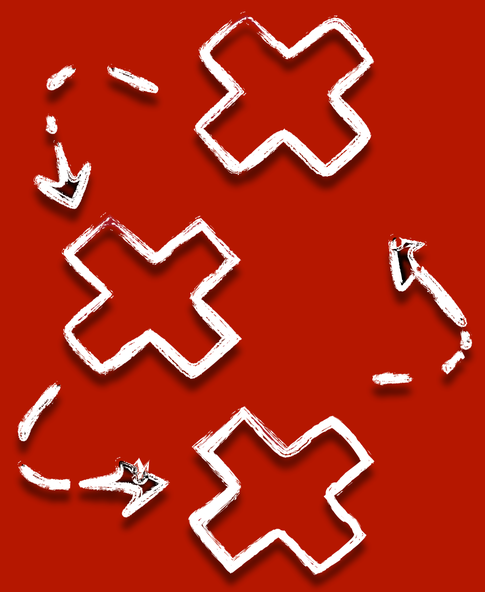


Causal Data Science

Lecture 2.1: Probability recap

Lecturer: Sara Magliacane

UvA - Spring 2024

Last class: correlation is not causation

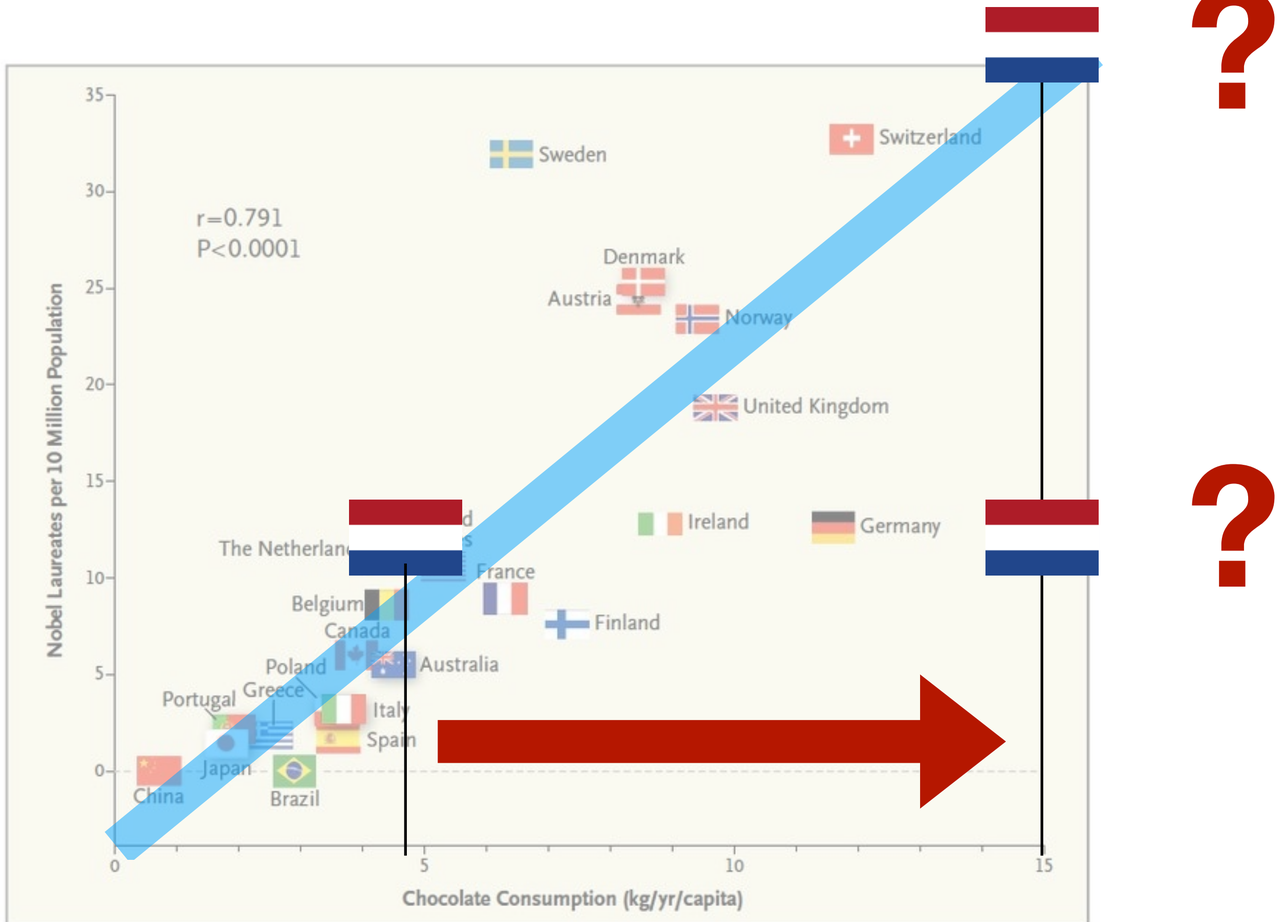


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

[Messerli, 2012] https://www.nejm.org/doi/full/10.1056/NEJMon1211064

Probability theory

- Provides a consistent framework for the quantification and manipulation of uncertainty

Why should we care?

