# Modelling, Uncertainty and Data for Engineers (MUDE)

# Week 1.4 : Univariate continuous distributions

Max Ramgraber

based on the slides of Patricia Mares Nasarre

TU Delft

Welcome to...

MUDE

Modelling, Uncertainty, and Data for Engineers

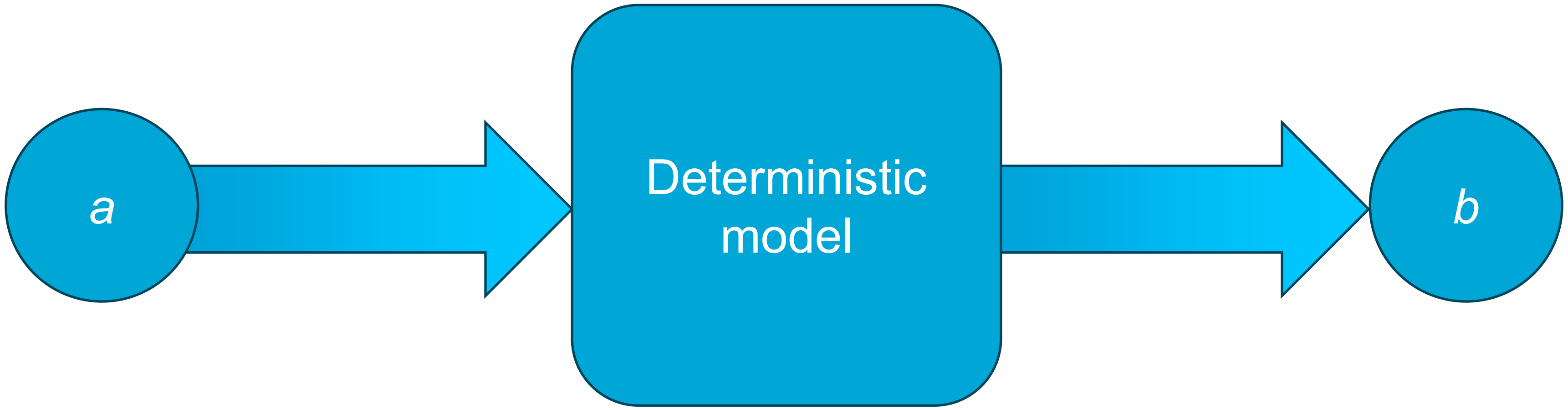
WEEK 4!

TU Delft

# Deterministic and stochastic models

## Deterministic models

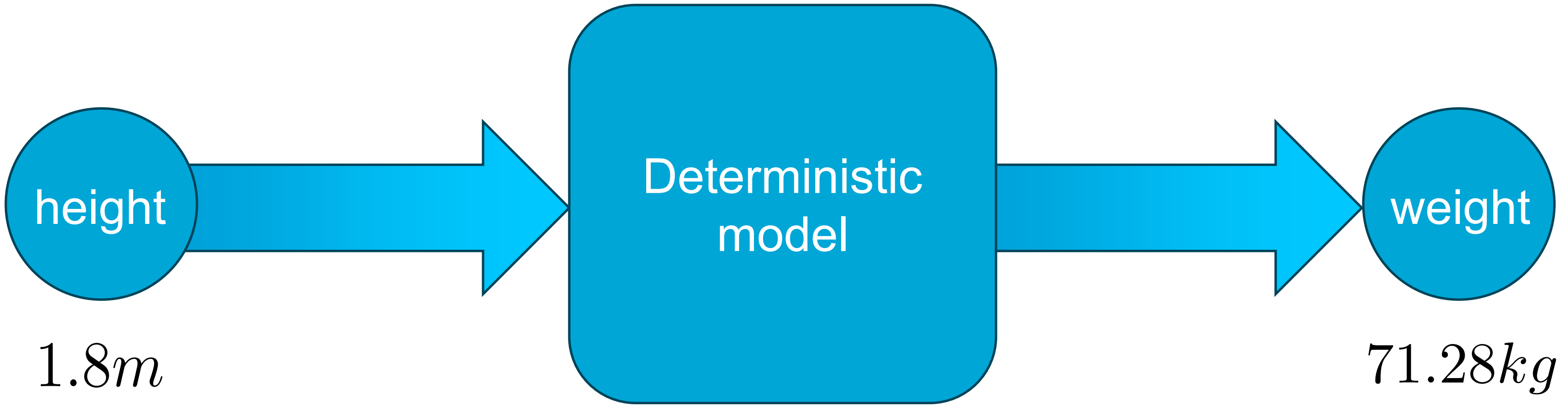
In a **deterministic** model: if the input is ‘*a*’, the output will always be ’*b*’.



# Deterministic and stochastic models

## Deterministic models

A deterministic model to predict weight from height: $\text{weight} = 22.0 \cdot \text{height}^2$



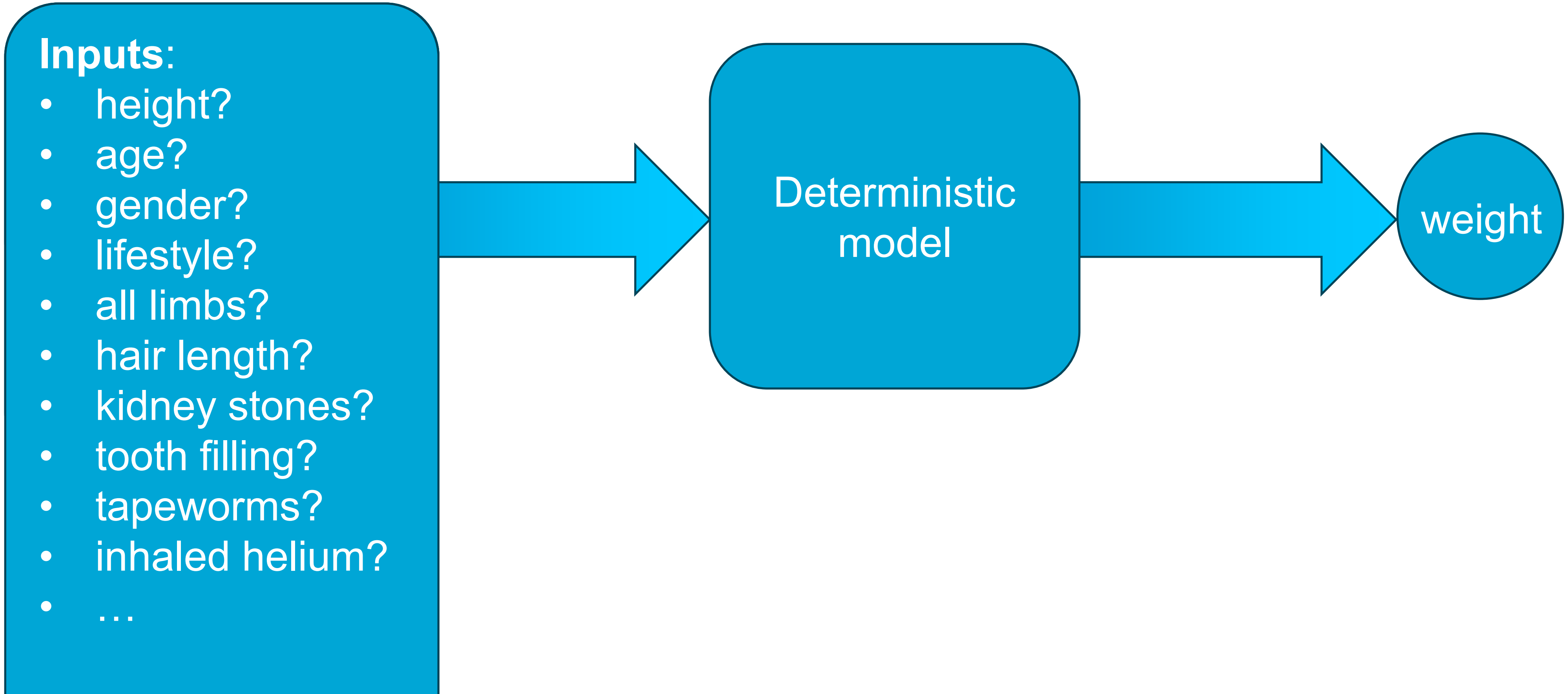

~~**Conclusion**: *if you are 1.80m tall, you weigh 71.28kg*~~

**Nonsense**

# Deterministic and stochastic models

## Deterministic models

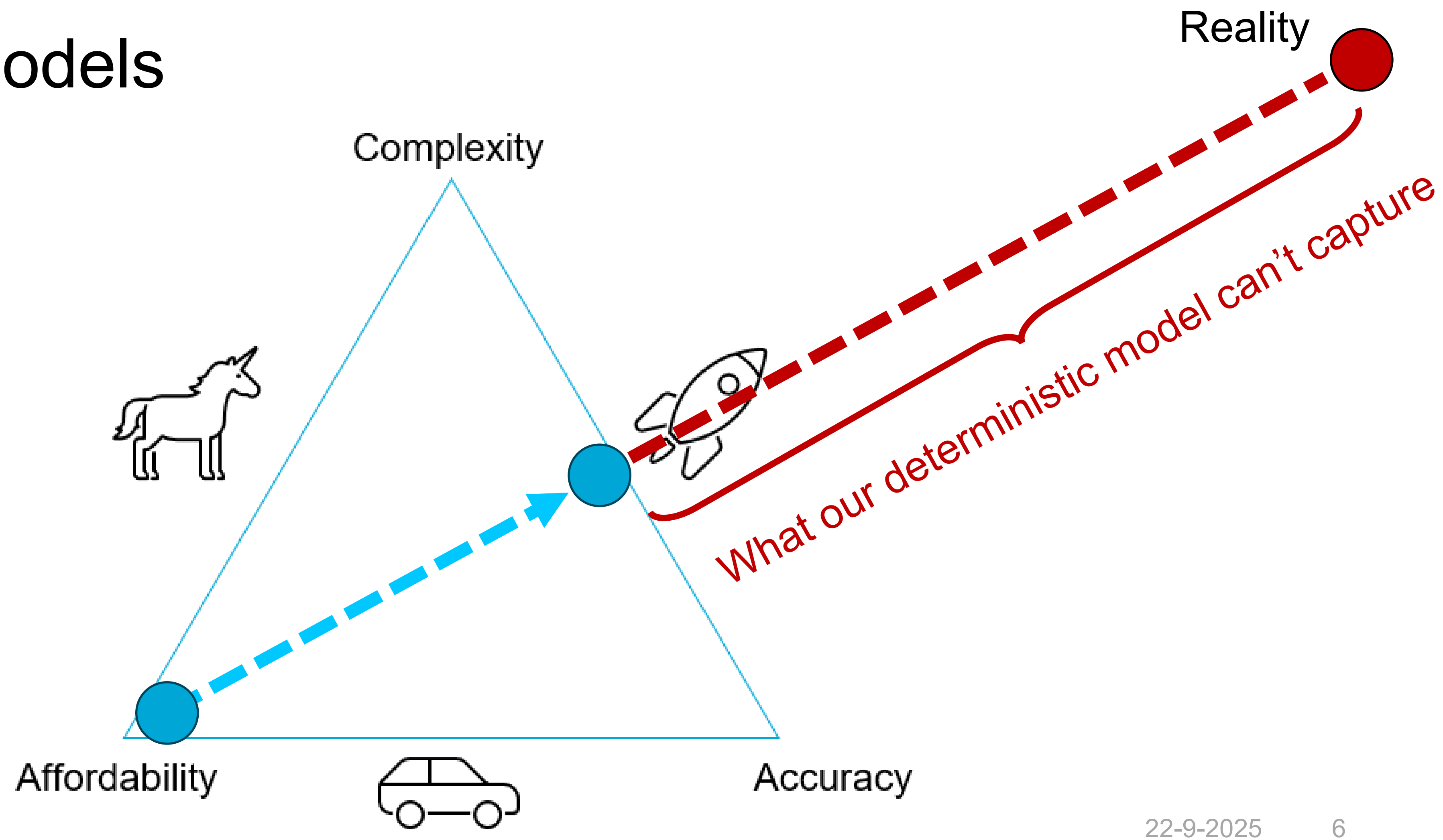
Is height really all we need?



# Deterministic and stochastic models

## Deterministic models

Reality is complicated.




TU Delft

# Deterministic and stochastic models

## Stochastic models

Stochastic models do not always return the same output for the same input.


$1.8m$

about $71.28kg$ -ish:

$82.73kg$

$65.21kg$

$107.60kg$

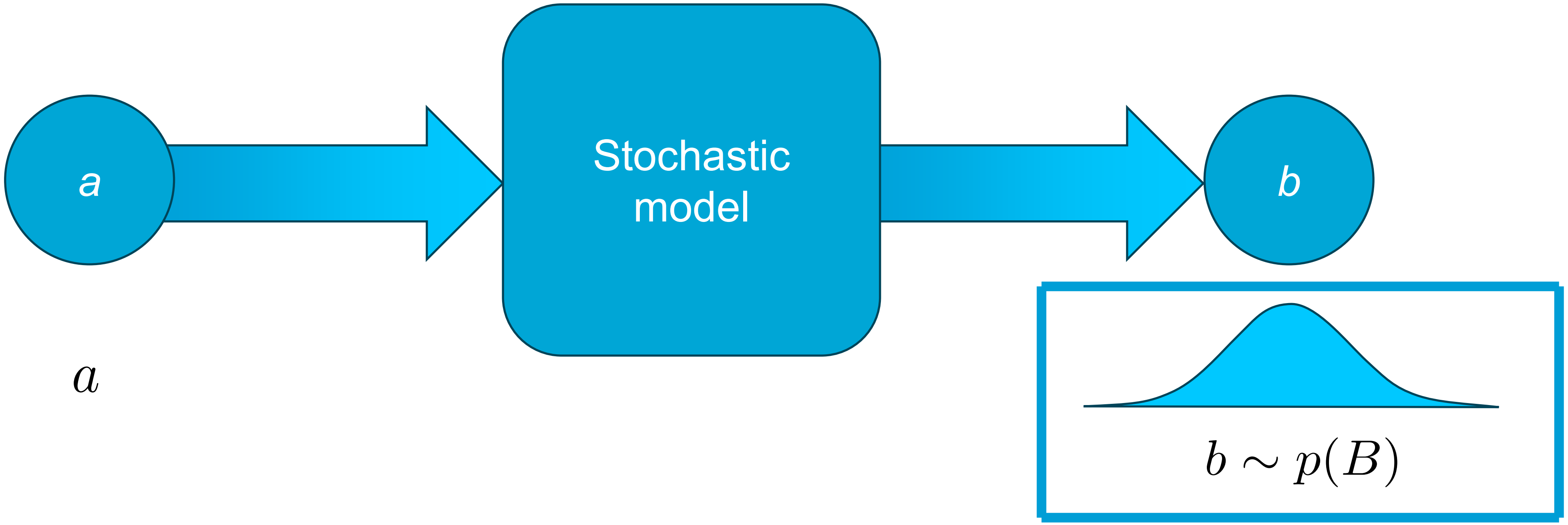
$\dots$

TU Delft

# Deterministic and stochastic models

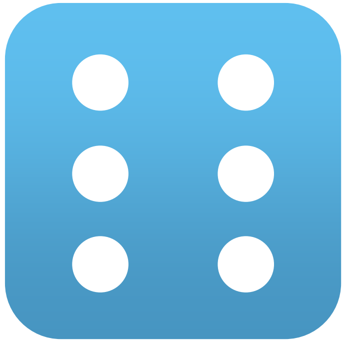
## Stochastic models

In a **stochastic** model: if the input is ‘*a*’, the output will be a random number from a distribution around ’*b*’.



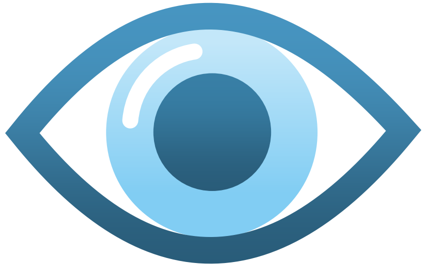

TU Delft

# Recall from lecture 1: three types of uncertainty



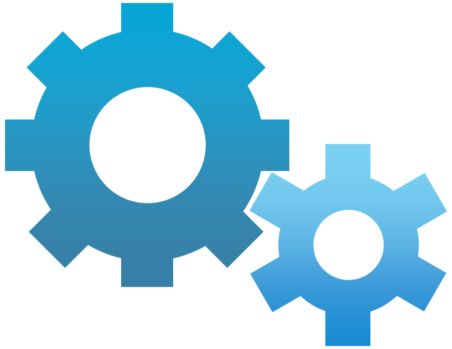
## Aleatoric

Inherent, irreducible uncertainty in the system



## Epistemic

Lack of knowledge, limits to what we can measure



## Error

Deficiency in modelling and simulation

No matter the source of the uncertainty, we represent it as a **probability distribution**.

