# Advanced Machine Learning

## *Probability*

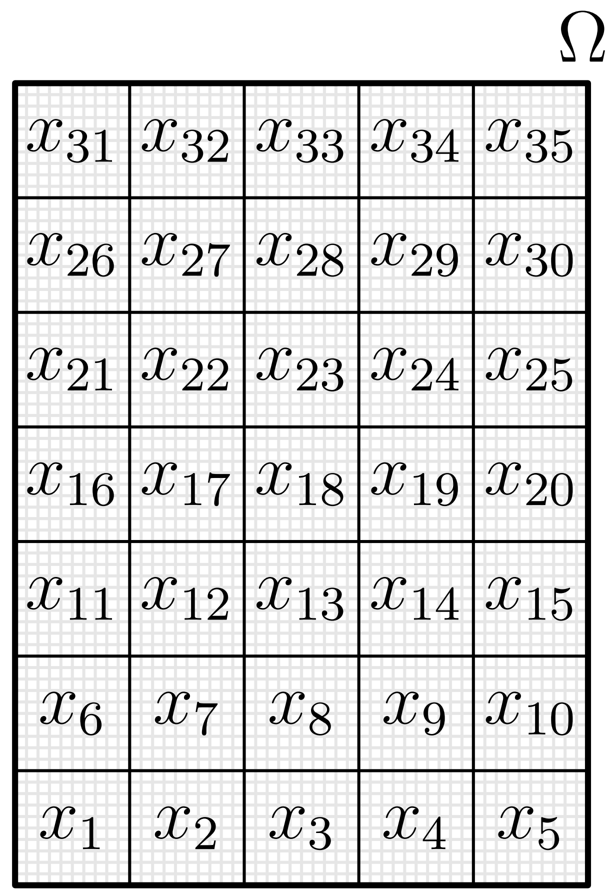
$$Y = g(X) \qquad \Omega$$

| | | | |
|---|---|---|---|
| $y_{13} = g(x_{13})$ | $y_{14} = g(x_{14})$ | $y_{15} = g(x_{15})$ | $y_{16} = g(x_{16})$ |
| $y_9 = g(x_9)$ | $y_{10} = g(x_{10})$ | $y_{11} = g(x_{11})$ | $y_{12} = g(x_{12})$ |
| $y_5 = g(x_5)$ | $y_6 = g(x_6)$ | $y_7 = g(x_7)$ | $y_8 = g(x_8)$ |
| $y_1 = g(x_1)$ | $y_2 = g(x_2)$ | $y_3 = g(x_3)$ | $y_4 = g(x_4)$ |

