Outline of the talk:

- Probability vs. statistics.
- Random events.
- Probability, joint, conditional.
- Bayes theorem.
- Distribution function, density.
- Characteristics of a random variable.
Recommended reading

- http://mathworld.wolfram.com/
Probability, motivating example

- A lottery ticket is sold for the price EUR 2.
- 1 lottery ticket out of 1000 wins EUR 1000. Other lottery tickets win nothing. This gives the value of the lottery ticket after the draw.
- For what price should the lottery ticket be sold before the draw?

Only a fool would buy the lottery ticket for EUR 2. (Or not?)

The value of the lottery ticket before the draw is \( \frac{1}{1000} \times 1000 = EUR 1 \) = the average value after the draw.

The probability theory is used here.

A lottery question: Why are the lottery tickets being bought? Why do lotteries prosper?
Statistics, motivating example

We have assumed so far that the parameters of the probability model are known. However, this is seldom fulfilled.

**Example – Lotto:** One typically looses while playing Lotto because the winnings are set according to the number of winners. It is of advantage to bet differently than others. For doing so, it is needed what model do the other use.

**Example – Roulette:** Both parties are interested if all the numbers occur with the same probability. More precisely said, what are the differences from the uniform probability distribution. How to learn it? What is the risk of wrong conclusions?

Statistics is used here.
Probability, statistics

- **Probability**: probabilistic model $\implies$ future behavior.
  - It is a theory (tool) for purposeful decisions when the outcome of future events depends on circumstances we know only partially and the randomness plays a role.
  - An abstract model of uncertainty description and quantification of the results.

- **Statistics**: behavior of the system $\implies$ probabilistic representation.
  - It is a tool for seeking a probabilistic description of real systems based on observing them and testing them.
  - It provides more: a tool for investigating the world, seeking and testing dependencies which are not apparent.
  - Two types: descriptive and inference statistics.
  - Collection, organization and analysis of data.
  - Generalization from restricted / finite samples.
Random events, concepts

An experiment with random outcome – states of nature, possibilities, experimental results, etc.

A sample space is an nonempty set $\Omega$ of all possible outcomes of the experiment.

An elementary event $\omega \in \Omega$ are elements of the sample space (outcomes of the experiment).

A space of events $\mathcal{A}$ is composed of the system of all subsets of the sample space $\Omega$.

A random event $A \in \mathcal{A}$ is an element of the space of events.

Note: The concept of a random event was introduced in order to be able to define the probability, probability distribution, etc.
• **Classic.** P.S. Laplace, 1812. It is not regarded to be the definition of the probability any more. It is merely an estimate of the probability.

\[ P(A) \approx \frac{N_A}{N} \]

• **Limit (frequency) definition**

\[ P(A) = \lim_{N \to \infty} \frac{N_A}{N} \]

• **Axiomatic definition** (Andrey Kolmogorov 1930)

- histogram vs. continuous probability distribution function
Axiomatic definition of the probability

- Ω - the sample space.
- A - the space of events.

1. \( P(A) \geq 0, \ A \in A. \)
2. \( P(\Omega) = 1. \)
3. If \( A \cap B = \emptyset \) then \( P(A \cup B) = P(A) + P(B), \ A \in A, \ B \in B. \)
Probability

is a function $P$, which assigns a number from the interval $[0, 1]$ to events and fulfils the following two conditions:

- $P(true) = 1$,

- $P \left( \bigcup_{n \in \mathbb{N}} A_n \right) = \sum_{n \in \mathbb{N}} P(A_n)$, if the events $A_n$, $n \in \mathbb{N}$, are mutually exclusive.

From these conditions, it follows:

$P(false) = 0$, \quad $P(\neg A) = 1 - P(A)$, \quad if $A \Rightarrow B$ then $P(A) \leq P(B)$.

Note: Strictly speaking, the space of events have to fulfil some additional conditions.
Derived relations

- If $A \subset B$ then $P(B \setminus A) = P(B) - P(A)$.
  The symbol $\setminus$ denotes the set difference.