# Probability distribution functions

# What is a univariate continuous distribution?

## Univariate

A univariate function only takes a single variable as input.

## Continuous

A continuous variable can be defined to an arbitrary precision.

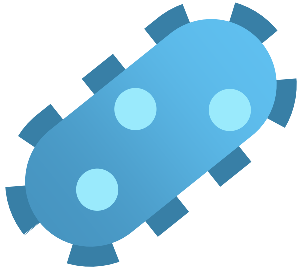
## Distribution

A function that assigns probability (densities) to a random variable.

# Continuous distribution functions - RVs

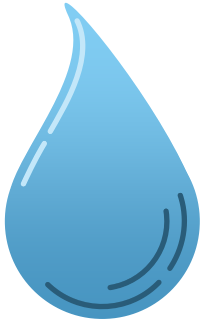
A **continuous random variable** $X$ assigns a continuous numerical value to an outcome of a random process.

Examples:

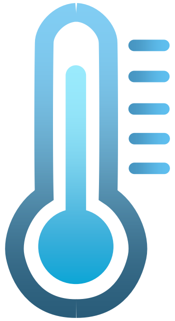


viral concentration

cloud coverage



precipitation



temperature

A **probability distribution** relates the numeric value of a random variable to a probability.

# Continuous distribution functions – Axioms

Recall the **axioms of probability** (informally summarized):

1. The probability of any outcome is non-negative.

$$P(X_i) \geq 0$$

2. The probability of all mutually exclusive outcomes must sum to one.

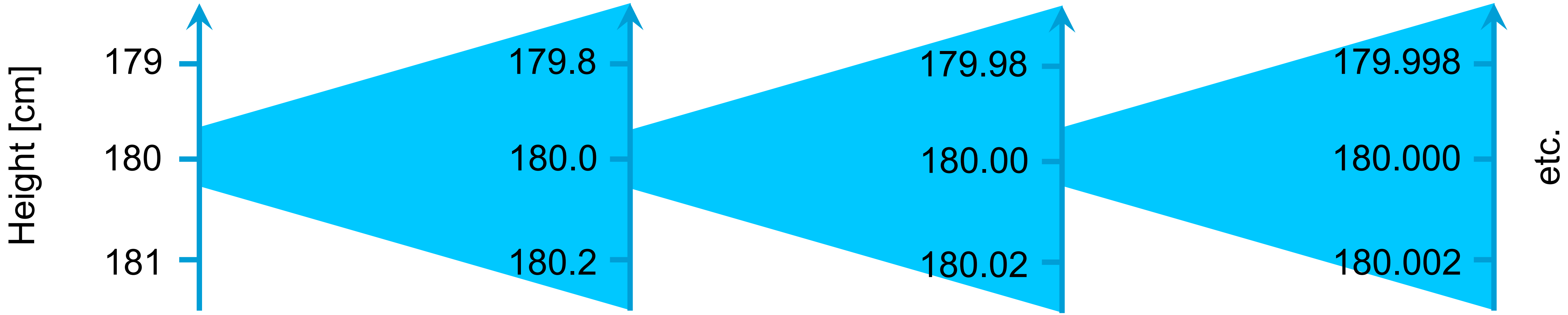
$$\sum_i P(X_i) = 1$$

3. The probability of a union of mutually exclusive outcomes is the sum of their probabilities.

$$P(X_1 \text{ or } X_2) = P(X_1) + P(X_2)$$

# Continuous distribution functions – PDF

For **continuous** random variables, there are **infinitely many** mutually exclusive outcomes.




Since the probability of each outcome must be non-negative (first axiom), and the sum of these infinitely many probabilities must be one (second axiom), the probability of each outcome must be **infinitesimally small**.

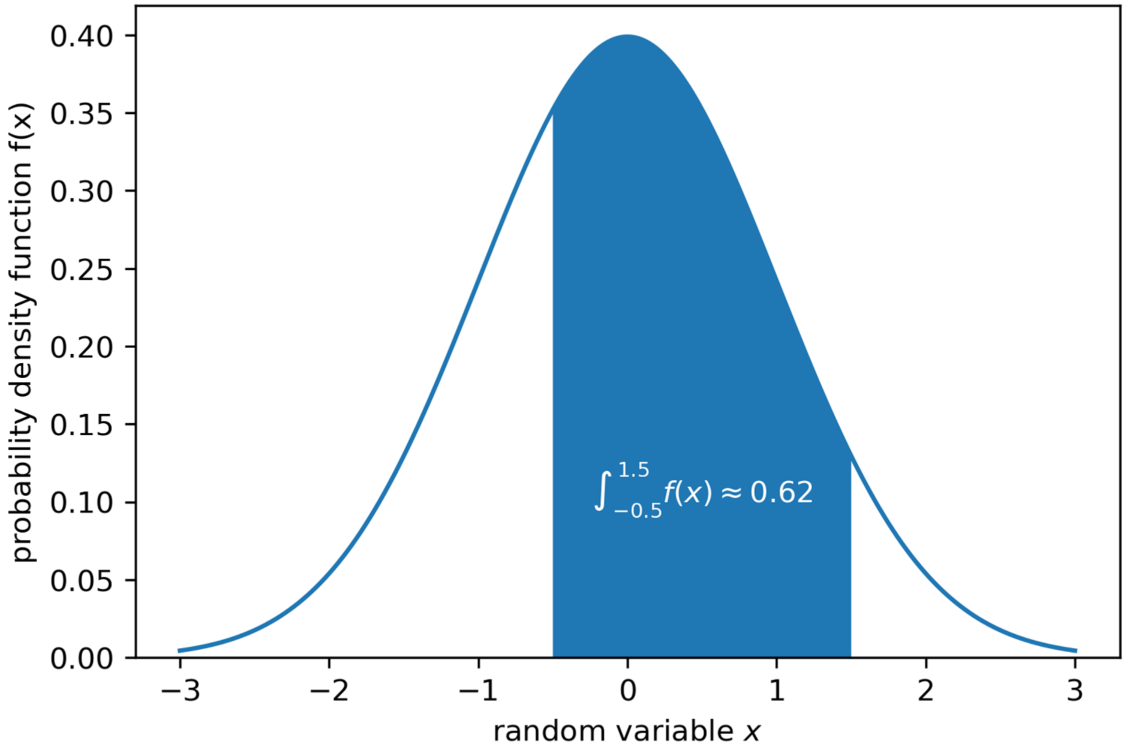
# Continuous distribution functions – PDF

In consequence, we define a **probability density function** (PDF). We can obtain probabilities from this PDF through integration.

This is a **Gaussian PDF**:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma_2}}$$

Integrating this PDF over a range of values returns the probability of x taking on values in this range.

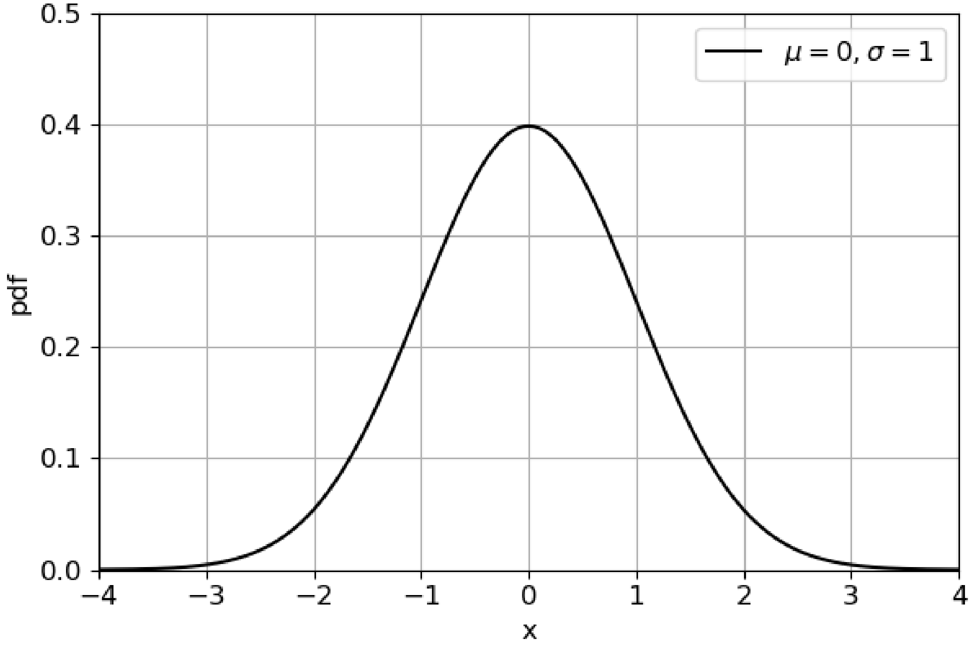


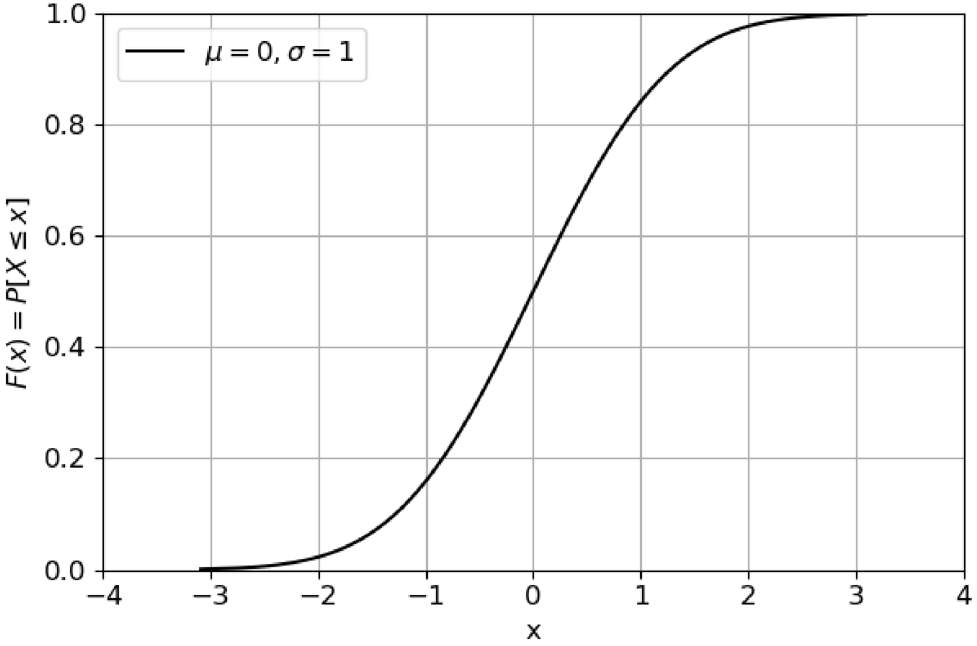
# From PDF to CDF

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

CDF of the Gaussian distribution

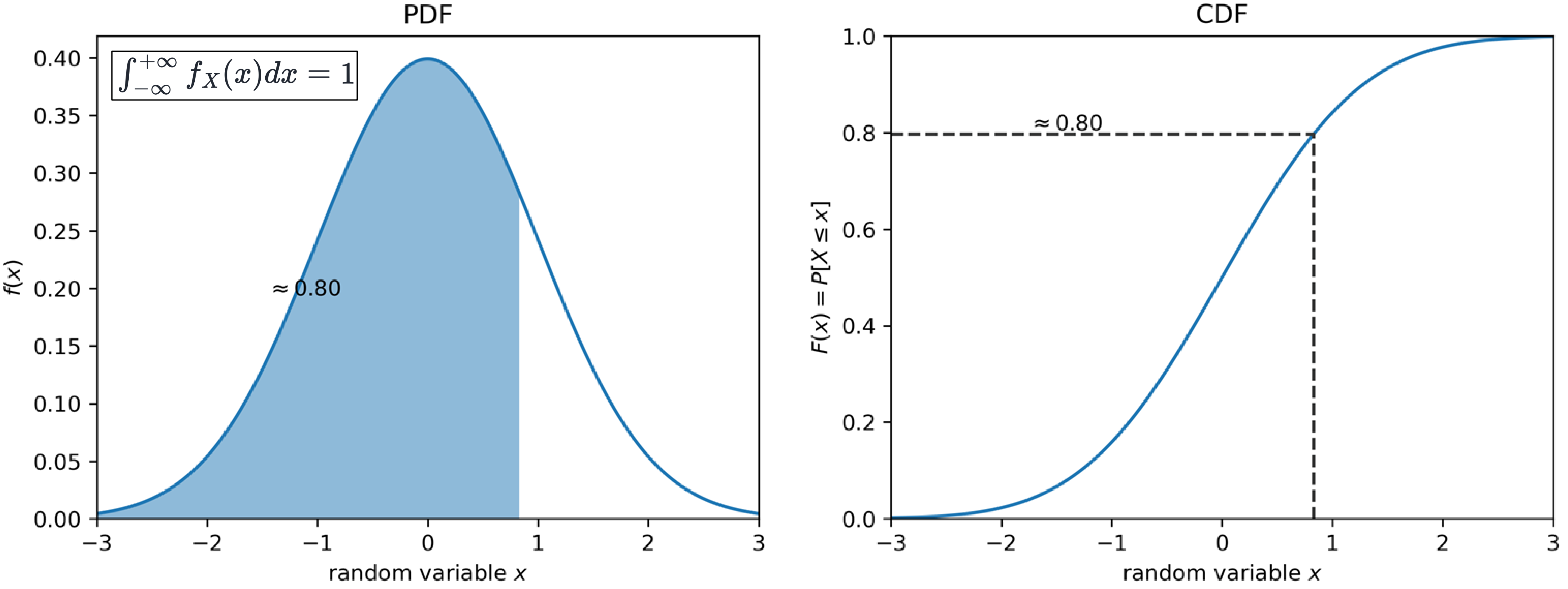
$$F(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$$




