Modelling, Uncertainty and Data for Engineers (MUDE)

Week 1.5 : Multivariate distributions

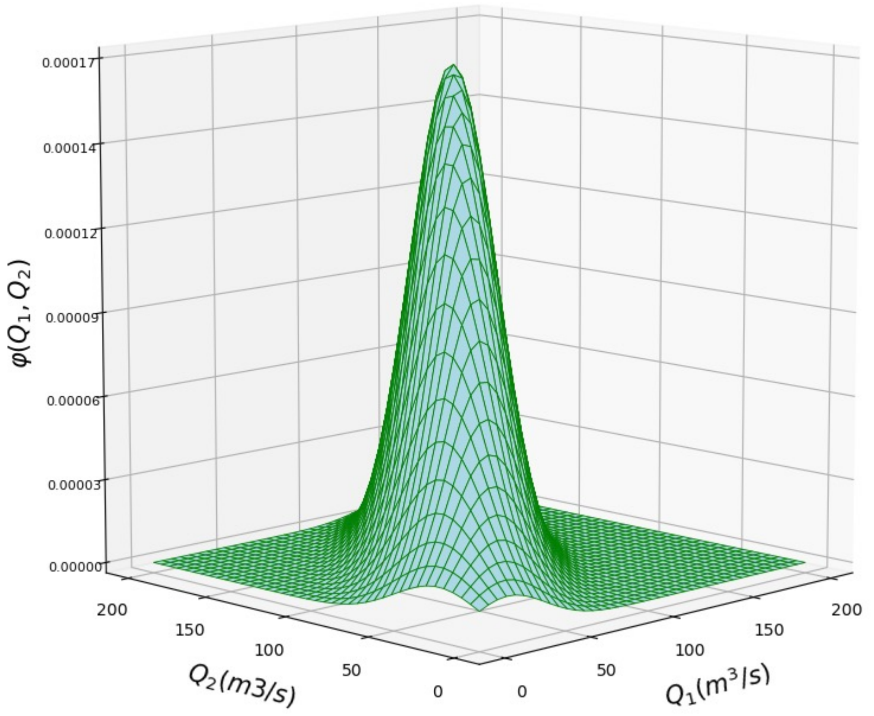
Patricia Mares Nasarre

TU Delft

Contents – Multivariate distributions

1. Recap & motivation
2. Set theory and basic operations
3. Continuous variables
   a. Extension to multivariate & empirical computations
   b. Measures of dependence
4. Multivariate Gaussian distribution

Bivariate Normal pdf (ρ=0.77)



Recap 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 $\mathrm{CO_2}$ in a city as function of the traffic. For a certain value of traffic, the model will always provide the same concentration of $\mathrm{CO_2}$. 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'

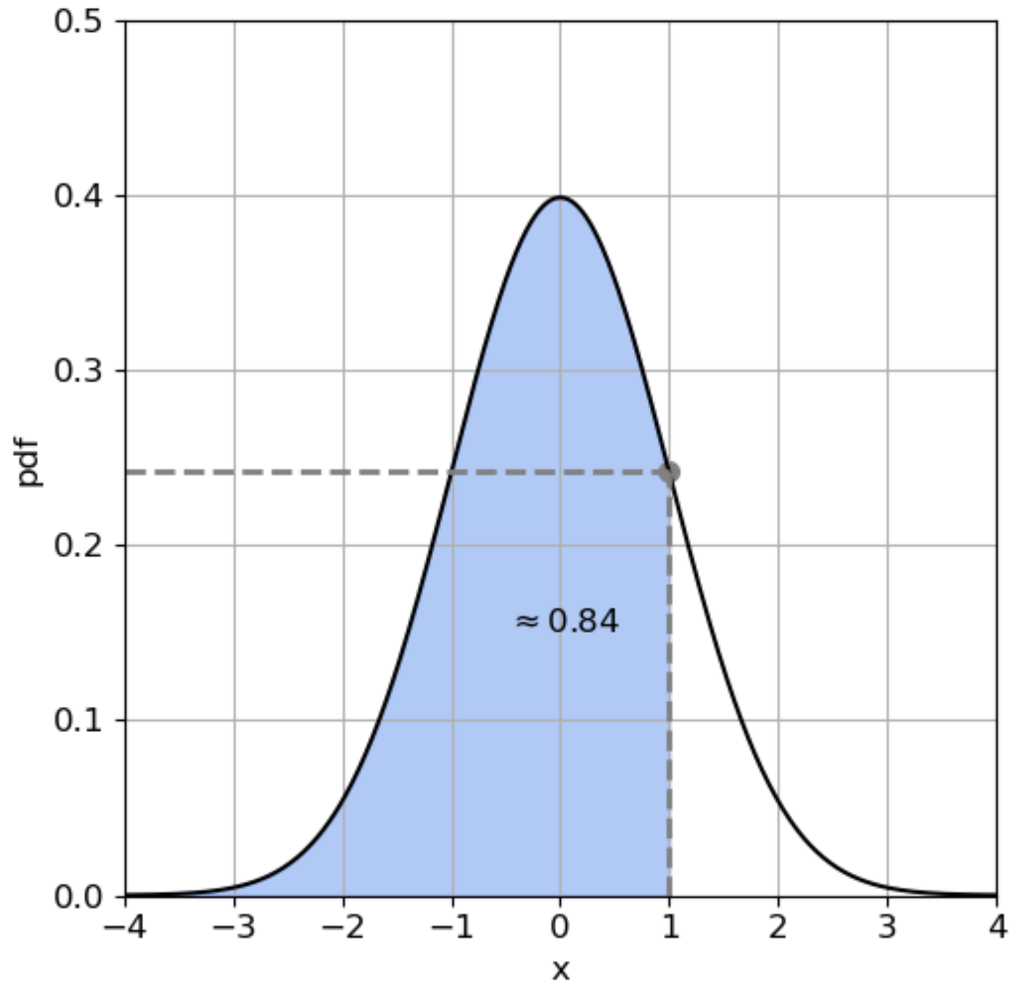
→ **Model uncertainty**

Recap Week 4

- The distribution of the random variable X has a PDF $f_X(x)$ and a CDF $F_X(x)$. The CDF is then defined as

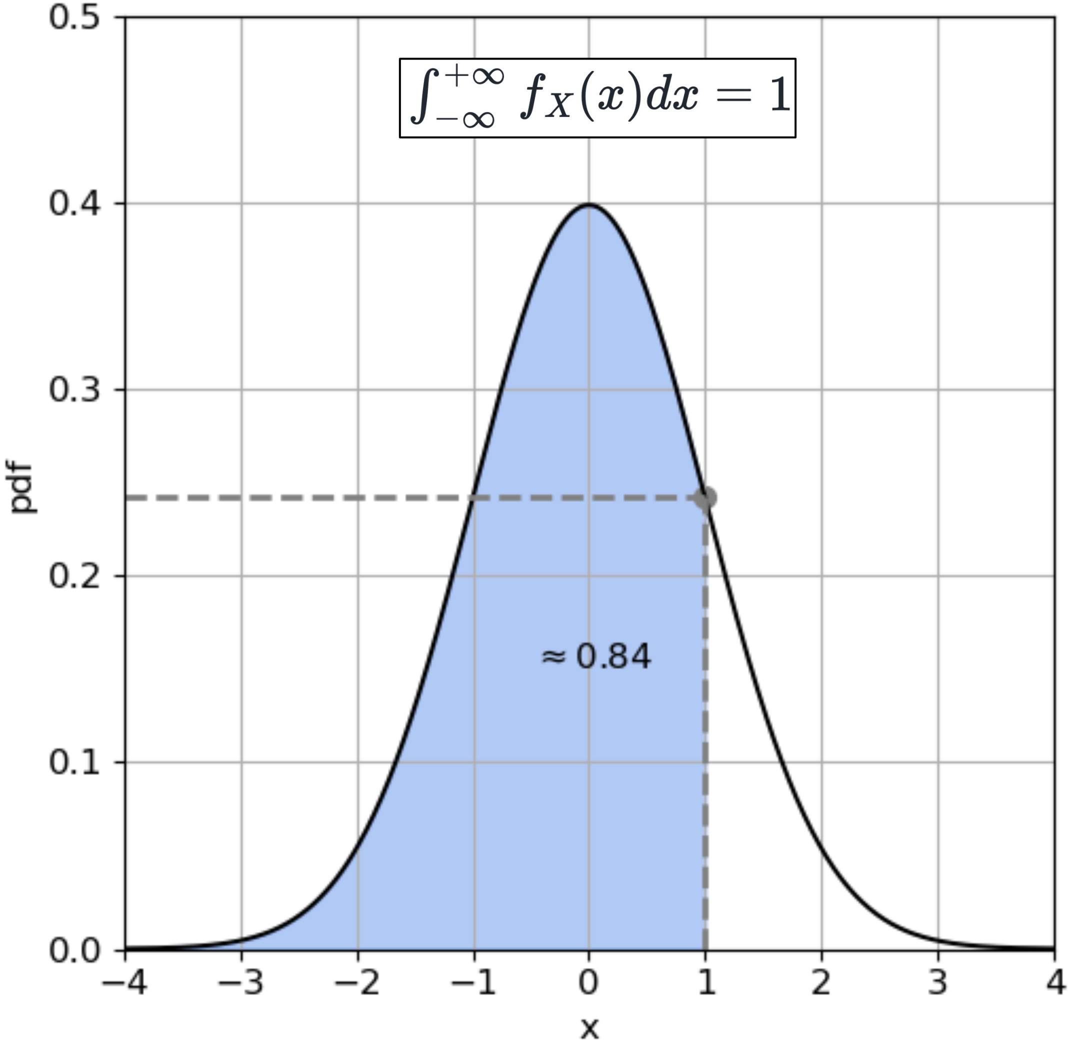
$$P(X \leq x) = F_X(x) = \int_{-\infty}^{x} f_X(x)dx$$

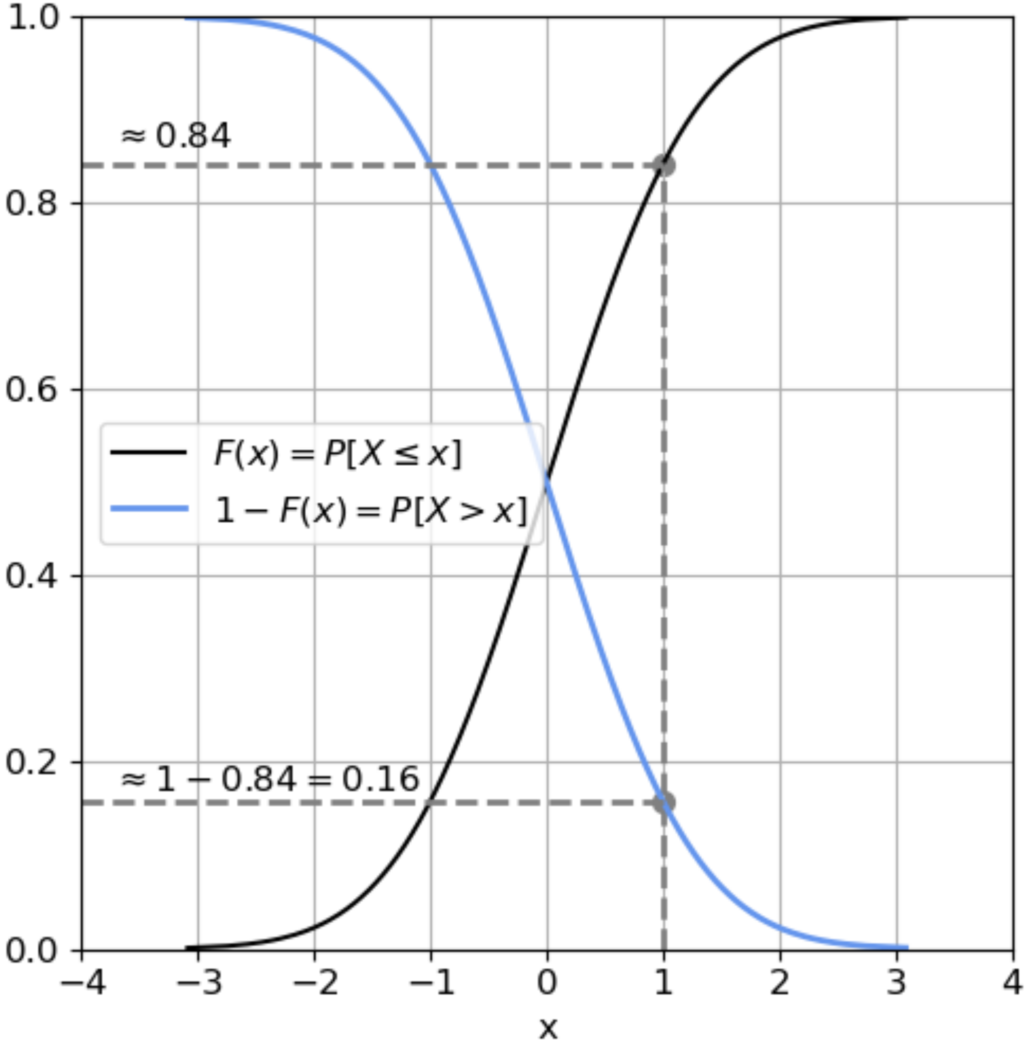
The expression above can be applied also over an interval, Ω, by changing the limits of the integral.



Recap Week 4

Model the uncertainty of one variable





What about Week 5?

Model the uncertainty of more than one random variable

Multivariate:

- Multi:
- Variate:

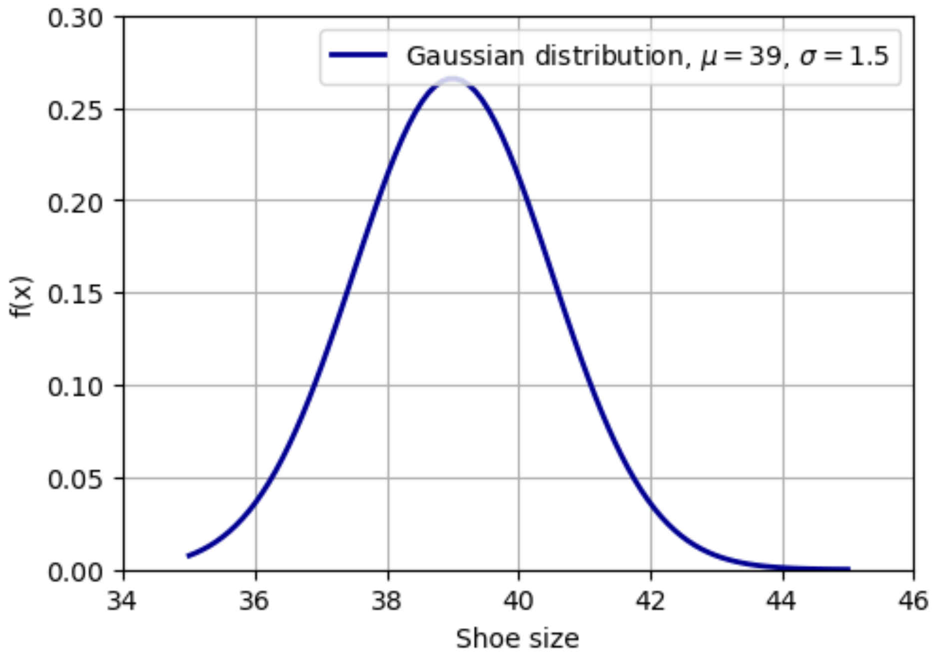
Distributions

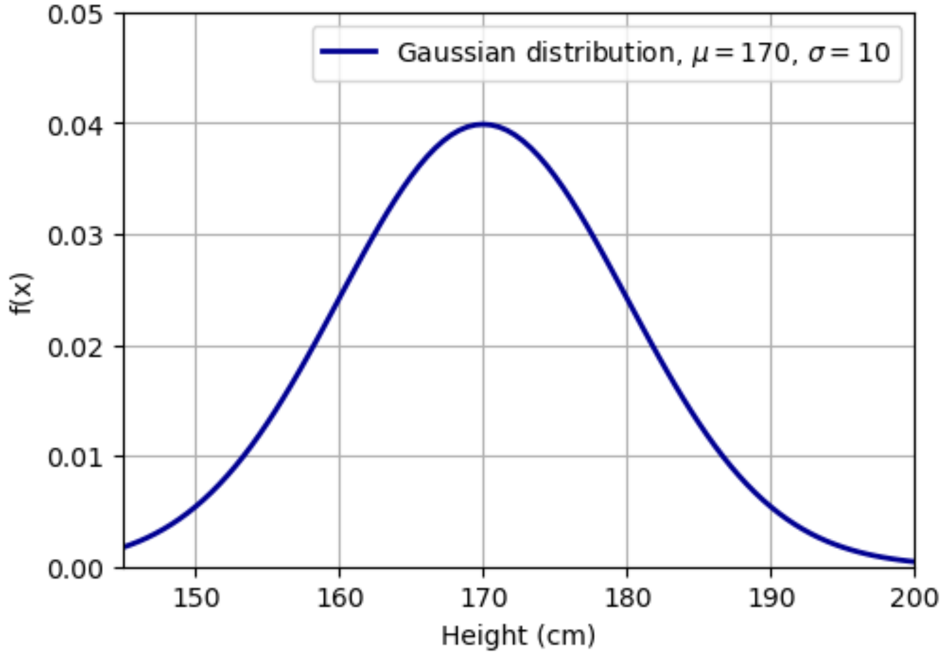
- We are going to keep having fun with uncertainty modelling!

Why is that different to Week 4?

We are studying the uncertainty in the relationship between the height and shoe size of people.

But what are we missing?





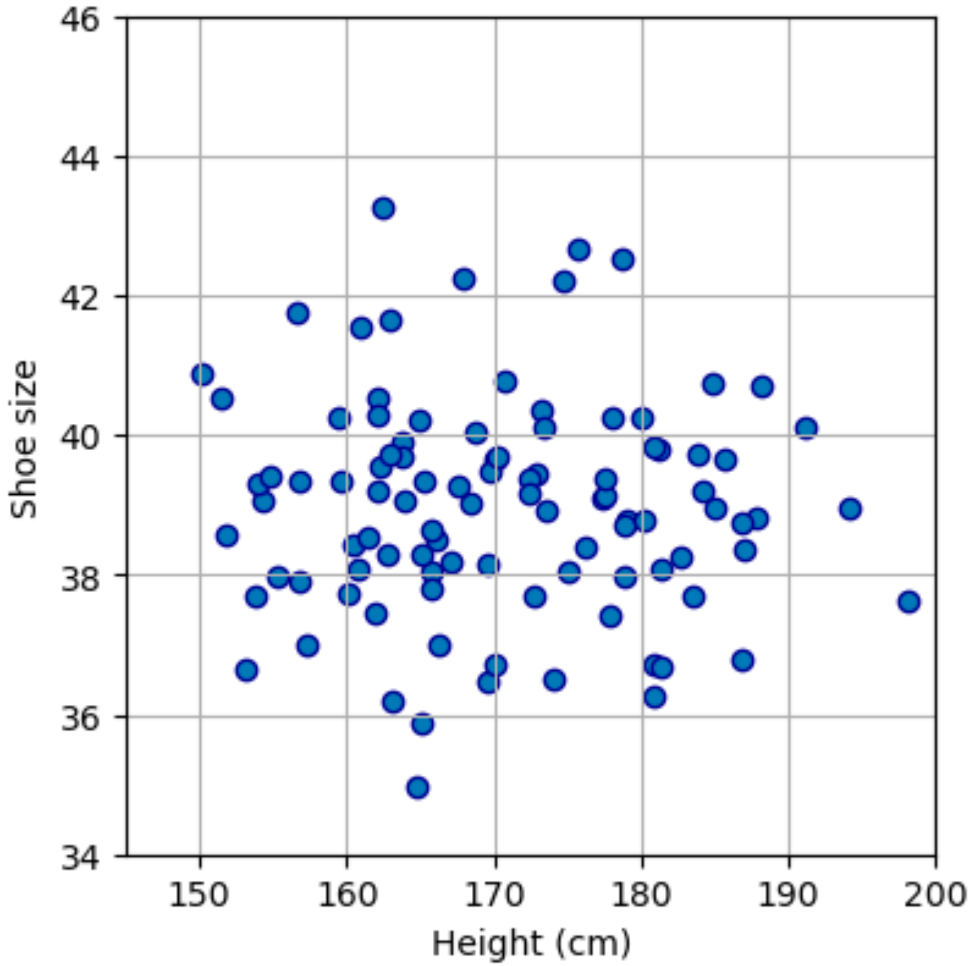
TU Delft

Why is that different to Week 4?

We are studying the uncertainty in the relationship between the height and shoe size of people.

But what are we missing?

1. Generate 100 random samples from the distribution of shoe size.
2. Generate 100 random samples from the distribution of height.
3. Pair the two sets of samples.



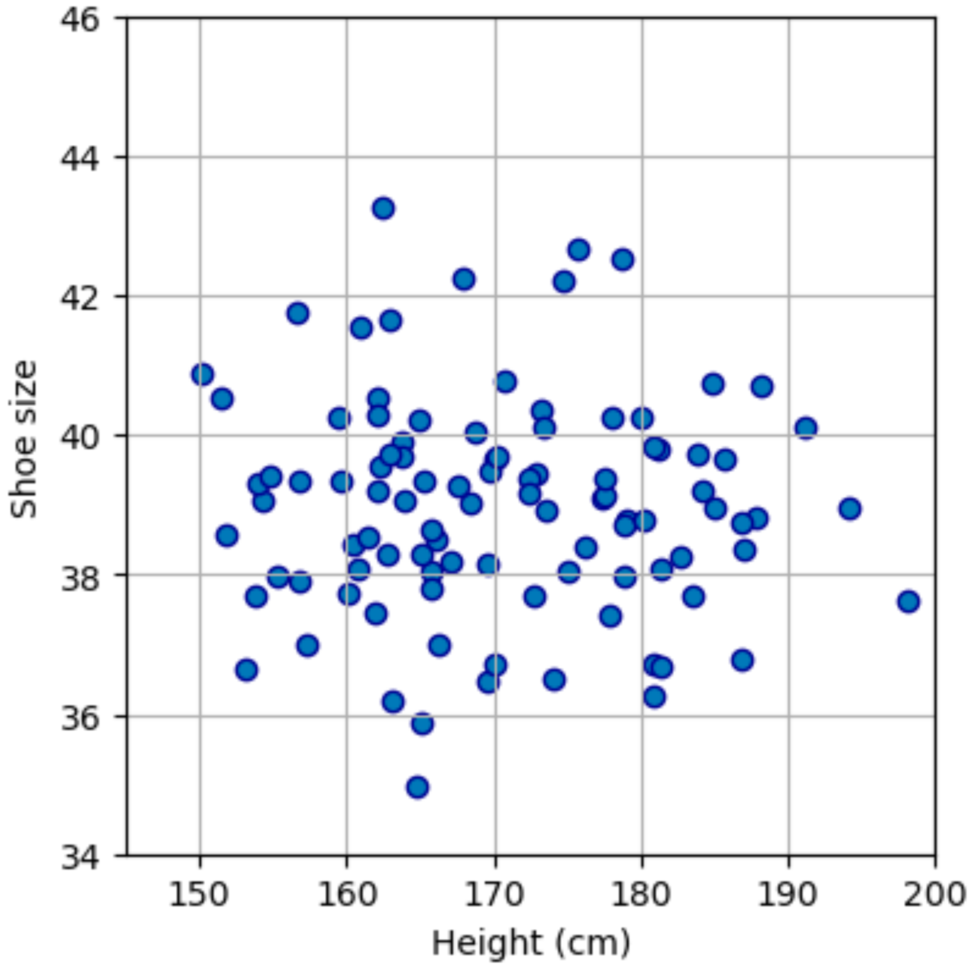
Why is that different to Week 4?