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. 
The **joint probability** \( P(A, B) \), also sometimes denoted \( P(A \cap B) \), is the probability that events \( A \), \( B \) co-occur.

The joint probability is symmetric: \( P(A, B) = P(B, A) \).

**Marginalization** (the sum rule): \( P(A) = \sum_B P(A, B) \) allows computing the probability of a single event \( A \) from the joint probability \( P(A, B) \) by summing \( P(A, B) \) over all possible events \( B \). The probability \( P(A) \) is called the marginal probability.
Contingency table, marginalization
Example: orienteering race

### Orienteering competition example, participants

<table>
<thead>
<tr>
<th>Age</th>
<th>&lt;= 15</th>
<th>16-25</th>
<th>26-35</th>
<th>36-45</th>
<th>46-55</th>
<th>56-65</th>
<th>66-75</th>
<th>&gt;= 76</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>22</td>
<td>36</td>
<td>45</td>
<td>33</td>
<td>29</td>
<td>21</td>
<td>12</td>
<td>2</td>
<td>200</td>
</tr>
<tr>
<td>Women</td>
<td>19</td>
<td>32</td>
<td>37</td>
<td>30</td>
<td>23</td>
<td>14</td>
<td>5</td>
<td>0</td>
<td>160</td>
</tr>
<tr>
<td>Sum</td>
<td>41</td>
<td>68</td>
<td>82</td>
<td>63</td>
<td>52</td>
<td>35</td>
<td>17</td>
<td>2</td>
<td>360</td>
</tr>
</tbody>
</table>

### Orienteering competition example, frequency

<table>
<thead>
<tr>
<th>Age</th>
<th>&lt;= 15</th>
<th>16-25</th>
<th>26-35</th>
<th>36-45</th>
<th>46-55</th>
<th>56-65</th>
<th>66-75</th>
<th>&gt;= 76</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>0.061</td>
<td>0.100</td>
<td>0.125</td>
<td>0.092</td>
<td>0.081</td>
<td>0.058</td>
<td>0.033</td>
<td>0.006</td>
<td>0.556</td>
</tr>
<tr>
<td>Women</td>
<td>0.053</td>
<td>0.089</td>
<td>0.103</td>
<td>0.083</td>
<td>0.064</td>
<td>0.039</td>
<td>0.014</td>
<td>0.000</td>
<td>0.444</td>
</tr>
<tr>
<td>Sum</td>
<td>0.114</td>
<td>0.189</td>
<td>0.228</td>
<td>0.175</td>
<td>0.144</td>
<td>0.097</td>
<td>0.047</td>
<td>0.006</td>
<td>1</td>
</tr>
</tbody>
</table>

Marginal probability $P(Age\_group)$

Marginal probability $P(sex)$
The conditional probability

- Let us have the probability representation of a system given by the joint probability $P(A, B)$.

- If an additional information is available that the event $B$ occurred then our knowledge about the probability of the event $A$ changes to

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

  which is the **conditional probability** of the event $A$ under the condition $B$.

- The conditional probability is defined only for $P(B) \neq 0$.

- **Product rule**: $P(A, B) = P(A|B) P(B) = P(B|A) P(A)$.

- From the symmetry of the joint probability and the product rule, the Bayes theorem can be derived (to come in a more general formulation for more than two events).
Properties of the conditional probability

- $P(true|B) = 1$, $P(false|B) = 0$.
- If $A = \bigcup_{n \in \mathbb{N}} A_n$ and events $A_1, A_2, \ldots$ are mutually exclusive then
  $$P(A|B) = \sum_{n \in \mathbb{N}} P(A_n|B).$$
- Events $A, B$ are independent, if and only if $P(A|B) = P(A)$.
- If $B \Rightarrow A$ then $P(A|B) = 1$.
- If $B \Rightarrow \neg A$ then $P(A|B) = 0$.