TUDelft

# From PDF to CDF – non-exeedance

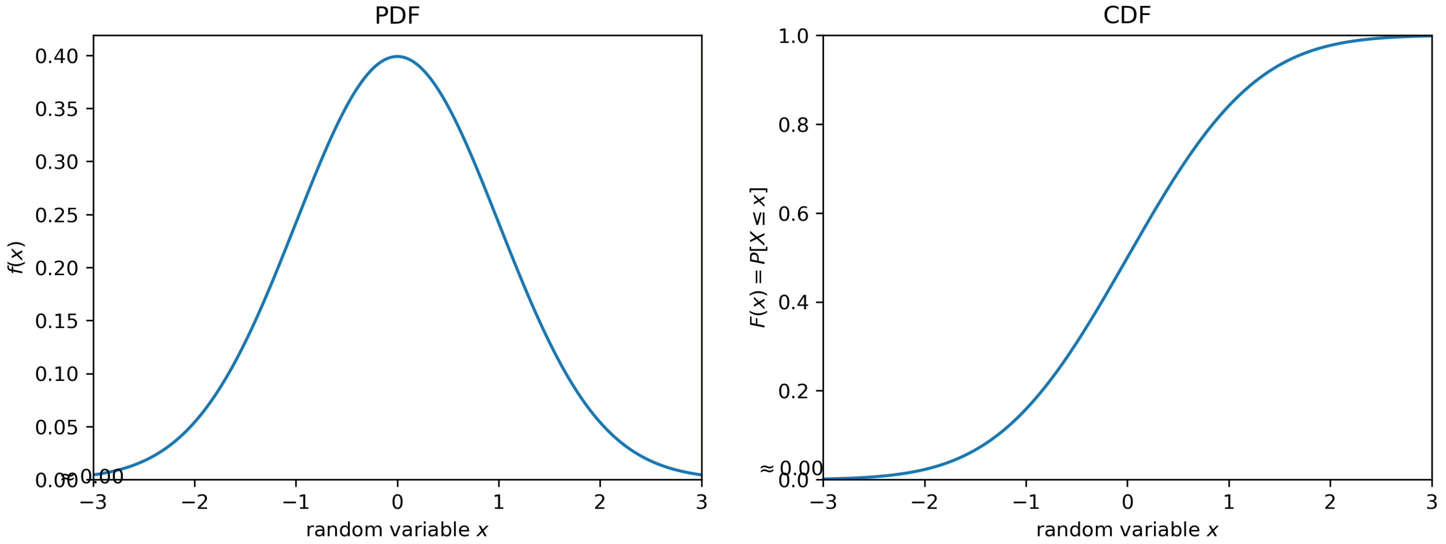


# From PDF to CDF

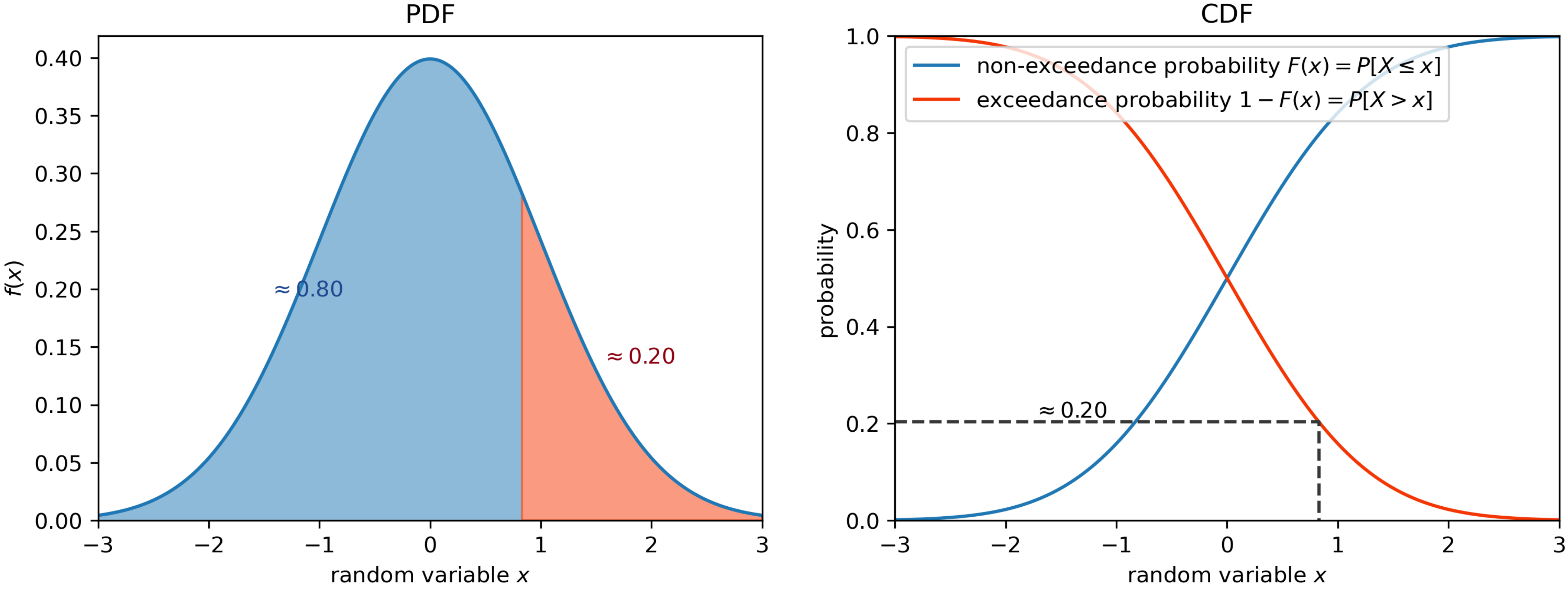
A CDF integrates pdf from $-\infty$ to $x$.

Integrating a PDF over an interval represents the probability that a value from the distribution will fall **within that interval**.

In consequence, the CDF returns the **non-exceedance probability** $P(X \leq x)$, which is the probability that a random sample $X$ will have a smaller or equal value to $x$.



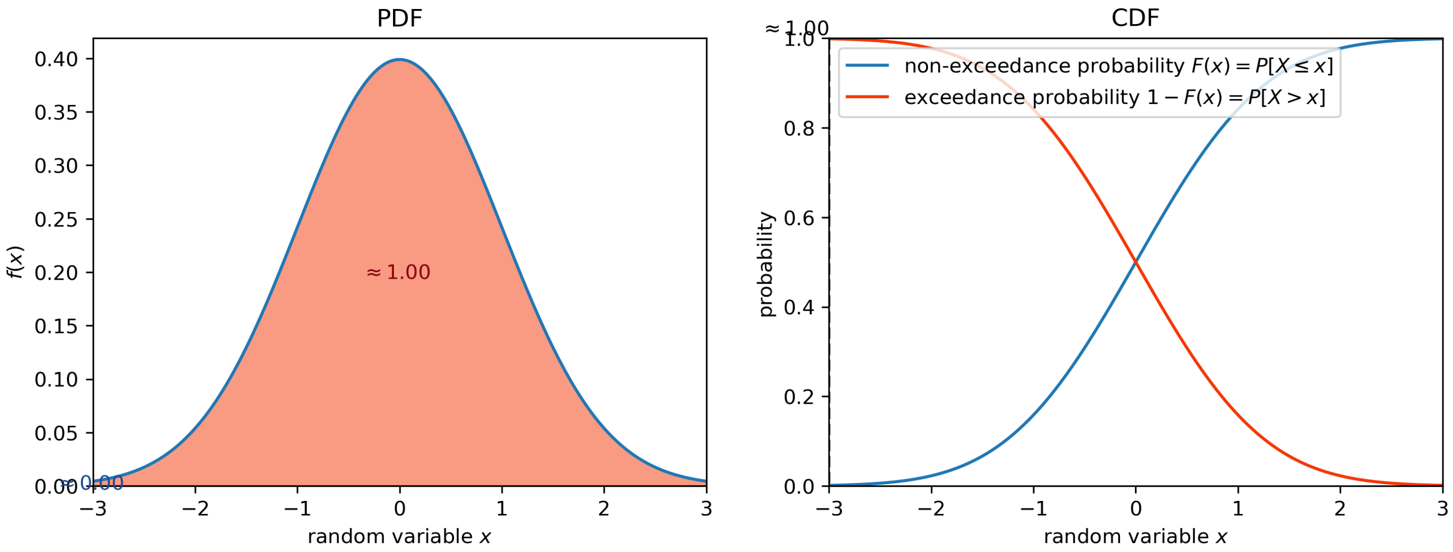
# From PDF to CDF - exceedance



# From PDF to CDF

The **exceedance probability** $P(X > x)$ is computed as $1 - F(x)$ and represents the probability that a random sample $X$ will have a larger value than $x$.

Exceedance probabilities are important for **safety design**. For instance, when designing a dike, we might want to ensure that the wave height exceedance probability $P(\text{wave height} > \text{wave height})$ is low.

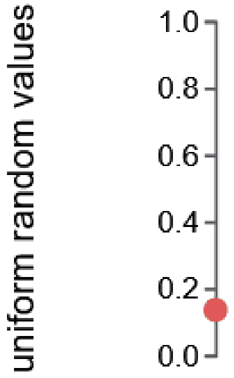


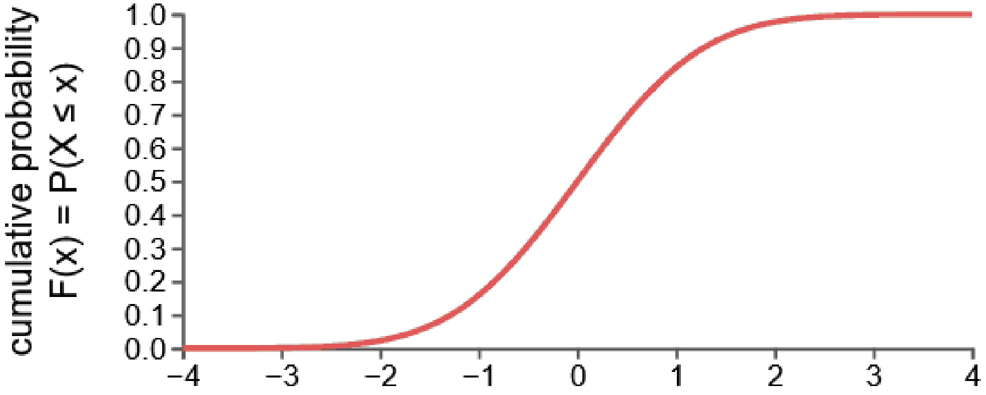
# How can we sample from a PDF?

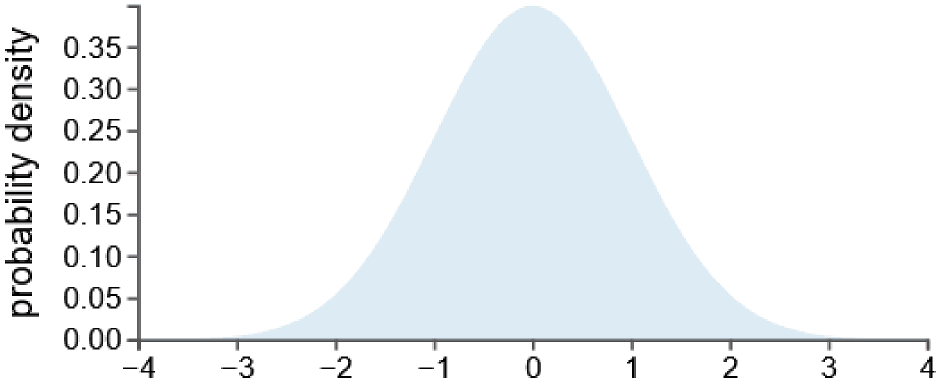
Computers only generate **pseudo-random** uniform samples.

Sending these samples through the **inverse CDF** generates samples from the corresponding PDF!

This interactive element is also available in the MUDE book.



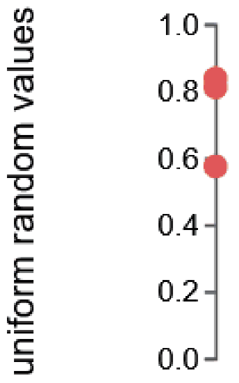



# How can we sample from a PDF?

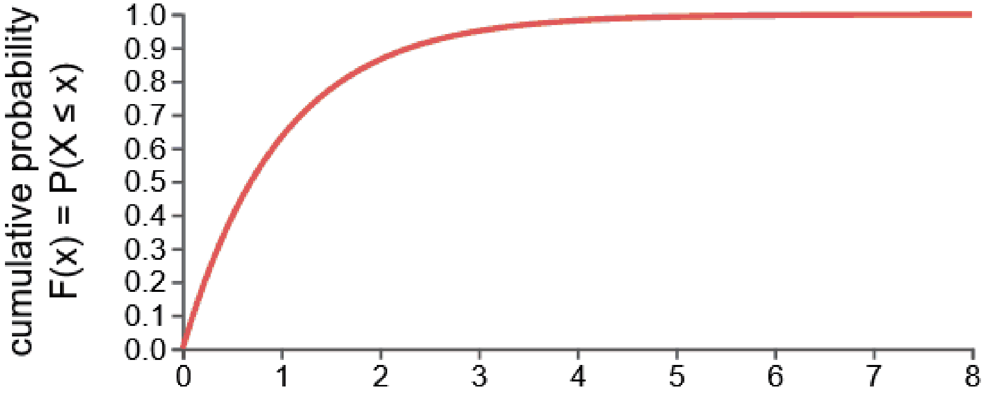
Computers only generate **pseudo-random** uniform samples.

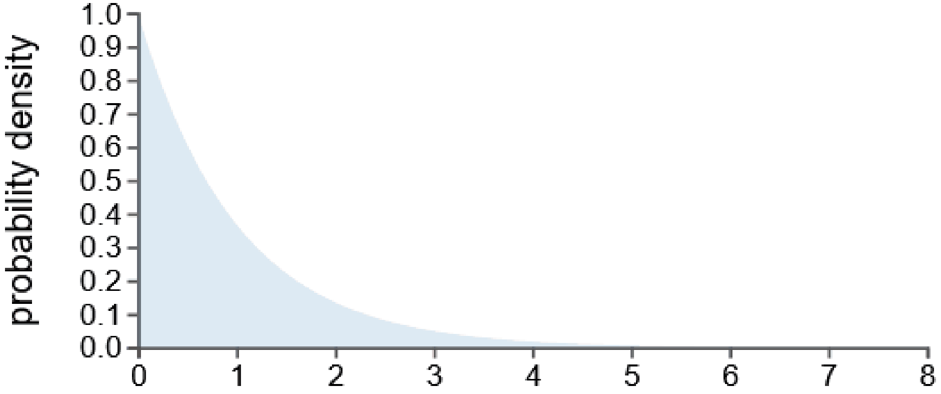
Sending these samples through the **inverse CDF** generates samples from the corresponding PDF!

This interactive element is also available in the MUDE book.




Different CDFs sample different distributions!




TU Delft

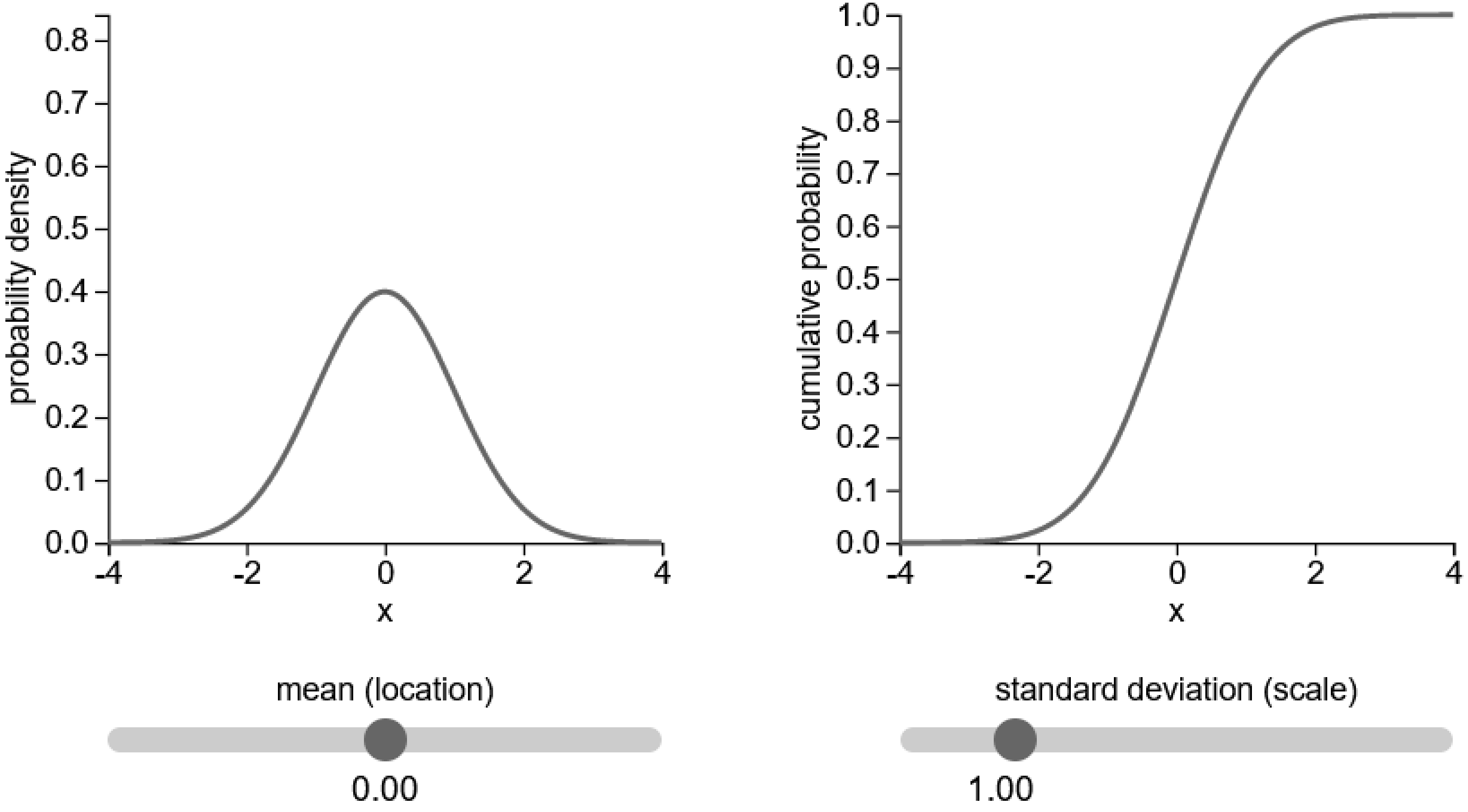
# Parameters in PDF and CDF – Gaussian distribution

PDF

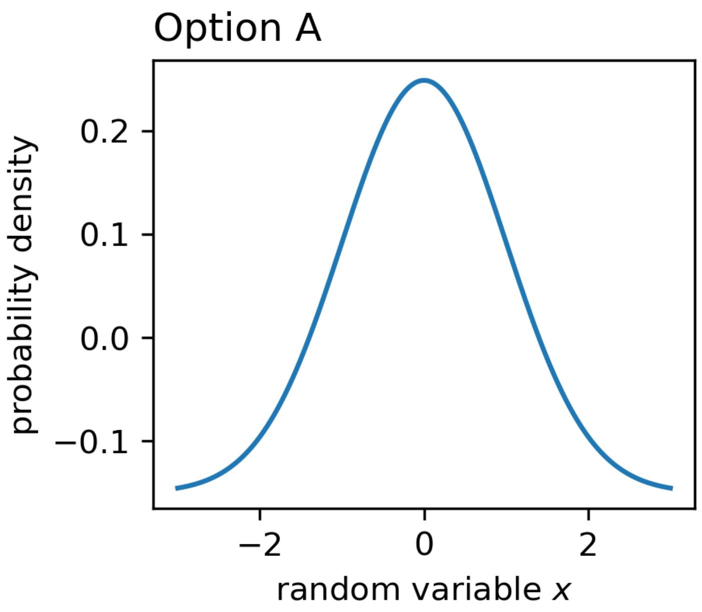
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