We are studying the uncertainty in the relationship between the height and shoe size of people.

But what are we missing?

Typically, **higher people use larger shoe sizes**.

Just by characterizing **independently** the distribution of the two random variables **is not enough to account for the relationship** between them.



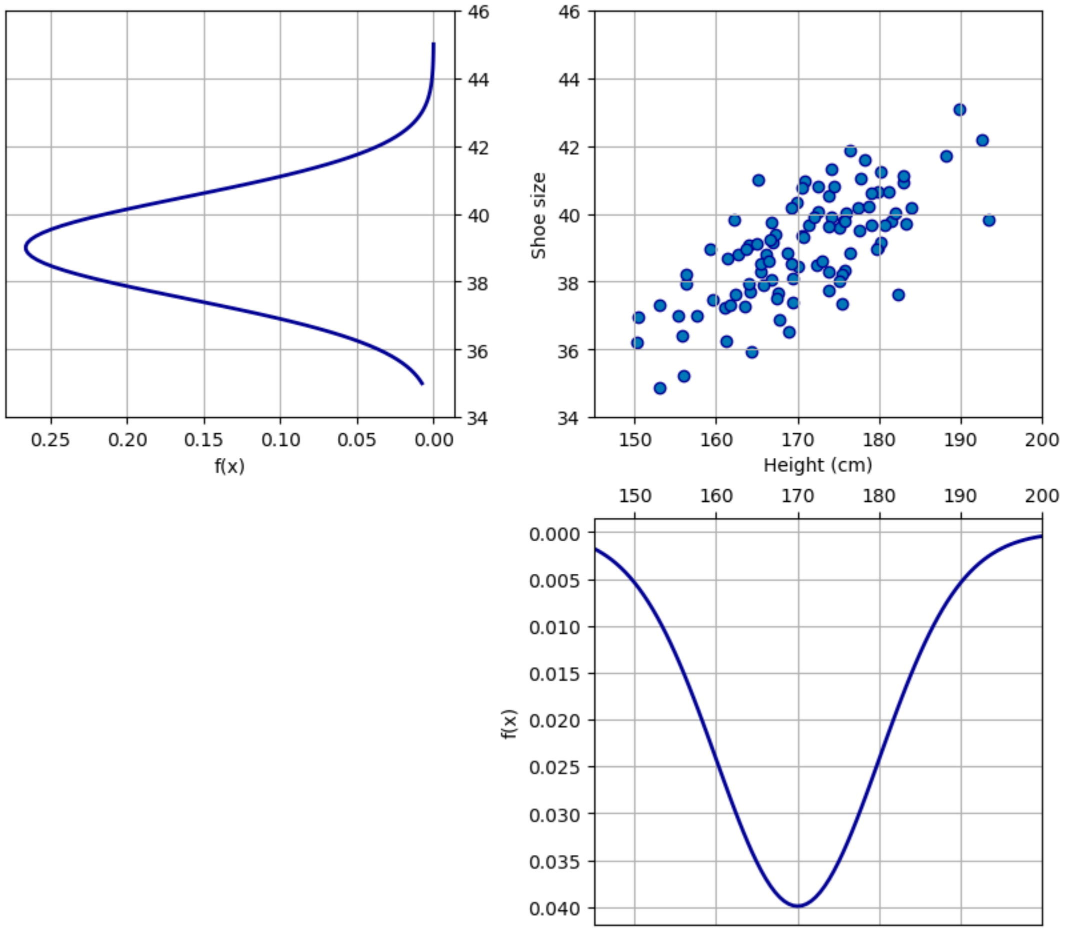
Why is that different to Week 4?

We are studying the uncertainty in the relationship between the height and shoe size of people.

But what are we missing?

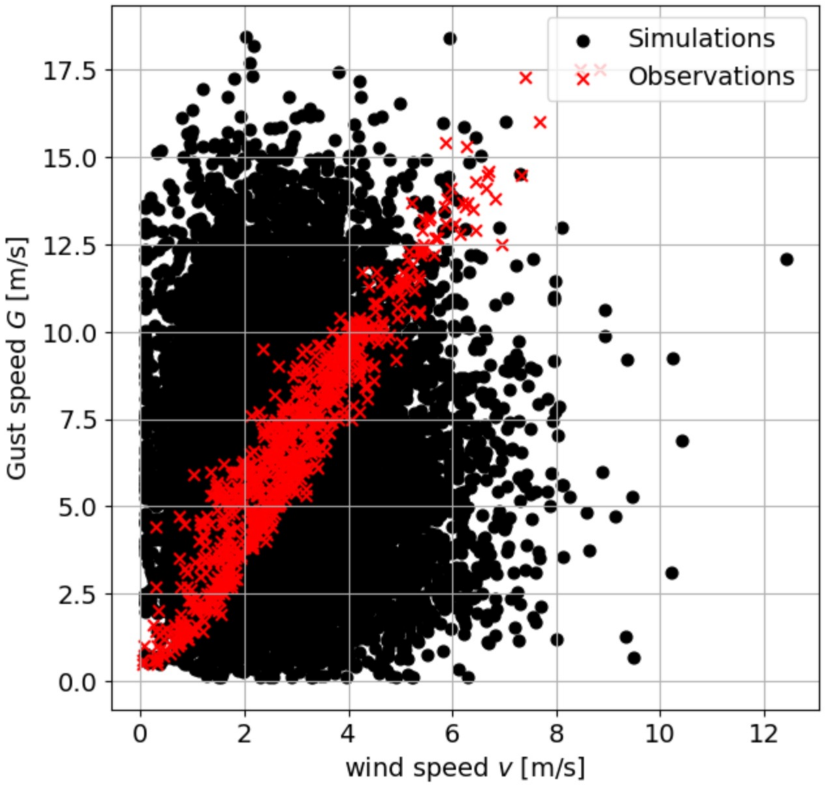
Typically, **higher people use larger shoe sizes**.

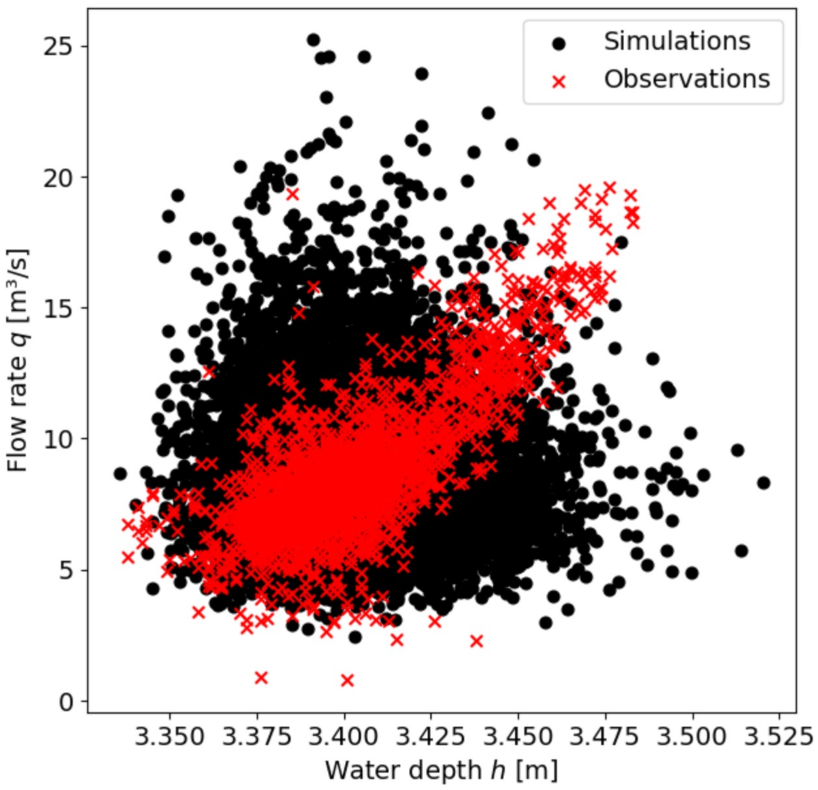
Just by characterizing **independently** the distribution of the two random variables **is not enough to account for the relationship** between them.

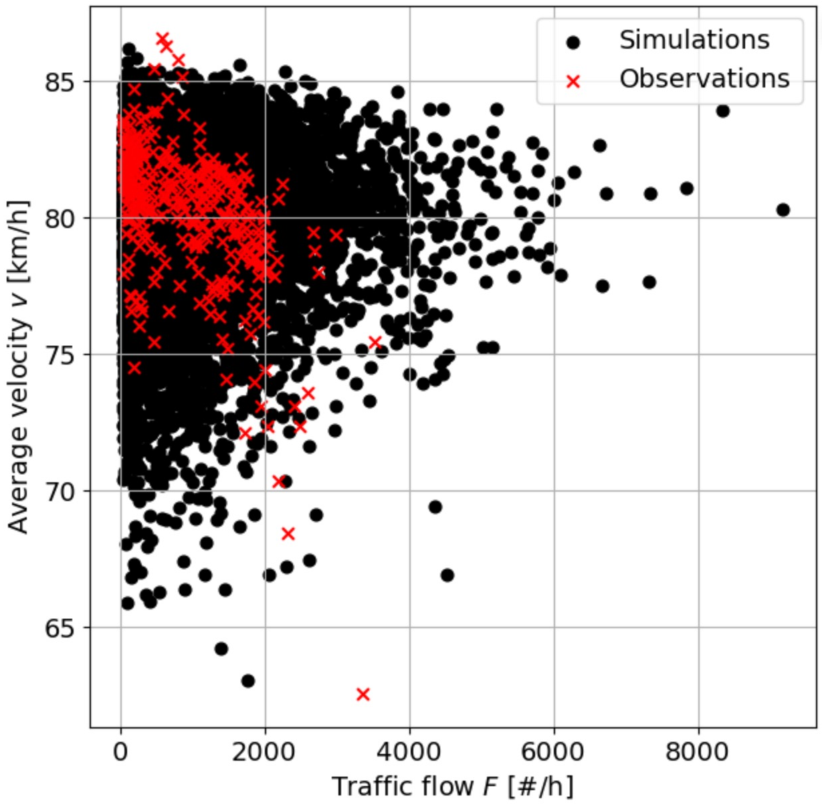


TU Delft

Why is that different to Week 4?







Why is that different to Week 4?

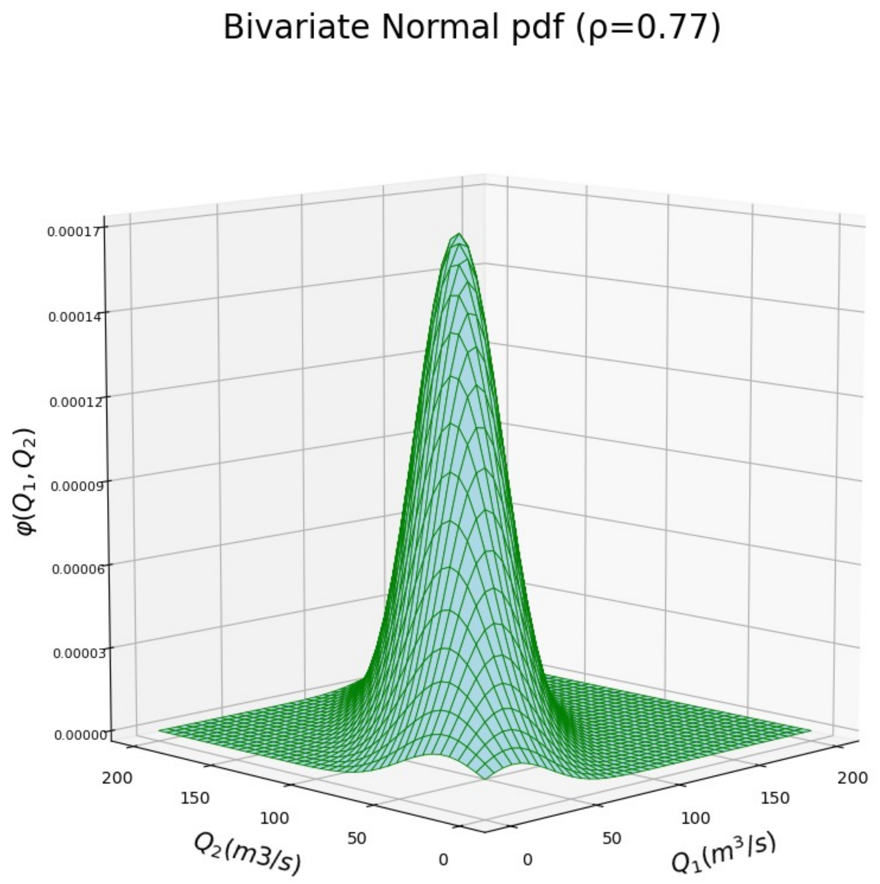
Variables of interest are often ‘tied’ to each other in engineering and geosciences problems, as they are generated by the same drivers.

Some examples:

- Wave height and wave period;
- Temperature, soil moisture and precipitation;
- Wind velocity and wind direction;
- Vehicle velocity and CO_2 emissions;
- Concentrations of nitrogen and phosphorus in water;
- Concrete compressive strength and tensile strength;
- And many more!

Contents – Multivariate distributions

1. Recap & motivation
2. Set theory and basic operations
3. Continuous variables
    a. Extension to multivariate & empirical computations
    b. Measures of dependence
4. Multivariate Gaussian distribution



Set theory and basic operations

TU Delft

Basic definitions

- **Sample space (S):** collection of all possible outcomes arising from an experiment that involves chance.

2 x

{{Heads, Heads},
{Heads, Tails},
{Tails, Heads},
{Tails, Tails}}

- **Event:** specific outcome or set of outcomes from the experiment

Basic definitions

- **Probability of an event:**

Sample space? P(1)?

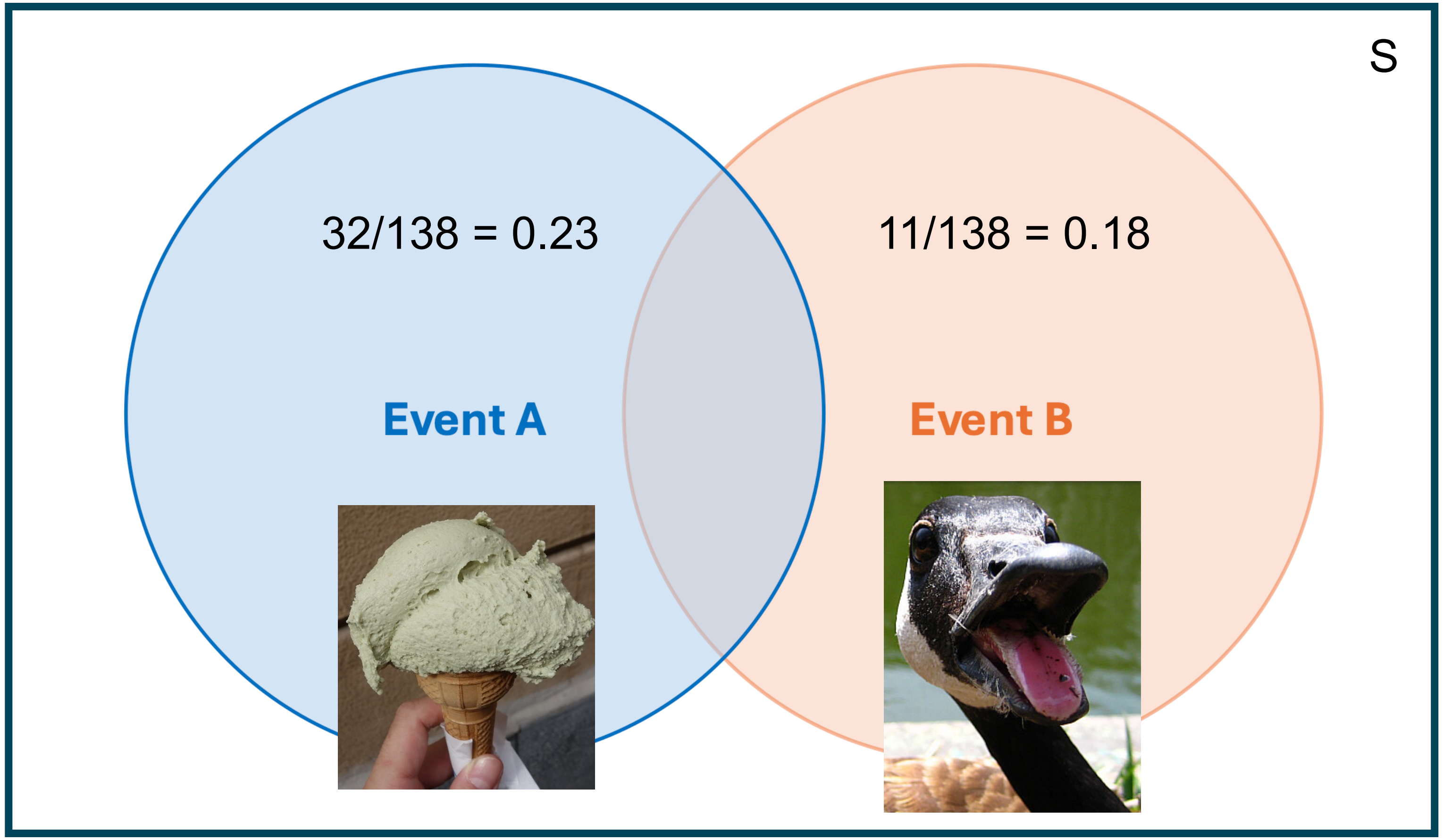
1 x  **{1, 2, 3, 4, 5, 6}** $P(1) = 1/6 \approx 0.17$

- **Probability of the complement:** $P(\bar{1})$?

$$P(\bar{1}) = 1 - P(1) \approx 0.83$$

TU Delft

Discrete events



"Pistachio ice cream" by pelican
"Angry Goose!" by This Incredible World

Probability of the sample space:

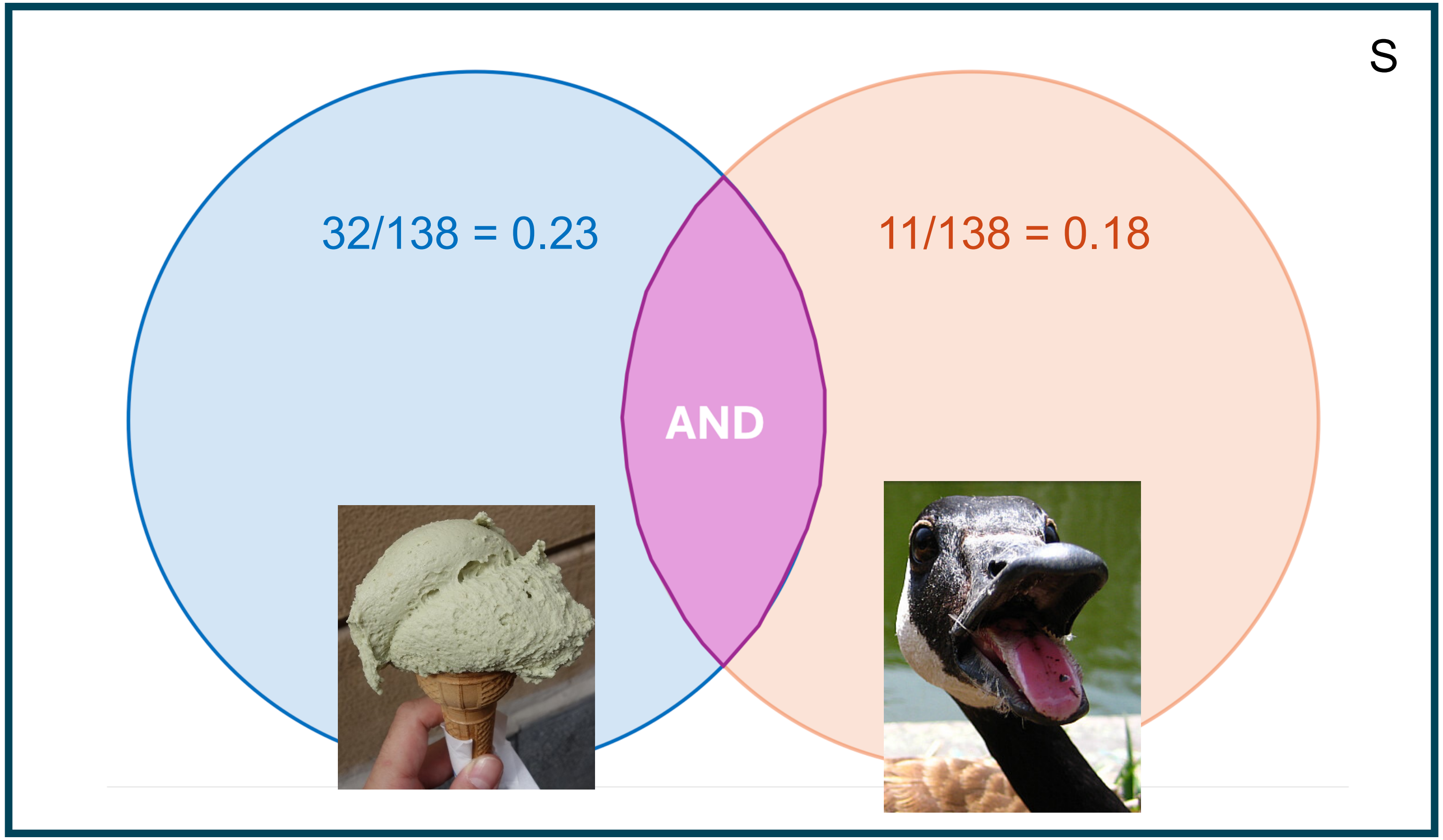
$$P(S) = 1$$

We can better know MUDE students thanks to Max!