- We mostly care about **probabilistic causes**, i.e. a certain gene increases the probability of developing a disease (but **not deterministic**)

- There might be **measurement noise,** e.g. a treatment that always works, but patients might not all be complying

Sample space Ω = all possible outcomes

- Consider flipping a coin, there are 2 different **outcomes**:

Heads **Tails**

H = “The outcome of flipping the coin is heads”

T = “The outcome of flipping the coin is tails”

Sample space Ω = all possible outcomes

- Consider flipping a coin, there are 2 different **outcomes**:

Heads **Tails**

H = “The outcome of flipping the coin is heads”

T = “The outcome of flipping the coin is tails”

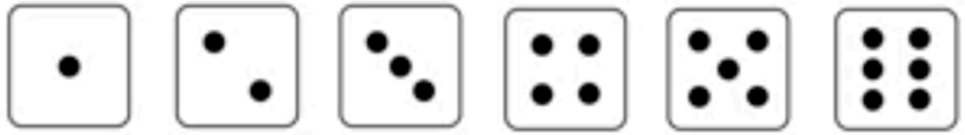
$\Omega = \{H, T\}$ All possible outcomes for flipping the coin

Sample space Ω = all possible outcomes

- Consider flipping a coin, there are 2 different **outcomes**:

$$\Omega = \{H, T\}$$

- Consider throwing a (six-sided) dice once, there are 6 different **outcomes**:



1 = “The outcome of throwing the dice is 1”

2 = “The outcome of throwing the dice is 2”

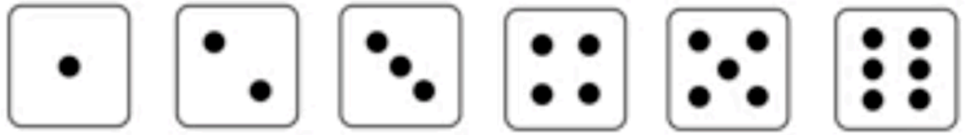
…

Sample space Ω = all possible outcomes

- Consider flipping a coin, there are 2 different **outcomes**:

$$\Omega = \{H, T\}$$

- Consider throwing a (six-sided) dice once, there are 6 different **outcomes**:



1 = “The outcome of throwing the dice is 1”

2 = “The outcome of throwing the dice is 2”

…

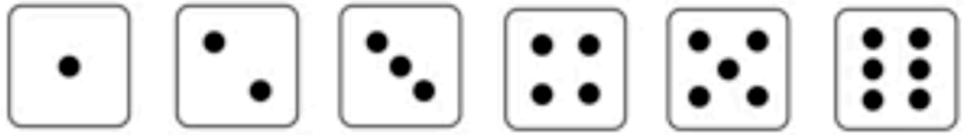
$$\Omega = \{1,\ldots,6\}$$

Sample space Ω = all possible outcomes

- Consider flipping a coin, there are 2 different **outcomes**:

$$\Omega = \{H, T\}$$

- Consider throwing a (six-sided) dice once, there are 6 different **outcomes**:



$$\Omega = \{1,\ldots,6\}$$

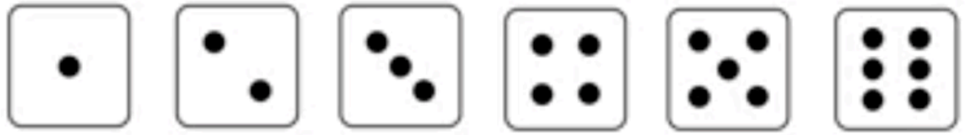
- Consider flipping **two coins**, what is the sample space?

Sample space Ω = all possible outcomes

- Consider flipping a coin, there are 2 different **outcomes**:

$$\Omega = \{H, T\}$$

- Consider throwing a (six-sided) dice once, there are 6 different **outcomes**:



$$\Omega = \{1,\ldots,6\}$$

- Consider flipping **two coins**, what is the sample space?