*Probability, Random Variables, Expectations*

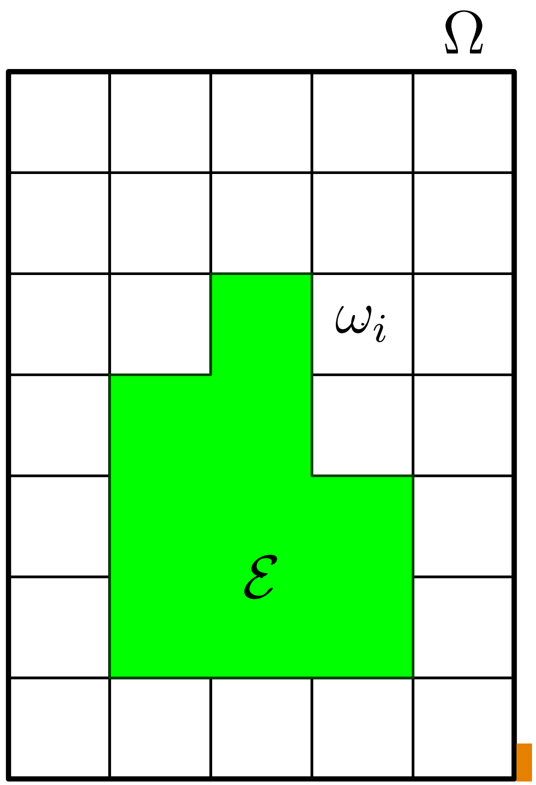
# Outline

1. **Random Variables**
2. Expectations
3. Calculus of Probabilities



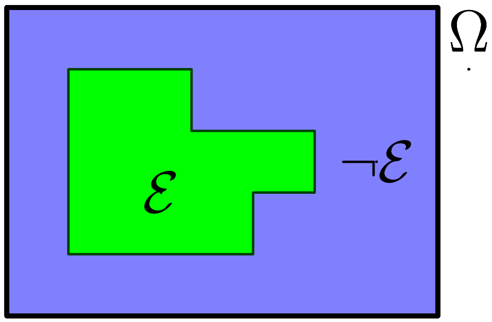
# Modelling Uncertainty

- To model a world with uncertainty we consider some set of **elementary events** or **outcomes** $\Omega$
- For the outcome of rolling a dice $\Omega = \{1,2,3,4,5,6\}$
- The elementary events $\omega_i$ are **mutually exclusive** $\omega_i \cap \omega_j = \emptyset$ and **exhaustive** $\bigcup_i \omega_i = \Omega$
- We consider **events** $\mathcal{E} = \bigcup_{i \in \mathcal{I}} \omega_i$
- E.g. For a dice throw $\mathcal{E} = \{2,4,6\}$



# Probabilities

- We attribute a **probability**, $\mathbb{P}(\mathcal{E})$, to an event, $\mathcal{E}$, with the requirements
  - $\star$ $0 \leq \mathbb{P}(\mathcal{E}) \leq 1$
  - $\star$ $\mathbb{P}(\mathcal{E}) + \mathbb{P}(\neg\mathcal{E}) = 1$ where $\neg\mathcal{E} = \Omega \setminus \mathcal{E}$




- In some cases we can interpret $\mathbb{P}(\mathcal{E})$ as the expected frequency of occurrence of a repetitive trial
- But $\mathbb{P}(\text{Pass COMP6208 exam})$ is something you do once
- Can think of probability as an informed belief that something might happen
- When our knowledge changes the probability changes

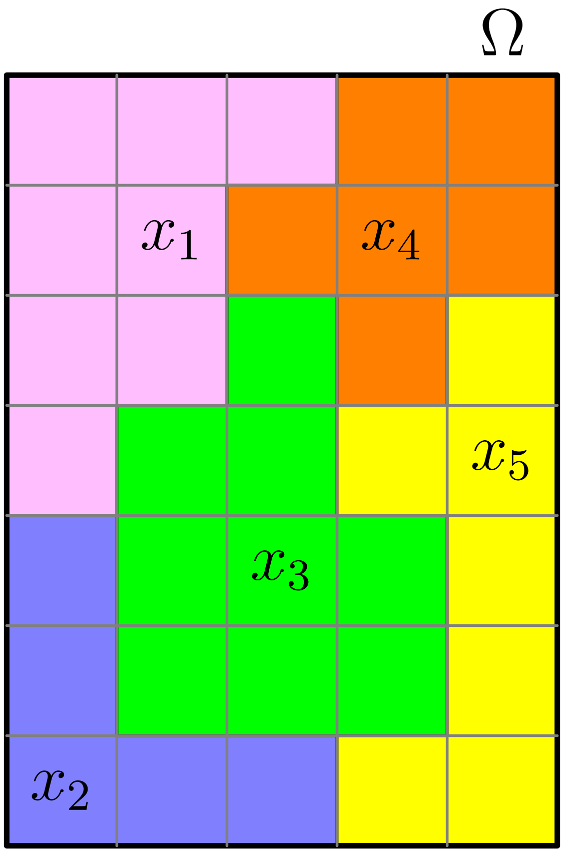
# Random Variables

- We can define a **random variable**, $X$, by partition the set of outcomes $\Omega$ and assign a numbers to each partition

- E.g. for a dice

$$X = \begin{cases} 0 & \text{if } \omega \in \{1,3,5\} \\ 1 & \text{if } \omega \in \{2,4,6\} \end{cases}$$

- $\mathbb{P}(X = x_i) = \mathbb{P}(\mathcal{E}_i)$ where $\mathcal{E}_i$ is the event that corresponding to the partition with value $x_i$



