# MATHS CRASH COURSE

German Research Center for Artificial Intelligence

# PROBABILITY THEORY

Julian Liersch
AG Robotik, Universität Bremen

Prof. Dr. Dr. h.c. Frank Kirchner
AG Robotik, Universität Bremen
https://robotik.dfki-bremen.de/
robotik@dfki.de
October 13, 2023 – Bremen, Deutschland

University of Bremen

# General setup



## General information

- Our goal is to have a common language and understanding of the relevant content.
- The crash course shall provide you an opportunity to identify knowledge gaps.
- If something is too fast, feel free to pause the video.

# Table of contents

# Probability spaces

# Content on probability spaces



## Content

- Definition of $\sigma$-algebras
- Definition of measure spaces
- Definition of probability spaces as measure spaces

German Research Center for Artificial Intelligence

**Definition**: $\sigma$-algebra

Let $\Omega \neq \emptyset$ and $\mathcal{A} \subseteq \mathfrak{P}(\Omega) := \{A | A \subseteq \Omega\}$. Then $\mathcal{A}$ is called a *$\sigma$-algebra over* $\Omega$ if the following conditions hold:

- $\Omega \in \mathcal{A}$
- $A \in \mathcal{A} \Rightarrow \Omega \backslash A \in \mathcal{A}$
- $\{A_n\}_{n \in \mathbb{N}} \subseteq \mathcal{A} \Rightarrow \bigcup_{n \in \mathbb{N}} A_n \in \mathcal{A}$

**Corollary**: Basic properties

Let $\mathcal{A}$ be a $\sigma$-algebra over $\Omega \neq \emptyset$, then the following conditions hold:

- $\emptyset \in \mathcal{A}$
- $A, B \in \mathcal{A} \Rightarrow A \cap B \in \mathcal{A} \wedge A \backslash B \in \mathcal{A}$

# Examples of $\sigma$-algebras

## General examples

The following sets are $\sigma$-algebras for every $\Omega \neq \emptyset$:

- The power set $\mathfrak{P}(\Omega) := \{A | A \subseteq \Omega\}$.
- The set $\{\emptyset, \Omega\}$.
- For any $\mathcal{M} \subseteq \mathfrak{P}(\Omega)$, the set $\sigma(\mathcal{M}) := \bigcap_{\mathcal{A} \in \mathcal{F}(\mathcal{M})} \mathcal{A}$ where $\mathcal{F}(\mathcal{M}) := \{\mathcal{A} \subseteq \mathfrak{P}(\Omega) | \mathcal{M} \subseteq \mathcal{A} \wedge \mathcal{A} \text{ is } \sigma\text{-algebra}\}$.This is called the *$\sigma$-algebra generated from $\mathcal{M}$*.

## Specific examples

- For $\Omega := \{1, 2\}$, $\mathcal{A} := \mathfrak{P}(\Omega) = \{\emptyset, \Omega, \{1\}, \{2\}\}$.
- For $\Omega := \mathbb{R}^m$, $m \in \mathbb{N}$ and $\tau(\Omega) := \{U \subseteq \Omega | U \text{ open}\}$, the *Borel $\sigma$-algebra* $\mathcal{B}(\Omega) := \sigma(\tau(\Omega))$.

University of Bremen

German Research Center for Artificial Intelligence

**Definition**: Measurable space

Let $\mathcal{A}$ be a $\sigma$-algebra over $\Omega \neq \emptyset$. Then the tuple $(\Omega, \mathcal{A})$ is called a *measurable space*.

**Definition**: Measures

Let $(\Omega, \mathcal{A})$ be a measurable space. Then $\mu : \mathcal{A} \to \mathbb{R}_0^+ \cup \{\infty\}$ is called a *measure*, if the following conditions hold:

- $\mu(\emptyset) = 0$
- $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{A}$ pairwise disjoint $\Rightarrow \mu\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mu(A_n)$ ($\sigma$-additivity)

University of Bremen

German Research Center for Artificial Intelligence

## Remark

- For the measurable space $(\mathbb{R}, \mathfrak{P}(\mathbb{R}))$ there is **no** measure, as the $\sigma$-additivity can not be fulfilled! This is a reason to use the Borel $\sigma$-algebra for $\Omega \subseteq \mathbb{R}^m$ instead.