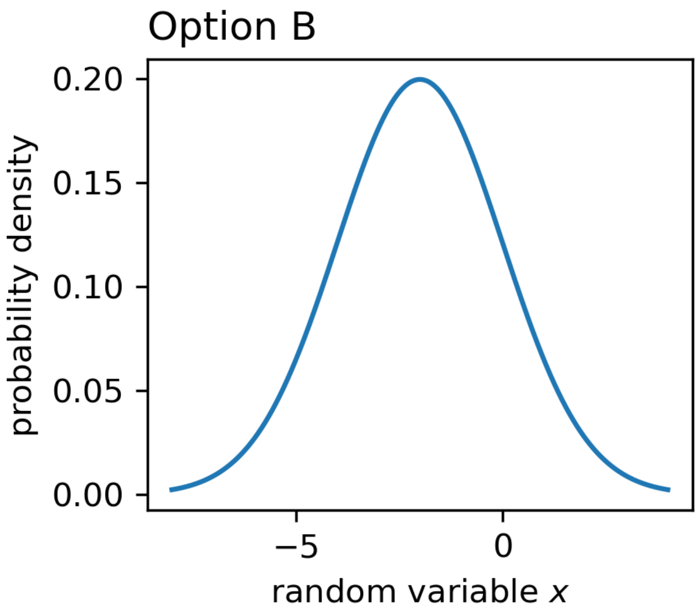
CDF:

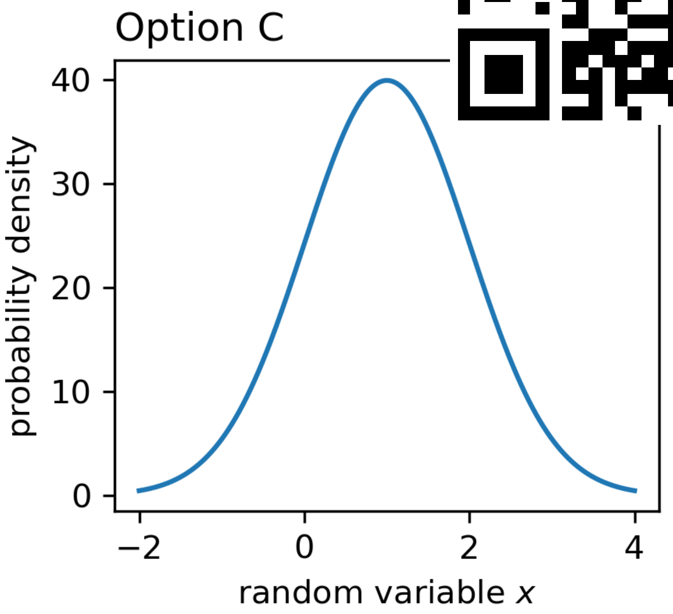
$$F(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$$



# Quiz: Which of these functions is a valid PDF?



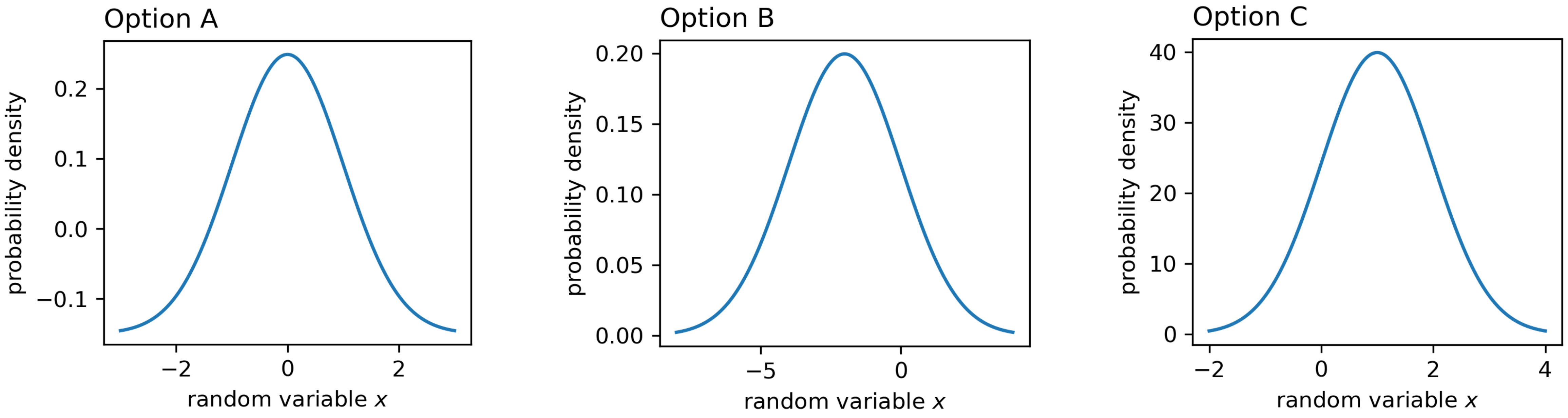




Option A

Option B

Option C

# Quiz: Which of these functions is a valid PDF?

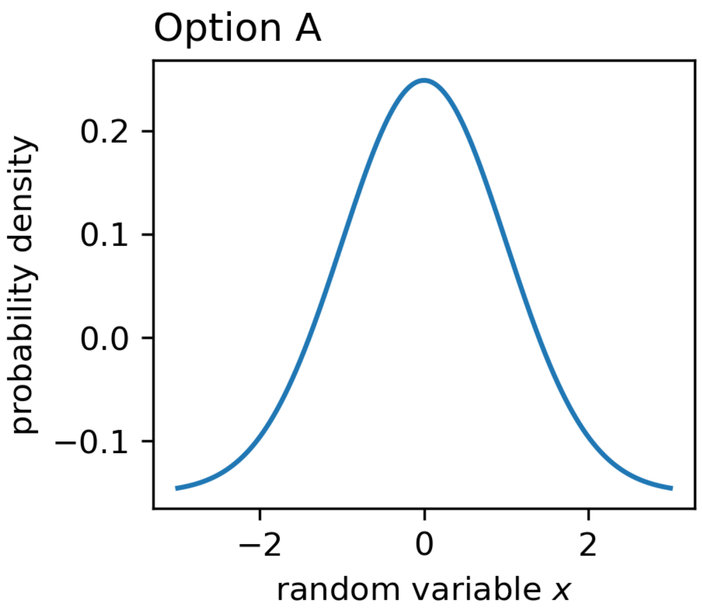



Option A 21.93%

Option B 73.68%

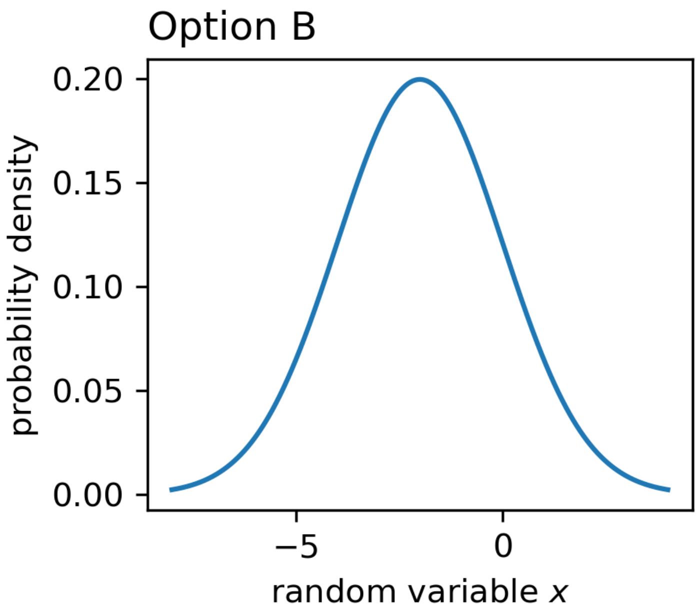
Option C 4.39%

# Quiz: Which of these functions is a valid PDF?



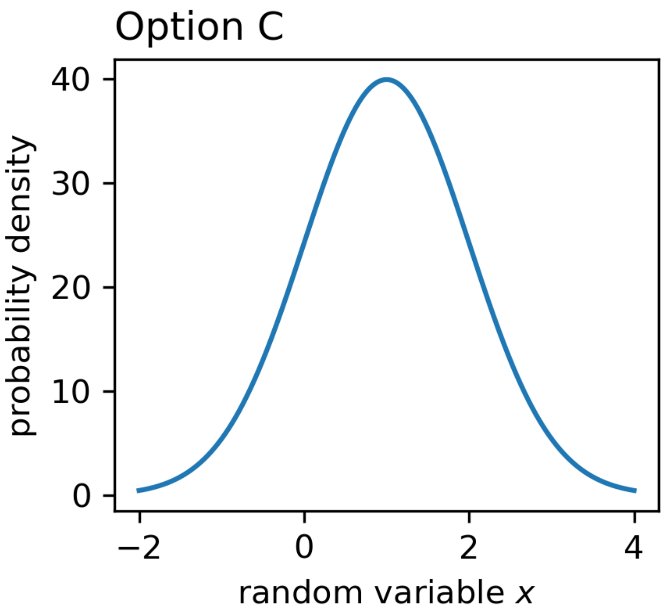

**Not valid.**

This function has negative densities.




**Valid.**

This function has only positive densities and integrates to one.




**Not valid.**

This function does not integrate to one.

# Empirical distribution functions

# Continuous distribution functions

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

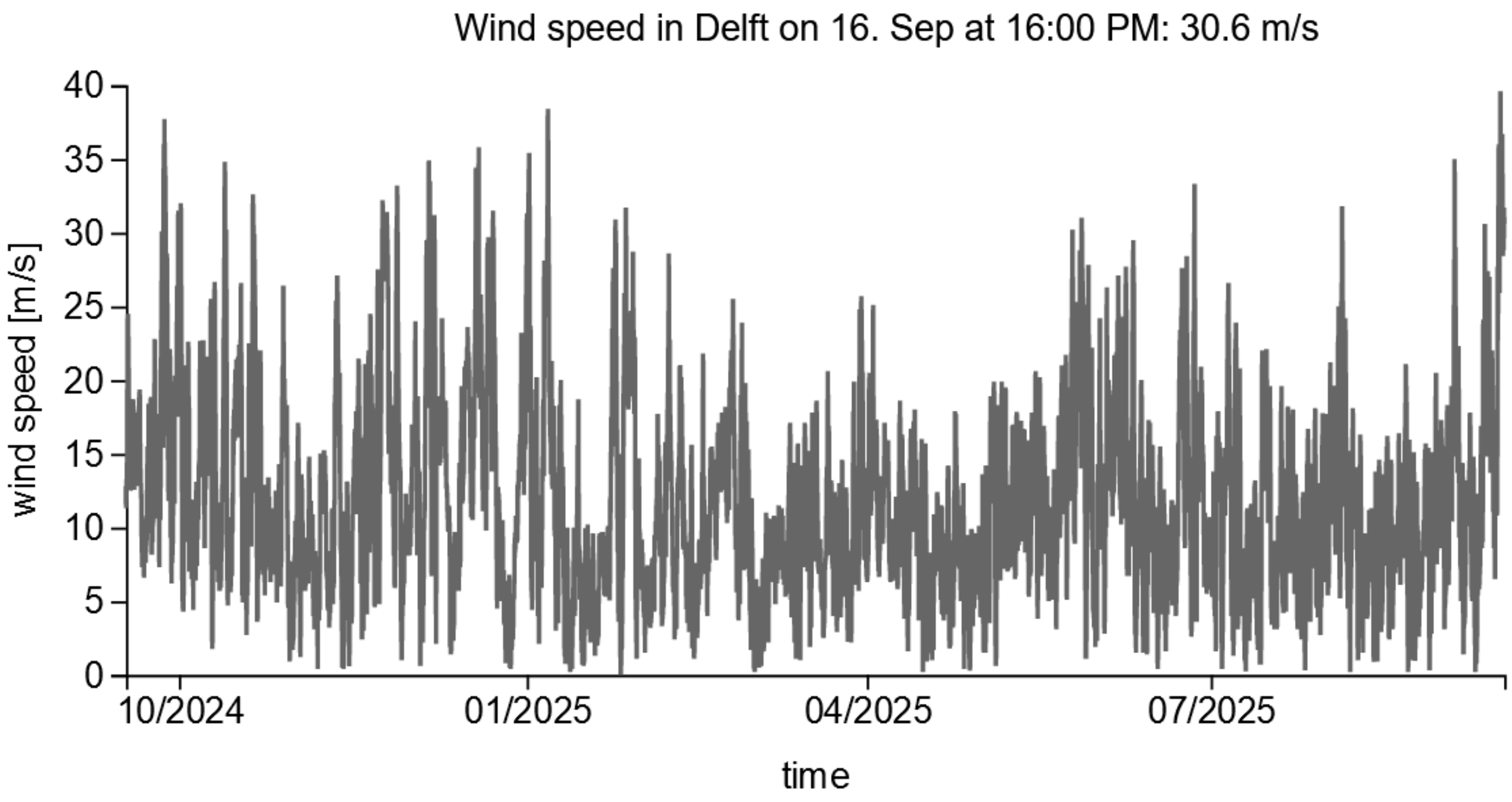
***But what do I want to model?***

Observations → Empirical distribution function

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

# Empirical distribution functions

We can define an empirical PDF and empirical CDF from our observations. How? Let's see it with an **example**!

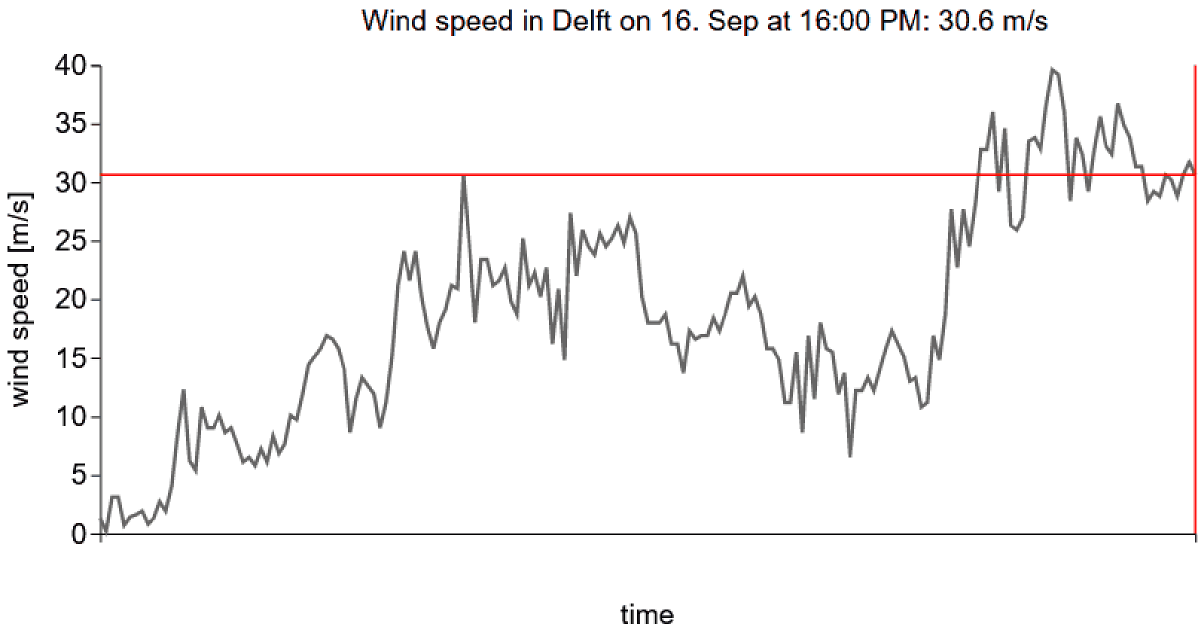



This is the wind speed in Delft at 10 m height over the past 365 days, taken from OpenMeteo.com.

You can access an **interactive real-time** version of this in the book!