# What's In A Name

- We denote random variables with capital letters, $X$, $Y$, $Z$, etc.
- The symbol denote an object that can take one of a number of different values, but which one is still to be decided by chance
- When we write $\mathbb{P}(X)$ we can view this as short-hand for

$$(\mathbb{P}(X=x) \mid x \in \mathcal{X}) = (\mathbb{P}(X=x_1), \mathbb{P}(X=x_2), \ldots \mathbb{P}(X=x_n))$$

  where $\mathcal{X}$ is the set of possible values that $X$ can take
- We treat random variables very differently to normal numbers (scalars) when we consider taking expectations

# Function of Random Variables

- Any function, $Y = g(X)$, of a random variable, $X$, is a random variable

$$Y = g(X) \qquad \Omega$$

| | | | |
|---|---|---|---|
| $y_{13} = g(x_{13})$ | $y_{14} = g(x_{14})$ | $y_{15} = g(x_{15})$ | $y_{16} = g(x_{16})$ |
| $y_9 = g(x_9)$ | $y_{10} = g(x_{10})$ | $y_{11} = g(x_{11})$ | $y_{12} = g(x_{12})$ |
| $y_5 = g(x_5)$ | $y_6 = g(x_6)$ | $y_7 = g(x_7)$ | $y_8 = g(x_8)$ |
| $y_1 = g(x_1)$ | $y_2 = g(x_2)$ | $y_3 = g(x_3)$ | $y_4 = g(x_4)$ |

# Continuous Spaces

- If the space of elementary events is continuous (e.g. for darts $\boldsymbol{x} = (x, y)$) then $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) = 0$

- But if we consider a region, $\mathcal{R}$, then we can assign a probability to landing in the region $\mathbb{P}(\boldsymbol{X} \in \mathcal{R})$

- It is useful to work with **probability densities function** (PDF)

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \lim_{\epsilon \to 0} \frac{\mathbb{P}(\boldsymbol{X} \in \mathcal{B}(\boldsymbol{x}, \epsilon))}{|\mathcal{B}(\boldsymbol{x}, \epsilon)|}$$

  where $\mathcal{B}(\boldsymbol{x}, \epsilon)$ is a ball of radius $\epsilon$ around the point $\boldsymbol{x}$ and $|\mathcal{B}(\boldsymbol{x}, \epsilon)|$ is the volume of the ball

- If we make a change of variable the volume $|\mathcal{B}(\boldsymbol{x}, \epsilon)|$ might change so $f_{\boldsymbol{X}}(\boldsymbol{x})$ will change

# Change of Variables

- Consider a region $\mathcal{R}$—we can describe this using different coordinate systems $x$ or $y = g(x)$

- But

$$\mathbb{P}(X \in \mathcal{R}) = \int_{\mathcal{R}} f_X(x)\mathrm{d}x = \mathbb{P}(Y \in \mathcal{R}) = \int_{\mathcal{R}} f_Y(y)\mathrm{d}y$$

- As this is true for any region $\mathcal{R}$: $f_X(x)|\mathrm{d}x| = f_Y(y)|\mathrm{d}y|$

- Or

$$f_X(x) = f_Y(y)\left|\frac{\mathrm{d}y}{\mathrm{d}x}\right| = f_Y(g(x))\,|g'(x)|$$

- The probability density measured in units of probability per cm is different to that measured in units of probability per inch

# Jacobian

- In high dimension if we make a change of variables $\boldsymbol{x} \to \boldsymbol{y}(\boldsymbol{x})$ (which can be seen as a change of random variables $\boldsymbol{X} \to \boldsymbol{Y}(\boldsymbol{X})$)
- Then

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{Y}}(\boldsymbol{y}) |\det(\mathbf{J})|$$

where $\mathbf{J}$ is the Jacobian matrix

$$\mathbf{J} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix}$$

- Ensures integrals over volumes are the same

# Meaning of Probability Densities

- Probability densities are not probabilities
- They are positive, but don't need to be less than 1
- Note that

$$f_X(x) = \lim_{\delta x \to 0} \frac{\mathbb{P}(x \leq X < x + \delta x)}{\delta x}$$

- We can think of $f_X(x)\delta x$ as $\mathbb{P}(x \leq X < x + \delta x)$
- Note that $f_X(x)\delta x \leq 1$

# Cumulative Distribution Functions

- We can define the **cumulative distribution function** (CDF)

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} \sum_{i:x_i \leq x} \mathbb{P}(X = x_i) \\ \int_{-\infty}^{x} f_X(y)\,\mathrm{d}y \end{cases}$$