## **Definition**: Measure space

Let $(\Omega, \mathcal{A})$ be a measurable space and $\mu : \mathcal{A} \to \mathbb{R}_0^+ \cup \{\infty\}$ be a measure. Then the triple $(\Omega, \mathcal{A}, \mu)$ is called a *measure space*.

## **Definition**: Probability space

A measure space $(\Omega, \mathcal{A}, \mu)$ with $\mu(\Omega) = 1$ is called a *probability space*. In this case, the measure is usually called $P$ instead of $\mu$ and its co-domain is defined to be $[0, 1]$.

German Research Center for Artificial Intelligence

### Remarks

- $P : \Omega \to [0,1]$ is well-defined, as $A \subseteq \Omega \Rightarrow P(A) \leqslant P(\Omega) = 1$.
- For probability spaces, elements of $\mathcal{A}$ are called *events*.

### Example: Dice throw

Consider two fair six sided dice. The possible values the dice can show are $Z := \{1,2,3,4,5,6\}$. By setting $\Omega := Z \times Z$, $\mathcal{A} := \mathfrak{P}(\Omega)$ and $\mu : \mathcal{A} \to [0,1], E \mapsto \frac{\#E}{\#\mathcal{A}}$ we are defining a probability space which models throwing the two dice.

University of Bremen

# Basic concepts

German
Research Center
for Artificial
Intelligence

## Content

- Conditional probability
- Visualization using Venn diagrams
- Bayes' theorem

University
of Bremen

German Research Center for Artificial Intelligence

**Lemma**: Conditional probability space

Let $(\Omega, \mathcal{A}, P)$ be a probability space and $B \in \mathcal{A}$ with $P(B) > 0$. Then defining $\mathcal{A}_B := \{A \cap B | A \in \mathcal{A}\}$ and $P(\,\cdot\,|B) : \mathcal{A}_B \to [0,1], P(A|B) := \frac{P(A \cap B)}{P(B)}$ yields a probability space $(B, \mathcal{A}_B, P(\,\cdot\,|B))$.

**Example**

As before, consider throwing two fair six sided dice. The probability to throw a pair of sixes is $P(\{(6,6)\}) = \frac{1}{36}$. But this probability changes, if we already have thrown the first six.

- $E_{6,6} := \{(6,6)\} \mathrel{\widehat{=}}$ both dice show a six
- $E_{6,x} := \{(6,w) | w \in \{1,2,3,4,5,6\}\} \mathrel{\widehat{=}}$ the first die shows a six
- $P(E_{6,6}|E_{6,x}) = \frac{P(E_{6,6})}{P(E_{6,x})} = \frac{1}{36} \cdot \frac{36}{6} = \frac{1}{6}$

University of Bremen

# Visualization of conditional probability via Venn diagrams



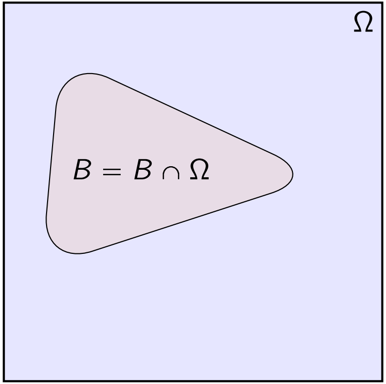


Figure: The area of the shapes represent their probabilities. As $P(\Omega) = 1$, it holds $P(B) = P(B|\Omega) = \frac{P(B \cap \Omega)}{P(\Omega)}$.

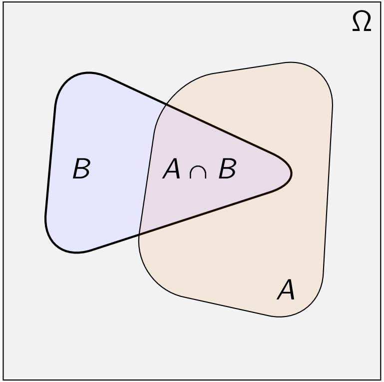


Figure: The conditional probability relates the "area" of $A \cap B$ with the "area" of $B$: $P(A|B) = \frac{P(A \cap B)}{P(B)}$ similar as $P(B)$ is relating the "area" of $B \cap \Omega$ with the "area" of $\Omega$.