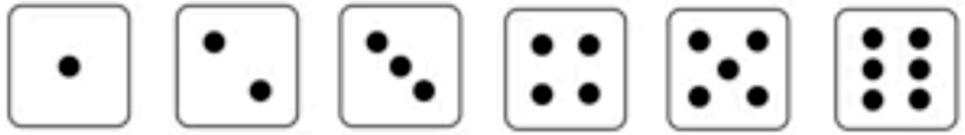
$$\Omega = \{(T,T),(T,H),(H,T),(H,H)\}$$

Sample space Ω = all possible outcomes

- Consider flipping a coin, there are 2 different **outcomes**:

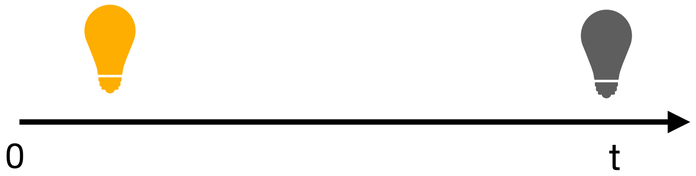
$$\Omega = \{H, T\}$$

- Consider throwing a (six-sided) dice once, there are 6 different **outcomes**:



$$\Omega = \{1,\ldots,6\}$$

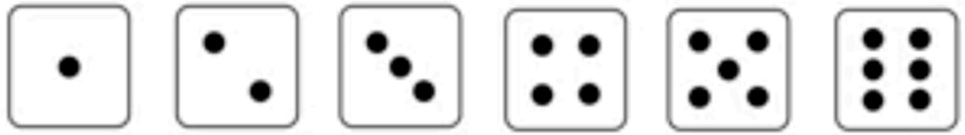
- Can be **infinitely large**: e.g. the lifetime of a lightbulb:



$$\Omega = [0,\infty)$$

Events ~= subsets of outcomes

- For **finite sample spaces**, we can consider subsets of outcomes as **events**



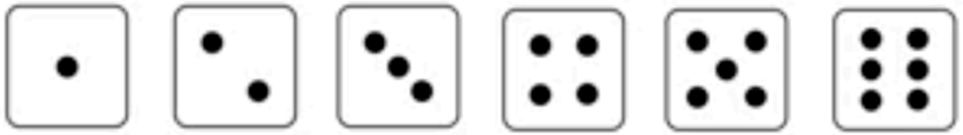
1 = “The outcome of throwing the dice is 1”

2 = “The outcome of throwing the dice is 2”

…

Events ~= subsets of outcomes

- For **finite sample spaces**, we can consider subsets of outcomes as **events**



1 = “The outcome of throwing the dice is 1”

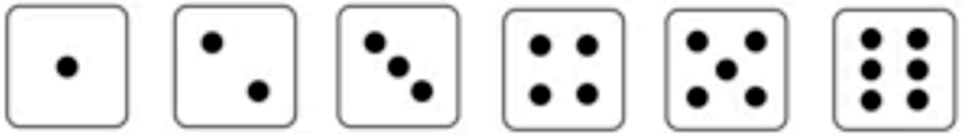
2 = “The outcome of throwing the dice is 2”

…

$\{2,4,6\}$ = **“The outcome of throwing the dice is an even number”**

Events ~= subsets of outcomes

- For **finite sample spaces**, we can consider subsets of outcomes as **events**



1 = “The outcome of throwing the dice is 1”

2 = “The outcome of throwing the dice is 2”

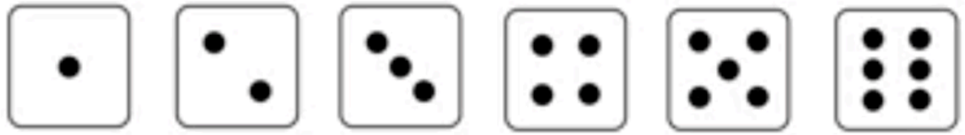
…

$\{2,4,6\} =$ **“The outcome of throwing the dice is an even number”**

$\{1,3,5\} =$ **“The outcome of throwing the dice is an odd number”**

Events ~= subsets of outcomes

- For **finite sample spaces**, we can consider subsets of outcomes as **events**



1 = “The outcome of throwing the dice is 1”

2 = “The outcome of throwing the dice is 2”

…

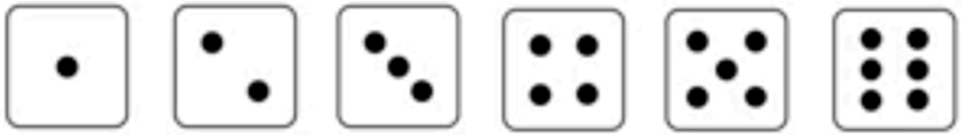
$\{2,4,6\}$ = **“The outcome of throwing the dice is an even number”**

$\{1,3,5\}$ = **“The outcome of throwing the dice is an odd number”**

$\{1,\ldots,6\}$ = **“The outcome of throwing the dice is any possible number”**

Events ~= subsets of outcomes

- For **finite sample spaces**, we can consider subsets of outcomes as **events**

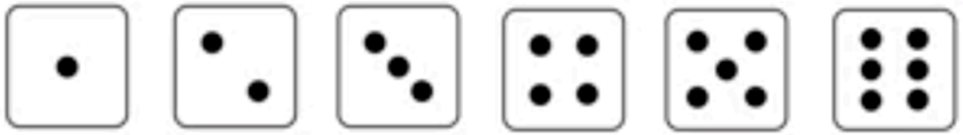


$\varnothing, \{2,4,6\}, \{1,3,5\}, \{1,\ldots,6\}$

- In this case, we can consider the set of all subsets of Ω:

Events ~= subsets of outcomes

- For **finite sample spaces**, we can consider subsets of outcomes as **events**



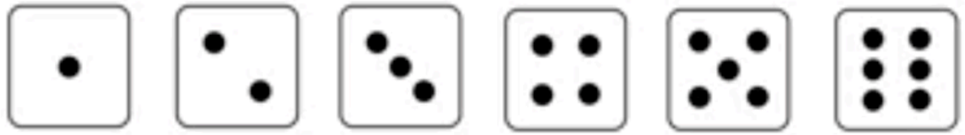
$\varnothing, \{2,4,6\}, \{1,3,5\}, \{1,\ldots,6\}$

- In this case, we can consider the set of all subsets of Ω:

$$\mathscr{F} = 2^{\Omega} = \{\varnothing, \{1\}, \ldots, \{6\}, \{1,2\}, \{1,3\}, \ldots, \{1,2,3,4,5,6\}\}$$

Events ~= subsets of outcomes

- For **finite sample spaces**, we can consider subsets of outcomes as **events**



$\varnothing, \{2,4,6\}, \{1,3,5\}, \{1,\ldots,6\}$

- In this case, we can consider the set of all subsets of Ω:

$$\mathscr{F} = 2^{\Omega} = \{\varnothing, \{1\}, \ldots, \{6\}, \{1,2\}, \{1,3\}, \ldots, \{1,2,3,4,5,6\}\}$$

- Since probability is defined as a measure over events (*spoiler*), we need additional properties to define events for **infinitely large sample spaces**
  - **Events =** elements of a **σ-algebra $\mathscr{F}$**

Events = elements of a σ-algebra

- A **σ-algebra** $\mathscr{F}$ on a set Ω is a collection of subsets of Ω
  - Ω is in $\mathscr{F}$

Events = elements of a σ-algebra

- A **σ-algebra** $\mathscr{F}$ on a set Ω is a collection of subsets of Ω
  - Ω is in $\mathscr{F}$
  - $\mathscr{F}$ is closed under complement (if $A \in \mathscr{F}$ then $(\Omega \backslash A) \in \mathscr{F}$)

Events = elements of a σ-algebra

- A **σ-algebra** $\mathscr{F}$ on a set Ω is a collection of subsets of Ω
  - Ω is in $\mathscr{F}$
  - $\mathscr{F}$ is closed under complement (if $A \in \mathscr{F}$ then $(\Omega \backslash A) \in \mathscr{F}$)
  - $\mathscr{F}$ is closed under countable unions

$$\text{(if } A_i \in \mathscr{F} \text{ for } \forall i = 1,2,\ldots \text{ then } \bigcup_{i=1}^{\infty} A_i \in \mathscr{F}\text{)}$$

Events = elements of a σ-algebra

- A **σ-algebra** $\mathscr{F}$ on a set Ω is a collection of subsets of Ω
  - Ω is in $\mathscr{F}$
  - $\mathscr{F}$ is closed under complement (if $A \in \mathscr{F}$ then $(\Omega \backslash A) \in \mathscr{F}$)
  - $\mathscr{F}$ is closed under countable unions

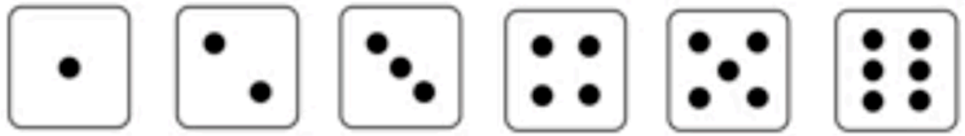
$$\text{(if } A_i \in \mathscr{F} \text{ for } \forall i = 1,2,\ldots \text{ then } \bigcup_{i=1}^{\infty} A_i \in \mathscr{F}\text{)}$$

  - (=> $\emptyset$ is in $\mathscr{F}$)
  - (=> $\mathscr{F}$ is closed under countable intersections)

Events = elements of a σ-algebra

- A **σ-algebra** $\mathscr{F}$ on a set Ω is a collection of subsets of Ω
  - Ω is in $\mathscr{F}$
  - $\mathscr{F}$ is closed under complement
  - $\mathscr{F}$ is closed under countable unions