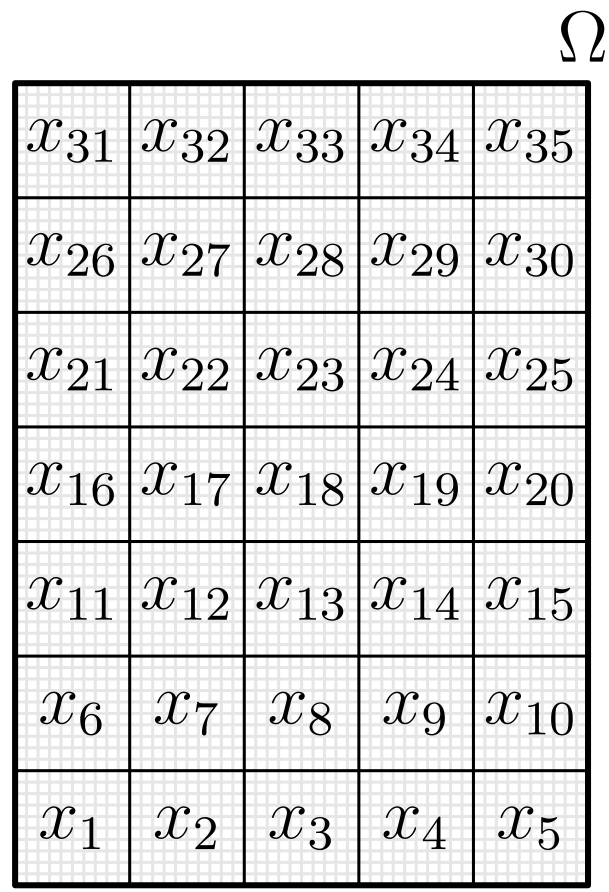
- This is a function that goes from 0 to 1 as $x$ goes from $-\infty$ to $\infty$

- We note that for continuous random variables

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}$$

# Outline

1. Random Variables
2. **Expectations**
3. Calculus of Probabilities



# Expectation

- We can define the expectation of $\boldsymbol{Y} = g(\boldsymbol{X})$ as

$$\mathbb{E}_{\boldsymbol{X}}[g(\boldsymbol{X})] = \begin{cases} \displaystyle\sum_{\boldsymbol{x} \in \mathcal{X}} g(\boldsymbol{x}) \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) \\ \displaystyle\int g(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \end{cases}$$

- The expectation of a constant $c$ is

$$\mathbb{E}_{\boldsymbol{X}}[c] = \begin{cases} \displaystyle\sum_{\boldsymbol{x} \in \mathcal{X}} c \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) = c \sum_{\boldsymbol{x} \in \mathcal{X}} \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) = c \\ \displaystyle\int c f_{\boldsymbol{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = c \int f_{\boldsymbol{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = c \end{cases}$$

- Note $\mathbb{E}_X[\mathbb{E}_X[g(X)]] = \mathbb{E}_X[g(X)]$

# Linearity of Expectation

- Because sums and integrals are linear operators

$$\sum_i (a x_i + b y_i) = a \left( \sum_i x_i \right) + b \left( \sum_i y_i \right)$$

$$\int (a f(\boldsymbol{x}) + b g(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x} = a \left( \int f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \right) + b \left( \int g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \right)$$

  then expectations are linear

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

- Beware usually $\mathbb{E}[X Y] \neq \mathbb{E}[X]\,\mathbb{E}[Y]$ (unless $X$ and $Y$ are independent)

# Indicator Functions

- An indicator function has the property

$$[\![predicate]\!] = \begin{cases} 1 & \text{if } predicate \text{ is True} \\ 0 & \text{if } predicate \text{ is False} \end{cases}$$

(sometimes written $\boldsymbol{I}_A(x)$ where $A(x)$ is the predicate)