German Research Center for Artificial Intelligence

**Corollary**: Chain rule

Let $(\Omega, \mathcal{A}, P)$ be a probability space and $A, B \in \mathcal{A}$. Then the following equality holds by definition of the conditional probability space:

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

**Corollary**: Bayes' theorem

Let $(\Omega, \mathcal{A}, P)$ be a probability space and $A, B \in \mathcal{A}$. Then the following equality holds:

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

# Random variables

German Research Center for Artificial Intelligence

## Content

- Definition random variables
- Discrete and continuous random variables
- Arity of random variables

University of Bremen

German Research Center for Artificial Intelligence

**Definition**: Measurable function

Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $(S, \Sigma)$ be a measurable space. Then a function $X : \Omega \to S$ is called *measurable*, if the following implication holds:

$$E \in \Sigma \Rightarrow X^{-1}[E] \in \mathcal{A}$$

**Definition**: Random variable

Let $(\Omega, \mathcal{A}, P)$ be a probability space and $(S, \Sigma)$ be a measurable space. Then a measurable function $X : \Omega \to S$ is called a *random variable*.

- For $s \in S$ we define the event, that $X$ equals $s$ as $X = s := \{\omega \in \Omega | X(\omega) = s\}$.
- Similarly we set $X \in E := \{\omega \in \Omega | X(\omega) \in E\} = X^{-1}[E]$ for $E \in \Sigma$.

**Definition**: Discrete random variable

Let $(\Omega, \mathcal{A}, P)$ be a probability space and $(S, \Sigma)$ be a measurable space. Then the random variable $X : \Omega \to S$ is called *discrete* if $S$ is a discrete set.

Example

Sum of the throw of two fair six-sided dice:

- $X : \{1, 2, 3, 4, 5, 6\}^2 \to S := \mathbb{N},\ X : (x, y) \mapsto x + y$
- $P(X = 3) = P(\{(x, y) \in \{1, 2, 3, 4, 5, 6\}^2 | x + y = 3\}) = \frac{\#\{(1,2),(2,1)\}}{\#\{1,2,3,4,5,6\}^2} = \frac{2}{36}$

German Research Center for Artificial Intelligence

**Definition**: Continuous random variable

Let $(\Omega, \mathcal{A}, P)$ be a probability space and $(S, \Sigma)$ be a measurable space. Then the random variable $X : \Omega \rightarrow S$ is called *continuous* if $S$ is a continuous space.

Examples

- Height of a randomly selected person: $S \subseteq \mathbb{R}^+$
- Runtime of the execution of a non-locking program in real time: $S = \mathbb{R}_0^+$

University of Bremen

German Research Center for Artificial Intelligence

**Definition**: Arity of random variables

Let $X : \Omega \rightarrow S$ be a random variable. The *arity of* $X$ refers to the dimension of the co-domain of $X$. Especially:

- If $S$ is "one-dimensional", $X$ is called *univariate*.
- If $S$ is "multi-dimensional", $X$ is called *multivariate*.

Examples

- Univariate random variable: Focussing on the height of a randomly selected person ($S \subseteq \mathbb{R}^+$)
- Multivariate random variable: Jointly focussing on the height and weight of a randomly selected person ($S \subseteq \mathbb{R}^+ \times \mathbb{R}^+$)

University of Bremen

# Probability distributions

## Content

- Definition probability distributions
- Mass, density functions and cumulative distribution function
- Common distributions

German Research Center for Artificial Intelligence

**Lemma**: Probability distribution

Let $X : \Omega \to S$ be a random variable for a probability space $(\Omega, \mathcal{A}, P)$ and a measurable space $(S, \Sigma)$. Then $P_X : \Sigma \to [0,1], E \mapsto P(X \in E)$ is a probability measure for the measurable space $(S, \Sigma)$, thus making $(S, \Sigma, P_X)$ a probability space. The probability measure $P_X$ is called *the distribution of* $X$.

Identity distribution

Let $(\Omega, \mathcal{A}, P)$ be a probability space. Then $\mathrm{id} : \Omega \to \Omega, \omega \mapsto \omega$ is a random variable, making $P_{\mathrm{id}}$ the distribution of id. Thus each probability measure can be associated with a distribution.