- Events $B_i, i \in I$, constitute a complete system of events if they are mutually exclusive and $\bigcup_{i \in I} B_i = true$.
- A complete system of events has such property that one and only one event of them occurs.
**Example: conditional probability**

Consider rolling a single dice.

What is the probability that the number $> 3$ comes up (event A) under the conditions that the odd number came up (event B)?

$$\Omega = \{1, 2, 3, 4, 5, 6\}, \quad A = \{4, 5, 6\}, \quad B = \{1, 3, 5\}$$

$$P(A) = P(B) = \frac{1}{2}$$

$$P(A, B) = P(\{5\}) = \frac{1}{6}$$

$$P(A|B) = \frac{P(A, B)}{P(B)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$
Joint and conditional probabilities, example

Courtesy T. Brox
The total probability theorem

Let \( B_i, i \in I \), be a complete system of events and let it hold \( \forall i \in I: P(B_i) \neq 0 \).

Then for every event \( A \) holds

\[
P(A) = \sum_{i \in I} P(B_i) P(A|B_i).
\]

Proof:

\[
P(A) = P \left( \left( \bigcup_{i \in I} B_i \right) \cap A \right) = P \left( \bigcup_{i \in I} (B_i \cap A) \right)
= \sum_{i \in I} P(B_i \cap A) = \sum_{i \in I} P(B_i) P(A|B_i).
\]
Bayes theorem

(Thomas Bayes *1702 - †1761)

Let \( B_i, i \in I \), be a complete system of events and \( \forall i \in I: P(B_i) \neq 0 \).

For each event \( A \) fulfilling the condition \( P(A) \neq 0 \) the following holds

\[
P(B_i|A) = \frac{P(B_i) P(A|B_i)}{\sum_{i \in I} P(B_i) P(A|B_i)},
\]

where \( P(B_i|A) \) is the **posterior** probability; \( P(B_i) \) is the **prior** probability; and \( P(A|B_i) \) are known **conditional probabilities** of \( A \) having observed \( B_i \).

**Proof** (exploring the total probability theorem):

\[
P(B_i|A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(B_i) P(A|B_i)}{\sum_{i \in I} P(B_i) P(A|B_i)}.
\]
The importance of the Bayes theorem

- Bayes theorem is a fundamental rule for machine learning (pattern recognition). Given $B_i, i \in I$ is the partitioning of the sample space. Suppose that event $A$ occurs. What is the probability of event $B_i$?

- The conditional probabilities (also likelihoods) $P(A|B_i)$ are estimated from experiments or from a statistical model.

- Having $P(A|B_i)$, the posterior (also posterior) probabilities $P(B_i|A)$ are determined serving as optimal estimates, which event from $B_i$ occurred.

- It is needed to know the prior (also apriori) probability $P(B_i)$ to determine posterior probability $P(B_i|A)$.

- Informally: posterior $\propto (prior \times conditional\ probability)$ of the event having some observations.

- In a similar manner, we define the conditional probability distribution, conditional density of the continuous random variable, etc.
ML and MAP

Bayes theorem from the slide 18 is copied here

\[
P(B_i|A) = \frac{P(B_i) P(A|B_i)}{\sum_{i \in I} P(B_i) P(A|B_i)}.\]

- The prior probability is the probability of \(P(B_i)\) without any evidence from observations (measurements).

- The likelihood (conditional probability of the event \(A\) under the condition \(B_i\)) evaluates a candidate output on the measurement. Seeking the output that maximizes the likelihood is known as the maximum likelihood (ML) approach.

- The posterior probability is the probability of \(B_i\) after taking the observation (measurement) into account. Its maximization leads to the maximum a-posteriori (MAP) approach.
Conditional independence

Random events $A, B$ are *conditionally independent* under the condition $C$, if

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

Similarly, a conditional independence of more events, random variables, etc. is defined.
Independent events

Event $A, B$ are independent $\iff P(A \cap B) = P(A) \cdot P(B)$.