- We can obtain probabilities from expectations

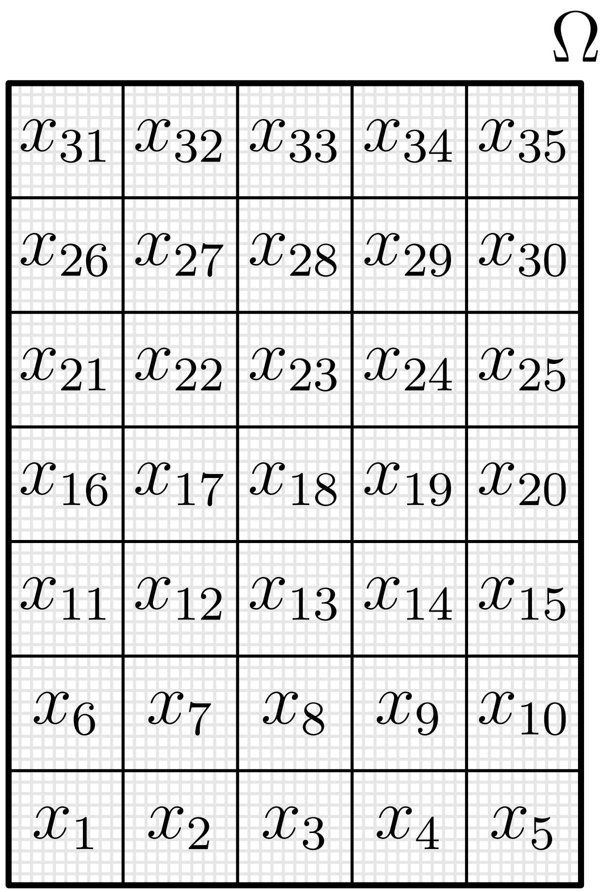
$$\mathbb{P}(predicate) = \mathbb{E}[[\![predicate]\!]]$$

- E.g. The CDF is given by

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{E}[[\![X \leq x]\!]]$$

# Outline

1. Random Variables
2. Expectations
3. **Calculus of Probabilities**



# Joint Probabilities

- Often we want to model complex processes where we have multiple random variables
- We can define the joint probability

$$p_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$$

  i.e. the probability of the event where both $X = x$ and $Y = y$
- Clearly $\mathbb{P}(X,Y) = \mathbb{P}(Y,X)$

# Marginalisation

- Probabilities are extremely easy to manipulate (although lots of people struggle)
- One of the most useful properties is known as **marginalisation**

$$\mathbb{P}(X) = \sum_{y \in \mathcal{Y}} \mathbb{P}(X, Y = y)$$

  where $\mathcal{Y}$ is the set of values that the random variable $Y$ takes
- Note that when we write $\mathbb{P}(X)$ we are saying this is true for all values that $X$ can take
- Although obvious and easy this is extremely useful

# Conditional Probability

- We can also define the probability of an event $X$ given that $Y = y$ has occurred

$$\mathbb{P}(X \mid Y = y) = \frac{\mathbb{P}(X, Y = y)}{\mathbb{P}(Y = y)}$$

- In constructing a model it is often much easier to specify conditional probabilities (because you know something) rather than joint probabilities

- When manipulating probabilities it is often easier to work with joint probabilities because we can simplify them by marginalising out random variables we are not interested in

# Basic Calculus

- To obtain the joint probability we can use

$$\mathbb{P}(X,Y) = \mathbb{P}(X|Y)\,\mathbb{P}(Y) = \mathbb{P}(Y|X)\,\mathbb{P}(X)$$

- This generalises to more random variables

$$\mathbb{P}(X,Y,Z) = \mathbb{P}(X,Y|Z)\,\mathbb{P}(Z) = \mathbb{P}(X|Y,Z)\,\mathbb{P}(Y|Z)\,\mathbb{P}(Z)$$

- We can do this in a number of different ways

$$\mathbb{P}(X,Y,Z) = \mathbb{P}(Y,Z|X)\,\mathbb{P}(X) = \mathbb{P}(Z|Y,X)\,\mathbb{P}(Y|X)\,\mathbb{P}(X)$$

- Note that $\mathbb{P}(A,B \mid X,Y)$ means the probability of random variables $A$ and $B$ given that $X$ and $Y$ take particular values

# Beware

- Conditional probabilities, $\mathbb{P}(X \mid Y)$ are probabilities for $X$, but not $Y$

$$\sum_{x \in \mathcal{X}} \mathbb{P}(X = x \mid Y) = 1$$

$$\sum_{y \in \mathcal{Y}} \mathbb{P}(X \mid Y = y) \neq 1$$

(in general)