A: The favorite ice cream flavor of a MUDE student is pistachio

B: A MUDE student believes that could win a fight against a Goose (very angry)

Discrete events



"Pistachio ice cream" by pelican
"Angry Goose!" by This Incredible World

Is it possible that A and B occur at the same time?

AND or intersection, $\boldsymbol{P(A \cap B)}$

Any idea on how to calculate?

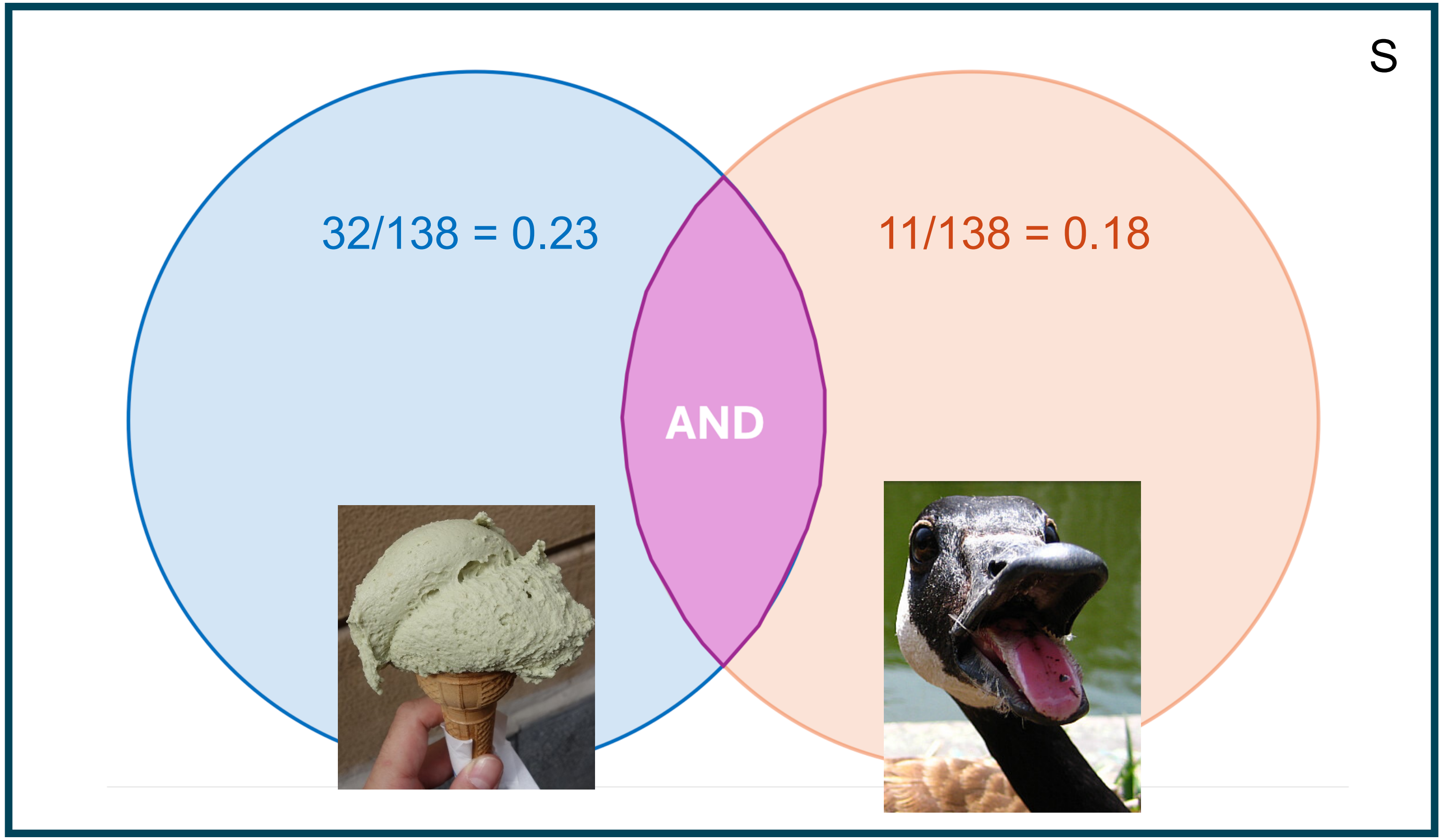
Independence

- **Independence:**

  A and B are considered independent **if and only if** the AND probability, $P(A \cap B)$, can be factorized into the product of their probabilities.

$$P(A \cap B) = P(A)P(B)$$

Discrete events



"Pistachio ice cream" by pelican
"Angry Goose!" by This Incredible World

Is it possible that A and B occur at the same time?

AND or intersection, $\boldsymbol{P(A \cap B)}$

Assuming independence:

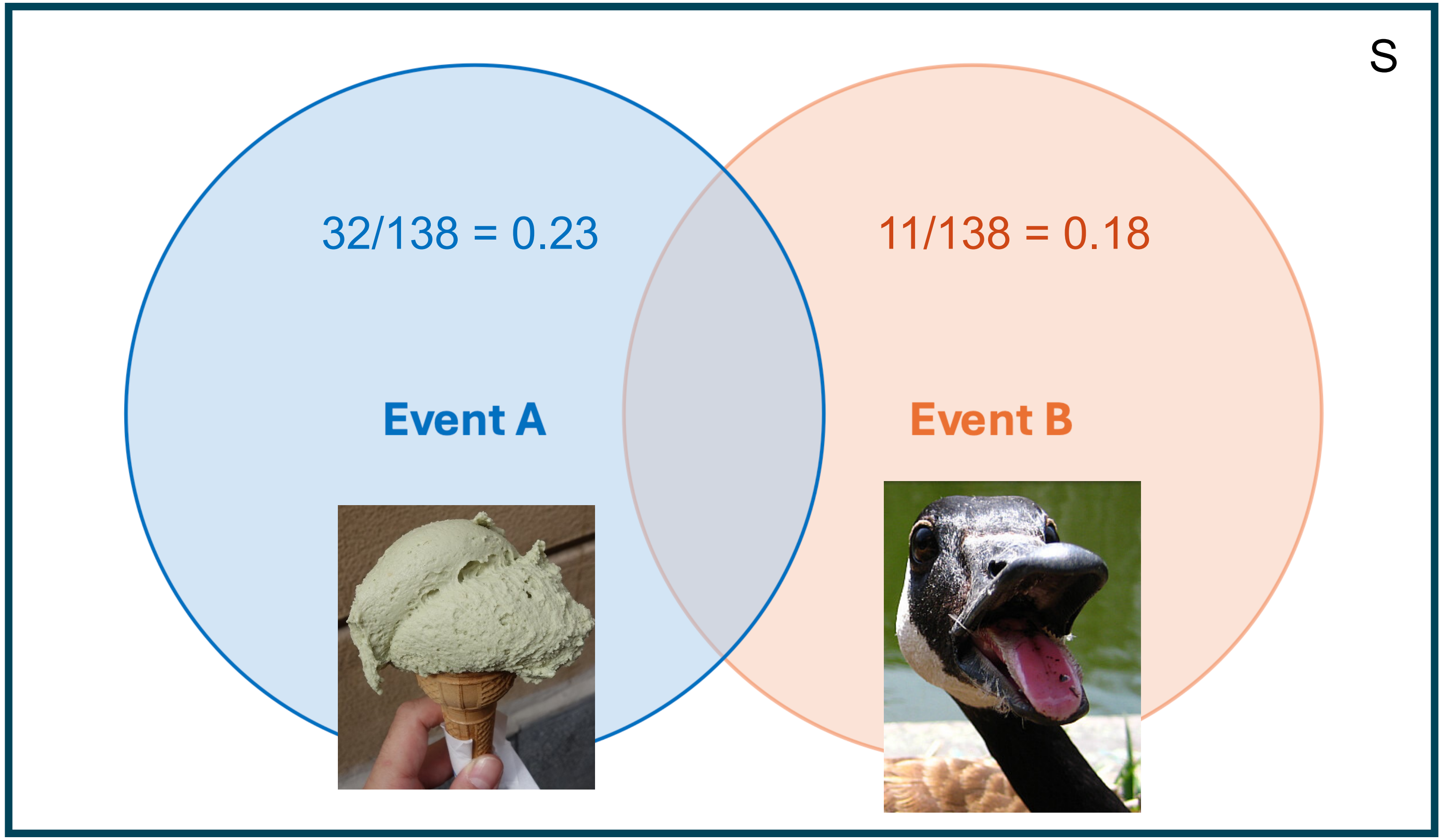
$$P(A \cap B) = 0.23 \times 0.18 \approx 0.04$$

According to MUDE student responses:

$$P(A \cap B) = 3/138 \approx 0.02$$

There may be some degree of dependence?

Discrete events



Is it possible that A and B occur at the same time?

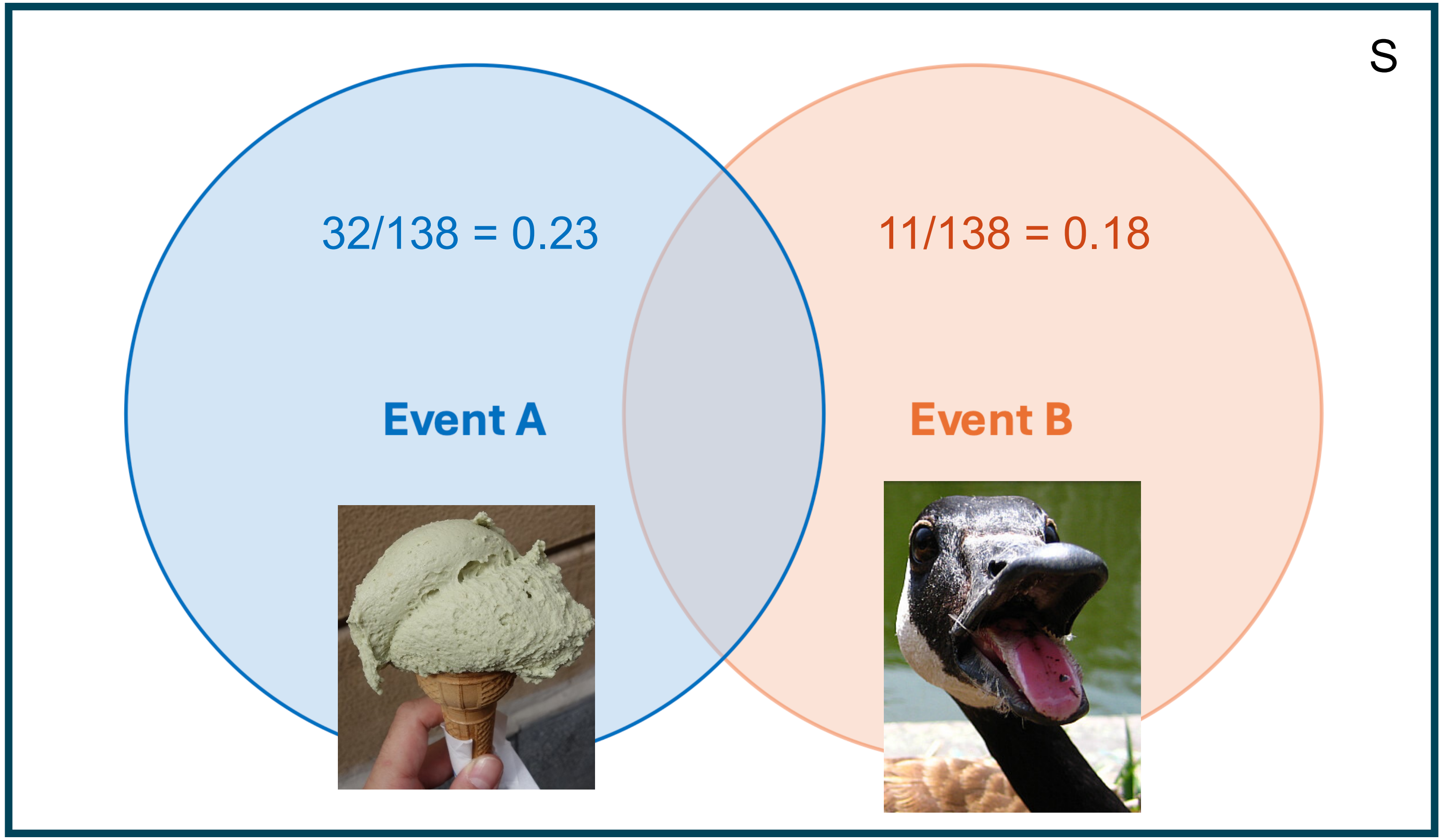
Is it possible that none of them occur?

Area outside both circles is not zero.

"Pistachio ice cream" by pelican
"Angry Goose!" by This Incredible World

Discrete events

$P(A \cap B) = 3/138 \approx 0.02$



"Pistachio ice cream" by pelican
"Angry Goose!" by This Incredible World

Is it possible that A and B occur at the same time?

Is it possible that none of them occur?

Is it possible that either of them occurs?

OR or union, $\boldsymbol{P(A \cup B)}$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B) = 0.23 + 0.18 - 0.02 = 0.39$$

TU Delft

Independence & conditional probability

- **Independence:**

  A and B are considered independent **if and only if** the AND probability, $P(A \cap B)$, can be factorized into the product of their probabilities.

$$P(A \cap B) = P(A)P(B)$$

- **Conditional probability:**

  Probability of an event given that another event is known to have occurred.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)\,\cancel{P(B)}}{\cancel{P(B)}} = P(A)$$

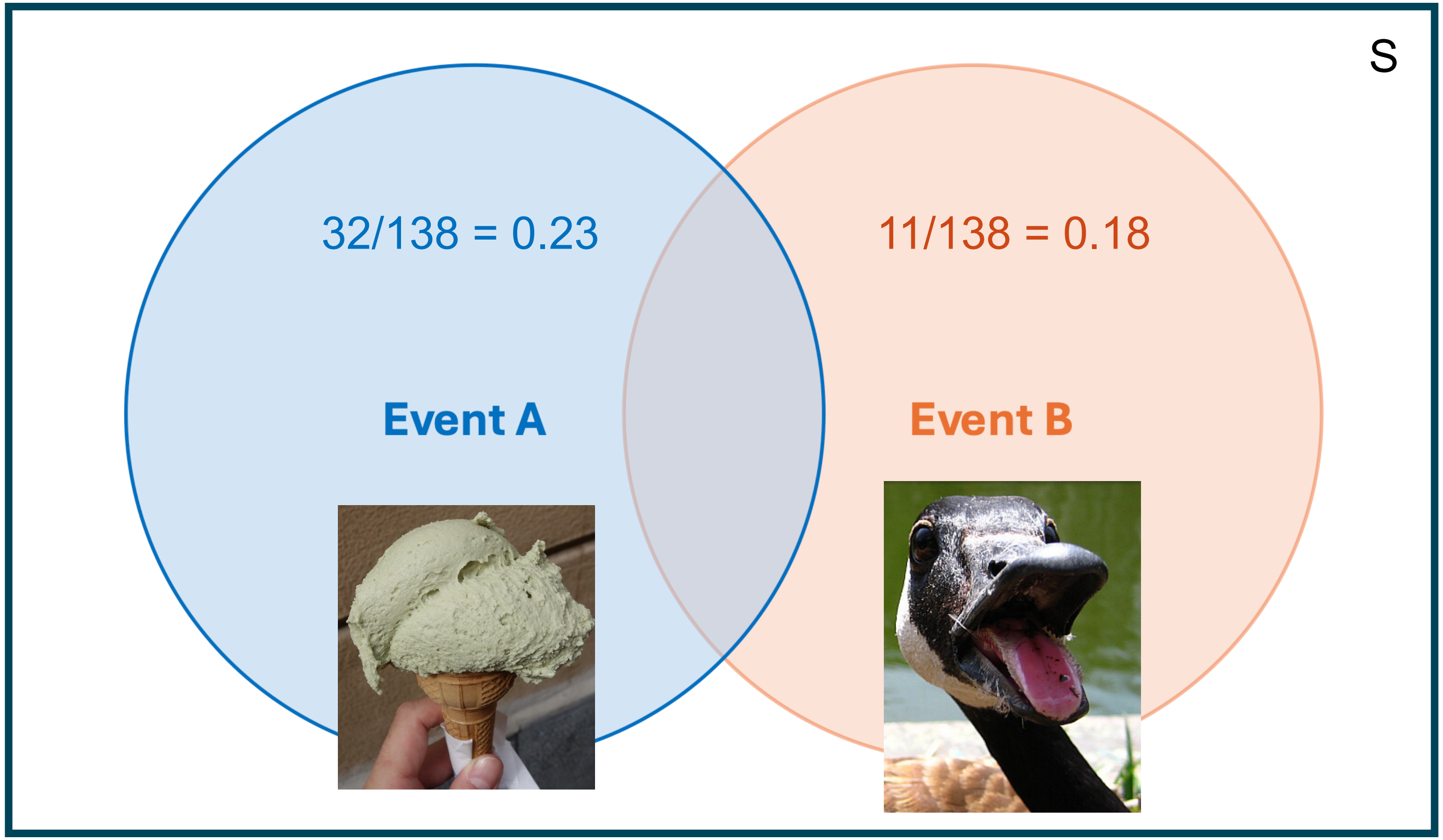
If independent

Today we will see how to compute P(A ∩ B) when random variables are not independent

Discrete events – conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = 3/138 \approx 0.02$$



What is the probability of a MUDE student that can win a fight with a very angry goose given that their favorite ice cream flavor is pistachio?

P(B|A) = 0.02/0.23 =0.09

MUDE students who like pistachio ice cream do not feel especially strong against a very angry goose 🤕

"Pistachio ice cream" by pelican
"Angry Goose!" by This Incredible World

Let’s move to continuous variables

Extension to multivariate & empirical computations

TU Delft

Continuous vs discrete

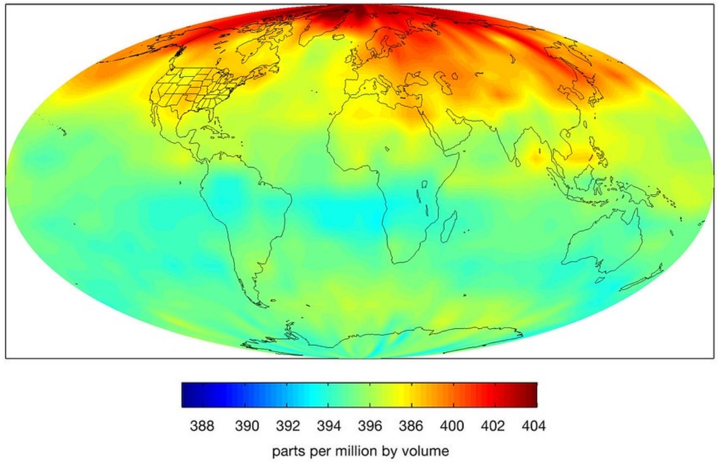
Random variables

- What is a discrete variable? ⇒ Finite number of possible values

"Coin Toss (3635981474)" by ICMA Photos.

- What is a continuous variable? ⇒ Infinite number of possible values

"Carbon Dioxide in Earth's Mid-Troposphere, April 2013 Monthly Average" by Atmospheric Infrared Sounder



"Naryn River, Kyrgyzstan" by Ninara is licensed under CC BY 2.0.

TU Delft

From events to intervals

- **Continuous variables**

  Take an infinite number of values so we evaluate them in an interval, Ω.

  $$\Omega = x \in \mathbb{R}: a \ \leq x < b$$

  Examples: exceedance probability P(X > x)

- **From discrete events to continuous** random variables: A $\rightarrow \Omega$

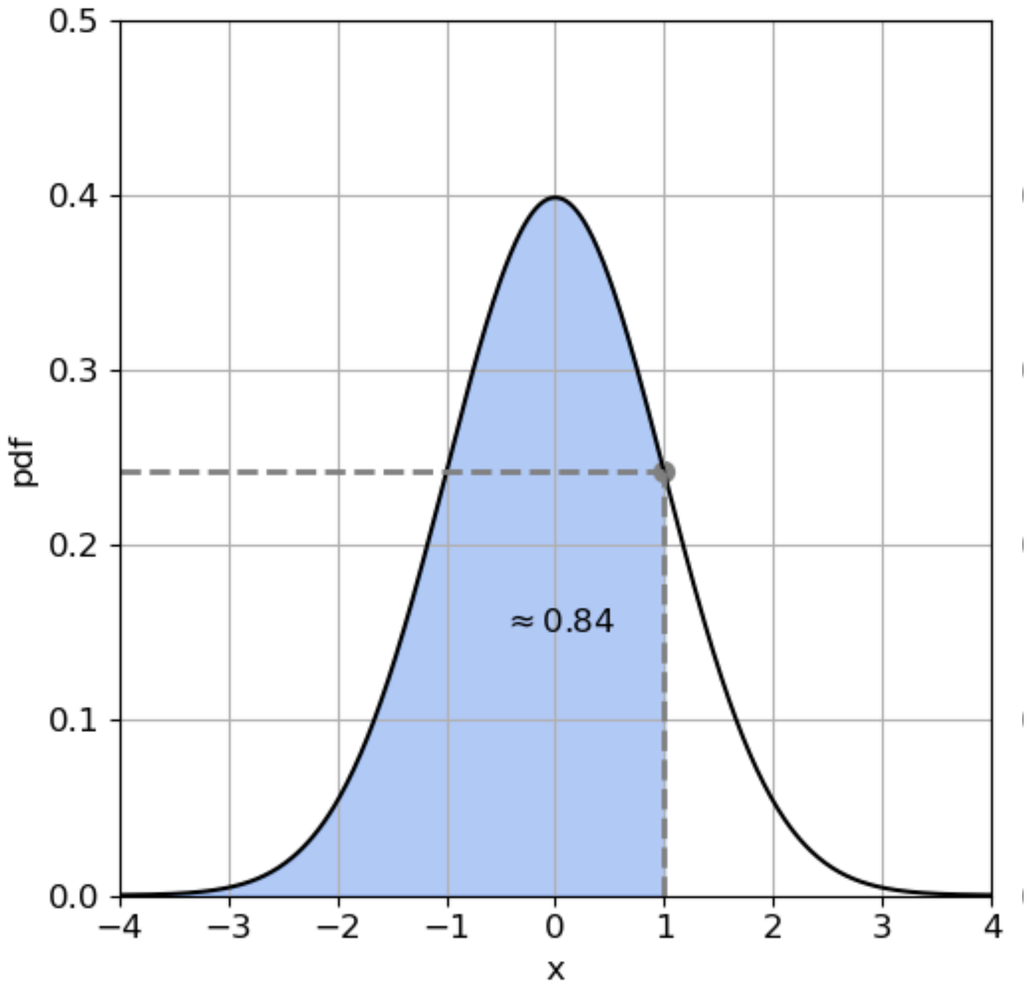
  Instead of P(A) $\rightarrow$ $\mathrm{P}(x \in \Omega)$

Refresher: one random variable

- The distribution of the random variable X has a PDF $f_X(x)$ and a CDF $F_X(x)$. The CDF is then defined as

$$P(X \leq x) = F_X(x) = \int_{-\infty}^{x} f_X(x)dx$$

The expression above can be applied also over an interval, Ω, by changing the limits of the integral.