University of Bremen

# Probability mass & density functions



**Definition**: Probability mass function

Let $S$ be countable and $p : S \to [0, 1]$ such that $P(X = x) = p(x)$. The function $p$ is called *probability mass function* and abbreviated as *pmf*.

**Lemma**: Probability density function

Let $S \subseteq \mathbb{R}^m$ be uncountable for $m \in \mathbb{N}$, $X : \Omega \to S$ be a uni- or multivariate random variable and $P_X : \Sigma \to [0, 1]$ be absolutely continuous relative to the $m$-dimensional Lebesgue-measure[a], which is denoted as $\lambda^m$. Then there is a function $p : S \to \mathbb{R}_0^+$ such that $P_X(E) = \int_E p d\lambda^m$ for all $E \in \Sigma$. This function $p$ is called *probability density function* and abbreviated as *pdf*.

[a] This means each $\lambda^m$-null-set $N \in \Sigma$ fulfills $P(X \in N) = 0$.

German Research Center for Artificial Intelligence

**Definition**: Cumulative distribution function

Let $(\Omega, \mathcal{A}, P)$ be a probability space, $S \subseteq \mathbb{R}$ and $X : \Omega \to S$ be a univariate random variable. Then the function $F : S \to [0,1], x \mapsto P(X \leqslant x)$ is called *cumulative distribution function* and abbreviated as *cdf*. Here $X \leqslant x := \{\omega \in \Omega | X(\omega) \leqslant x\}$.

**Remark**

- The cdf is isotone.
- If $X$ is discrete with pmf $p : S \to \mathbb{R}$, then $F(x) = \sum_{y \in S, y \leqslant x} p(y)$.
- If $X$ is continuous with pdf $p : S \to \mathbb{R}$, then $F(x) = \int_{y \in S, y \leqslant x} p(y) dy$.

German Research Center for Artificial Intelligence

## Uniform distribution

For finite $S$ with $\#S = n$, the function $p : x \mapsto \frac{1}{n}$ is the probability mass function of the uniform distribution on $S$.
Notation: $X \sim U(S)$

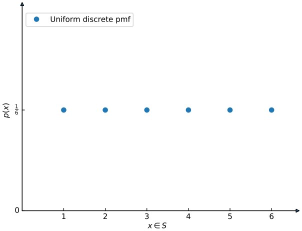


Figure: Probability mass function of uniform distribution for throwing a fair six sided die.

University of Bremen

German Research Center for Artificial Intelligence

## Bernoulli distribution

For $S = \{0, 1\}$ and parameter $p \in [0, 1]$, the function $f(\cdot\,|p) : x \mapsto p^x(1-p)^{1-x}$ is the probability mass function of the Bernoulli distribution with probability $p$.
Notation: $X \sim \mathcal{B}(p)$

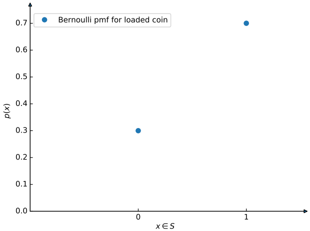


Figure: Probability mass function of Bernoulli distribution for flipping a loaded coin. "Head" may for example be represented by 0, and "tails" by 1.

University of Bremen

German Research Center for Artificial Intelligence

## Uniform distribution

For $S \subseteq \mathbb{R}$ with $\lambda^1(S) \in \mathbb{R}^+$, the function $p : x \mapsto \frac{1}{\lambda(S)}$ is the probability density function of the uniform distribution on $S$.

Notation: $X \sim U(S)$

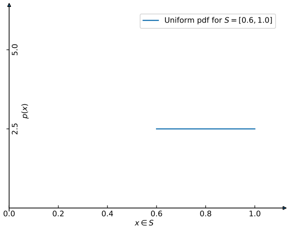


Figure: Probability density function of uniform distribution in $S = [0.6, 1.0]$.