Example: six-sided dice

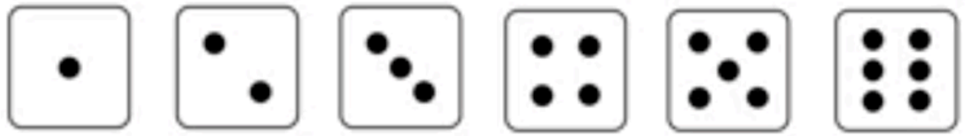


$$\Omega = \{1,\ldots,6\}$$

Events = elements of a σ-algebra

- A **σ-algebra** $\mathscr{F}$ on a set Ω is a collection of subsets of Ω
  - Ω is in $\mathscr{F}$
  - $\mathscr{F}$ is closed under complement
  - $\mathscr{F}$ is closed under countable unions

Example: six-sided dice

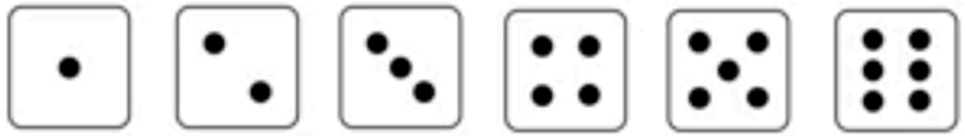


$$\Omega = \{1,\ldots,6\} \qquad \mathscr{F} = \{\varnothing, \{1,\ldots,6\}\}$$

Events = elements of a σ-algebra

- A **σ-algebra** $\mathscr{F}$ on a set Ω is a collection of subsets of Ω
  - Ω is in $\mathscr{F}$
  - $\mathscr{F}$ is closed under complement
  - $\mathscr{F}$ is closed under countable unions

Example: six-sided dice



$$\Omega = \{1,\ldots,6\}$$

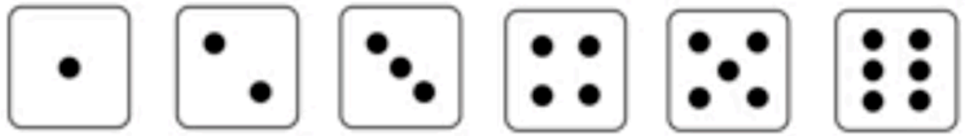
$$\mathscr{F} = \{\varnothing, \{1,\ldots,6\}\}$$

$$\mathscr{F} = \{\varnothing, \{2\}, \{4\}, \{1,\ldots,6\}\}?$$

Events = elements of a σ-algebra

- A **σ-algebra** $\mathscr{F}$ on a set Ω is a collection of subsets of Ω
  - Ω is in $\mathscr{F}$
  - $\mathscr{F}$ is closed under complement
  - $\mathscr{F}$ is closed under countable unions

Example: six-sided dice



$$\Omega = \{1,\ldots,6\}$$

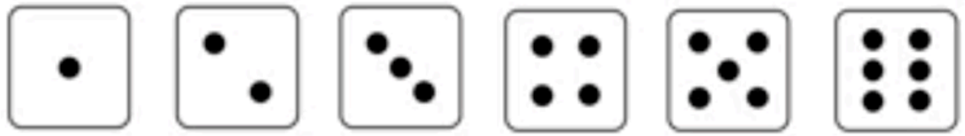
$$\mathscr{F} = \{\varnothing, \{1,\ldots,6\}\}$$

$$\cancel{\mathscr{F} = \{\varnothing, \{2\}, \{4\}, \{1,\ldots,6\}\}}? \quad \{2\} \cup \{4\} = \{2,4\} \notin \mathscr{F}$$

Events = elements of a σ-algebra

- A **σ-algebra** $\mathscr{F}$ on a set Ω is a collection of subsets of Ω
  - Ω is in $\mathscr{F}$
  - $\mathscr{F}$ is closed under complement
  - $\mathscr{F}$ is closed under countable unions

Example: six-sided dice



$$\Omega = \{1,\ldots,6\}$$

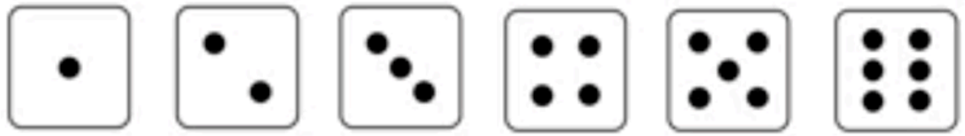
$$\mathscr{F} = \{\varnothing, \{1,\ldots,6\}\}$$

$$\mathscr{F} = \{\varnothing, \{2,4,6\}, \{1,3,5\}, \{1,\ldots,6\}\}$$

Events = elements of a σ-algebra

- A **σ-algebra** $\mathscr{F}$ on a set Ω is a collection of subsets of Ω
  - Ω is in $\mathscr{F}$
  - $\mathscr{F}$ is closed under complement
  - $\mathscr{F}$ is closed under countable unions

Example: six-sided dice



$$\Omega = \{1,\ldots,6\}$$

$$\mathscr{F} = \{\varnothing, \{1,\ldots,6\}\}$$

$$\mathscr{F} = \{\varnothing, \{2,4,6\}, \{1,3,5\}, \{1,\ldots,6\}\}$$

$$\mathscr{F} = 2^{\Omega} = \{\varnothing, \{1\}, \ldots, \{6\}, \{1,2\}, \{1,3\}, \ldots, \{1,2,3,4,5,6\}\}$$

