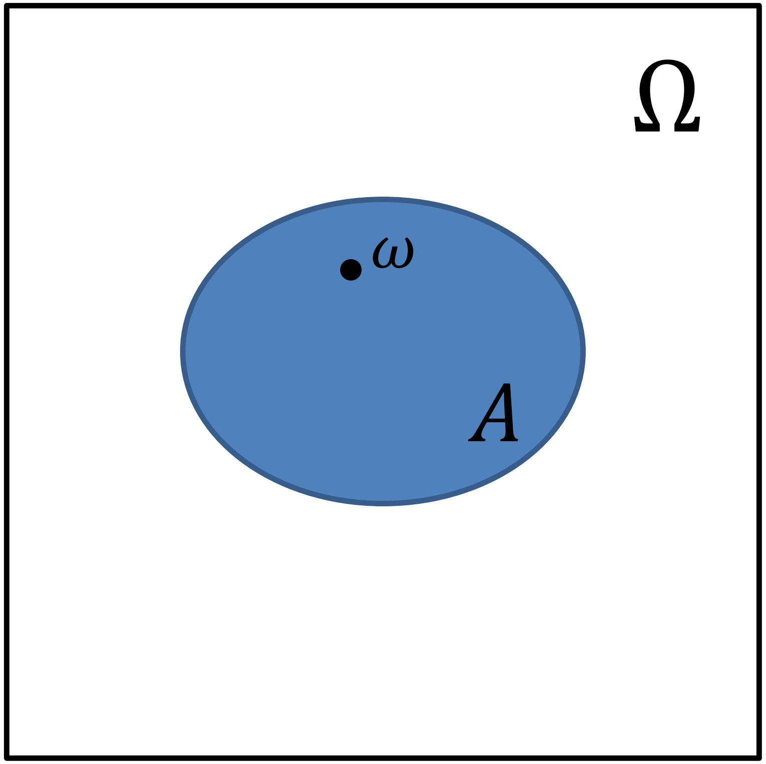
Machine Learning (521289S) Quick Probability Math Recap

M.Sc. Antti Isosalo
Physiological Signal Analysis Team
Center for Machine Vision and Signal Analysis (CMVS)
University of Oulu

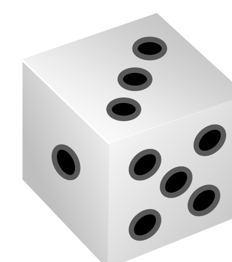
Probability Space

- Sample space / Probability space Ω
- Single event ω
  - An element of prob. space $\omega \in \Omega$
- Event A
  - Any subset of prob. space, $A \subseteq \Omega$
- Example: Coin toss
  - ω_1 = "heads"
  - ω_2 = "tails"
  - $\Omega = \{\omega_1, \omega_2\}$
- All events = the power set of Ω
  - $P(\Omega) = \{\emptyset, \{\omega_1\}, \{\omega_2\}, \{\omega_1, \omega_2\}\}$