# Empirical CDF

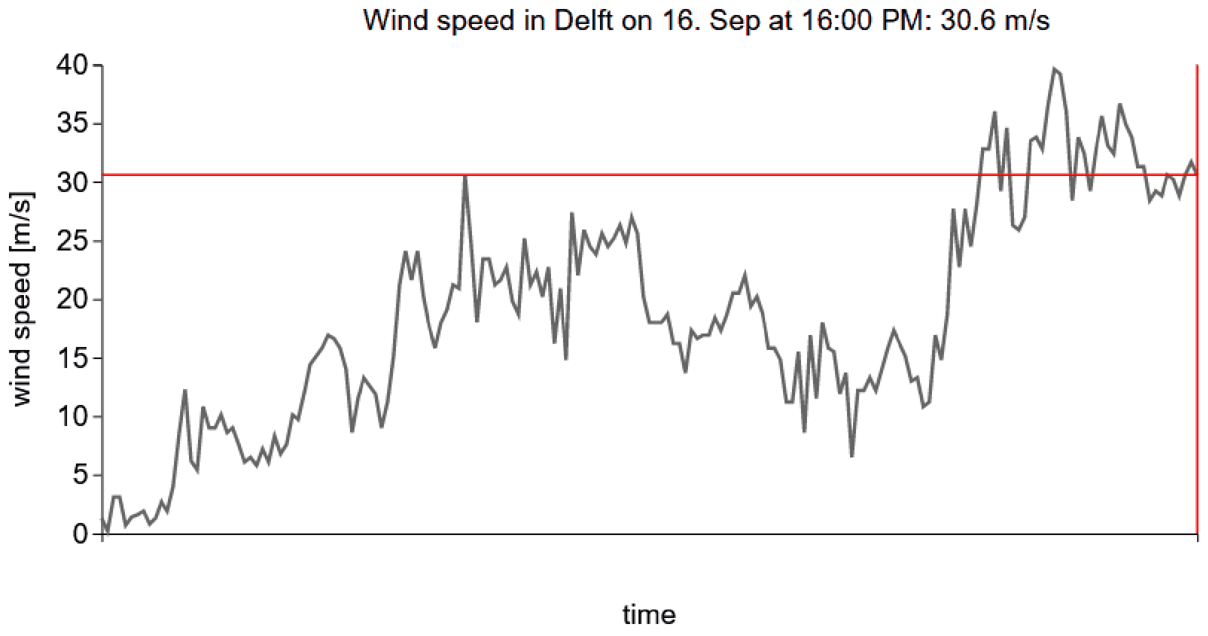
We need to assign a **non-exceedance probability** to each observation.




```
>> read observations
```

# Empirical CDF

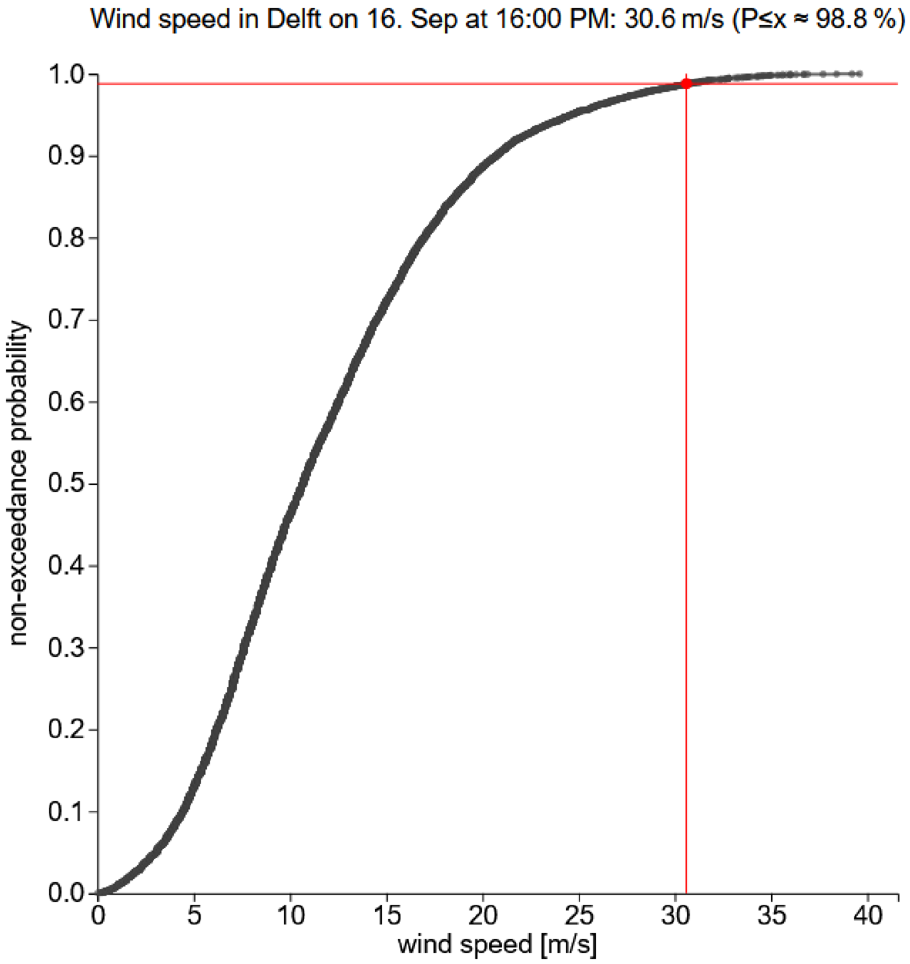
We need to assign a **non-exceedance probability** to each observation.




```
>> read observations

>> x = sort observations in ascending order

>> length = the number of observations

>> non-exceedance probability = (range of integer values from 1 \ to length) / (length + 1)
```

TU Delft

# 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

>> **non-exceedance probability** = (range of integer values from 1 \ to length) / (length + 1)

>> **Plot** x versus non-exceedance probability

# Empirical CDF

Let's do it slowly!

**Length = 5**

| x |
|---|
| 3.2 |
| 4.5 |
| 3.8 |
| 7.5 |
| 2 |

```
>> read observations

>> x = sort observations in ascending order

>> length = the number of observations

>> non-exceedance probability = (range of integer values from 1 \ to length) / (length + 1)

>> Plot x versus non-exceedance probability
```

TUDelft

# Empirical CDF

Let's do it slowly!

**Length = 5**

| x | Sort(x) | Rank |
|---|---|---|
| 3.2 | 2 | 1 |
| 4.5 | 3.2 | 2 |
| 3.8 | 3.8 | 3 |
| 7.5 | 4.5 | 4 |
| 2 | 7.5 | 5 |

```
>> read observations

>> x = sort observations in ascending order

>> length = the number of observations

>> non-exceedance probability = (range of integer values from 1 \ to length) / (length + 1)

>> Plot x versus non-exceedance probability
```

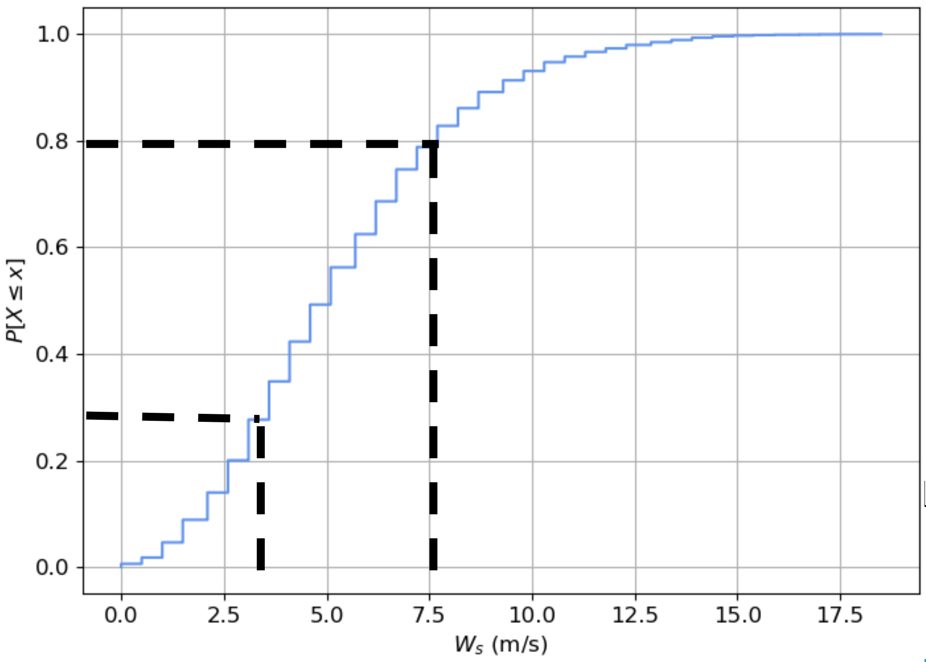
TU Delft

# Empirical CDF

Let's do it slowly!

**Length = 5**

| x | Sort(x) | Rank | Rank/length + 1 |
|---|---|---|---|
| 3.2 | 2 | 1 | 1/6 = 0.17 |
| 4.5 | 3.2 | 2 | 2/6 = 0.33 |
| 3.8 | 3.8 | 3 | 3/6 = 0.5 |
| 7.5 | 4.5 | 4 | 4/6 = 0.67 |
| 2 | 7.5 | 5 | 5/6 = 0.83 |



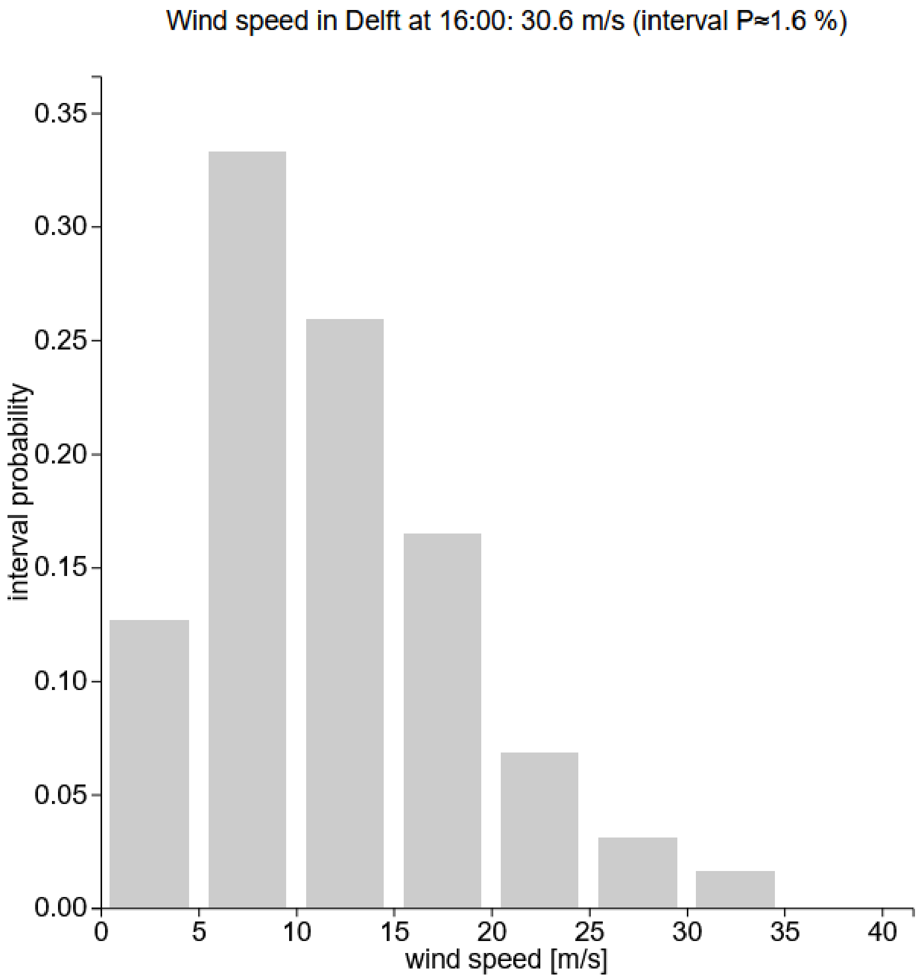

```
>> re
>> x                                                    ler
>> le
```

```
>> non-exceedance probability = (range of integer values from 1 \ to length) / (length + 1)
```

```
>> Plot x versus non-exceedance probability
```

# Empirical PDF

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




```
>> read observations

>> bin_size = 5  #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
```

# Let’s collect some data!

## We want to know you!

We would like to collect data about our students that we can use for teaching in future years. If you want to support us in this, please fill in this **anonymous poll** and tell us a little bit more about yourself.

Direct link:
https://forms.office.com/e/2yxYwYrrjQ

## Let’s have a break!

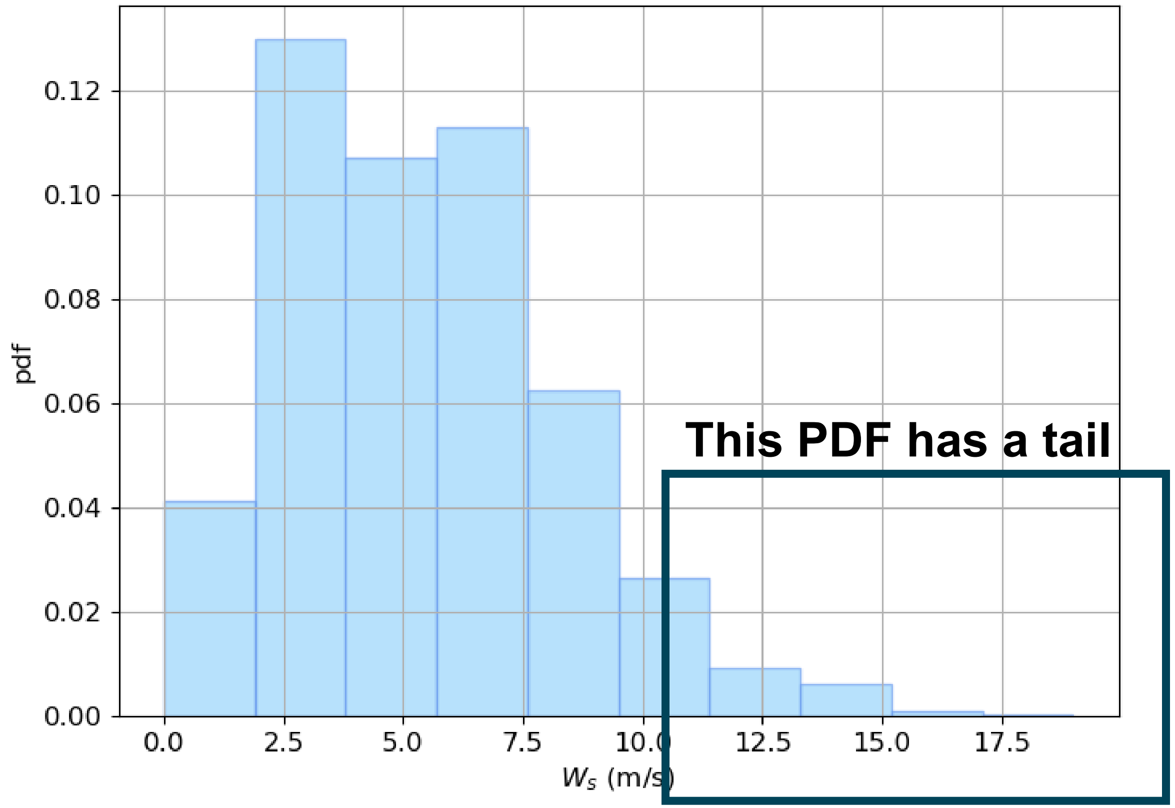
# Let’s continue

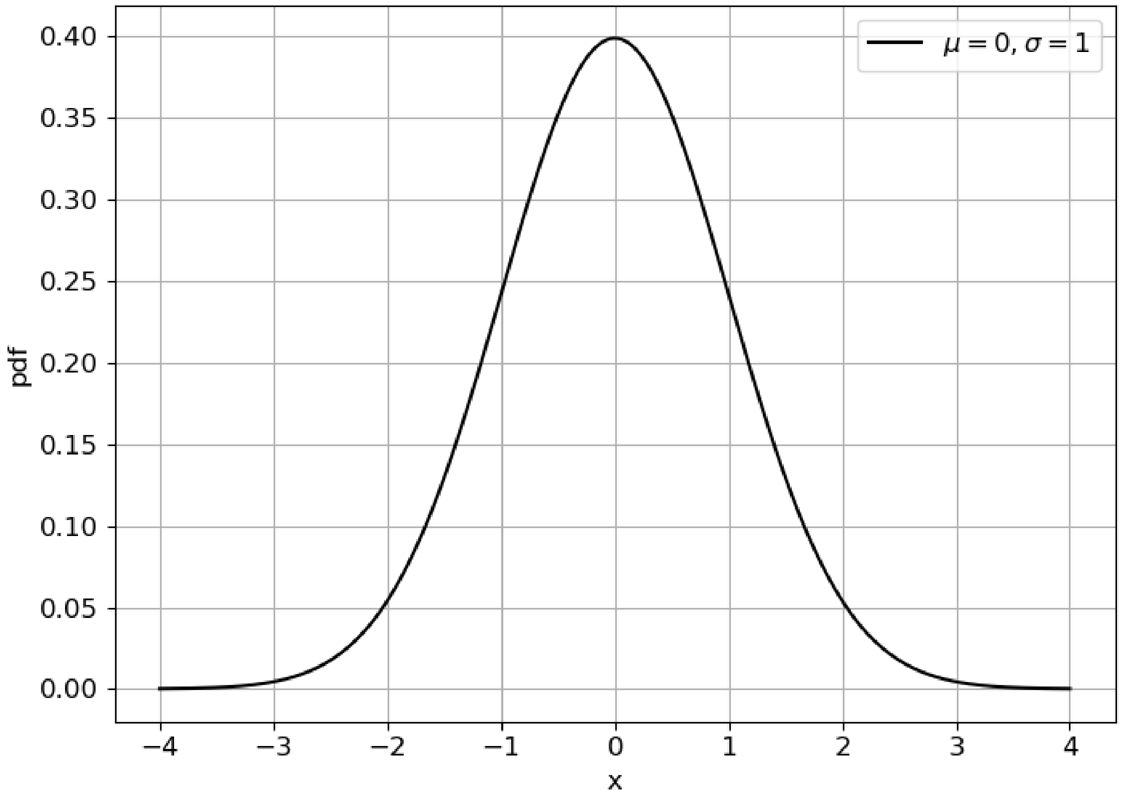
# Why non-Gaussian?