Example

The dice is rolled once. Events are: $A > 3$, event $B$ is an odd number. Are these events independent?

$$\Omega = \{1, 2, 3, 4, 5, 6\}, \quad A = \{4, 5, 6\}, \quad B = \{1, 3, 5\}$$

$$P(A) = P(B) = \frac{1}{2}$$

$$P(A \cap B) = P(\{5\}) = \frac{1}{6}$$

$$P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

$$P(A \cap B) \neq P(A) \cdot P(B) \iff \text{The events are dependent.}$$
The random variable

- The random variable is an arbitrary function $X : \Omega \to \mathbb{R}$, where $\Omega$ is a sample space.

- Why is the concept of the random variable introduced? It allows to work with concepts as the probability distribution function, probability density function, expectation (mean value), etc.

- There are two basic types of random variables:
  
  - **Discrete** – a countable number of values. Examples: rolling a dice, the count of number of cars passing through a street in a hour.
    
    The discrete probability is given as $P(X = a_i) = p(a_i)$, $i = 1, \ldots$, $\sum_i p(a_i) = 1$.
  
  - **Continuous** – values from some interval, i.e. infinite number of values. Example: the height persons.
    
    The continuous probability is given by the distribution function or the probability density function.
Distribution function of a random variable

The probability distribution function of the random variable $X$ is a function $F: X \to [0, 1]$ defined as $F(x) = P(X \leq x)$, where $P$ is a probability.

Properties:

1. $F(x)$ is a non-decreasing function, i.e. $\forall$ pair $x_1 < x_2$ it holds $F(x_1) \leq F(x_2)$.

2. $F(X)$ is continuous from the right, i.e. it holds $\lim_{h \to 0^+} F(x + h) = F(x)$.

3. It holds for every distribution function $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$. Written more concisely: $F(-\infty) = 0$, $F(\infty) = 1$.
   - If the possible values of $F(x)$ are from the interval $(a, b)$ then $F(a) = 0$, $F(b) = 1$.

Any function fulfilling the above three properties can be understood as a distribution function.
Continuous distribution and density functions

- The distribution function $F$ is called (absolutely) continuous if a nonnegative function $f$ (probability density) exists and it holds

$$F(x) = \int_{-\infty}^{x} f(u) \, du \quad \text{for every } x \in X.$$ 

- The probability density function fulfills

$$\int_{-\infty}^{\infty} f(x) \, dx = 1.$$ 

- If the derivative of $F(x)$ exists in the point $x$ then $F'(x) = f(x)$.

- For $a, b \in \mathbb{R}, \ a < b$, it holds

$$P(a < X < b) = \int_{a}^{b} f(x) \, dx = F(b) - F(a).$$
Example, normal distribution

\[ F(x) \]

\[ f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \]

Distribution function  Probability density function
Example: the difference between the probability and the probability density function

Q1: What is the probability that the measured temperature is exactly 31.5°C?
A1: This probability is zero in a limit.

Q2: What is the probability that the measured temperature is in the interval between 30°C and 31°C?
A2: The probability is given by the area of under the probability density (also the probability distribution function), i.e. approximately 0.09 as estimated from the figure above.
The law of large numbers

The law of large numbers says that if very many independent experiments can be made then it is almost certain that the relative frequency will converge to the theoretical value of the probability density.

Jakob Bernoulli, Ars Conjectandi: Usum & Applicationem Praecedentis Doctrinae in Civilibus, Moralibus & Oeconomicis, 1713, Chapter 4.
Expectation

- (Mathematical) expectation = the average of a variable under the probability distribution.

- Continuous definition: \( E(x) = \mu = \int_{-\infty}^{\infty} x \, f(x) \, dx \).

- Discrete definition: \( E(x) = \mu = \sum_x x \, P(x) \).

- The expectation can be estimated from a number of samples by \( E(x) \approx \frac{1}{N} \sum_i x_i \). The approximation becomes exact for \( N \to \infty \).