Probability space definition

- A **probability space** is a tuple $(\Omega, \mathscr{F}, P)$ where
  - Ω is the **sample space** (the set of all possible outcomes)
  - $\mathscr{F}$ is the **event space** (the set of events, which is a σ-algebra)

Probability space definition

- A **probability space** is a tuple $(\Omega, \mathscr{F}, P)$ where
  - Ω is the **sample space** (the set of all possible outcomes)
  - $\mathscr{F}$ is the **event space** (the set of events, which is a σ-algebra)
  - P is the **probability function** $P : \mathscr{F} \to [0,1]$

Probability space definition

- A **probability space** is a tuple $(\Omega, \mathscr{F}, P)$ where
  - Ω is the **sample space** (the set of all possible outcomes)
  - $\mathscr{F}$ is the **event space** (the set of events, which is a σ-algebra)
  - P is the **probability function** $P : \mathscr{F} \to [0,1]$ such that:
    - $P(\Omega) = 1$

Probability space definition

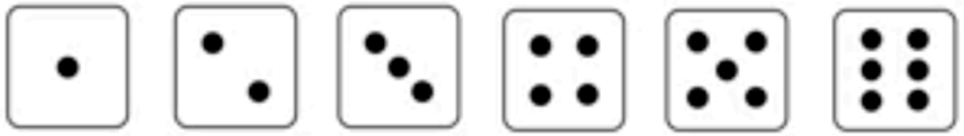
- A **probability space** is a tuple $(\Omega, \mathscr{F}, P)$ where
  - Ω is the **sample space** (the set of all possible outcomes)
  - $\mathscr{F}$ is the **event space** (the set of events, which is a σ-algebra)
  - P is the **probability function** $P : \mathscr{F} \to [0,1]$ such that:
    - $P(\Omega) = 1$
    - P is countably additive, for pairwise disjoint $A_i \in \mathscr{F}$ for $i = 1,2,\ldots$

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Example of probability space - unbiased dice

- A probability space is a tuple $(\Omega, \mathscr{F}, P)$ where $P : \mathscr{F} \to [0,1]$

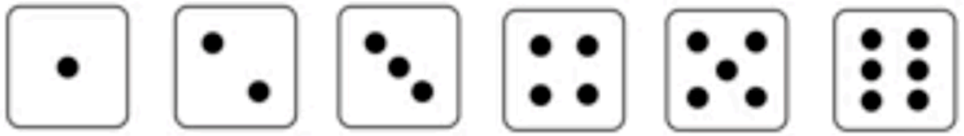
$\Omega = \{1,2,3,4,5,6\}$



Example of probability space - unbiased dice

- A probability space is a tuple $(\Omega, \mathscr{F}, P)$ where $P : \mathscr{F} \to [0,1]$

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

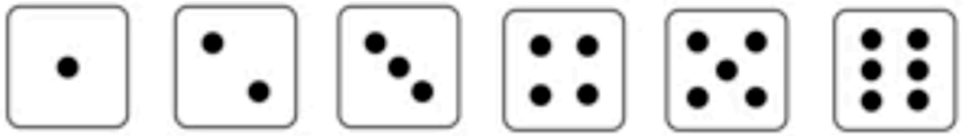


$\mathscr{F} = 2^{\Omega} = \{\varnothing, \{1\}, \ldots, \{6\}, \{1,2\}, \{1,3\}, \ldots, \{1,2,3,4,5,6\}\}$

Example of probability space - unbiased dice

- A probability space is a tuple $(\Omega, \mathscr{F}, P)$ where $P : \mathscr{F} \to [0,1]$

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



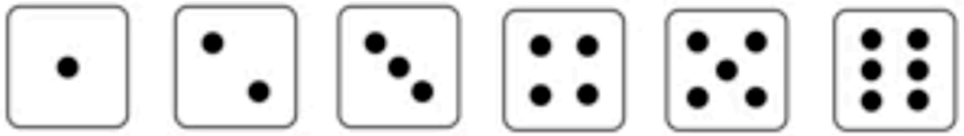
$\mathscr{F} = 2^{\Omega} = \{\varnothing, \{1\}, \ldots, \{6\}, \{1,2\}, \{1,3\}, \ldots, \{1,2,3,4,5,6\}\}$

$$P(\{1\}) = \frac{1}{6} \quad P(\{2\}) = \frac{1}{6} \quad \cdots \quad P(\{6\}) = \frac{1}{6}$$

Example of probability space - unbiased dice

- A probability space is a tuple $(\Omega, \mathscr{F}, P)$ where $P : \mathscr{F} \to [0,1]$