University of Bremen

German Research Center for Artificial Intelligence

## Normal distribution

For $S = \mathbb{R}$, $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$, the function $p : x \mapsto \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$ is the probability density function of the Gaussian / normal distribution with mean $\mu$ and standard deviation $\sigma$. For $\mu = 0$ and $\sigma = 1$, $P_X$ is called *standard normal distribution*.
Notation: $X \sim \mathcal{N}(\mu, \sigma)$

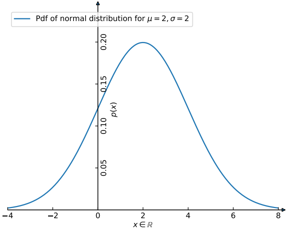


Figure: Probability densitiy function of normal distribution with $\mu = 2$ and $\sigma = 2$.

University of Bremen

# Additional concepts

# Content on additional concepts



## Content

- Transformations
- Moments
- Joint and marginal distributions
- Multivariate normal distribution
- Independence

German Research Center for Artificial Intelligence

**Lemma**: Transformation

Let $X : \Omega \to \mathbb{R}$ be a continuous random variable and $g : X[\Omega] \to S$ be measurable with a measurable space $(S, \Sigma)$. Then $Y := g \circ X : \Omega \to S$ is a continuous random variable with $P(Y \in B) = P(X \in g^{-1}[B])$ for $B \in \Sigma$.

Examples

- Let $X \sim U([0,1])$, $a, b \in \mathbb{R}$ and $g : x \mapsto (b-a)x + a$, then $g \circ X \sim U([a,b])$
- Let $X \sim \mathcal{N}(\mu, \sigma)$ and $g : x \mapsto \frac{x-\mu}{\sigma}$, then $g \circ X \sim \mathcal{N}(0,1)$

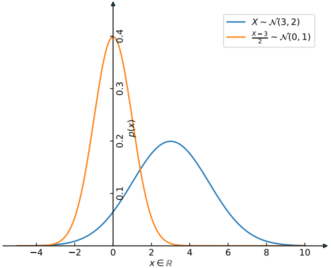


Figure: Transforming the random variable $X \sim \mathcal{N}(3,2)$ using $g : x \mapsto \frac{x-3}{2}$ *normalizes* it and $g \circ X \sim \mathcal{N}(1,0)$. The corresponding pdfs can be seen in the figure.

University of Bremen

German Research Center for Artificial Intelligence

**Definition**: First moment

Let $S$ be a vector space and $X : \Omega \to S$ be a random variable. The *first moment of* $X$ is defined as

$$\mathbb{E}(X) := \begin{cases} \sum_{x \in S} xp(x) & \text{if } X \text{ is discrete and the series converges absolutely} \\ \int_S xp(x)dx & \text{if } X \text{ is continuous and has a probability density function } p \end{cases}$$

The first moment of a random variable is also called its *expected value* and noted as $\mu := \mathbb{E}(X)$.

**Example**

- For $X \sim \mathcal{N}(\mu_0, \sigma_0)$ is $\mathbb{E}(X) = \mu = \mu_0$.

University of Bremen

**Definition**: *$n$-th central moment*

Let $X : \Omega \to \mathbb{R}$ be a random variable with expected value $\mu := \mathbb{E}(X) \in \mathbb{R}$ and $n \in \mathbb{N}$. If it exists, the *$n$-th central moment of $X$* is defined as

$$\mu_n := \mathbb{E}\left[(X - \mu)^n\right]$$

The second central moment of a random variable is also called its *variance* and denoted as $\mathrm{Var}(X) := \mu_2$.

**Definition**: Standard deviation

Let $X : \Omega \to \mathbb{R}$ be a random variable with variance $\mathrm{Var}(X)$. Then $\sigma := \sqrt{\mathrm{Var}(X)}$ is called the *standard deviation* of $X$.

Example

- For $X \sim \mathcal{N}(\mu_0, \sigma_0)$ is $\sigma = \sigma_0$.

German Research Center for Artificial Intelligence

**Definition**: Joint distribution

Let $X_i : \Omega \to S_i$ be random variables mapping into probability spaces $(S_i, \Sigma_i, P_{X_i})$ for $i \in \{1, 2\}$. Then we can define the *joint probability space* $(S, \Sigma, P_{X_1,X_2})$ with $S := S_1 \times S_2$, $\Sigma := \sigma(\Sigma_1 \times \Sigma_2)$ and $P_{X_1,X_2} : (E_1, E_2) \mapsto P(X_1 \in E_1 \cap X_2 \in E_2)$. The probability measure $P_{X_1,X_2} : \Sigma \to [0, 1]$ is called the *joint distribution of* $X_1$ *and* $X_2$.