- Expectation over multiple variables: \( E_x(x, y) = \int_{-\infty}^{\infty} (x, y) \, f(x) \, dx \).

- Conditional expectation: \( E(x|y) = \int_{-\infty}^{\infty} x \, f(x|y) \, dx \).
### Basic characteristics of a random variable

<table>
<thead>
<tr>
<th>Continuous distribution</th>
<th>Discrete distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expectation</strong></td>
<td><strong>Expectation</strong></td>
</tr>
<tr>
<td>$E(x) = \mu = \int_{-\infty}^{\infty} x f(x) , dx$</td>
<td>$E(x) = \mu = \sum x , P(x)$</td>
</tr>
<tr>
<td><strong>$k$-th (general) moment</strong></td>
<td><strong>$k$-th (general) moment</strong></td>
</tr>
<tr>
<td>$E(x^k) = \int_{-\infty}^{\infty} x^k f(x) , dx$</td>
<td>$E(x^k) = \sum x^k , P(x)$</td>
</tr>
<tr>
<td><strong>$k$-th central moment</strong></td>
<td><strong>$k$-th central moment</strong></td>
</tr>
<tr>
<td>$\mu_k = \int_{-\infty}^{\infty} (x - E(x))^k f(x) , dx$</td>
<td>$\mu_k = \sum (x - E(x))^k , P(x)$</td>
</tr>
<tr>
<td><strong>Dispersion, variance, 2nd central moment</strong></td>
<td><strong>Dispersion, variance, 2nd central moment</strong></td>
</tr>
<tr>
<td>$D(x) = \int_{-\infty}^{\infty} (x - E(x))^2 f(x) , dx$</td>
<td>$D(x) = \sum (x - E(x))^2 , P(x)$</td>
</tr>
<tr>
<td>Standard deviation $\sigma(x) = \sqrt{D(x)}$</td>
<td></td>
</tr>
</tbody>
</table>
Central limit theorem (1)

The Central limit theorem describes the probability characteristics of the ‘population of the means’, which has been created from the means of an infinite number of random population samples of size $N$, all of them drawn from a given ‘parent population’. The Central limit theorem predicts characteristics regardless of the distribution of the parent population.

1. The mean of the population of means (i.e., the means of many times randomly drawn samples of size $N$ from the parent population) is always equal to the mean of the parent population.

2. The standard deviation of the population of means is always equal to the standard deviation of the parent population divided by the square root of the sample size $N$.

3. The distribution of sample means will increasingly approximate a normal (Gaussian) distribution as the size $N$ of samples increases.
Central limit theorem (2)

- A consequence of the Central limit theorem is that if we average measurements of a particular quantity, the distribution of our average tends toward a normal (Gaussian) one.

- In addition, if a measured variable is actually a combination of several other uncorrelated variables, all of them ‘contaminated’ with a random error of any distribution, our measurements tend to be contaminated with a random error that is normally distributed as the number of these variables increases.

- Thus, the Central limit theorem explains the ubiquity of the bell-shaped ‘Normal distribution’ in the measurements domain.
Central limit theorem (3), the application view

- It is important for applications that there is no need to generate a big amount of population samples. It suffices to obtain one big enough population sample. The Central limit theorem teaches us what is the distribution of population means without the need to generate these population samples.

- What can be considered a big enough population sample? It is application dependent. Trespassing the lower bound of 30-50 random observation is not allowed by statisticians. Recall samples with about 1000 observations serving to estimate outcomes of elections.

- The confidence interval in statistics indicates the reliability of the estimate. It gives the degree of uncertainty of a population parameter. We have talked about sample mean only so far. See a statistics textbook for details.
Statistical principal of noise filtration

Let us consider almost the simplest image statistical model.

Assume that each image pixel is contaminated by the additive noise:

- which is statistically independent of the image function,
- has a zero mean $\mu$, and
- has a standard deviation $\sigma$.