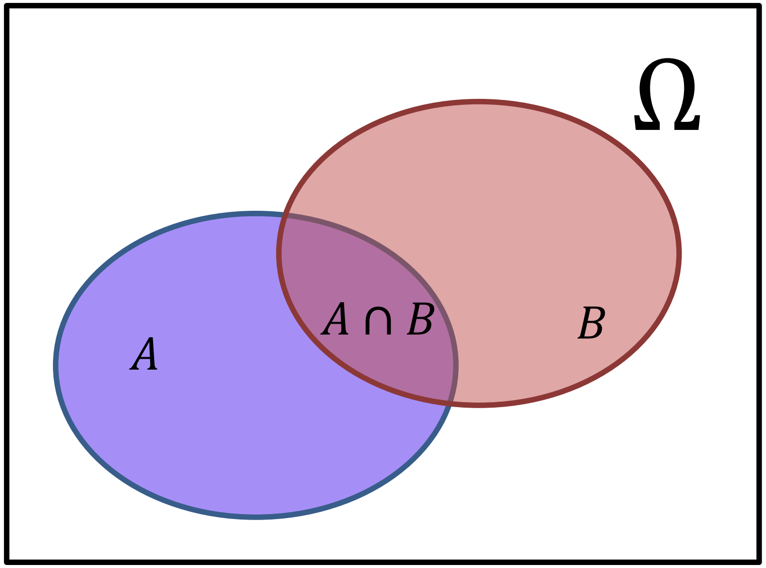
Probability of an Event

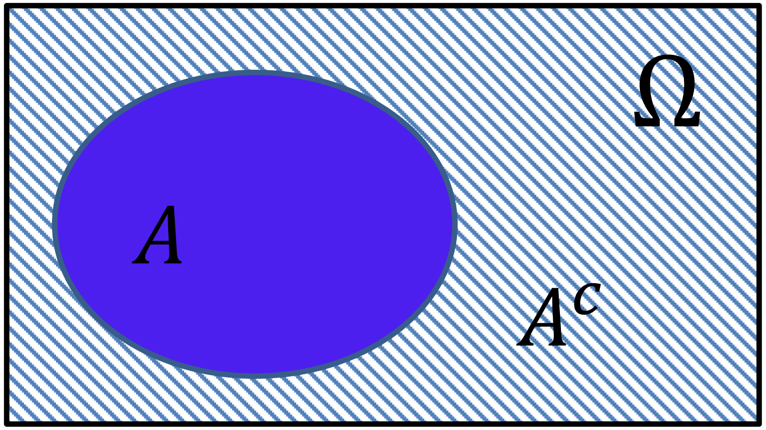
- Probability of an event A, denoted as $\mathbf{Pr}(\boldsymbol{A})$
- Classical geometric interpretation
  - The measure (area/count) of A divided by that of Ω
- Example: Rolling dice
  - ω_1 = “dice gives 1”, …, ω_6 = “dice gives 6”
  - $\Omega = \{\omega_1, \ldots, \omega_6\}$
  - A = "Dice showing even face value" = $\{\omega_2, \omega_4, \omega_6\}$
  - $\Pr(A) = \frac{\#A}{\#\Omega} = \frac{3}{6} = \frac{1}{2}$



Properties of Probability

- Let $A, B \subseteq \Omega$
- $0 \leq \Pr(A) \leq 1$
- $\Pr(\emptyset) = 0$
- $\Pr(\Omega) = 1$
- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$
  - If $A \cap B = \emptyset$, A and B are disjoint
    - $\Pr(A \cup B) = \Pr(A) + \Pr(B)$
  - **Shorthand**: $\Pr(A, B) = \Pr(A \cap B)$
- Complement events
  - $\Pr(A^c) = \Pr(\Omega \backslash A) = 1 - \Pr(A)$
    - ($\Omega\backslash$A, all elements of Ω that are **not** in A.)
  - $\Pr(A) + \Pr(A^c) = 1$
  - Sometimes: A' or $\bar{A}$ instead of A^c



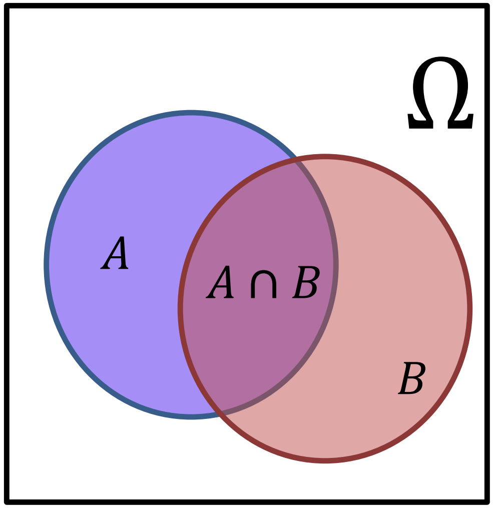


Independence and Conditional Probability

- Definition 1:
  - A and B are **independent** IF AND ONLY IF

$$\Pr(A, B) = \Pr(A)\Pr(B)$$

  - Otherwise A and B are **dependent**



- Definition 2: (Important!)
  - Probability of A given the occurrence of event B is

$$\Pr(A|B) = \frac{\Pr(A,B)}{\Pr(B)}, \text{ when } \Pr(B) > 0$$

  - If $\Pr(B) = 0$, $\Pr(A \mid B)$ is undefined

Random Variable

- A random variable is a **function** from sample space to the real line
  - $X: \Omega \to \mathbb{R}$

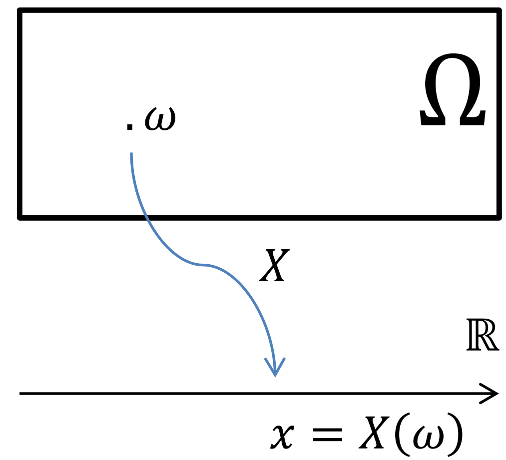
- Example: Dice rolling
  - $X(\omega) = 1$ when $\omega = 2, 4, 6$
  - $X(\omega) = 0$ when $\omega = 1, 3, 5$
  - 1 now represents *even*, and 0 *odd*