Remarks

- This definition can be extended canonically to an arbitrary finite number of random variables.
- The expression $P(X_1 \in E_1 \cap X_2 \in E_2)$ is often written as $P(X_1 \in E_1 \wedge X_2 \in E_2)$ or abbreviated as $P(X_1 \in E_1, X_2 \in E_2)$.

German Research Center for Artificial Intelligence

**Definition**: Marginal distribution

For $S = \mathbb{R}^m$ with $m \in \mathbb{N} \backslash \{1\}$ and a multivariate continuous random variable $X : \Omega \to S$ with pdf $p : S \to \mathbb{R}_0^+$ we can define *marginal distributions* $P_{X^i}(E) := \int_E p^i d\lambda^{m-1}$ using *marginal pdfs*

$$p^i(\tilde{x}^i) := \int_{-\infty}^{\infty} p((x_j)_{j=1}^m) d\lambda^1(x_i)$$

for all $i \in \{1, \ldots, m\}$ and $\tilde{x}^i \in \mathbb{R}^{m-1}$.

Remark

- By changing the integration to a summation, the definition can also by applied to discrete random variables.

German Research Center for Artificial Intelligence

## Multivariate normal distribution

For $\mu \in \mathbb{R}^m$ and $C \in \mathbb{R}^{m \times m}$ symmetric and positive definite we can define the multivariate normal distribution through its pdf

$$p(x) := (2\pi)^{-\frac{m}{2}} \det(C)^{-\frac{1}{2}} \exp(-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu))$$

This corresponds to a continuous random variable $X : \Omega \to \mathbb{R}^m$ with $P(X \in E) = \int_E p d\lambda^m$ for $E \in \Sigma$.

## Remark

- Usually the matrix $C$ is notated as $\Sigma$ instead.

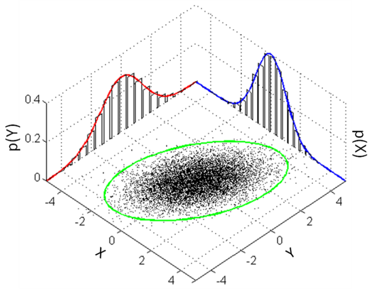


Figure: Realizations of multivariate normal distribution and corresponding marginal distributions.

University of Bremen

German Research Center for Artificial Intelligence

## Definition: Independence

Let $(\Omega, \mathcal{A}, P)$ be a probability space, $n \in \mathbb{N}\backslash\{1\}$ and for $i \in \{1, \ldots, n\}$ the tuples $(S_i, \Sigma_i)$ be measurable spaces and $X_i : \Omega \to S_i$ be random variables. Then the $X_i$ are called

- *Pairwise independent* if each pair $X_j \neq X_k$ is independent, i. e. $P((X_j \in E_j) \cap (X_k \in E_k)) = P(X_j \in E_j) \cdot P(X_k \in E_k)$ for all $E_j \in \Sigma_j$, $E_k \in \Sigma_k$.
- *Mutually independent* if for all $(E_i)_{i=1}^n \in \times_{i=1}^n \Sigma_i$ the equation $P(\bigcap_{i=1}^n (X_i \in E_i)) = \prod_{i=1}^n P(X_i \in E_i)$ holds.

## Collorary

- If a set of random variables is mutually independent, it is also pairwise independent.

## Examples for independence

- Consider throwing two fair six sided dice. This can be modeled using two random variables $X_1, X_2 \sim U(\{1, 2, 3, 4, 5, 6\})$ which can be assumed to be independent.

University of Bremen

German Research Center for Artificial Intelligence

## We recapped

- Basic definitions of probability theory,
- Conditional probability & Bayes theorem,
- Random variables, distributions & density functions,
- Transformation & independence of random variables,
- (Central) moments,
- Joint & marginal distributions,
- Important examples of distributions, including multivariate normal distributions.

University of Bremen

Thank You!

Feel free to ask questions in the forums!