# Concept of tail

TU Delft

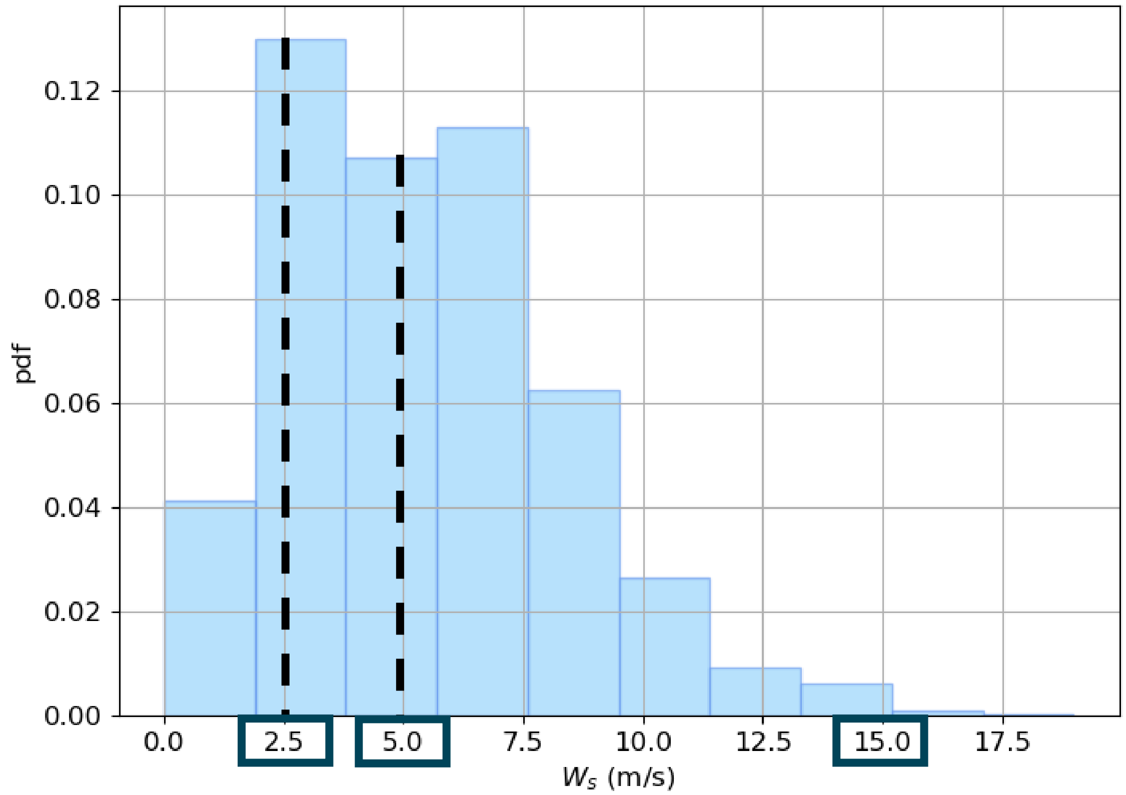
# Does this look Gaussian?




TU Delft

# Why is the tail important?



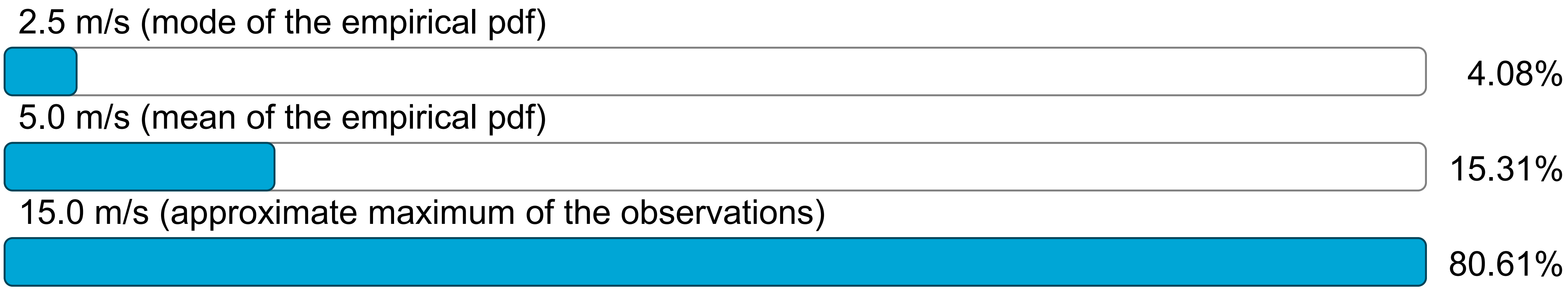

**Task:** You are designing a building against wind loading. Which value would you base your design on?

You vote!

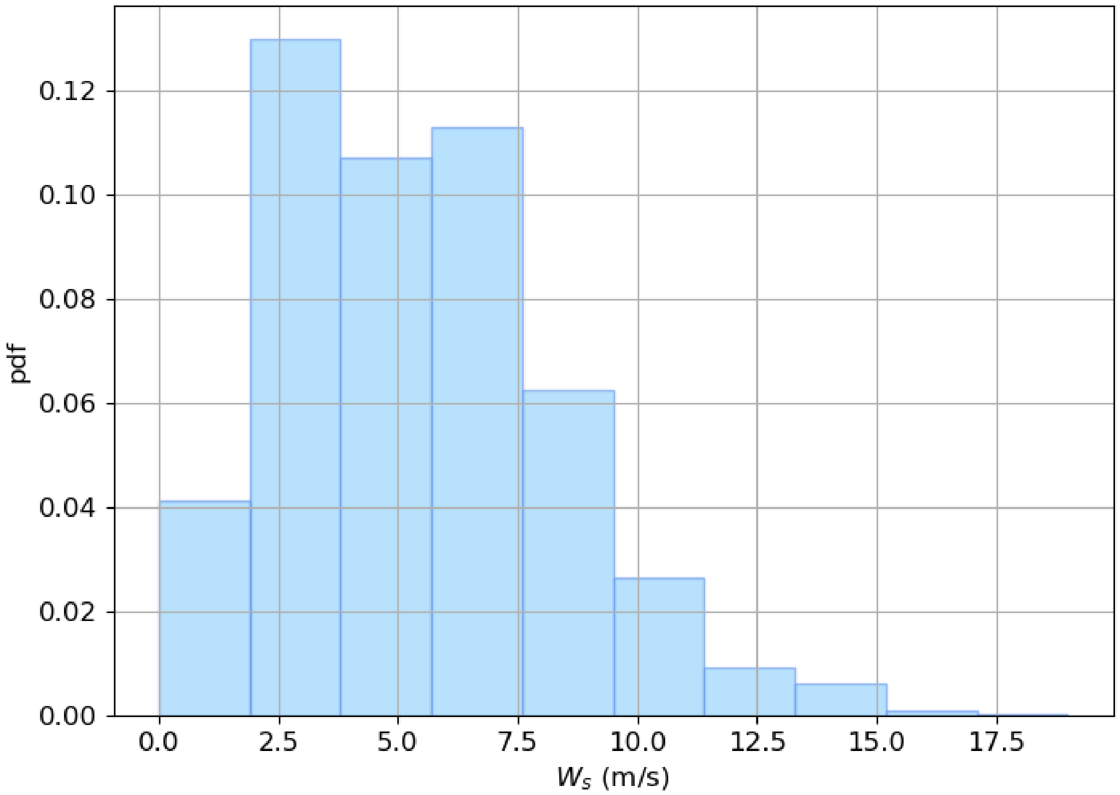
# Which design value would you choose?

| | |
|---|---|
| 2.5 m/s (mode of the empirical pdf) | 0% |
| 5.0 m/s (mean of the empirical pdf) | 0% |
| 15.0 m/s (approximate maximum of the observations) | 0% |

# Which design value would you choose?



# We typically design to withstand extreme values




We want the building to perform under **ordinary conditions** (i.e., around the central moments).

However, we also want the building to **withstand storms** – hence it is critical to plan for extremes!

Mind that 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

Textbook



## Gaussian distribution

Most of you should have already encountered the **Gaussian distribution** (also sometimes known as the **Normal distribution**) during your studies. This distribution is one of the most widely-used PDFs since it occurs commonly in nature and engineering and has many elegant mathematical properties. The PDF of the Normal distribution is given by

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

where $x$ is the value of the random variable and $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$ are the two parameters of the distribution, the mean and standard deviation. If we integrate the PDF, we obtain the CDF. In the case of the Normal distribution, there is no closed form of the CDF, but it can be expressed as

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

where $\mathrm{erf}$ denotes the **error function** given by

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

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

**There exist many more PDFs in the literature.**

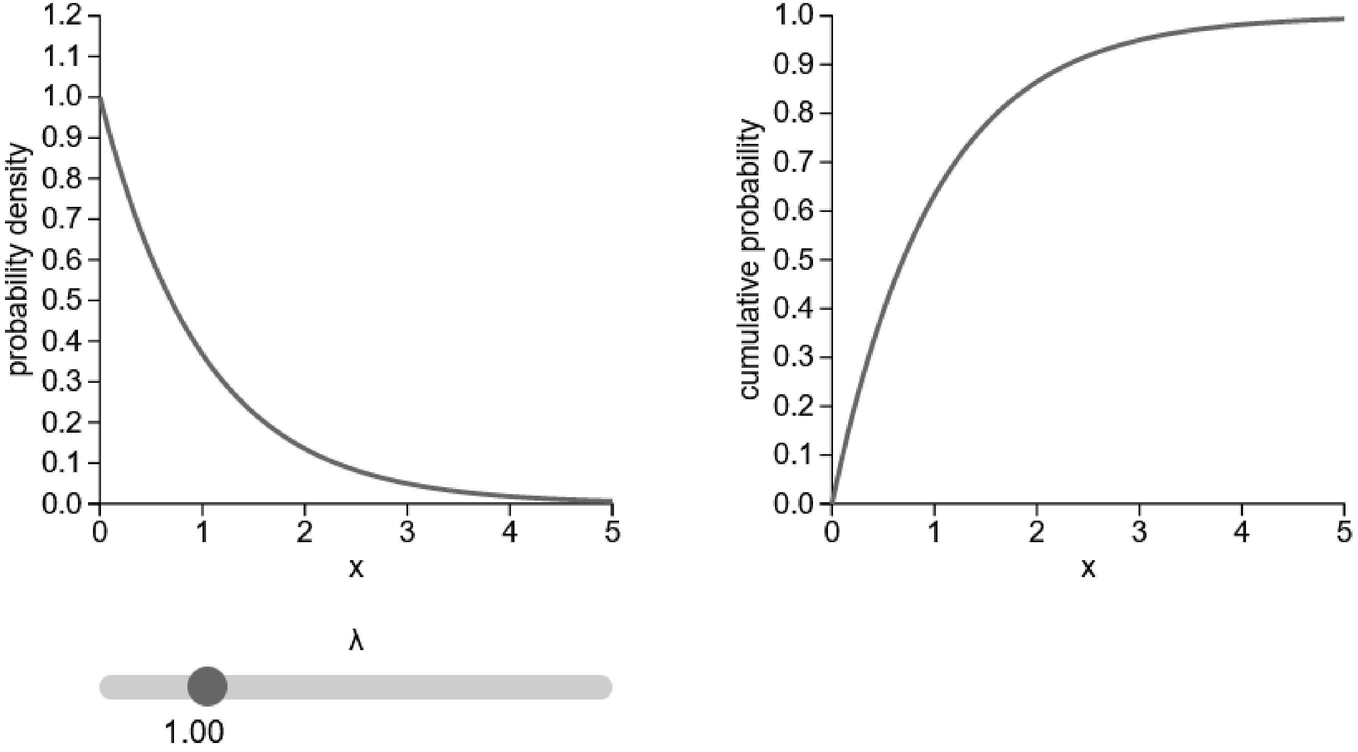
Read about the other PDFs in the book.

### What do I need to know?

- how the distribution looks (PDF/CDF),
- how it responds to changes in the parameters, and
- some basic properties like symmetry and bounds.

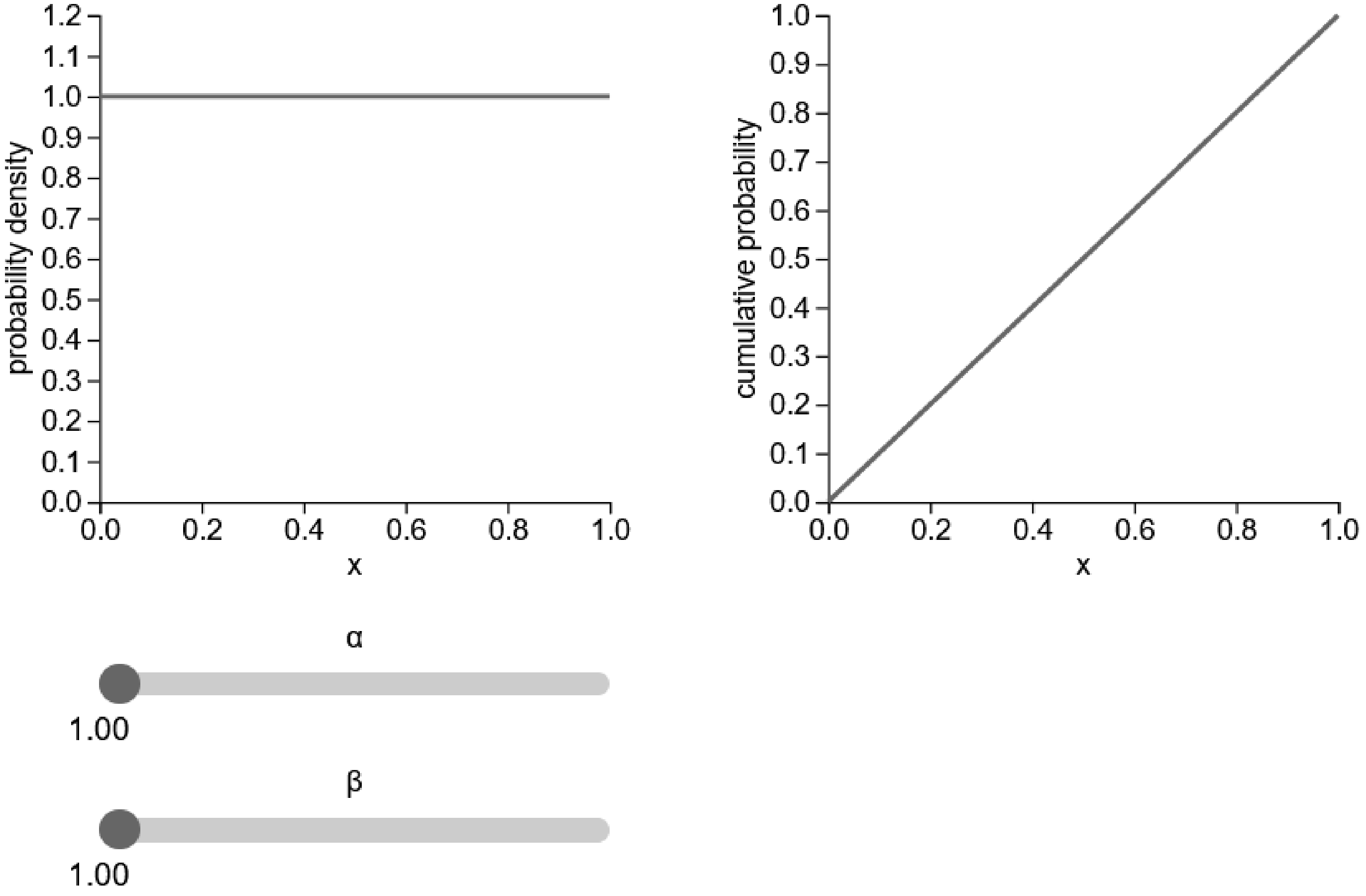
# Exponential distribution

- An exponential distribution has a single **rate parameter** $\lambda$.
- The distribution has a **left bound** and a **right tail**.
- The exponential PDF describes Poisson processes, which are **memoryless**. This means the chance of future events does not depend on the past.
- **Examples**: survival rate of a species, radioactive decay