Extending to a higher number of random variables

- Extending to **two random variables**, the *bivariate distribution* of the random variables X and Y has a PDF $f_{XY}(x,y)$ and a CDF $F_{XY}(x,y)$. The CDF is then defined as

$$\boxed{P(X \leq x, Y \leq y) = F_{XY}(x,y)} = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{XY}(x,y)dxdy$$

AND!

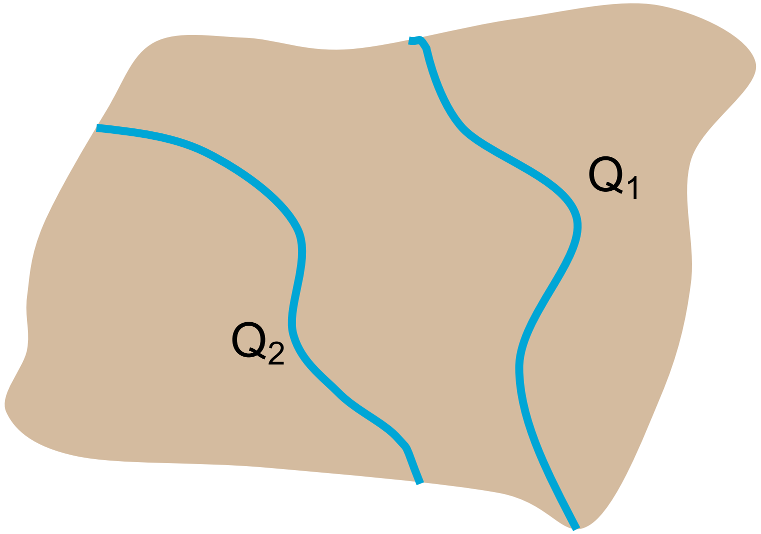
Joint probability/joint non-exceedance probability

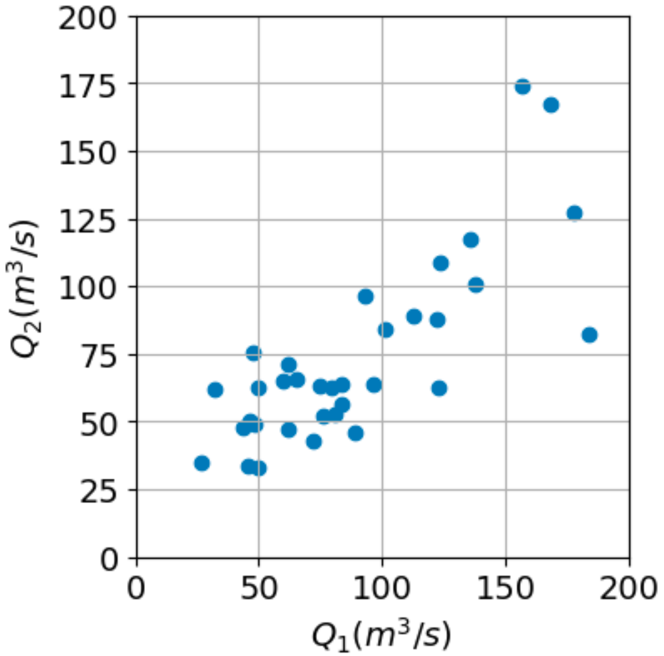
The expression above can be extended to a **higher number of variables** defining *multivariate distributions.* They can also be called *joint distributions*. It can also be applied on an interval, Ω.

We call *marginal distribution* to the univariate distribution associated with a single random variable that is part of a multivariate distribution. E.g.: in the previous slide $P(Q_1 > 100\ m^3/s)$

Empirical computations – one variable

Example case: discharge of two rivers located close by



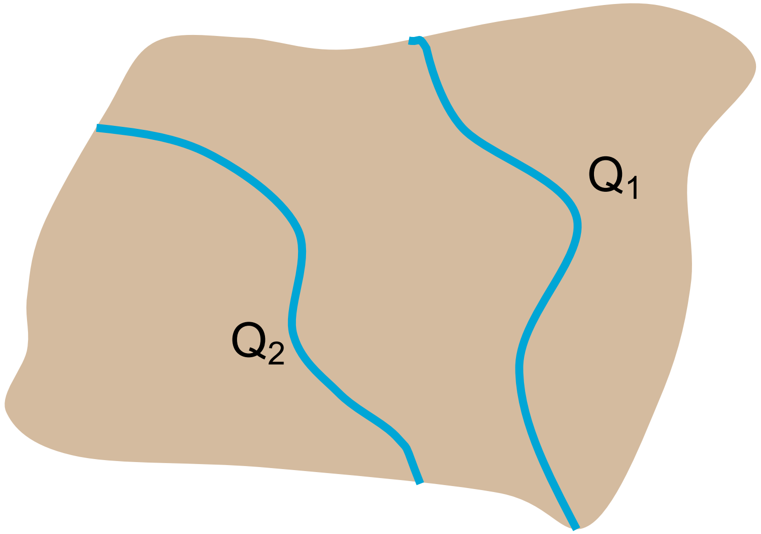


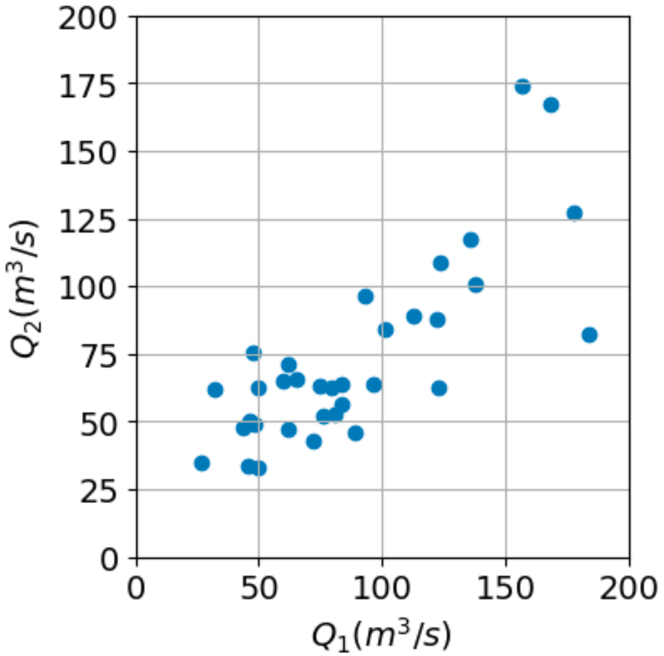
$P(Q_1 > 100\ m^3/s)$ =

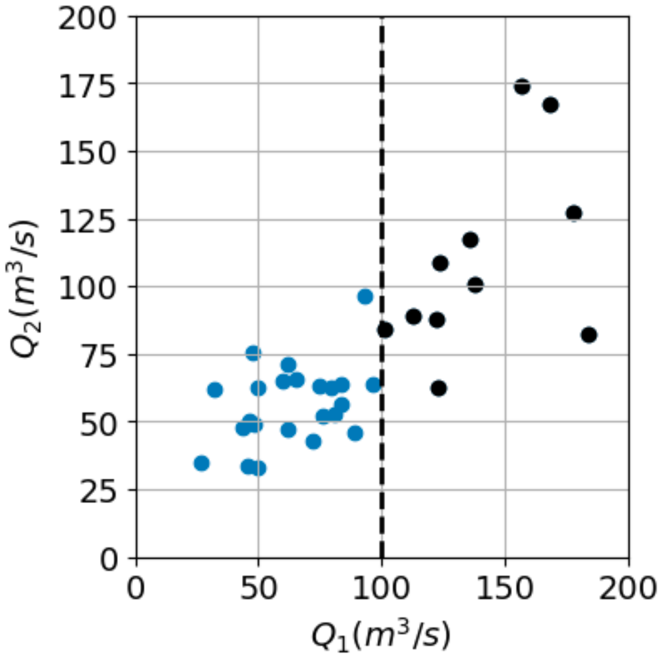
N = 34 observations

Empirical computations – one variable

Example case: discharge of two rivers located close by





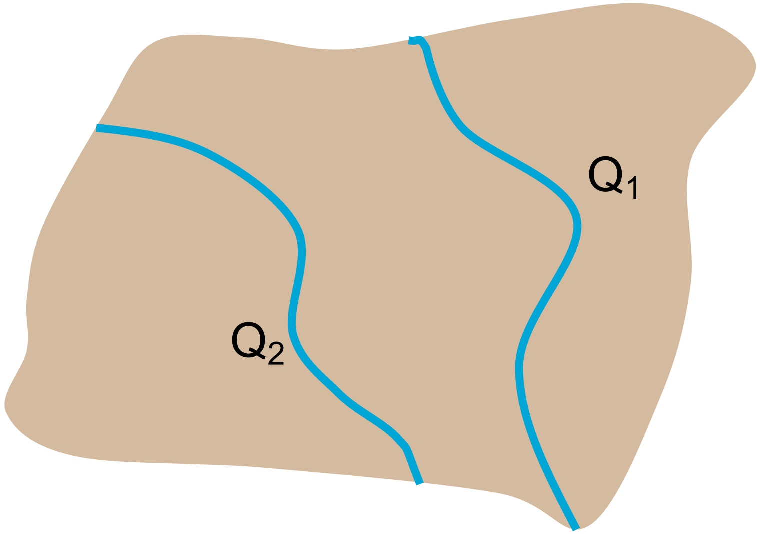


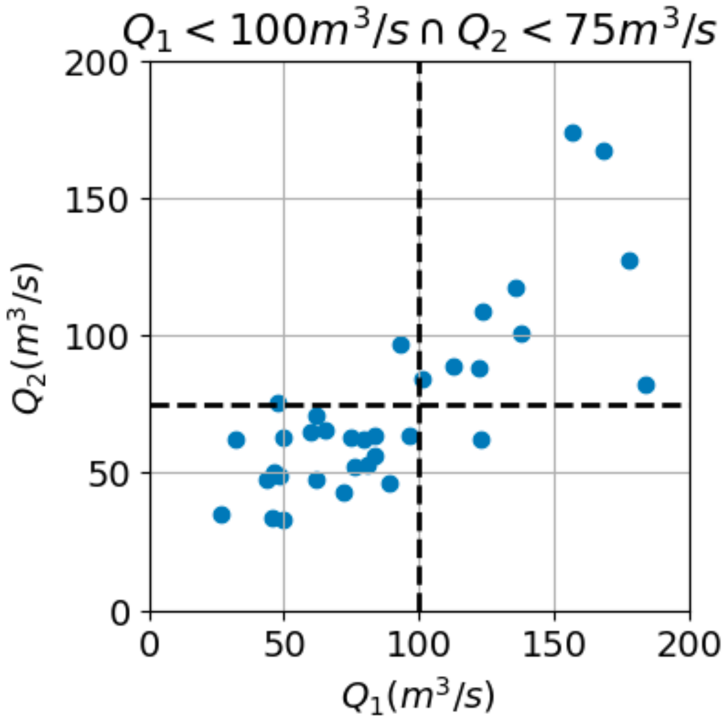
$P(Q_1 > 100\ m^3/s)$ = 11/34 = 0.32

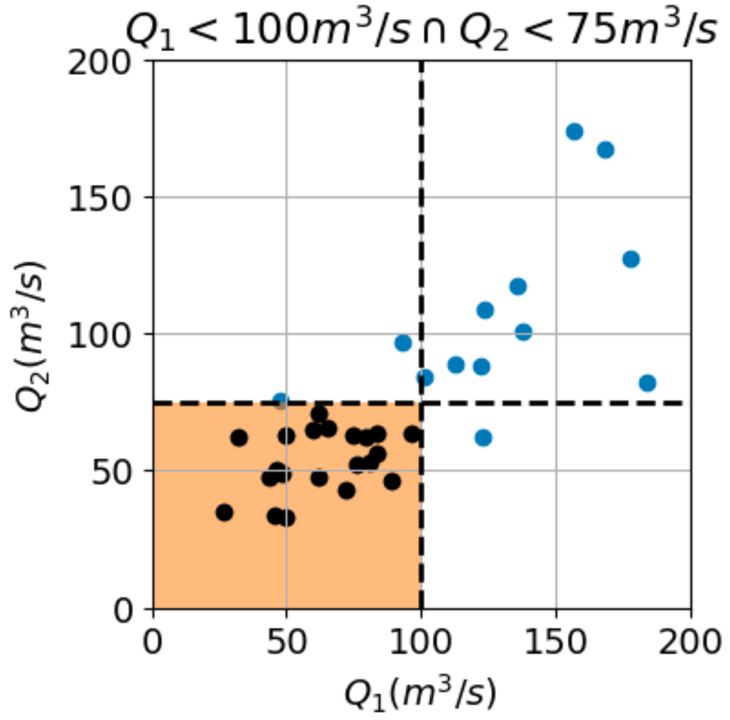
N = 34 observations

Empirical computations – two variables, joint prob

Example case: discharge of two rivers located close by





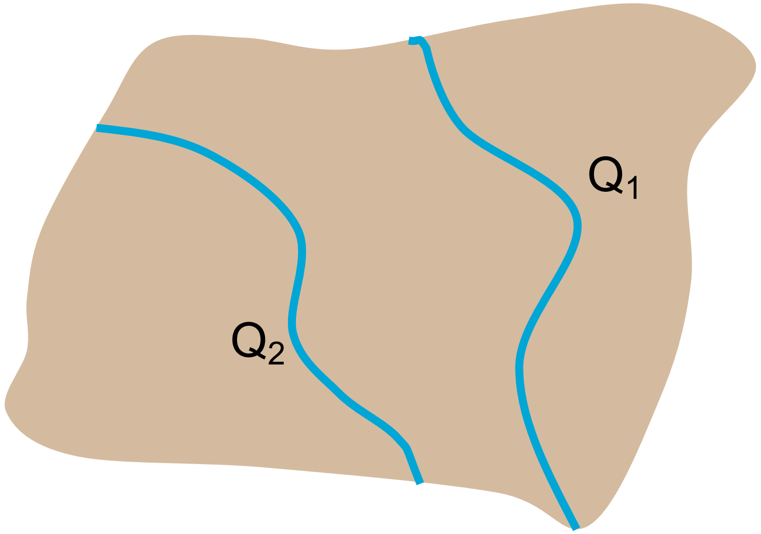


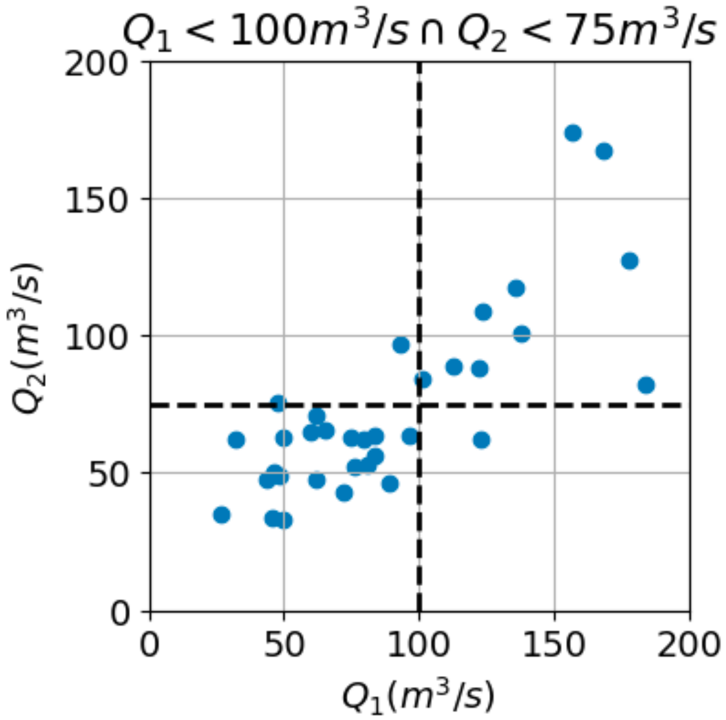
$P(Q_1 < 100\ m^3/s\,, Q_2 < 75\ m^3/s)$ =