Let have $i$ realizations of the image, $i = 1, \ldots n$. The estimate of the correct value is

$$
\frac{g_1 + \cdots + g_n}{n} + \frac{\nu_1 + \cdots + \nu_n}{n}.
$$

The outcome is a random variable with $\mu' = 0$ and $\sigma' = \sigma/\sqrt{n}$.

The thought above is anchored in the probability theory in its powerful Central limit theorem.
Random vectors

- The concept random vector extends the random number concept. A (column) vector random variable $X$ is a function that assigns a vector of random variables to each outcome in the sample space.

- Given the random vector $X = (x_1, x_2, \ldots, x_n)^\top$, the probability distribution function and the probability density function are extended as the

  - joint probability distribution function

  $$F_X(x) = P_X ((X_1 \leq x_1) \cap (X_2 \leq x_2) \cap \ldots \cap (X_n \leq x_n))$$

  - joint probability density function

  $$f_X(x) = \frac{\partial^n F_X(x)}{\partial x_1 \partial x_2 \ldots \partial x_n}$$
Simpler characterizations of random vectors, mean vector, covariance matrix

- We keep in mind that a random vector is fully characterized by its joint probability distribution function or its joint probability density function.

- Analogically as we did with random variables, it is practical to use simpler descriptive characteristics of random vectors as:
  
  - **Mean (expectation) vector**
    
    \[ E(X) = (E(x_1), E(x_2), \ldots, E(x_n))^\top = \mu = (\mu_1, \mu_2, \ldots, \mu_n)^\top \]
  
  - **Covariance matrix**
    
    \[ \Sigma_X(i, k) = \text{cov}(X) = E((X - \mu)(X - \mu)^\top) = \begin{bmatrix} \sigma_1^2 & \cdots & c_{1n} \\ \cdots & \ddots & \cdots \\ c_{n1} & \cdots & \sigma_n^2 \end{bmatrix} \]
**Covariance matrix, properties**

- The covariance matrix indicates the tendency of each pair of features (elements of the random vector) to vary together (co-vary).

- The covariance matrix has several important properties
  
  - The covariance matrix is symmetric (i.e. $\Sigma = \Sigma^\top$) and positive-semidefinite, which means that $x^* M x \geq 0$ for all $x \in \mathbb{C}$. The notation $x^*$ means a complex conjugate of $x$.
  
  - If $x_i$ and $x_k$ tend to increase together then $c_{ik} > 0$.
  
  - If $x_i$ tends to decrease when $x_k$ increases then $c_{ik} < 0$.
  
  - If $x_i$ and $x_k$ are uncorrelated then $c_{ik} = 0$.
  
  - $|c_{ik}| \leq \sigma_i^2$, where $\sigma_i$ is the standard deviation of $x_i$.
  
  - $c_{ii} = \sigma_i^2 = D(x_i)$.

- The covariance terms can be expressed as $c_{ii} = \sigma_i^2$ and $c_{ik} = \rho_{ik} \sigma_i \sigma_k$, where $\rho_{ik}$ is called the correlation coefficient.
Covariance terms, graphical illustration

\[ C_{lk} = \sigma_l \sigma_k \]
\[ \rho_{lk} = -1 \]

\[ C_{lk} = -\frac{1}{2} \sigma_l \sigma_k \]
\[ \rho_{lk} = -\frac{1}{2} \]

\[ C_{lk} = 0 \]
\[ \rho_{lk} = 0 \]

\[ C_{lk} = \frac{1}{2} \sigma_l \sigma_k \]
\[ \rho_{lk} = \frac{1}{2} \]

\[ C_{lk} = \sigma_l \sigma_k \]
\[ \rho_{lk} = 1 \]
Quantiles, median

- The $p$-quantile $Q_p$: $P(X < Q_p) = p$.
- The median is the $p$-quantile for $p = \frac{1}{2}$, i.e. $P(X < Q_p) = \frac{1}{2}$.

Note: Median is often used as a replacement for the mean value in robust statistics.