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



$\mathscr{F} = 2^{\Omega} = \{\varnothing, \{1\}, \ldots, \{6\}, \{1,2\}, \{1,3\}, \ldots, \{1,2,3,4,5,6\}\}$

$$P(\{1\}) = \frac{1}{6} \quad P(\{2\}) = \frac{1}{6} \quad \cdots \quad P(\{6\}) = \frac{1}{6}$$

$$\forall A \in \mathscr{F} : P(A) = \sum_{\omega \in A} P(\omega) \qquad P(\{1,2\}) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

Random variables

- A random variable X is a (measurable) function

$$X : \Omega \to E \qquad E \text{ measurable space (equipped with a } \sigma\text{-algebra } T)$$

Random variables

- A random variable X is a (***measurable***) function

$X : \Omega \to E$ E measurable space (equipped with a σ-algebra T)

$X : (\Omega, \mathscr{F}) \to (E, T)$ $\forall S \in T : X^{-1}(S) = \{\omega \in \Omega : X(\omega) \in S\} \in \mathscr{F}$

Random variables

- A random variable X is a (***measurable***) function

$$X : \Omega \to E \qquad E \text{ measurable space}$$

- We can consider **discrete random variables**

$$X : \Omega \to \mathbb{N} \qquad \text{E.g. age}$$
$$X : \Omega \to \{0,1\} \qquad \text{E.g. treated}$$

- We can consider **continuous random variables**

$$X : \Omega \to \mathbb{R} \qquad \text{E.g. height}$$

Random variables

- A random variable X is a (***measurable***) function

$$X : \Omega \to E \qquad E \text{ measurable space}$$

- We can consider **discrete random variables**

$$X : \Omega \to \mathbb{N}$$
$$X : \Omega \to \{0,1\}$$

$$P(\{\omega \in \Omega : X(\omega) \in A\}) = P(X \in A)$$

- We can consider **continuous random variables**

$$X : \Omega \to \mathbb{R}$$

Random variables

- A random variable X is a (***measurable***) function

$$X : \Omega \to E \qquad E \text{ measurable space}$$

- We can consider **discrete random variables**

$$X : \Omega \to \mathbb{N}$$
$$X : \Omega \to \{0,1\}$$

Probability mass function (PMF)

$$P(X \in A) = \sum_{x \in A} p(x)$$

- We can consider **continuous random variables**

$$X : \Omega \to \mathbb{R}$$

Random variables

- A random variable X is a (***measurable***) function

$$X : \Omega \to E \qquad E \text{ measurable space}$$

- We can consider **discrete random variables**

$$X : \Omega \to \mathbb{N}$$
$$X : \Omega \to \{0,1\}$$

Probability mass function (PMF)

$$P(X \in A) = \sum_{x \in A} p(x)$$

- We can consider **continuous random variables**

$$X : \Omega \to \mathbb{R}$$

Probability density function (PDF)

$$P(X \in A) = \int_{x \in A} p(x)dx$$

Random variables

- A random variable X is a (***measurable***) function

$$X : \Omega \to E \qquad E \text{ measurable space}$$

Warning: in ML we often use $p(X)$ and $P(X)$ inconsistently, so this will probably also happen in the slides

- We can consider **discrete random variables**

$$X : \Omega \to \mathbb{N}$$
$$X : \Omega \to \{0,1\}$$

Probability mass function (PMF)

$$P(X \in A) = \sum_{x \in A} p(x)$$

- We can consider **continuous random variables**

$$X : \Omega \to \mathbb{R}$$

Probability density function (PDF)

$$P(X \in A) = \int_{x \in A} p(x) dx$$

Example: two discrete random variables

- X is dice 1, Y is dice 2. If X <=3, then we fix Y=1, otherwise we cast dice 2.

example by Fredrik Johansson

Example: two discrete random variables

- X is dice 1, Y is dice 2. If X <=3, then we fix Y=1, otherwise we cast dice 2.