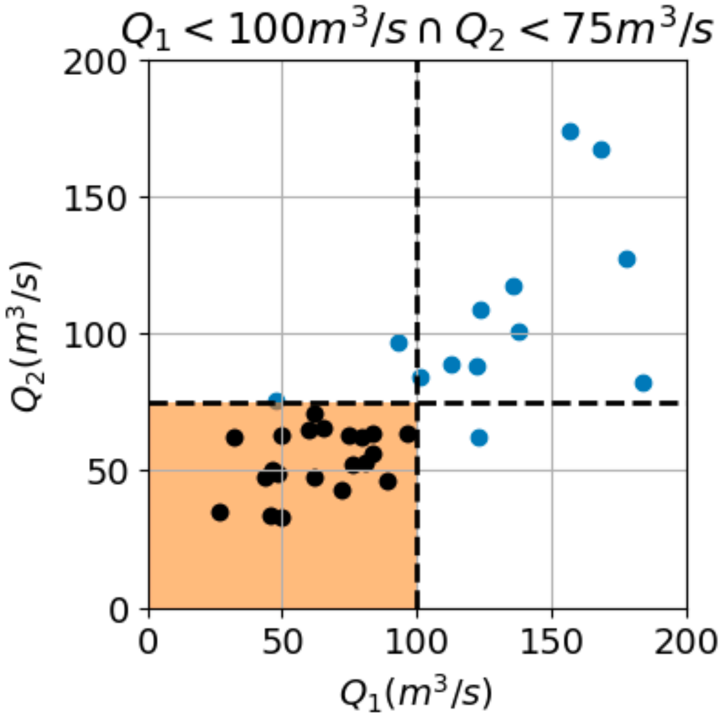
N = 34 observations

Empirical computations – two variables, joint prob

Example case: discharge of two rivers located close by





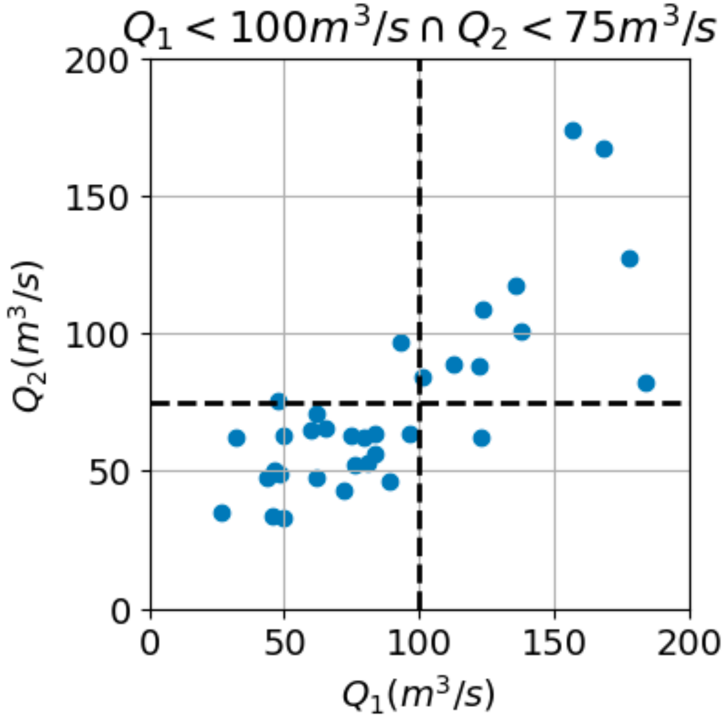


$P(Q_1 < 100\ m^3/s\,, Q_2 < 75\ m^3/s)$ = 21/34 = 0.62

N = 34 observations

Empirical computations – two variables, OR prob

N = 34 observations

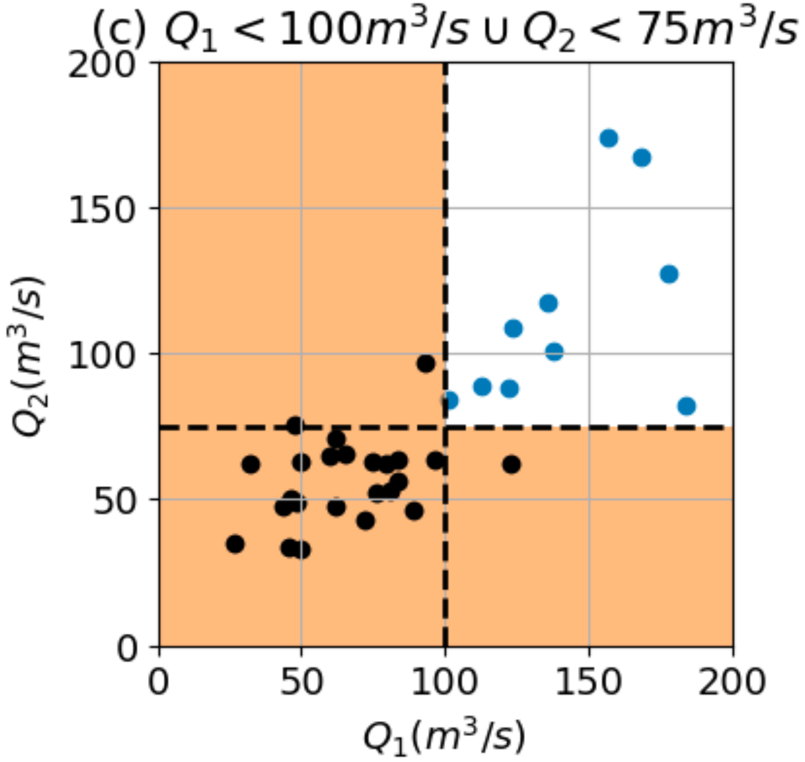


1) $P(Q_1 < 100\ m^3/s\ \ OR\ Q_2 < 75\ m^3/s)$ =

2) $P(Q_1 < 100\ m^3/s\ \ OR\ Q_2 < 75\ m^3/s)$ =

Empirical computations – two variables, OR prob

N = 34 observations

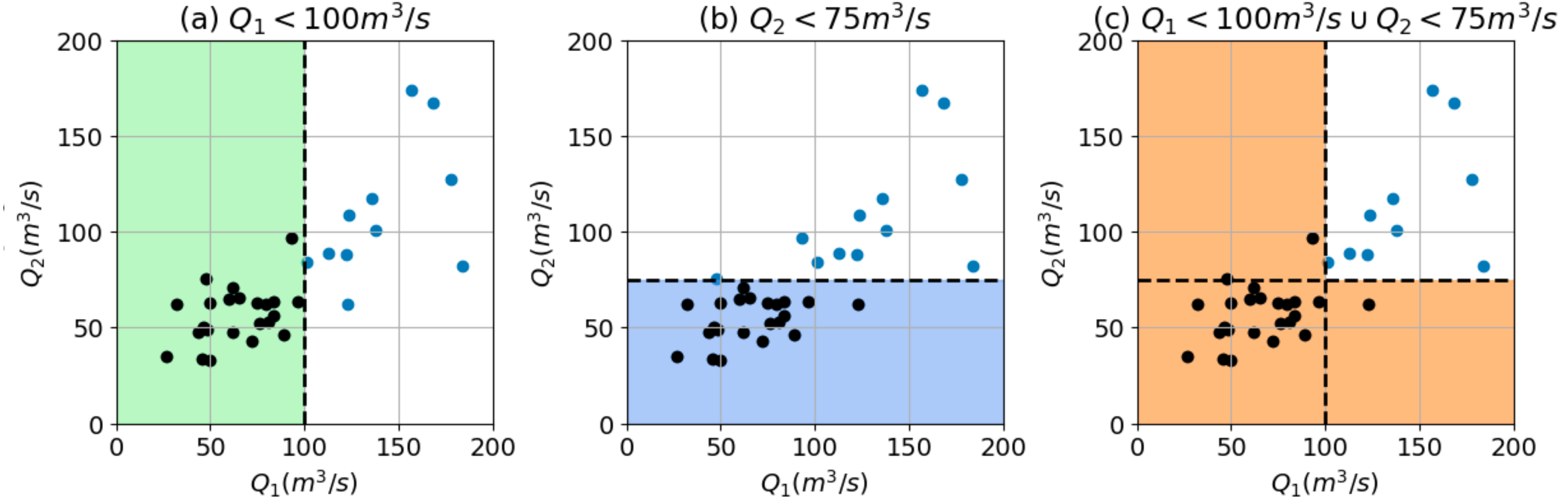


1) $P(Q_1 < 100\ m^3/s\ \ OR\ Q_2 < 75\ m^3/s)$ = 24/34 = 0.62

2) $P(Q_1 < 100\ m^3/s\ \ OR\ Q_2 < 75\ m^3/s)$ =

Empirical computations – two variables, OR prob

N = 34 observations



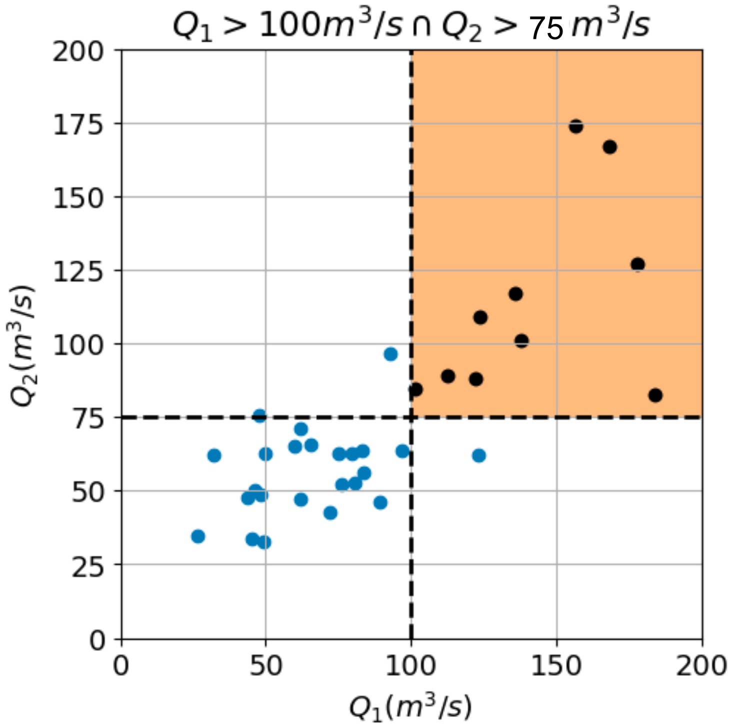
1) $P(Q_1 < 100\ m^3/s\ \ OR\ Q_2 < 75\ m^3/s)$ = 24/34 = 0.62

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

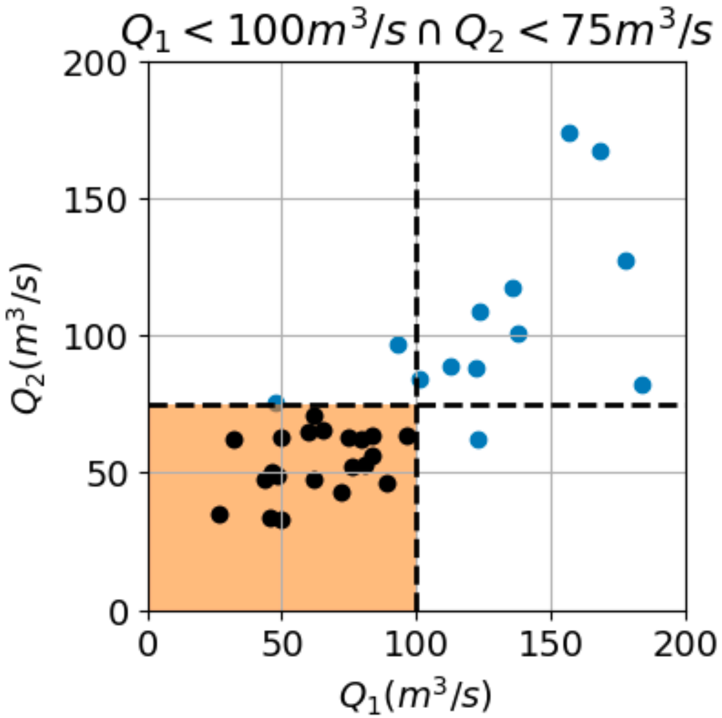
2) $P(Q_1 < 100\ m^3/s\ \ OR\ Q_2 < 75\ m^3/s)$ = $P(Q_1 < 100\ m^3/s) + P(Q_2 < 75\ m^3/s)$
$-P(Q_1 < 100\ m^3/s\,, Q_2 < 75\ m^3/s)$ =
=23/34+22/34− 21/34= 24/34

TU Delft

Empirical computations – two variables, joint exceed



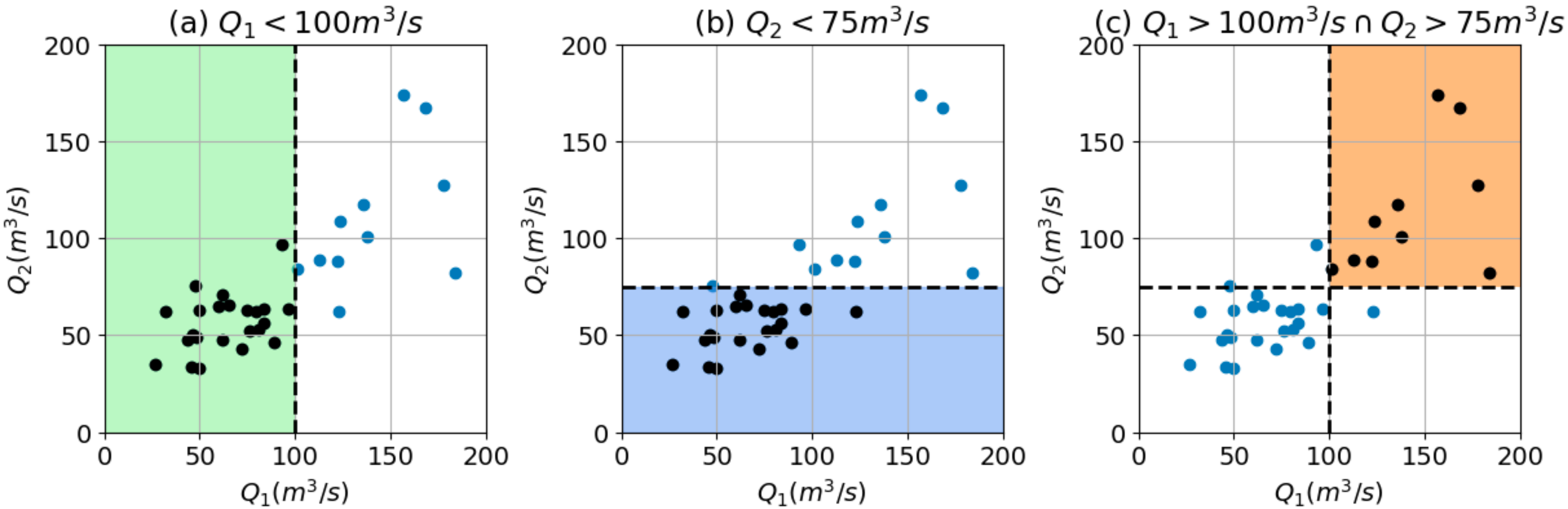
$\neq$ 1 -



$P(Q_1 > 100\, m^3/s\,, Q_2 > 75\, m^3/s)$ =

$$P(X > x) = 1 - P(X \leq x) = 1 - F(x)$$
$$P(X > x, Y > y) \neq 1 - P(X \leq x, Y \leq y)$$
$$\neq 1 - F_{XY}(x, y)$$

Empirical computations – two variables, joint exceed



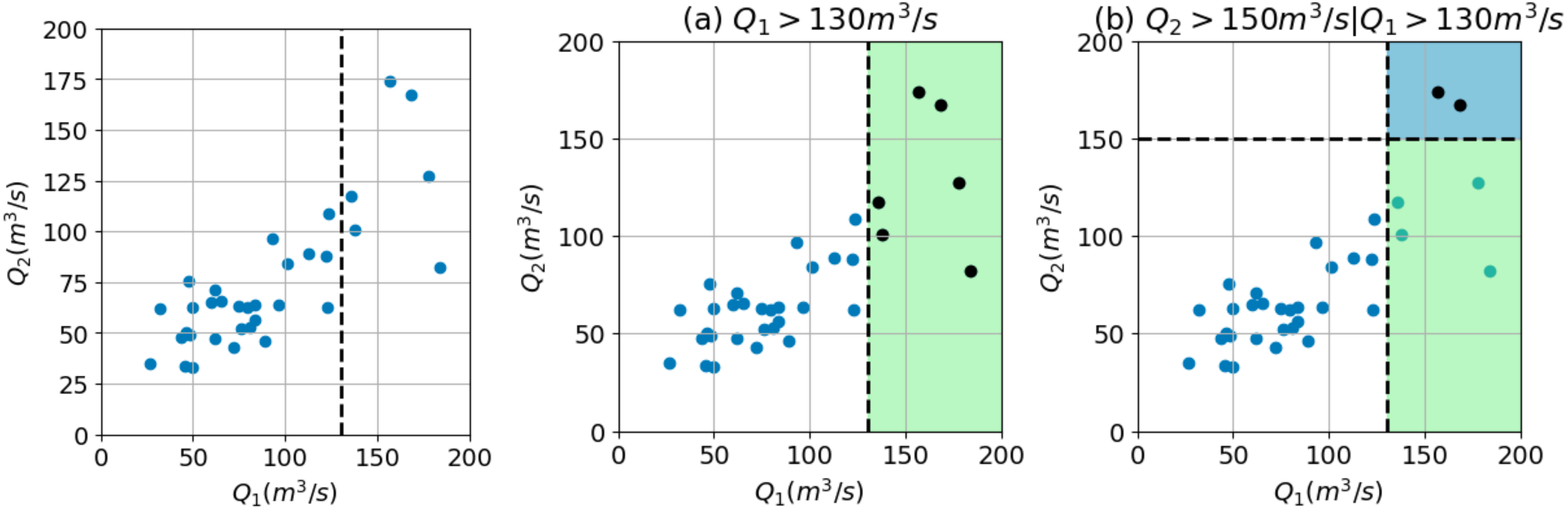
$P(Q_1 > 100\ m^3/s\ , Q_2 > 75\ m^3/s)$ = 1 - $P(Q_1 < 100\ m^3/s\ \ OR\ Q_2 < 75\ m^3/s)$ =

= 1 $- [P(Q_1 < 100\ m^3/s) + P(Q_2 < 75\ m^3/s) - P(Q_1 < 100\ m^3/s\ , Q_2 < 75\ m^3/s)]$ =

= 1 $- P(Q_1 < 100\ m^3/s) - P(Q_2 < 75\ m^3/s) + P(Q_1 < 100\ m^3/s\ , Q_2 < 75\ m^3/s) =$

= 1−23/34 − 22/34 + 21/34= 10/34

Empirical computations – two variables, conditional

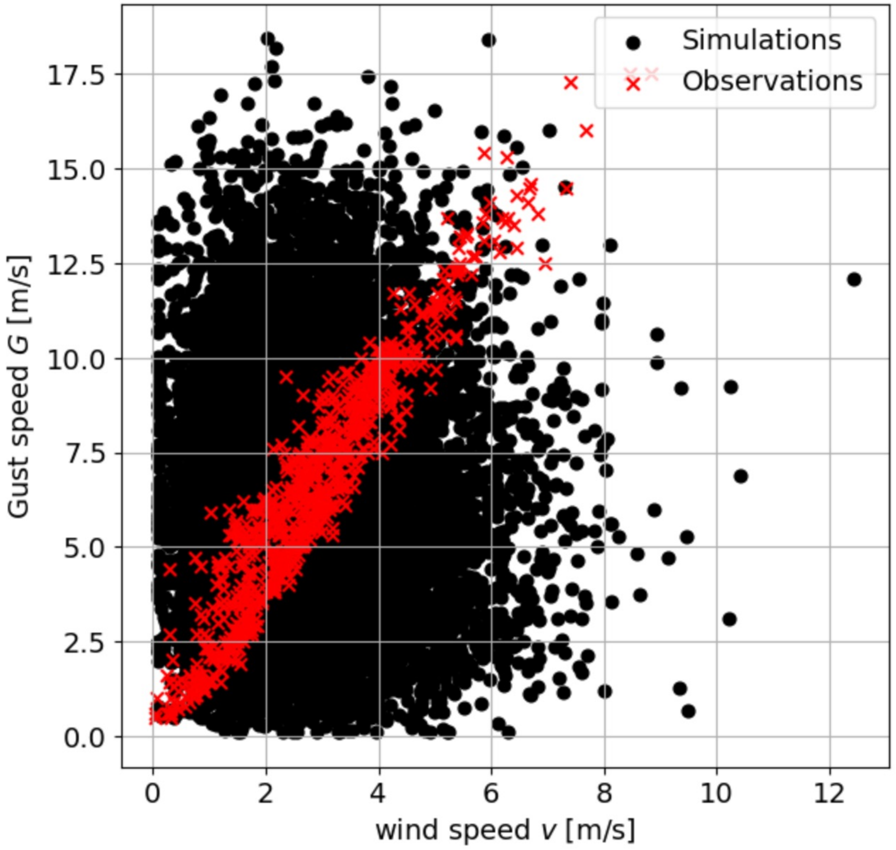


$P(Q_2 > 150\ m^3/s\ | Q_1 > 130\ m^3/s)$ = 2 / 6

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Contents – Multivariate distributions

1. Recap & motivation
2. Set theory and basic operations
3. Continuous variables
   a. Extension to multivariate & empirical computations
   b. Measures of dependence
4. Multivariate Gaussian distribution



Measures of dependence

Measures of dependence

Are two variables related? How much?

Covariance: measure of the joint variability of two variables

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}[(X_{1,i} - \mathbb{E}(X_1))(X_{2,i} - \mathbb{E}(X_2))] = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2)$$

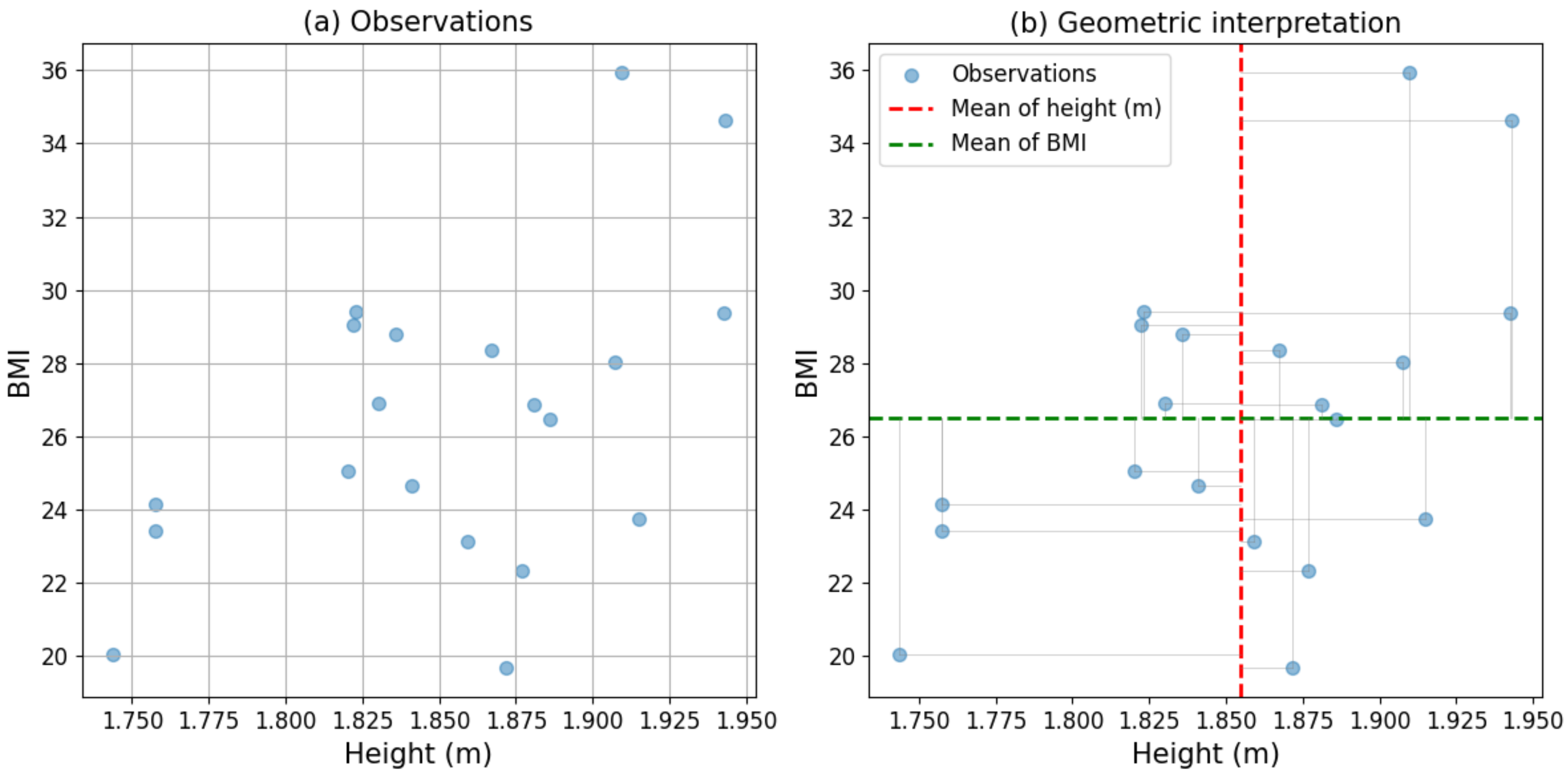
- Can take both positive and negative values
- Units equal to the product of the units of the analyzed variables
- High absolute values of covariance imply a strong relationship between variables.
- If $\mathrm{Cov}(X_1,X_2)>0$, high values of X_1 typically occur together with high values of X_2
- If $\mathrm{Cov}(X_1,X_2)<0$, high values of X_1 typically occur together with low values of X_2

TU Delft

Measures of dependence

Covariance. Geometric interpretation

$$Cov(X_1, X_2) = \mathbb{E}[(X_{1,i} - \mathbb{E}(\boldsymbol{X_1}))(X_{2,i} - \mathbb{E}(\boldsymbol{X_2}))]$$