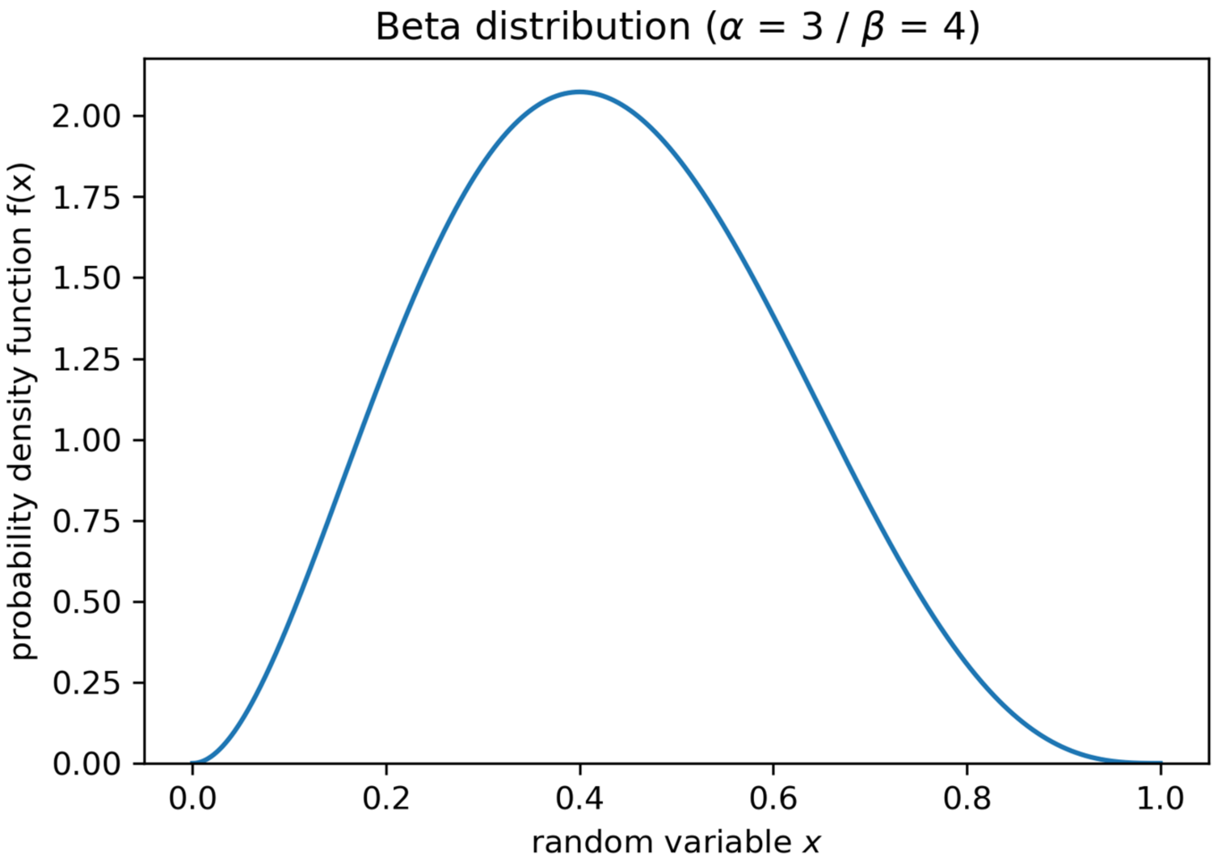
# Beta distribution

- A beta distribution is defined by **two parameters**: $\alpha$ and $\beta$
- A beta distribution has a **left bound** and a **right bound**. Depending on the parameters, it can be symmetric or skew in either direction.



# Beta distribution

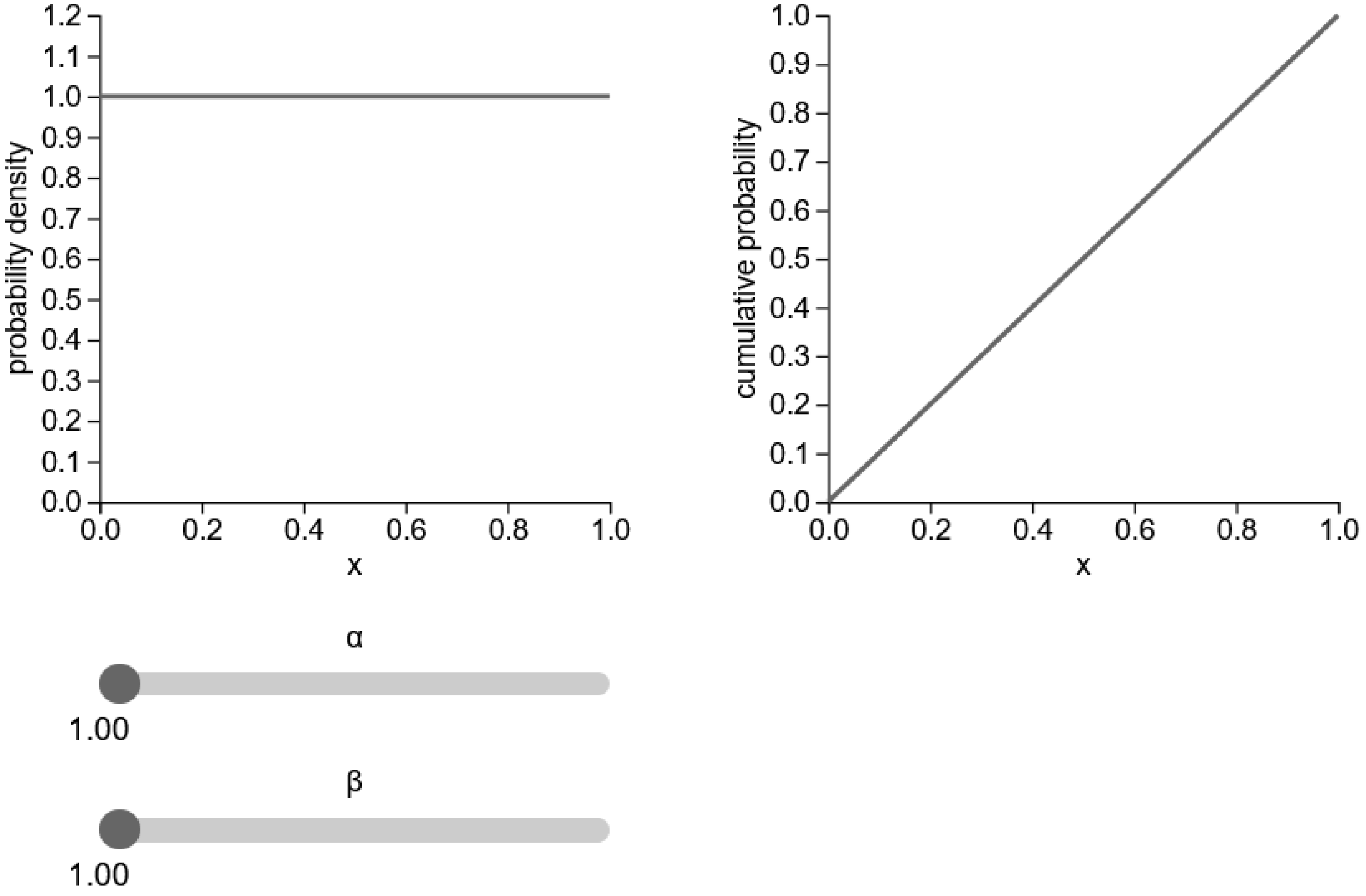
- A beta distribution is defined by **two parameters**: $\alpha$ and $\beta$
- A beta distribution has a **left bound** and a **right bound**. Depending on the parameters, it can be symmetric or skew in either direction.
- The beta PDF describes a distribution for the **expected value** of a Bernoulli process.




Coin flips: 0 0 1 0 1

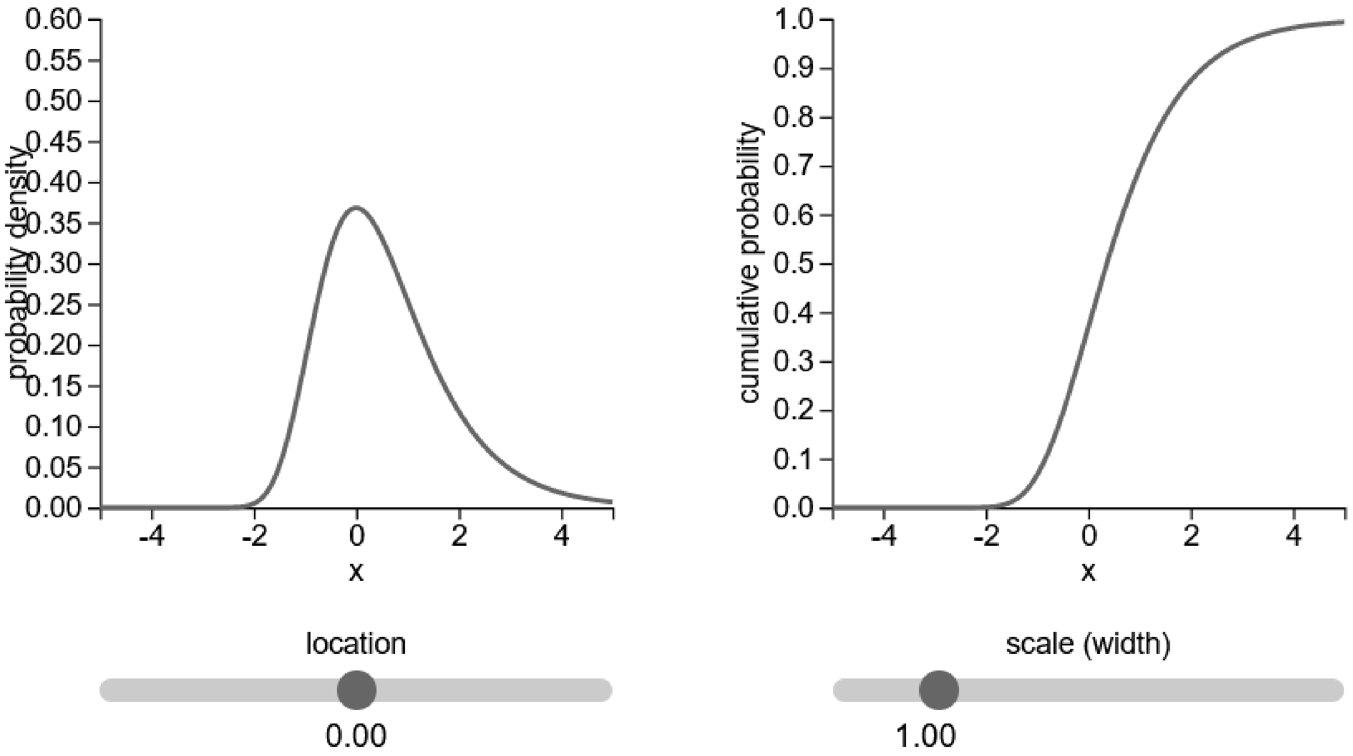
# Beta distribution

- A beta distribution is defined by **two parameters**: $\alpha$ and $\beta$
- A beta distribution has a **left bound** and a **right bound**. Depending on the parameters, it can be symmetric or skew in either direction.
- The beta PDF describes a distribution for the **expected value** of a Bernoulli process.
- **Examples**: coin fairness, chance of instrument failure



# Gumbel distribution

- A Gumbel distribution has a location and a scale parameter
- A Gumbel distribution has **no bounds** and either **a left tail** or a **right tail**. The left and right-tailed Gumbel are different distributions.
- The Gumbel PDF is important in **extreme value analysis**
- **Examples**: annual maximum daily rainfall, maximum material load before failure



# Fitting distribution functions

# Fitting distributions

**Setup**:

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

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

**Question**:

Which **parameter values** generate the distribution that best fits our data?

**How to choose the parametric distribution function:**

**next part of the lecture!**

**Different fitting methods**:

**Method of moments** and **MLE**.

# Method of moments

**Basic idea**:

Equate the **moments of the observations** to **those of the distribution function**, then solve for the parameters.

**Example**: Moments for the Gumbel distribution

$$E[X] = \mu + \gamma\beta \qquad \gamma \approx 0.577 \longrightarrow \text{Mean of the observations}$$

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

# Method of moments - Example

**Setup**:
The **intensity of earthquakes** in Rome (Italy) is a random process.

Using the *'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.

**Data moments**:

- Mean intensity = 3.02
- Variance of intensity = 0.99

**Gumbel distribution:**

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

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

**Substitute in the data moments:**

$$3.02 = \mu + 0.577\beta \longrightarrow \text{Solve for } \mu$$

$$0.99 = \frac{\pi^2}{6}\beta^2 \longrightarrow \text{Solve for } \beta$$

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

# Assessing the goodness of fit

# How do I choose a distribution?

**Decision factors**:

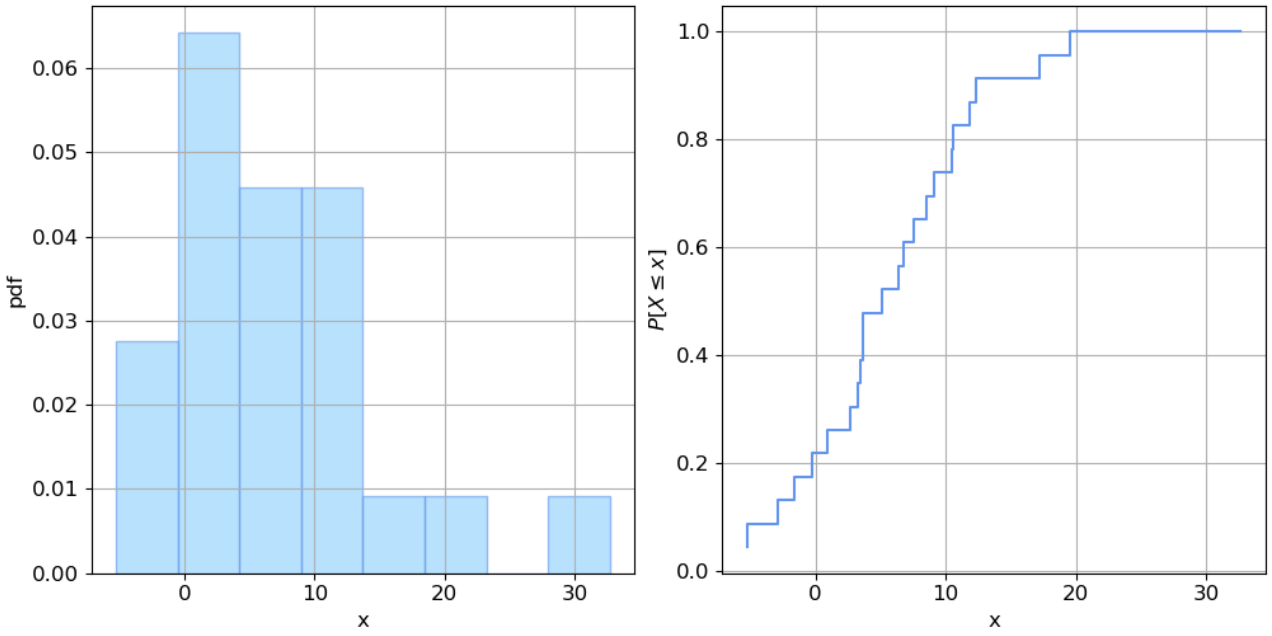
- Physical constrains (e.g., non-negativity)
- Statistics of the observations

**Goodness of fit** techniques can support the decision in a quantifiable way:

  - Objective way to compare models
  - You may obtain contradictory results!

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

EXAMPLE:
- Toy dataset
- Exponential or Gaussian?

