of parameters (e.g.: λ for Exponential distribution). Those parameters can be fitted to model real-world data as accurately as possible and, thus, use the distribution for predicting future events. Along the sections devoted to present a selection of distribution functions, the equations for the PDF and CDF as you can usually find them in text books were presented. However, they are just equations! That means that we can play with them and parameterize the distribution the way that fits best to our purposes.

In this section the parameterization loc-scale-shape will be addressed in the context of `scipy` Python package. This parameterization is very convenient due to the consistency it provides (all distributions with the same parameters), the ease of the interpretation of those parameters and the advantages of their implementation in computer code. That is why `scipy` package (between others) uses this parameterization for continuous distribution functions.

Definition of location, scale and shape

The location (μ) parameter shifts the distribution along the x-axis without changing its shape. The scale parameter (β) determines the width of the distribution. Finally, the shape parameter (ξ) is any extra parameter (if any) in the distribution function which is not μ or β and describes the form of the distribution. Let's see it better with a couple of examples!

You have already been introduced to the (right-tailed) Gumbel distribution, whose PDF is given by

$$f(x) = \frac{1}{\beta} e^{-\left(\frac{x-\mu}{\beta} + e^{-\left(\frac{x-\mu}{\beta}\right)}\right)}$$

TU Delft

And enjoy the journey!