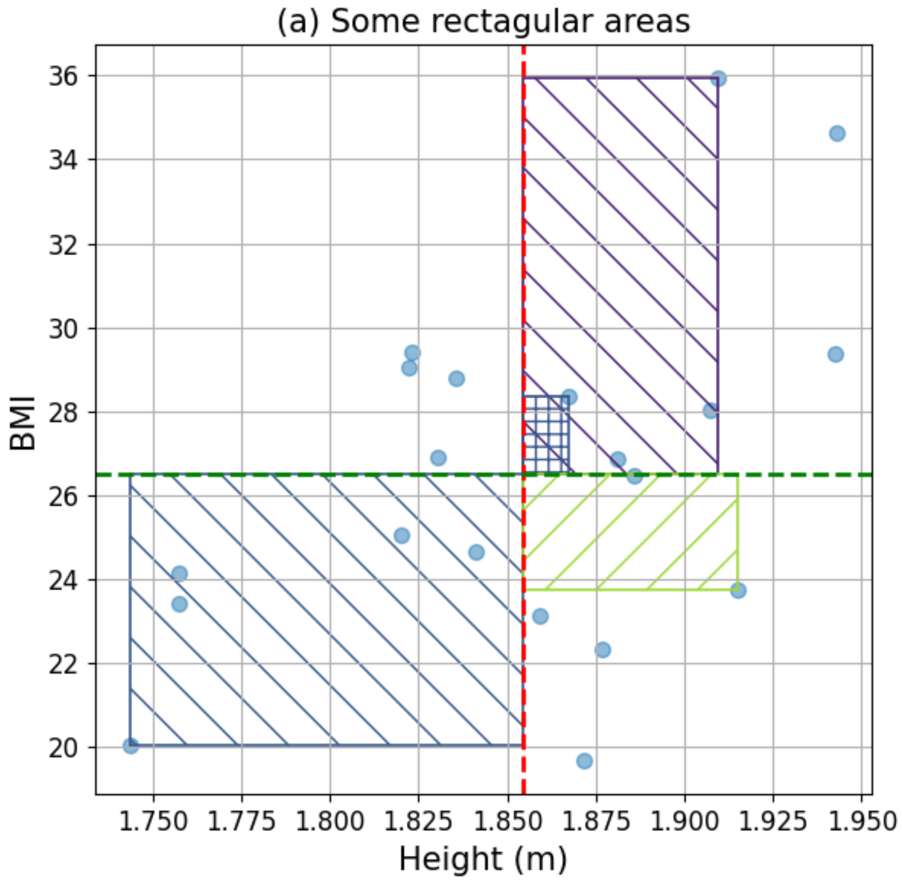


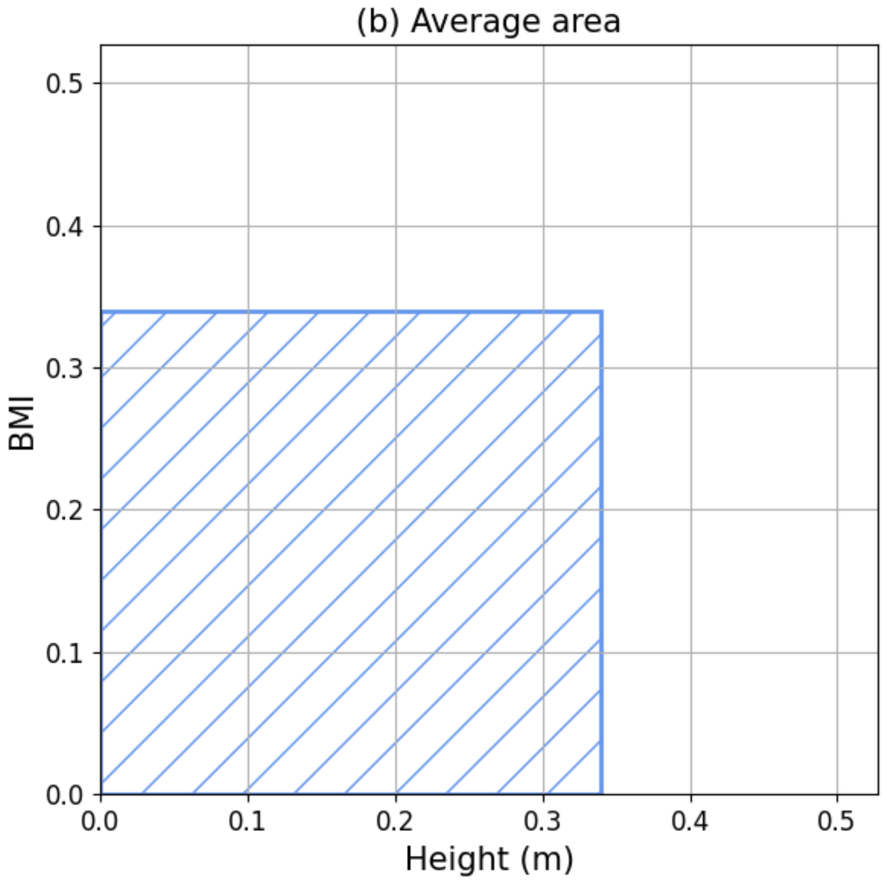
TU Delft

Measures of dependence

Covariance. Geometric interpretation

$$Cov(X_1, X_2) = \mathbb{E}[(X_{1,i} - \mathbb{E}(\boldsymbol{X_1}))(X_{2,i} - \mathbb{E}(\boldsymbol{X_2}))]$$





TU Delft

Measures of dependence

Covariance: measure of the joint variability of two variables

$$\text{Cov}(X_1, X_2) = \mathbb{E}[(X_{1,i} - \mathbb{E}(X_1))(X_{2,i} - \mathbb{E}(X_2))]$$

Let's do an example!

X_1: Wind speed (m/s)	X_2: Wave height (m)
12	2.5
0	0.4
36	1.7
3	0.2
14	2.6
25	3.1

$E(X_1)$=15 m/s

$E(X_1)$=1.75 m

Measures of dependence

Covariance: measure of the joint variability of two variables

$$\text{Cov}(X_1, X_2) = \mathbb{E}[(X_{1,i} - \mathbb{E}(X_1))(X_{2,i} - \mathbb{E}(X_2))]$$

Let's do an example!

X_1: Wind speed (m/s)	X_2: Wave height (m)	T1: X_{1i}-E(X_1) (m/s)	T2: X_{2i}-E(X_2) (m)
12	2.5		
0	0.4		
36	1.7		
3	0.2		
14	2.6		
25	3.1		
$E(X_1)=15$ m/s	$E(X_1)=1.75$ m		

$E(T_1 x T_2)=8.03$

Measures of dependence

Covariance: measure of the joint variability of two variables

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}[(X_{1,i} - \mathbb{E}(X_1))(X_{2,i} - \mathbb{E}(X_2))]$$

Let's do an example!

X_1: Wind speed (m/s)	X_2: Wave height (m)	T1: X_{1i}-E(X_1) (m/s)	T2: X_{2i}-E(X_2) (m)	T1 x T2 (m/sxm)
12	2.5	-3	0.75	-2.25
0	0.4	-15	-1.35	20.25
36	1.7	21	-0.05	-1.05
3	0.2	-12	-1.55	18.6
14	2.6	-1	0.85	-0.85
25	3.1	10	1.35	13.5

$E(X_1)$=15 m/s

$E(X_1)$=1.75 m

$E(T_1xT_2)$=8.03

Measures of dependence

$Cov(H_s, W_s)=8.03m^2/s$

Covariance: measure of the joint variability of two variables

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}[(X_{1,i} - \mathbb{E}(X_1))(X_{2,i} - \mathbb{E}(X_2))] = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2)$$

- Can take both positive and negative values
- Units equal to the product of the units of the analyzed variables
- High absolute values of covariance imply a strong relationship between variables.
- If $Cov(X_1,X_2)>0$, high values of X_1 typically occur together with high values of X_2
- If $Cov(X_1,X_2)<0$, high values of X_1 typically occur together with low values of X_2

Measures of dependence

$Cov(H_s, W_s) = 8.03 m^2/s$

Covariance: measure of the joint variability of two variables

$$Cov(X_1, X_2) = \mathbb{E}[(X_{1,i} - \mathbb{E}(X_1))(X_{2,i} - \mathbb{E}(X_2))] = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2)$$

- Can take both positive and negative values
- Units equal to the product of the units of the analyzed variables
- High absolute values of covariance imply a strong relationship between variables.
- If $Cov(X_1,X_2)>0$, high values of X_1 typically occur together with high values of X_2
- If $Cov(X_1,X_2)<0$, high values of X_1 typically occur to

Drawback: not standardized measure; difficult to compare different pairs of random variables with different units

Measures of dependence

There are other correlation coefficients!

Pearson's correlation coefficient: assesses the linear correlation between two random variables.

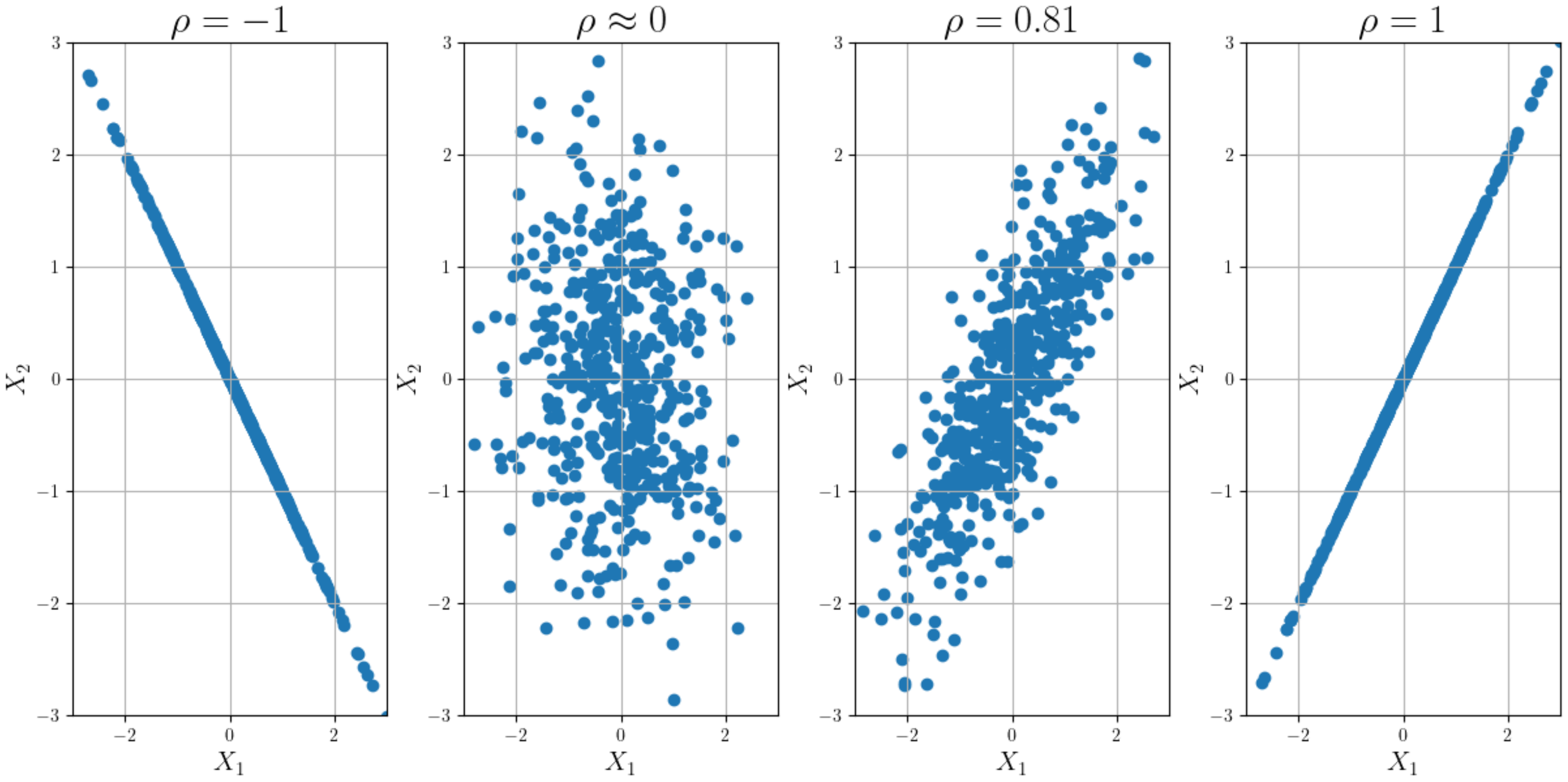
$$\rho = \frac{Cov(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}} = \frac{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sqrt{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2 \sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2}}$$

- Values between -1 and 1, regardless the units of the random variables.
- $\rho \approx 0$: random variables are uncorrelated or independent.
- $\rho > 0$: if one variable increases, the other one tends to increase.
- $\rho \approx 1$ or -1: knowing one variable implies I know the other variable through a linear relationship.

Measures of dependence

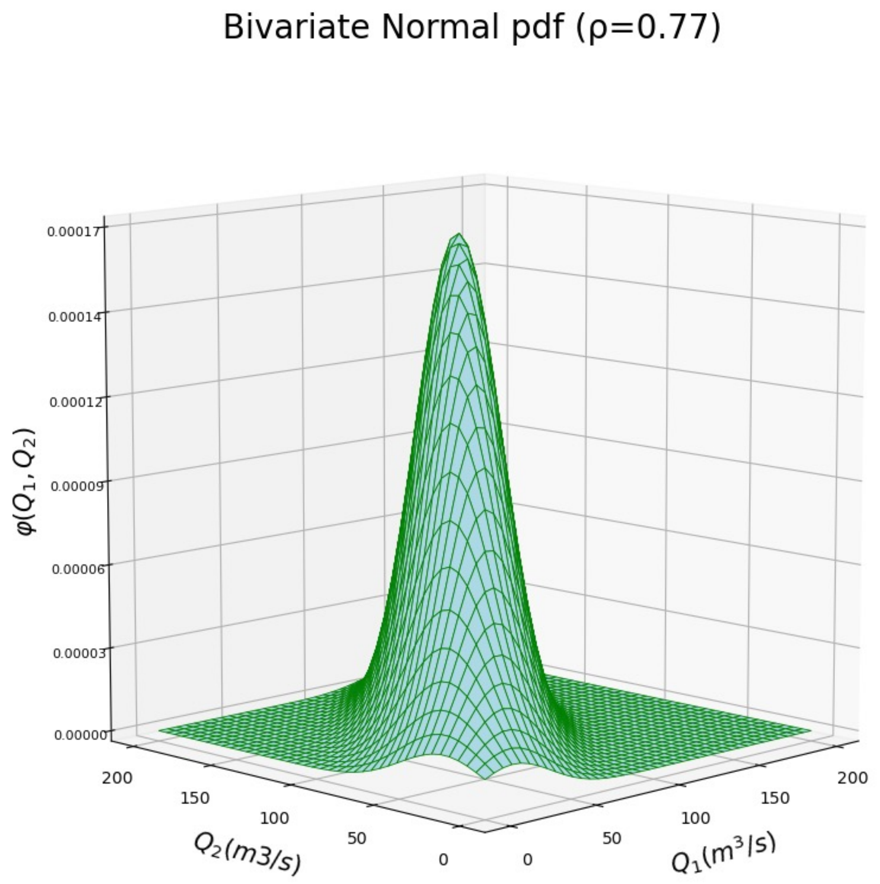
There are other correlation coefficients!

Pearson's correlation coefficient: assesses the linear correlation between two random variables.



Contents – Multivariate distributions

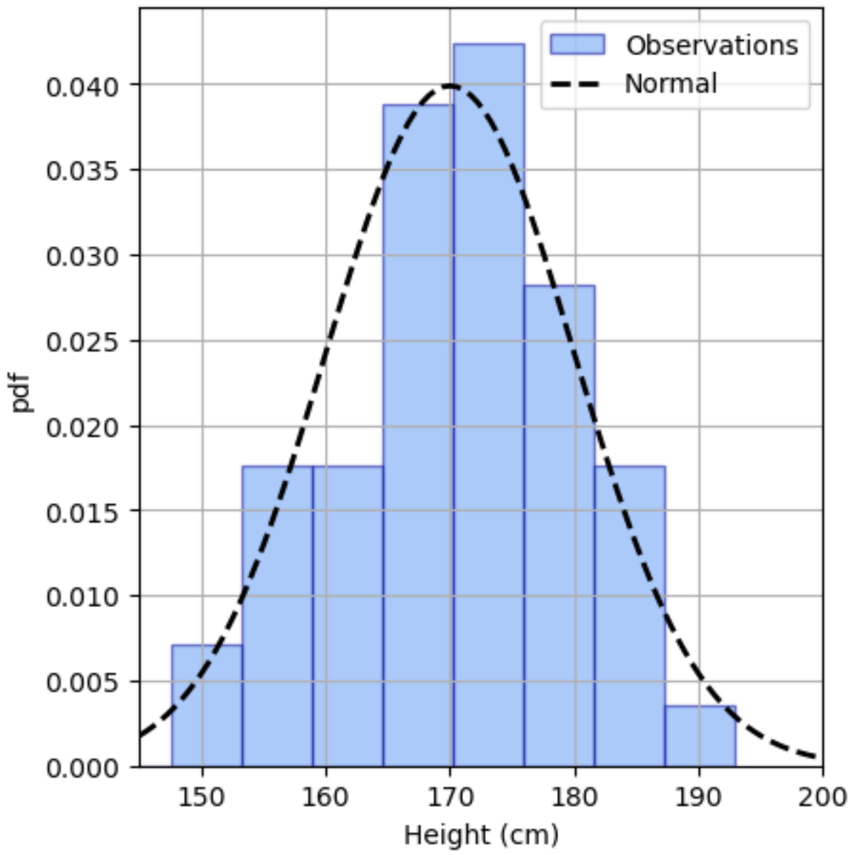
1. Recap & motivation
2. Set theory and basic operations
3. Continuous variables
    a. Extension to multivariate & empirical computations
    b. Measures of dependence
4. Multivariate Gaussian distribution



Multivariate Gaussian distribution

Modelling the joint distribution

Why did we use parametric distributions in week 4?



- Infer probabilities we have not observed
- Make easier/faster computations

Same holds for multivariate distributions!

Which distributions?

Here, only the multivariate Gaussian distribution.

Do you want to know **more**? Go to **MORE**!

Cross Over on Probabilistic modelling

Why?

- It can be manipulated analytically → wide range of applications
  - Reliability
  - Propagation of uncertainty (week 6)
  - Observation theory (weeks 7 and 8)
  - Between others
- First approach to model dependence

The model: multivariate Gaussian distribution

Bivariate Gaussian distribution PDF:

$$\phi_\rho(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{\left(\frac{x_1-\mu_1}{\sigma_1}\right) - \left(\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)}{2(1-\rho^2)} \right)$$

where

- $\phi_\rho(x_1, x_2)$ denotes the PDF of the bivariate Gaussian distribution of X_1 and X_2
- μ_1, μ_2 are the means of X_1 and X_2
- σ_1, σ_2 are the standard deviations of X_1 and X_2
- ρ is the Pearson's correlation coefficient of X_1 and X_2

The model: multivariate Gaussian distribution

Bivariate Gaussian distribution PDF:

$$\phi_\rho(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{\left(\frac{x_1-\mu_1}{\sigma_1}\right) - \left(\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)}{2(1-\rho^2)} \right)$$

I can be rewritten in matricial form as

$$\phi_\rho(x_1, x_2) = \frac{1}{\sqrt{(2\pi)^2 \begin{vmatrix} \sigma_1^2 & Cov(X_1, X_2) \\ Cov(X_1, X_2) & \sigma_2^2 \end{vmatrix}}} \exp\left(-\frac{1}{2}(x_1 - \mu_1 \; x_2 - \mu_2) \begin{pmatrix} \sigma_1^2 & Cov(X_1, X_2) \\ Cov(X_1, X_2) & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \right)$$

The model: multivariate Gaussian distribution

Bivariate Gaussian distribution PDF:

$$\phi_\rho(x_1, x_2) = \frac{1}{\sqrt{(2\pi)^2 \begin{vmatrix} \sigma_1^2 & Cov(X_1, X_2) \\ Cov(X_1, X_2) & \sigma_2^2 \end{vmatrix}}} \exp\left(-\frac{1}{2} \begin{pmatrix} x_1 - \mu_1 & x_2 - \mu_2 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & Cov(X_1, X_2) \\ Cov(X_1, X_2) & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \right)$$

Covariance matrix

Mean vector

Covariance matrix

Mean vector

Compressed fashion

$$\phi_\rho(x_1, x_2) = \frac{1}{\sqrt{(2\pi)^2 |\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right)$$

The model: multivariate Gaussian distribution

Bivariate Gaussian distribution CDF:

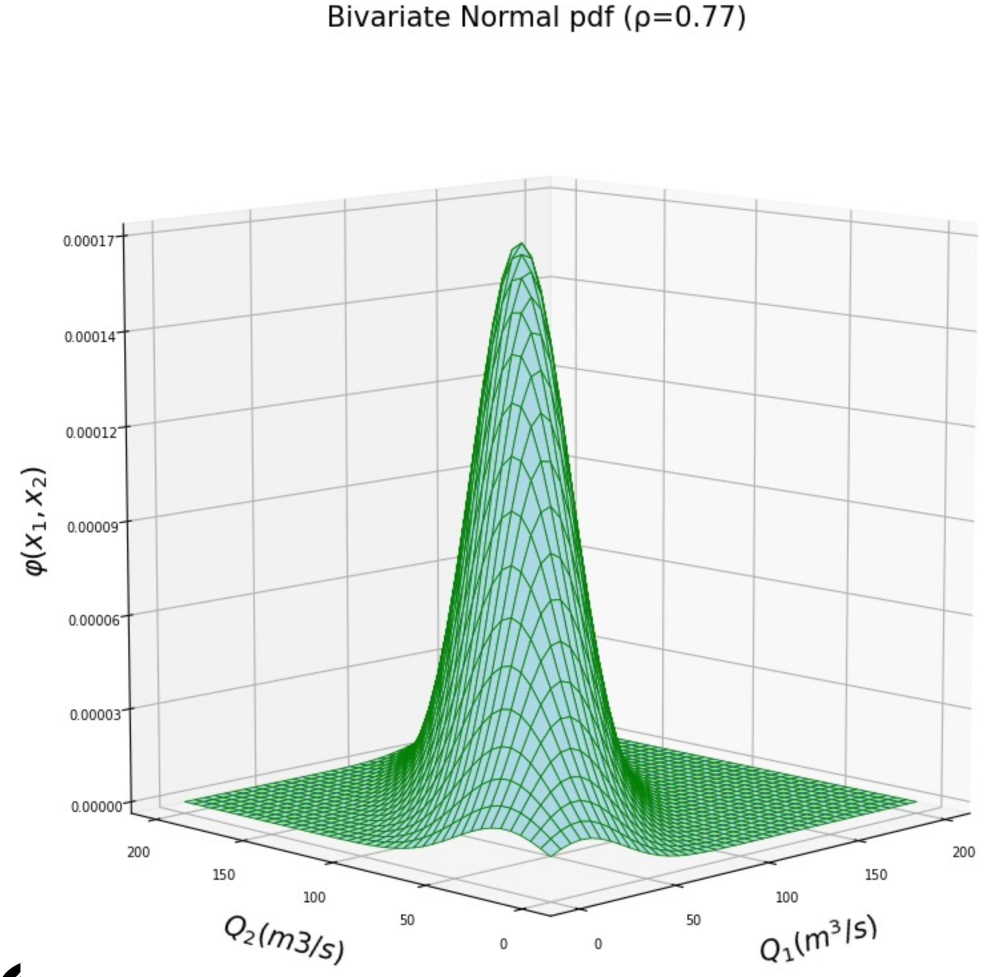
$$\Phi_{X_1,X_2}(x_1,x_2) = P(X_1 \leq x_1, X_2 \leq x_2) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} \phi_\rho(s_1,s_2)ds_1ds_2$$

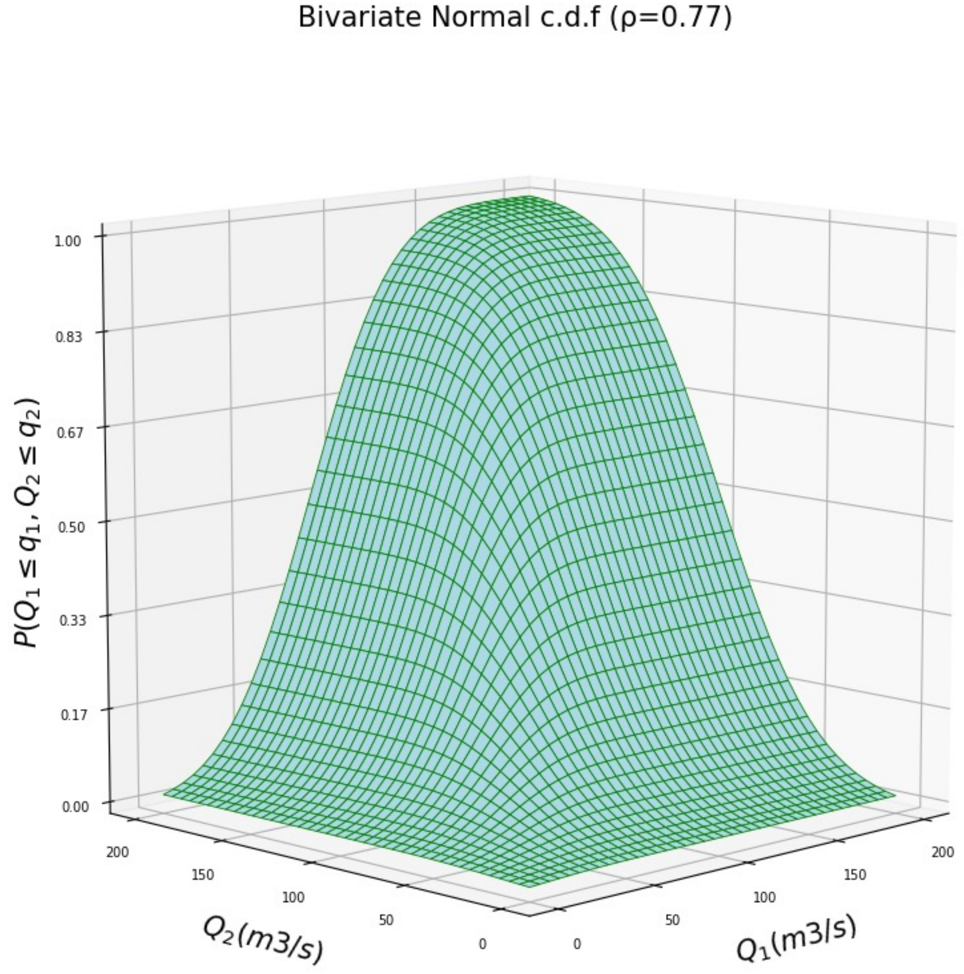
where

- $\Phi_{X_1,X_2}(x_1,x_2)$ denotes the CDF of the bivariate Gaussian distribution of X_1 and X_2

No closed form of the integral!