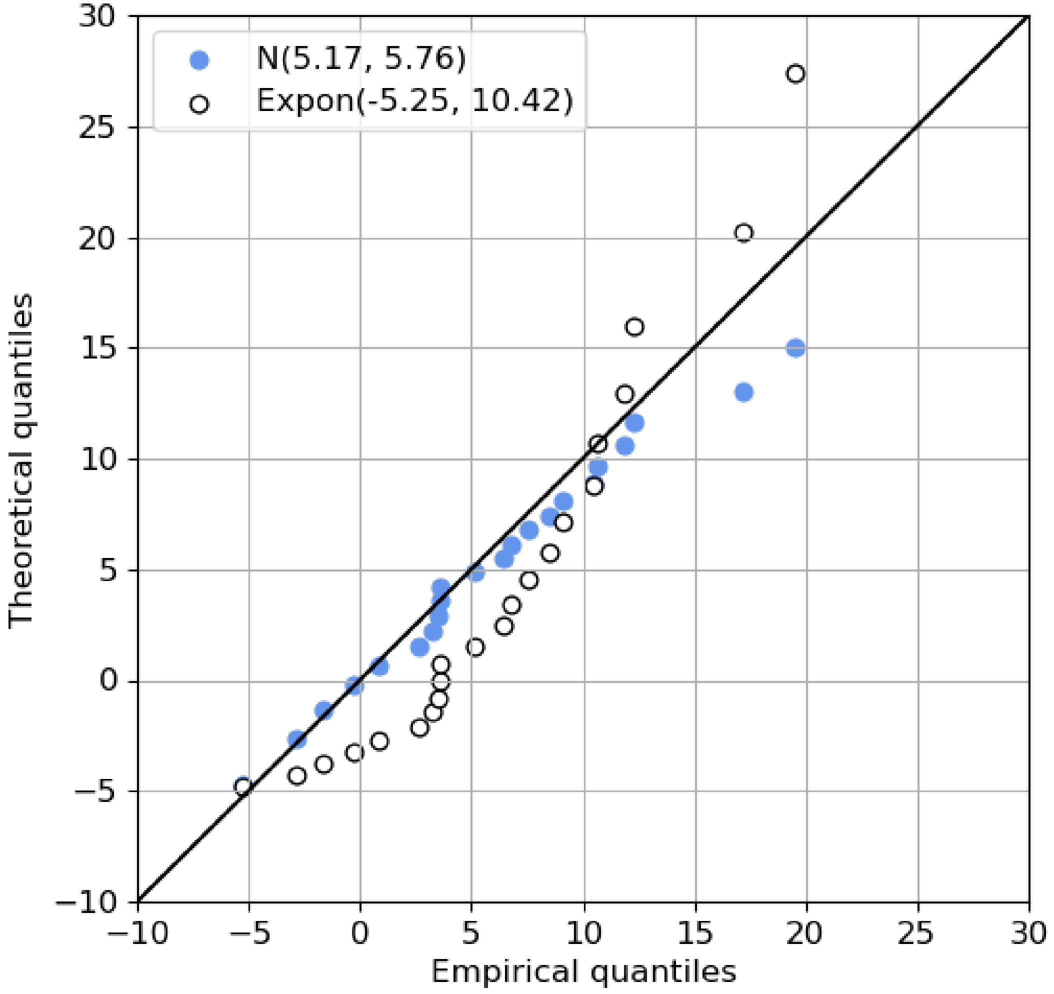



TU Delft

# Graphical methods – QQ plot

A **Quantile-Quantile plot** (QQ plot) plots the empirical against the predicted quantiles of the fitted distribution.

1. Compute the quantiles/non-exceedance probabilities of the observations
2. Evaluate the corresponding values from the fitted distribution via $F^{-1}(x)$
3. Plot observations against predictions; 45 degree-line is the perfect fit




**Advantages**:

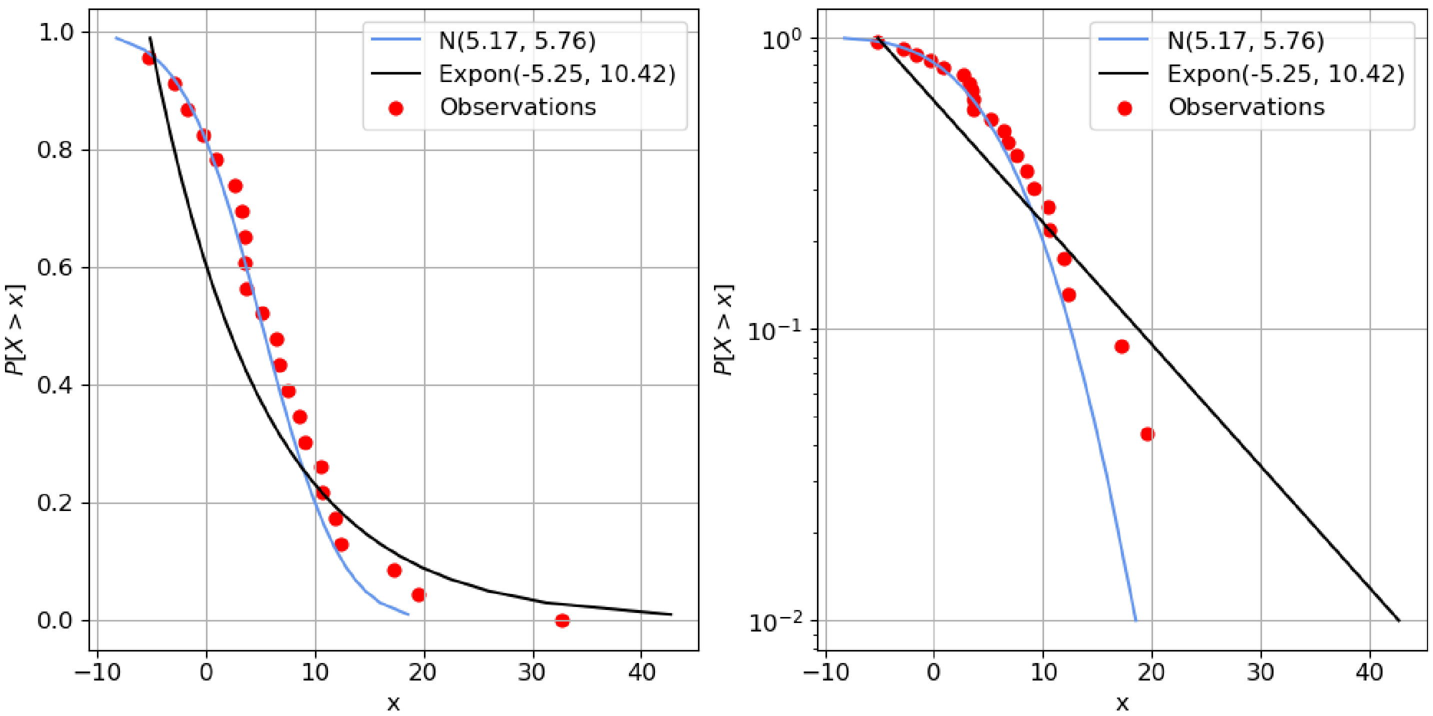
simple | fast to implement | central moments + tail

TU Delft

# Graphical methods – log-scale

In a **log-scale plot**, we compare the exceedance probabilities $P(X > x)$ of the observations and fitted distribution in log scale.

1. Compute empirical exceedance probabilities via $1 - F(x)$
2. Plot and compare empirical and fitted exceedance probabilities in semi-log scale




**Advantages**:
simple | fast to implement | focus on the tail: key element!

TUDelft

# Formal hypothesis tests – Kolmogorov-Smirnov

The **Kolmogorov-Smirnov test** (also known as the *KS-test*) is a widely used nonparametric **hypothesis test** based on comparing two CDFs.

It comes in one of **two variants**:

- Two sets of samples: same population?
- One set of samples: **goodness-of-fit** of a fitted distribution

**Hypothesis tests:**

$H_0$: null hypothesis
$H_1$: alternative hypothesis

Statistic ~ distribution → p-value
**p-value**: $P(\text{data as extreme as observed} \mid H_0)$

Significance (typically $\alpha = 0.05$)

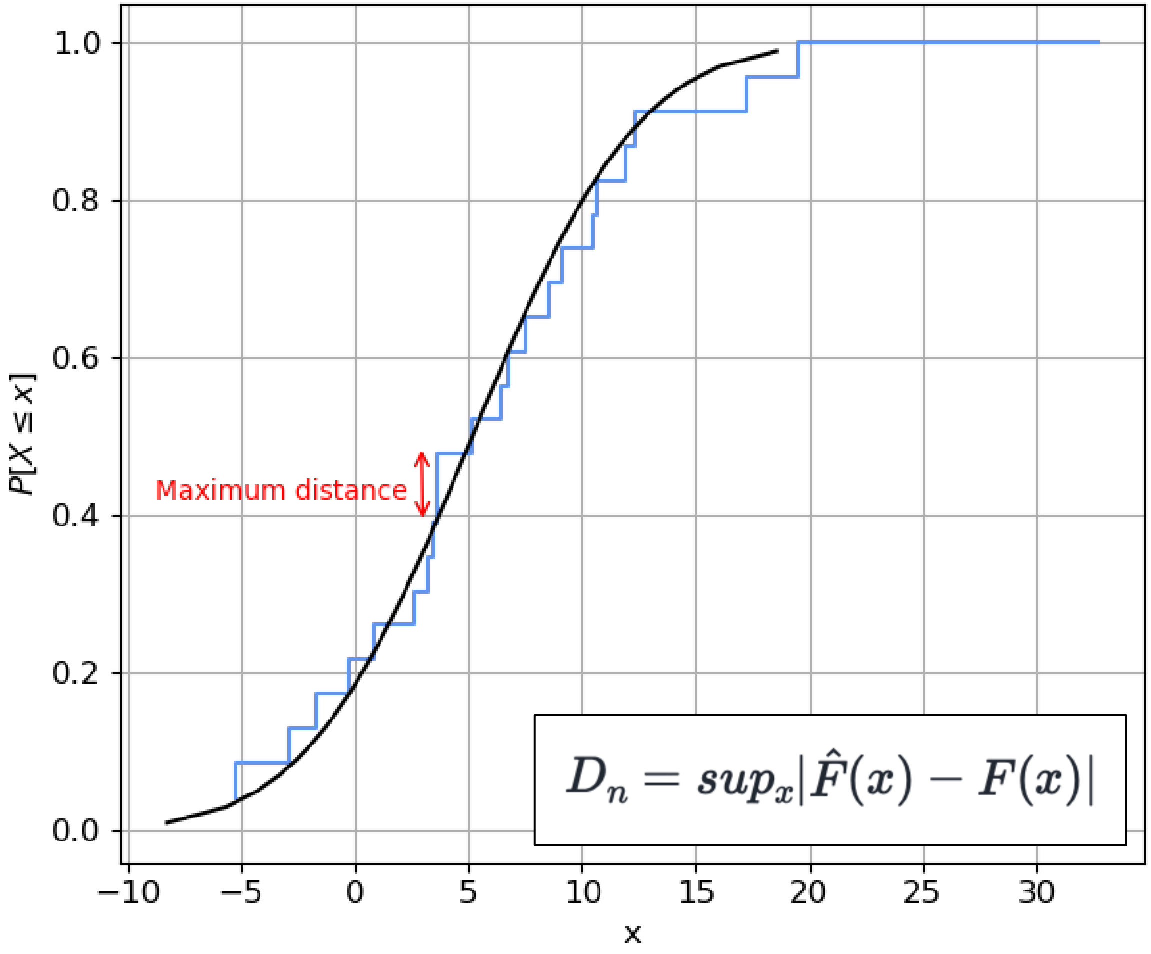
If p-value < $\alpha$: we reject $H_0$
If p-value > $\alpha$: we cannot reject $H_0$

TU Delft

# Formal hypothesis tests – Kolmogorov-Smirnov

**One set of samples**:
goodness-of-fit of a fitted distribution

- **Null hypothesis $H_0$** :
  our samples follow the fitted distribution
  $H_0 : \hat{F} \sim F$
- **KS statistic**: (roughly) the maximum distance between the ECDF and the fitted CDF
- **P-value > α = 0.05** ⟶ We cannot reject $H_0$, i.e., that the observations follow the distribution



# Formal hypothesis tests – Kolmogorov-Smirnov

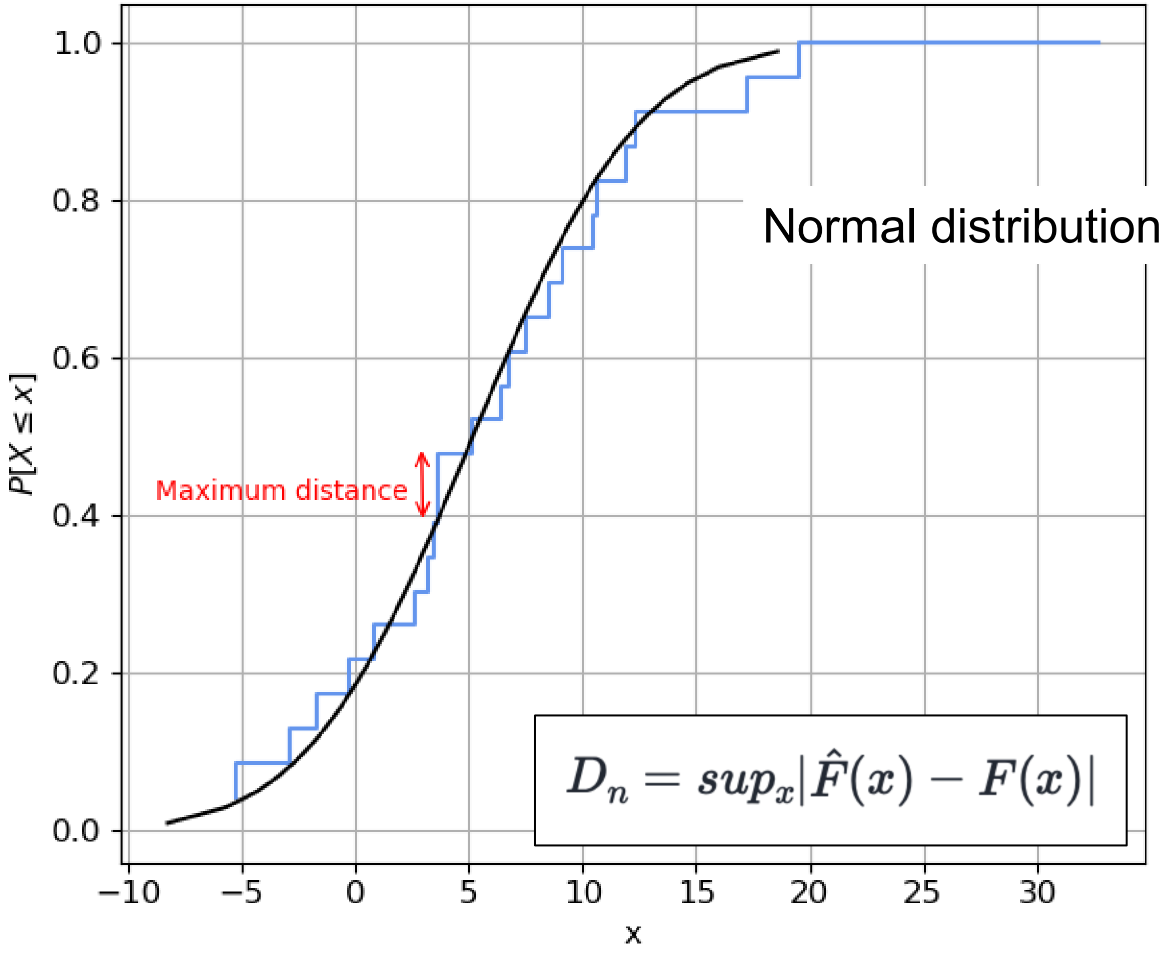
**Example**:

Here, our null hypothesis is

$H_0: \hat{F} \sim F_{\mathcal{N}}$

(observations follow a normal distribution)

- P-value = 0.93
- P-value = 0.93 > $\alpha$ = 0.05 ⟶ We cannot reject $H_0$, i.e., that the observations follow a Normal distribution



# What's next?

- There is more in the **textbook**:
  - Interactive elements!
  - Parameterization!
- **Wednesday workshop**: concrete or air temperature
- **Friday project**: your choice!
  - Wind gust factor in Delft
  - Traffic density in Finland
  - Flow velocity of the river Thames

## Summary of parametric distributions

Here a summary of the main equations for each of the presented distirbution functions is presented.

**MUDE exam information** Click to show

## Choosing a distribution

If you need help to choose a distribution type for your data, the table below may help you make a choice:

| Distribution | left bound | right bound | left-tailed | symmetric | right-tailed | scipy name |
|---|---|---|---|---|---|---|
| Uniform | yes | yes | no | yes | no | uniform |
| Gaussian | no | no | no | yes | no | norm |
| Lognormal | yes | no | no | no | yes | lognorm |
| Gumbel (right-tailed) | no | no | no | no | yes | gumbel_r |
| Gumbel (left-tailed) | no | no | yes | no | no | gumbel_l |
| exponential | yes | no | no | no | yes | expon |
| beta | yes | yes | possible | possible | possible | beta |

**Notation**

One challenge when dealing with distributions is notation, for two main reasons: 1) the symbols used to represent random variables and parameters vary across different fields (and even *within* a given

TU Delft

# Let’s collect some data!

## We want to know you!

We would like to collect data about our students that we can use for teaching in future years. If you want to support us in this, please fill in this **anonymous poll** and tell us a little bit more about yourself.

Direct link:
https://forms.office.com/e/2yxYwYrrjQ


## Enjoy the journey!

And enjoy the journey!