- Note: **Nothing random about it!**
  - Randomness comes from the "selection" of ω in Ω

- **Probability**: we can assign a probability to this random variable function using the preimage (inverse image) of the set in the codomain:
  - $\Pr(X(\omega) = 1) = \Pr\left(X^{-1}(\{1\})\right)$

$$= \Pr(\{\omega \in \Omega | X(\omega) = 1\})$$

$$= \Pr(\{2,4,6\}) = \frac{3}{6} = \frac{1}{2}$$



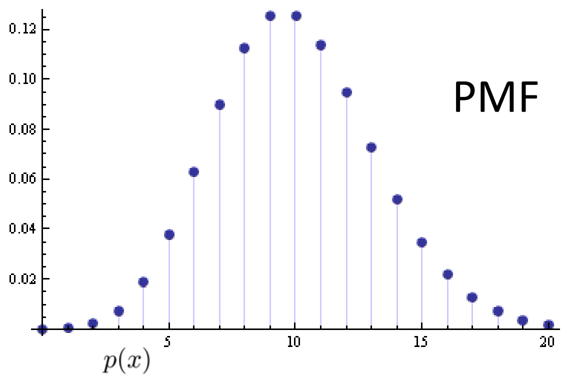
Notation:
- Random variable in capital letters
- It's values in small letters
- Often ω is omitted:

$$X(\omega) = X = x$$

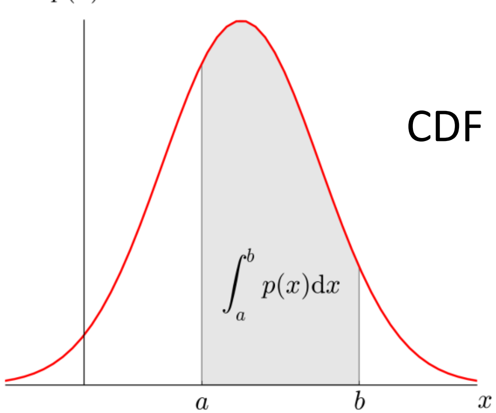
Probability Distributions

Discrete case

- The range of X, $X(\Omega)$, is *finite* or at most *countably infinite*
- **Probability mass function** (PMF) P
  - Notation: $P(x)$ or $P(X = x)$ or $P_X(x)$
- As before
  - $P(x) = \Pr(X(\omega) = x)$
    $= \Pr\left(X^{-1}(\{x\})\right) = \Pr(\{\omega \in \Omega | X(\omega) = x\})$



PMF



CDF

Continuous case

- The range of X, $X(\Omega)$, is *uncountable*
- Simple events have zero probability
  - $\Pr(X(\omega) = x) = 0$
- Makes better sense to consider intervals on the real line (subsets of $\mathbb{R}$)
  - $\Pr(X(\omega) \leq c) = \Pr\left(X^{-1}(]-\infty, c])\right)$
    $= \Pr(\{\omega \in \Omega | X(\omega) \leq c\})$
- We consider only the case where there exists a continuous function $p: \mathbb{R} \to \mathbb{R}$ such that
  - $\Pr(X(\omega) \leq c) = \int_{-\infty}^{c} p(x)\, dx$, where the function p is called **probability density function** (PDF)
- $P(x) = \Pr(X(\omega) \leq c)$ is called **cumulative (probability) density function** (CDF)
- Now $\Pr(a \leq X(\omega) \leq b) = P(b) - P(a)$

Multiple Random Variables

- Given two **discrete** random variables X_1 and X_2we may consider the probability of the joint event, i.e.

$$\Pr(\{X_1 = x_1 \text{and } X_2 = x_2\})$$

- We denote this with the **joint PMF** $P(x_1, x_2)$
- We could also use notation $P(X_1 = x_1, X_2 = x_2)$ or $P_{X_1,X_2}(x_1, x_2)$ if otherwise unclear
  - E.g. $P(X_1 = 3, X_2 = 2)$ or $P_{X_1,X_2}(3, 2)$ instead of $P(3, 2)$

- We can also combine random variables into a **random vector** $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$
  - Define its PMF as $\mathrm{P}(\mathbf{x}) = P(x_1, x_2)$

Marginalization

- Consider two random variables X_1 and X_2.
- Their joint distribution is known, and is $P(x_1, x_2)$.
- Now, the **marginal distribution** of X_1 is

$$P(x_1) = \sum_{x_2} P(x_1, x_2)$$

- We get the PMF of X_1 without reference to the values of the other variable X_2
  - The variable X_2 has been marginalized out

Marginalization (cont.)