The model: multivariate Gaussian distribution





Multivariate Gaussian distribution: marginals

Bivariate Gaussian distribution PDF:

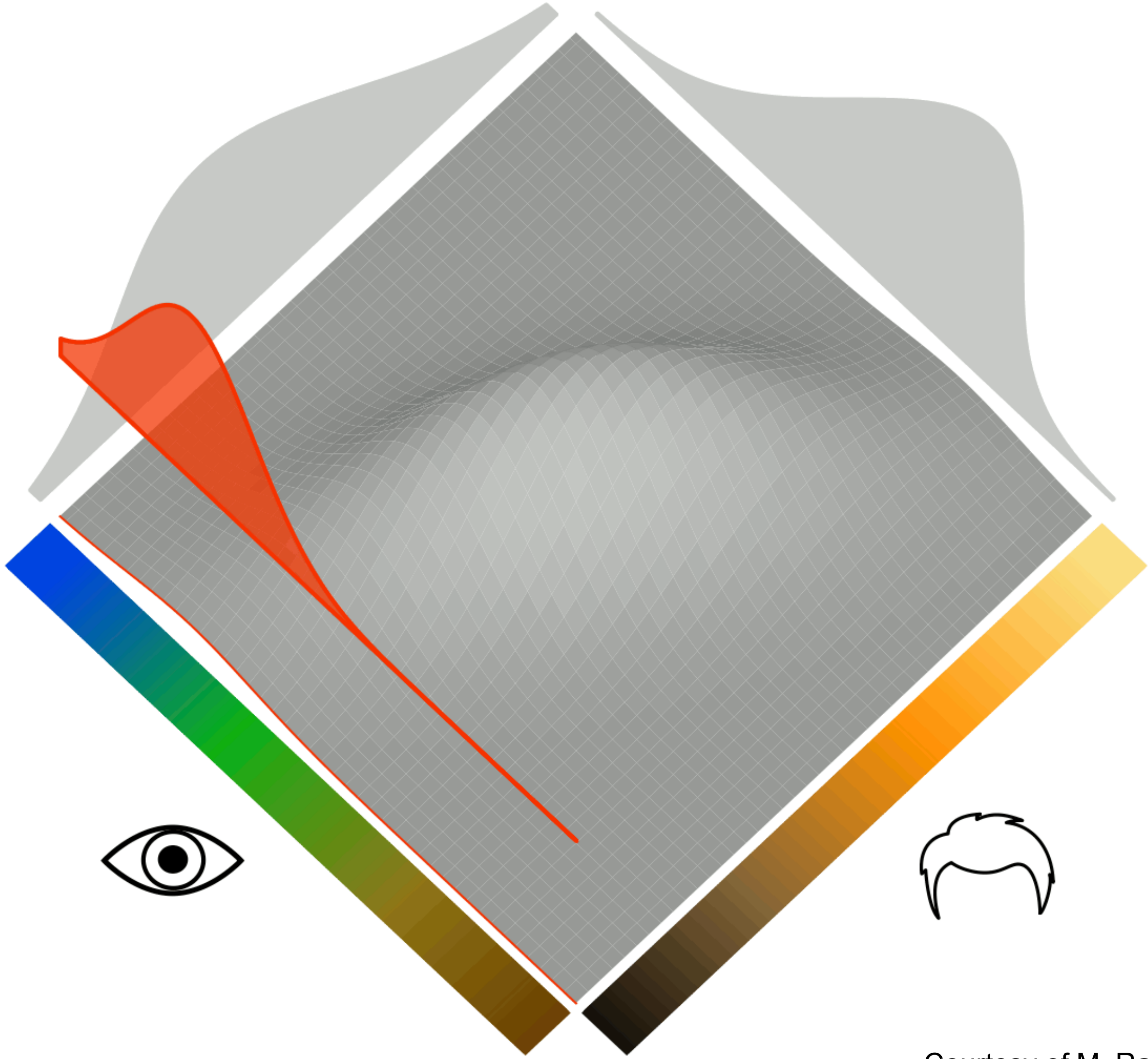
$$\phi_\rho(x_1, x_2) = \frac{1}{\sqrt{(2\pi)^2 \begin{vmatrix} \sigma_1^2 & Cov(X_1, X_2) \\ Cov(X_1, X_2) & \sigma_2^2 \end{vmatrix}}} \exp\left(-\frac{1}{2}(x_1 - \mu_1 \; x_2 - \mu_2) \begin{pmatrix} \sigma_1^2 & Cov(X_1, X_2) \\ Cov(X_1, X_2) & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}\right)$$

If I model the distribution of X_1 and X_2 using a bivariate Gaussian distribution, how is the univariate distribution of X_2?

Gaussian!

TU Delft

Multivariate Gaussian distribution: marginals



Courtesy of M. Ramgraber

Multivariate Gaussian distribution: fitting

Bivariate Gaussian distribution PDF:

$$\phi_\rho(x_1, x_2) = \frac{1}{\sqrt{(2\pi)^2 \begin{vmatrix} \sigma_1^2 & Cov(X_1, X_2) \\ Cov(X_1, X_2) & \sigma_2^2 \end{vmatrix}}} \exp\left(-\frac{1}{2} (x_1 - \mu_1 \; x_2 - \mu_2) \begin{pmatrix} \sigma_1^2 & Cov(X_1, X_2) \\ Cov(X_1, X_2) & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \right)$$

Given data of X_1 and X_2, how can I fit the above distribution?

By moments!

I can compute $\mu_1, \mu_2, \sigma_1, \sigma_2, Cov(X_1, X_2)$ to define the covariance matrix and the vector of means.

Multivariate Gaussian distribution: conditionalizing

What happens to the distribution of X_1 given that I know the value of X_2?

This is, the *conditional distribution of X_1 given X_2*

Property of Gaussian distribution:

If two sets of random variables are jointly Gaussian, then the conditional distribution of one set conditioned to other is again Gaussian

- Gaussian distribution is "friendly": we have analytical expressions

Multivariate Gaussian distribution: conditionalizing

$(x_1|x_2 = a) \sim N(\hat{\mu}, \hat{\Sigma})$

Using the definition of conditional density, replacing x_2 by *a* and doing the (unpleasant) algebra, we obtain

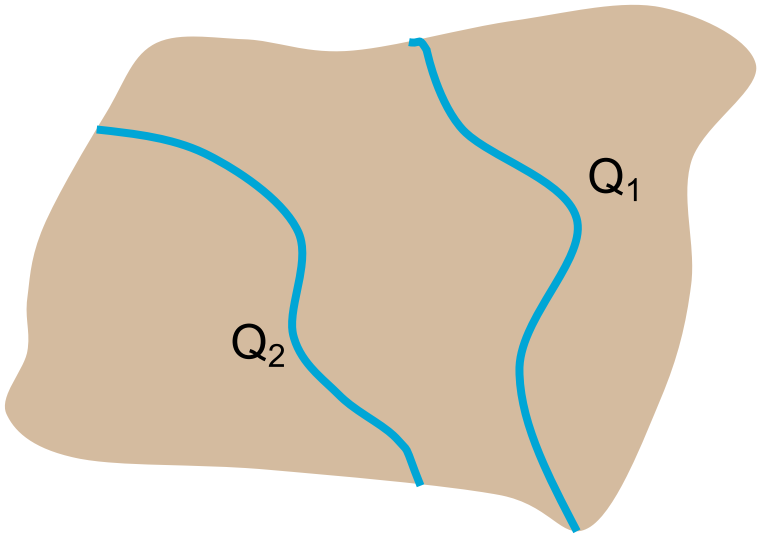
$$\hat{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)$$

$$\hat{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

where a is the known value of X_2 (here, Q_2) and $\mathbf{\Sigma} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$

TUDelft

Conditional *bivariate* Gaussian distribution - example



$$\boldsymbol{\mu} = \begin{pmatrix} 94 \\ 78 \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 41^2 & 1000 \\ 1000 & 35^2 \end{pmatrix}$$

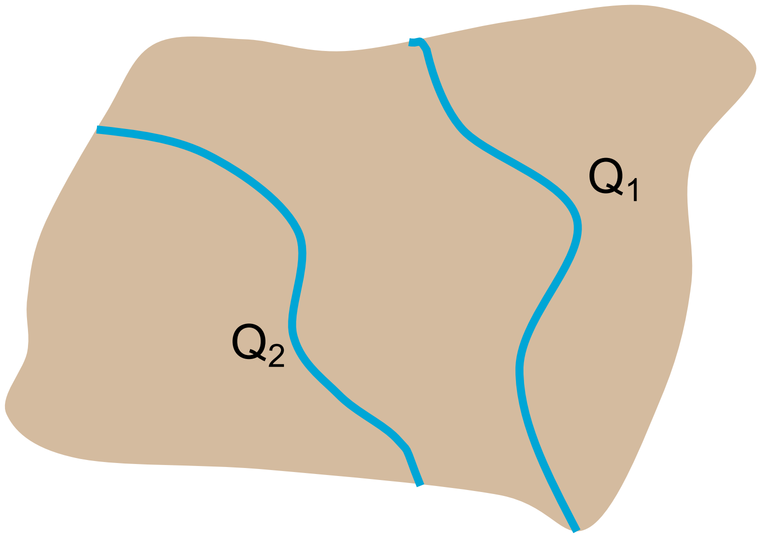
Given Q_2=100m^3/s, what is the distribution of Q_1?

$$\hat{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)$$

$$\hat{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

TU Delft

Conditional *bivariate* Gaussian distribution - example



$$\boldsymbol{\mu} = \begin{pmatrix} 94 \\ 78 \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 41^2 & 1000 \\ 1000 & 35^2 \end{pmatrix}$$

Given Q_2=100m^3/s, what is the distribution of Q_1?

$$\hat{\mu} = 94 + 1000 \times (35^2)^{-1} \times (100 - 78) \approx 112.0$$

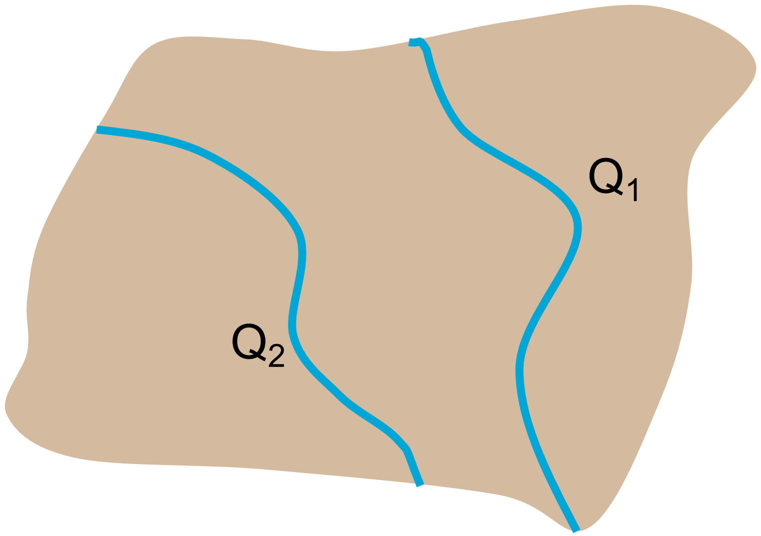
$$\hat{\Sigma} = 41^2 - 1000 \times (35^2)^{-1} \times 1000 \approx 864.7 \rightarrow \sigma_{x_1|x_2=a} = 29.4$$

$$\hat{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)$$

$$\hat{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

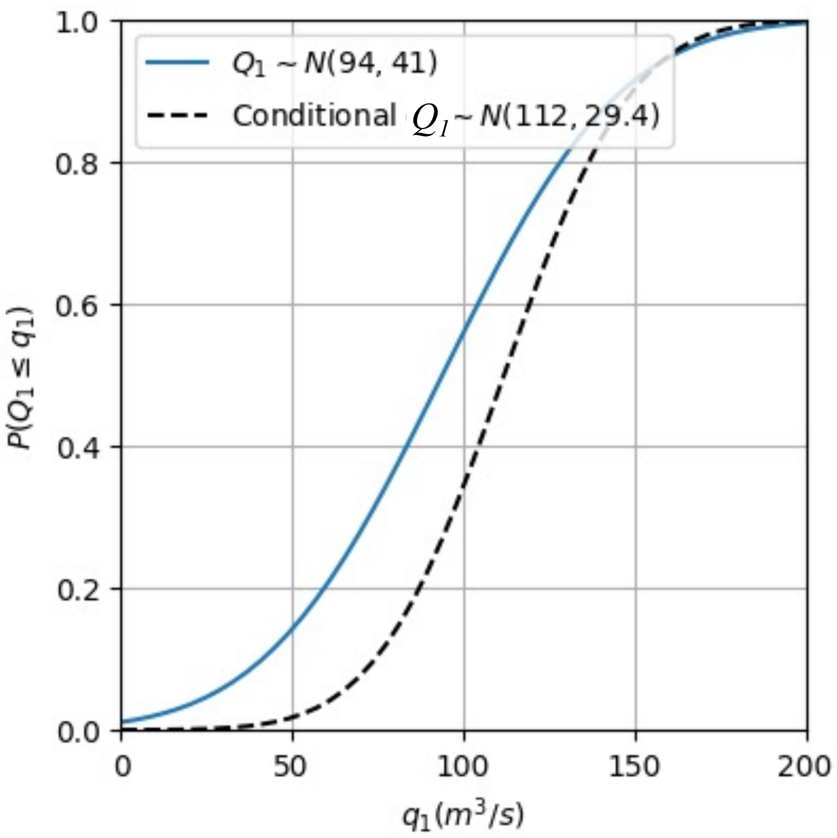
TU Delft

Conditional *bivariate* Gaussian distribution - example



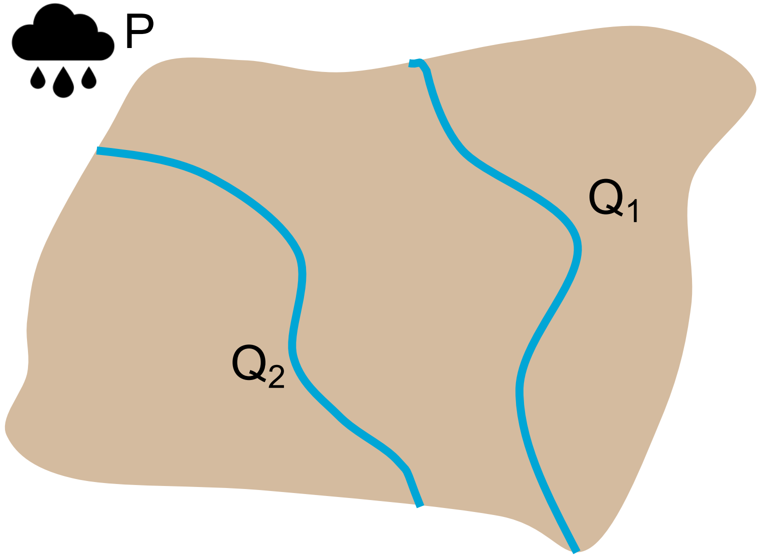
$$\boldsymbol{\mu} = \begin{pmatrix} 94 \\ 78 \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 41^2 & 1000 \\ 1000 & 35^2 \end{pmatrix}$$

Given Q_2=100m^3/s, what is the distribution of Q_1?



TU Delft

Conditional *multivariate* Gaussian distribution - example



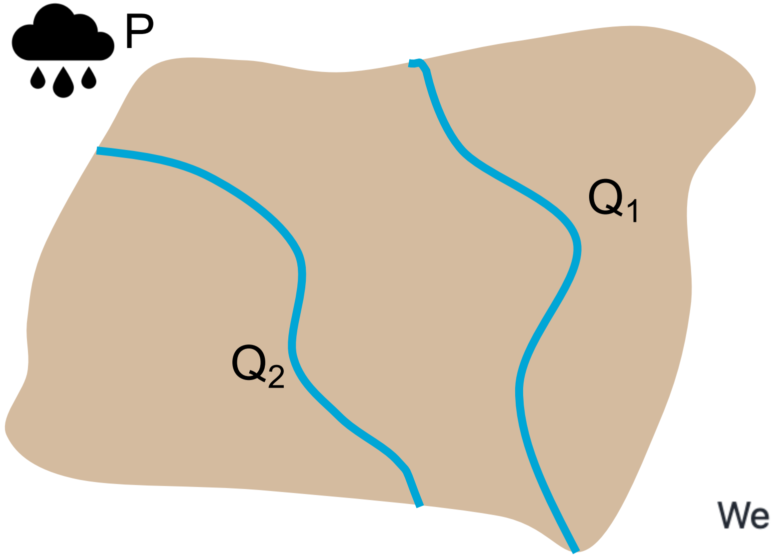
We can extend the covariance matrix on a higher number of dimensions.

$$\boldsymbol{\mu} = \begin{pmatrix} 94 \\ 78 \\ 12 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 41^2 & 1000 & 475 \\ 1000 & 35^2 & 520 \\ 475 & 520 & 27^2 \end{pmatrix}$$

Given p=22mm/h, what is the distribution of Q_1 and Q_2?

Note: it is a bivariate margin!!

Conditional *multivariate* Gaussian distribution - example



We can extend the covariance matrix on a higher number of dimensions.

$$\boldsymbol{\mu} = \begin{pmatrix} 94 \\ 78 \\ 12 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 41^2 & 1000 & 475 \\ 1000 & 35^2 & 520 \\ 475 & 520 & 27^2 \end{pmatrix}$$

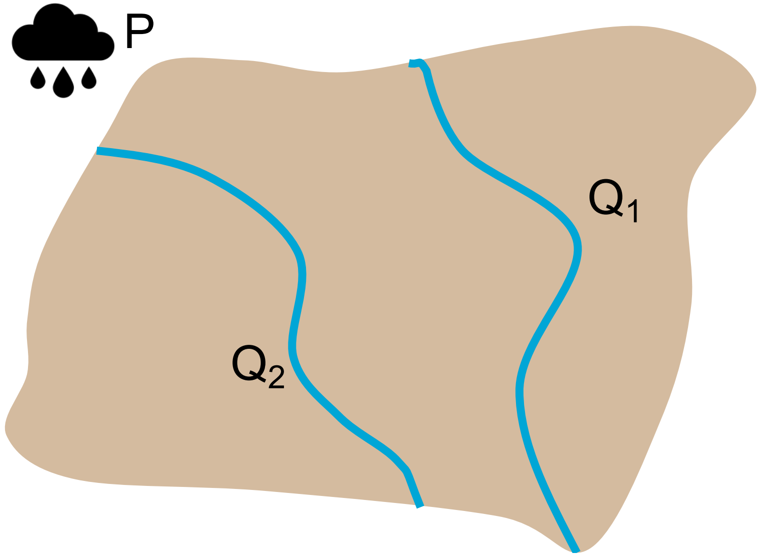
We can extend the analytical equation to conditionalize the bivariate Gaussian distribution to the 3D multivariate Gaussian distribution to compute $(x_1, x_2 | x_3 = a) \sim N(\hat{\mu}, \hat{\Sigma})$ as

$$\hat{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \Sigma_{13} \\ \Sigma_{23} \end{pmatrix} \Sigma_{33}^{-1} (a - \mu_3)$$

$$\hat{\Sigma} = \begin{pmatrix} \Sigma_{11} \Sigma_{12} \\ \Sigma_{21} \Sigma_{22} \end{pmatrix} - \begin{pmatrix} \Sigma_{13} \\ \Sigma_{23} \end{pmatrix} \Sigma_{33}^{-1} \left(\Sigma_{13} \; \Sigma_{23} \right)$$

TU Delft

More than two variables?



$$\boldsymbol{\mu} = \begin{pmatrix} 94 \\ 78 \\ 12 \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 41^2 \; 1000 \; 475 \\ 1000 \; 35^2 \; 520 \\ 475 \; 520 \; 27^2 \end{pmatrix}$$

Given p=22mm/h, what is the distribution of Q_1 and Q_2?