- Note that

$$\begin{aligned} \mathbb{E}_Y[\mathbb{P}(X \mid Y)] &= \sum_{y \in \mathcal{Y}} \mathbb{P}(Y = y)\, \mathbb{P}(X|Y = y) \\ &= \sum_{y \in \mathcal{Y}} \mathbb{P}(X, Y = y) = \mathbb{P}(X) \end{aligned}$$

# Causality

- Conditional probabilities does not imply causality
- We might have causal relationships

$$\mathbb{P}(\text{pass} \mid \text{study}) = 0.9 \qquad \mathbb{P}(\text{pass} \mid \neg\text{study}) = 0.2$$

- But if we know $\mathbb{P}(\text{study}) = 0.8$ then we can compute

$$\mathbb{P}(\text{pass}, \text{study}) = \mathbb{P}(\text{pass} \mid \text{study})\, \mathbb{P}(\text{study}) = 0.9 \times 0.8 = 0.72$$

$$\mathbb{P}(\text{pass}, \neg\text{study}) = \mathbb{P}(\text{pass} \mid \neg\text{study})\, \mathbb{P}(\neg\text{study}) = 0.2 \times 0.2 = 0.04$$

and

$$\begin{aligned}\mathbb{P}(\text{study} \mid \text{pass}) &= \frac{\mathbb{P}(\text{pass}, \text{study})}{\mathbb{P}(\text{pass})} \\ &= \frac{\mathbb{P}(\text{pass}, \text{study})}{\mathbb{P}(\text{pass}, \text{study}) + \mathbb{P}(\text{pass}, \neg\text{study})} = \frac{0.72}{0.72 + 0.04} \approx 0.947\end{aligned}$$

# Independence

- Random variables $X$ and $Y$ are said to be **independent** if

$$\mathbb{P}(X,Y) = \mathbb{P}(X)\,\mathbb{P}(Y)$$

- Because $\mathbb{P}(X,Y) = \mathbb{P}(X|Y)\,\mathbb{P}(Y)$ and $\mathbb{P}(X,Y) = \mathbb{P}(Y|X)\,\mathbb{P}(X)$ independence implies

$$\mathbb{P}(X|Y) = \mathbb{P}(X) \qquad\qquad \mathbb{P}(Y|X) = \mathbb{P}(Y)$$

- Probabilistic independence implies a mathematical co-incident not necessarily causal independence
- However causal independence implies probabilistic independence
- If $X \in \{0,1\}$ represents the outcome of tossing a coin and $Y \in \{1,2,3,4,5,6\}$ the outcome of rolling a dice then $X$ and $Y$ are independent

# Well Conducted Experiments

- In well conducted experiments we expect the results we obtain are independent
- Let $\mathcal{D} = (X_1, X_2, \ldots, X_m)$ represents possible outcomes from a set of $m$ well conducted experiments then

$$\mathbb{P}(\mathcal{D}) = \prod_{i=1}^{m} \mathbb{P}(X_i)$$

- Denoting a possible sentence I might say by $\mathcal{S} = (W_1, W_2, \ldots, W_m)$ then

$$\mathbb{P}(\mathcal{S}) \neq \prod_{i=1}^{m} \mathbb{P}(W_i)$$

otherwise it's time I retired

# Conditional Independence

- Let $K(d)$ be a random variable measuring the amount you know about ML on day $d$ of your revision

- From you revision schedule you can write down your belief

$$\mathbb{P}(K(d) \mid K(d-1), K(d-2), \ldots K(1))$$

- But a very reasonable model is

$$\mathbb{P}(K(d) \mid K(d-1), K(d-2), \ldots K(1)) = \mathbb{P}(K(d) \mid K(d-1))$$

  what you are going to know today will just depend on what you knew yesterday

- We say that $K(d)$ is **conditionally independent** on $K(d-2)$, $K(d-3)$, etc. given $K(d-1)$

# Conclusion

- To work with probabilities you need to know
  - ⋆ How to go back and forward between joint probabilities and conditional probabilities
  - ⋆ How to marginalise out variables
- You need to understand that for continuous outcomes, it makes sense to talk about the probability density
- You need to know that expectations are linear operators and the expectation of a constant is the constant
- You need to understand independence