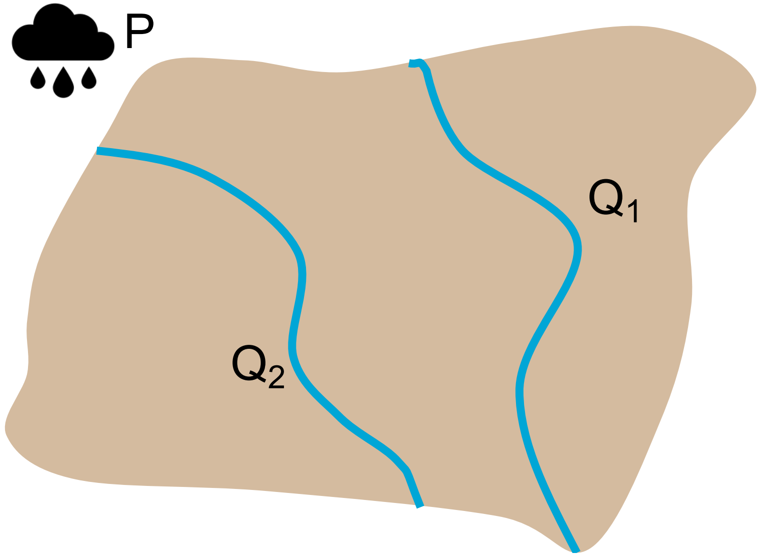
$$\hat{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \Sigma_{13} \\ \Sigma_{23} \end{pmatrix} \Sigma_{33}^{-1}(a - \mu_3)$$

$$\hat{\Sigma} = \begin{pmatrix} \Sigma_{11}\Sigma_{12} \\ \Sigma_{21}\Sigma_{22} \end{pmatrix} - \begin{pmatrix} \Sigma_{13} \\ \Sigma_{23} \end{pmatrix} \Sigma_{33}^{-1} \left(\Sigma_{13} \; \Sigma_{23}\right)$$

How do you expect the univariate distributions to be?

TU Delft

More than two variables?



$$\boldsymbol{\mu} = \begin{pmatrix} 94 \\ 78 \\ 12 \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 41^2 & 1000 & 475 \\ 1000 & 35^2 & 520 \\ 475 & 520 & 27^2 \end{pmatrix}$$

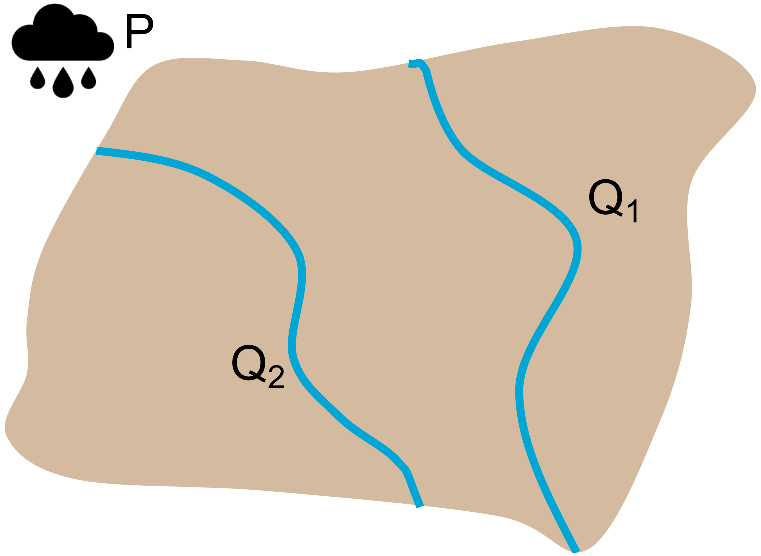
Given p=22mm/h, what is the distribution of Q_1 and Q_2?

$$\hat{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \Sigma_{13} \\ \Sigma_{23} \end{pmatrix} \Sigma_{33}^{-1} (a - \mu_3)$$

$$\hat{\Sigma} = \begin{pmatrix} \Sigma_{11} \Sigma_{12} \\ \Sigma_{21} \Sigma_{22} \end{pmatrix} - \begin{pmatrix} \Sigma_{13} \\ \Sigma_{23} \end{pmatrix} \Sigma_{33}^{-1} \left(\Sigma_{13} \; \Sigma_{23} \right)$$

TU Delft

More than two variables?



$$\boldsymbol{\mu} = \begin{pmatrix} 94 \\ 78 \\ 12 \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 41^2 & 1000 & 475 \\ 1000 & 35^2 & 520 \\ 475 & 520 & 27^2 \end{pmatrix}$$

Given p=22mm/h, what is the distribution of Q_1 and Q_2?

$$\hat{\mu} = \begin{pmatrix} 94 \\ 78 \end{pmatrix} + \begin{pmatrix} 475 \\ 520 \end{pmatrix} (27^2)^{-1}(22-12) = \begin{pmatrix} 100.5 \\ 85.1 \end{pmatrix}$$

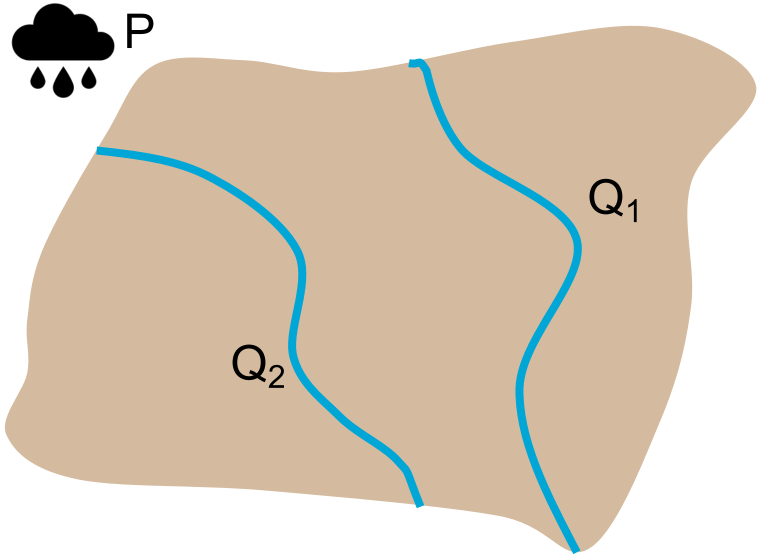
$$\hat{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \Sigma_{13} \\ \Sigma_{23} \end{pmatrix} \Sigma_{33}^{-1}(a-\mu_3)$$

$$\hat{\Sigma} = \begin{pmatrix} 41^2 & 1000 \\ 1000 & 35^2 \end{pmatrix} - \begin{pmatrix} 475 \\ 520 \end{pmatrix} (27^2)^{-1} (475 \; 520) = \begin{pmatrix} 41^2 & 1000 \\ 1000 & 35^2 \end{pmatrix} - \begin{pmatrix} 309.5 & 338.8 \\ 338.8 & 370.9 \end{pmatrix} = \begin{pmatrix} 1372.5 & 661.2 \\ 661.2 & 854.1 \end{pmatrix}$$

$$\hat{\Sigma} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} - \begin{pmatrix} \Sigma_{13} \\ \Sigma_{23} \end{pmatrix} \Sigma_{33}^{-1} (\Sigma_{13} \; \Sigma_{23})$$

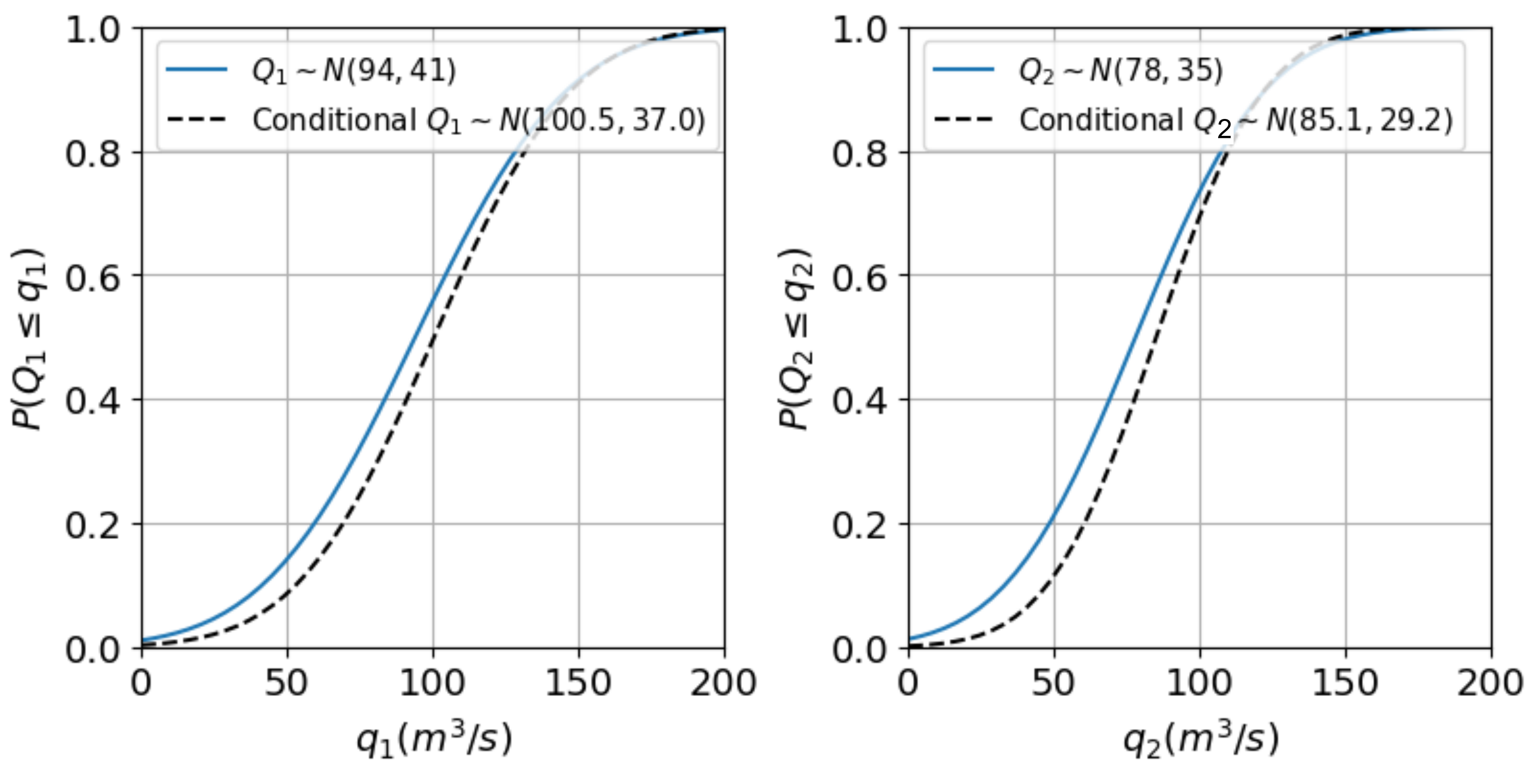
TU Delft

More than two variables?



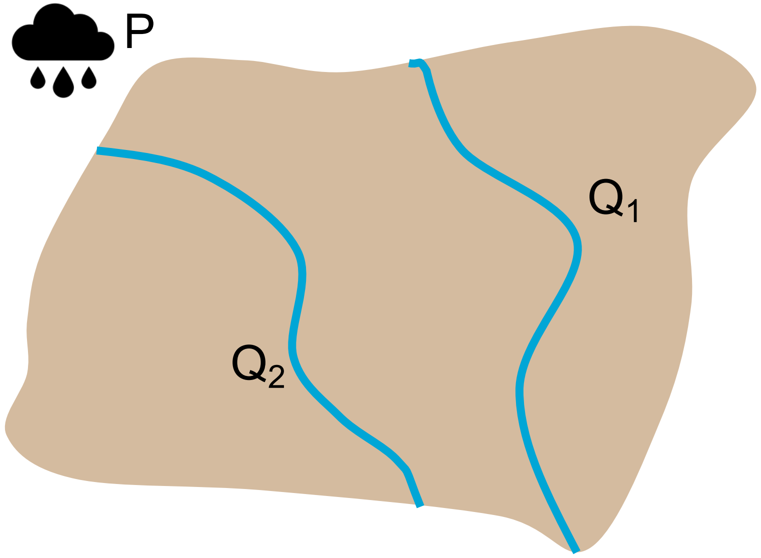
$$\boldsymbol{\mu} = \begin{pmatrix} 94 \\ 78 \\ 12 \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 41^2 & 1000 & 475 \\ 1000 & 35^2 & 520 \\ 475 & 520 & 27^2 \end{pmatrix}$$

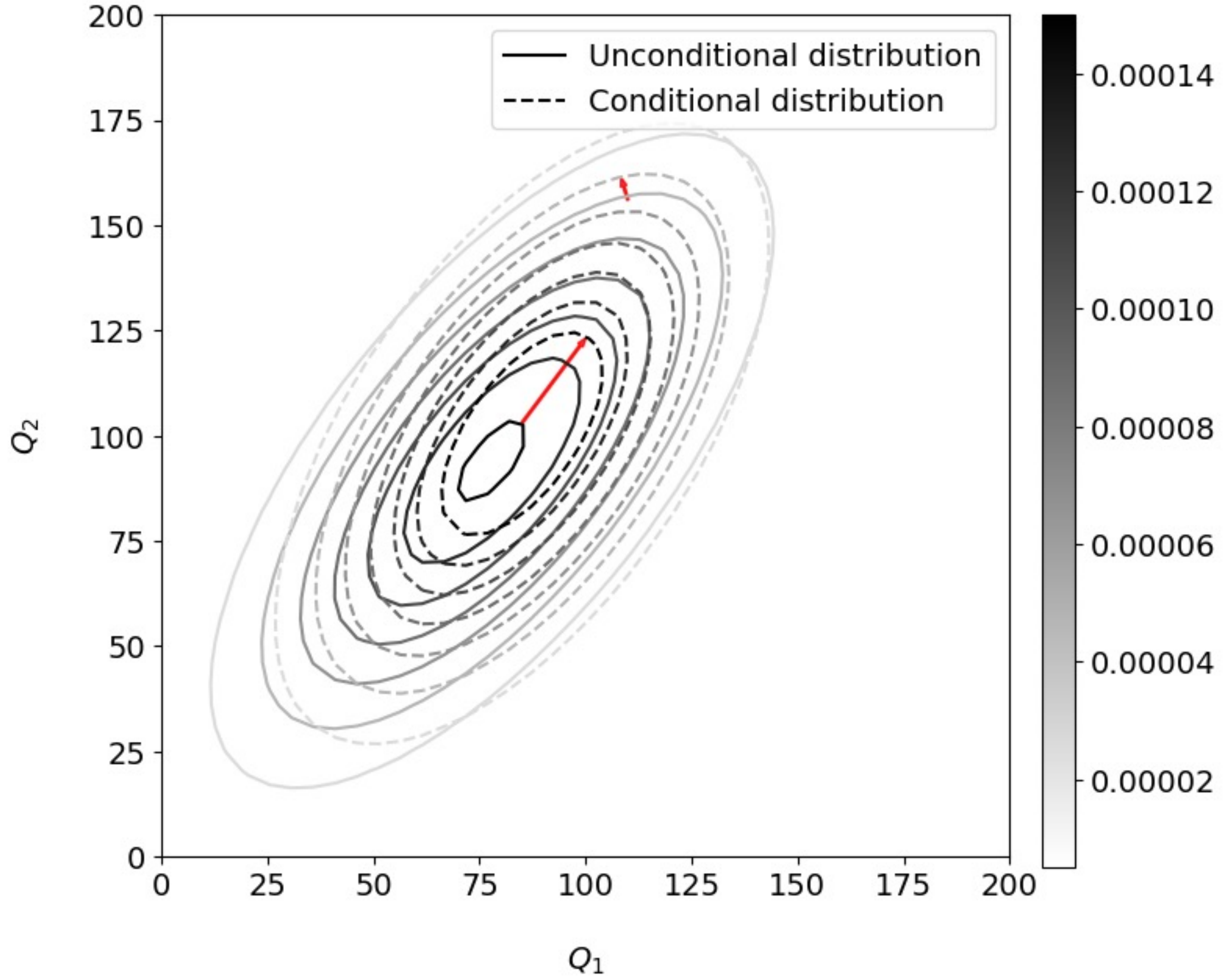
Given p=22mm/h, what is the distribution of Q_1 and Q_2?



TU Delft

More than two variables?



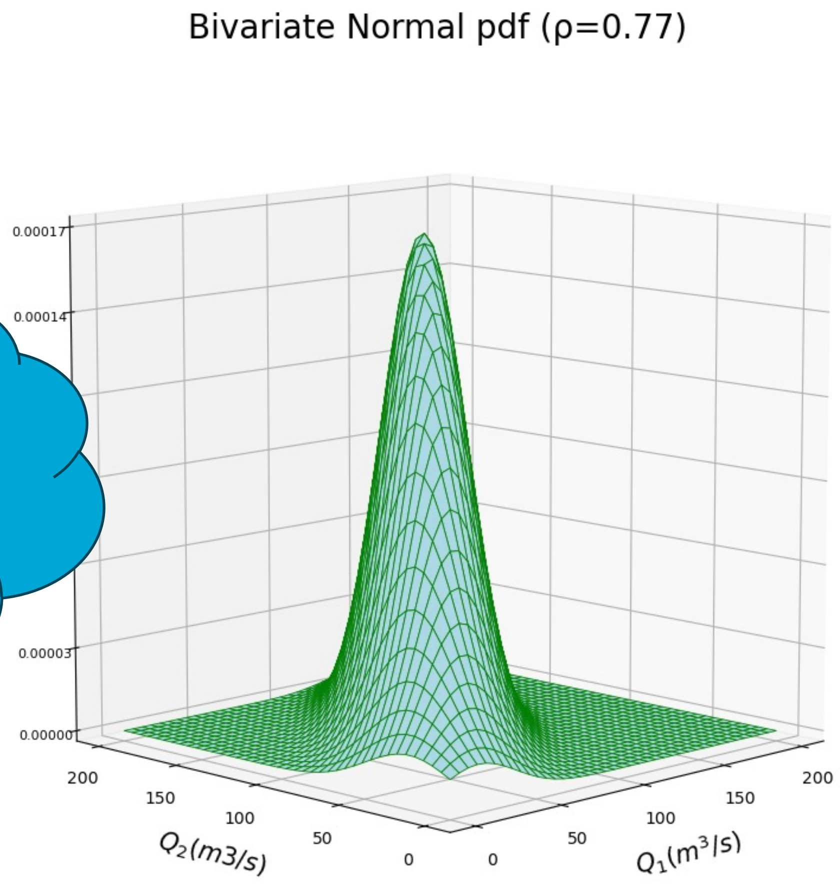


TU Delft

Contents – Multivariate distributions

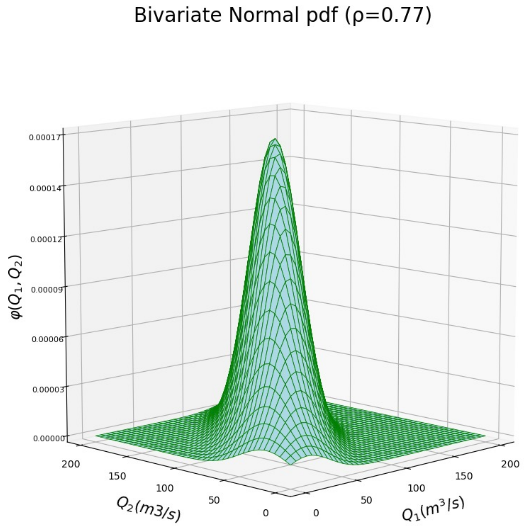
1. Recap & motivation
2. Set theory and basic operations
3. Continuous variables
    a. Extension to multiv computations
    b. Measures of dependence
4. Multivariate Gaussian distribution

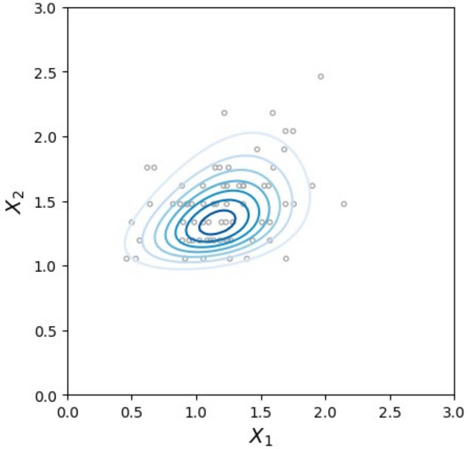
What's next?

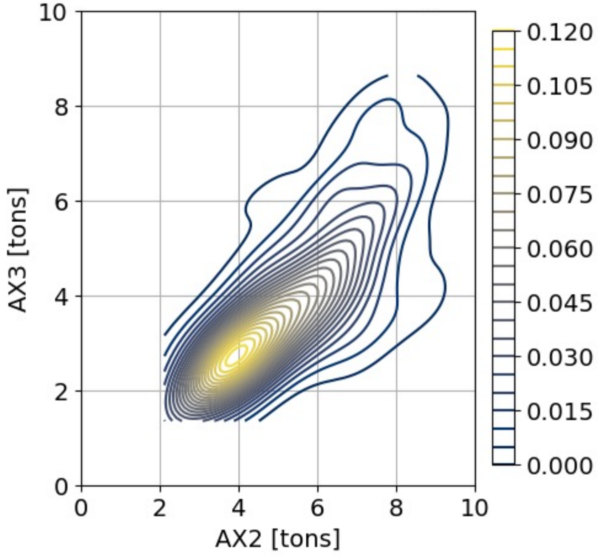


What's next?

- There is more in the textbook!
- PA: let's do cool 3D plots and contours
- Wednesday workshop: Starve or not to starve?
- Friday project: Gaussian & Furious!








"Seaweed" by CasparGirl


"Sky - clouds - mountain - sea - cars - road" by emran.

TU Delft

And enjoy the journey!