Y \ X	1	2	3	4	5	6
6	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

$p(X, Y)$

Joint probability distribution

$p(X = 6, Y = 2) = 1/36$

example by Fredrik Johansson

Marginal distributions

Y \ X	1	2	3	4	5	6
6	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

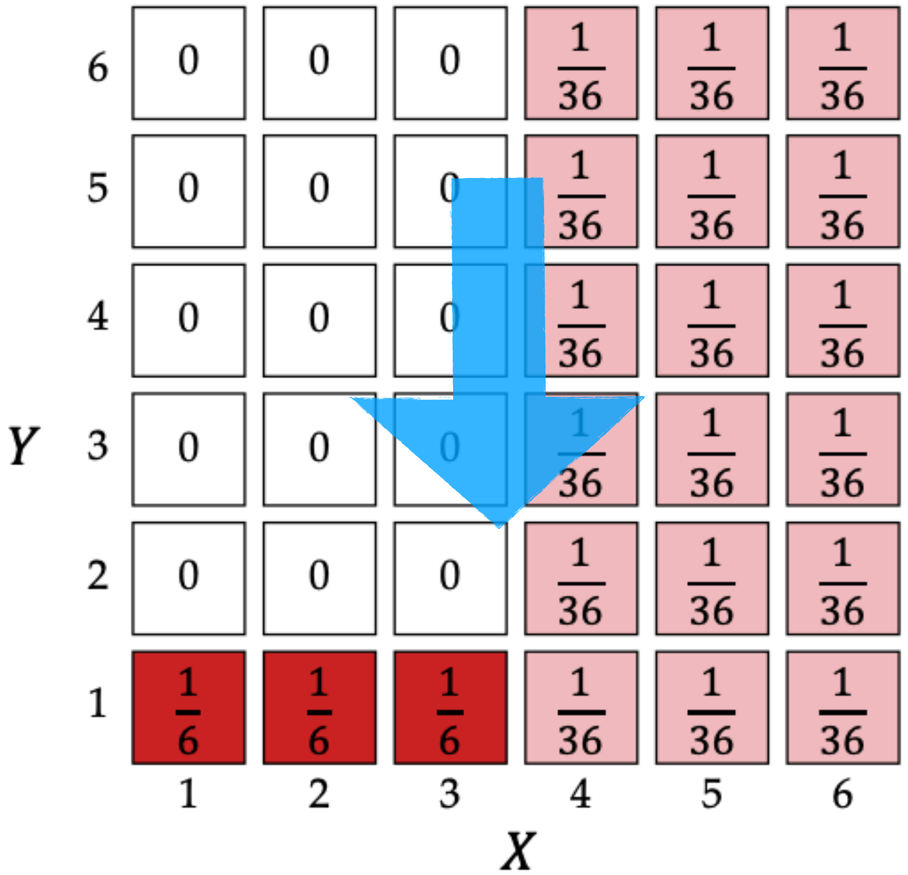
X is the random variable, x is an observation

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

X	1	2	3	4	5	6

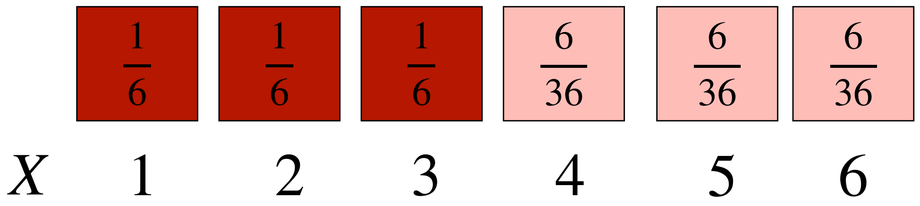
example by Fredrik Johansson

Marginal distributions



X is the random variable, x is an observation

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



example by Fredrik Johansson

Marginal distributions

Y \ X	1	2	3	4	5	6
6	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

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

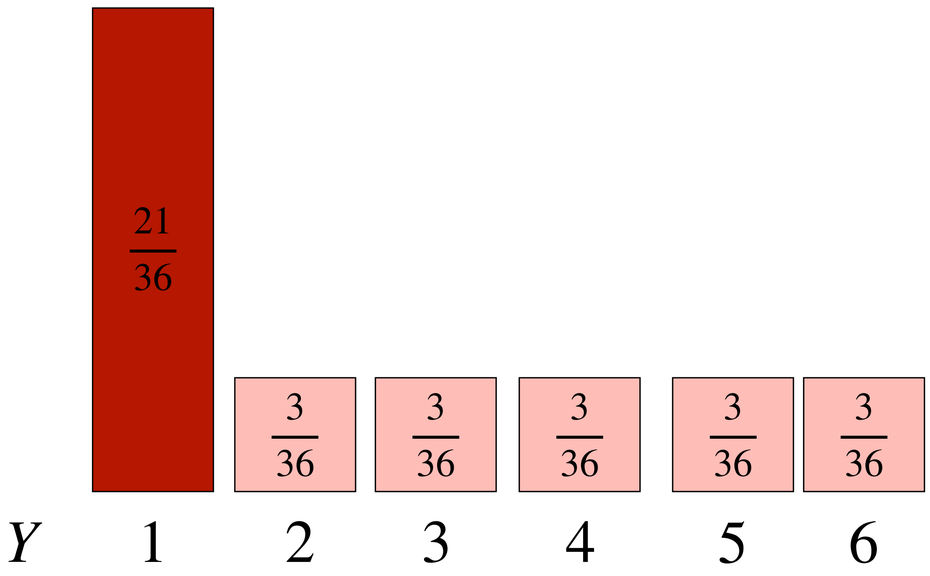
Y	1	2	3	4	5	6

example by Fredrik Johansson

Marginal distributions

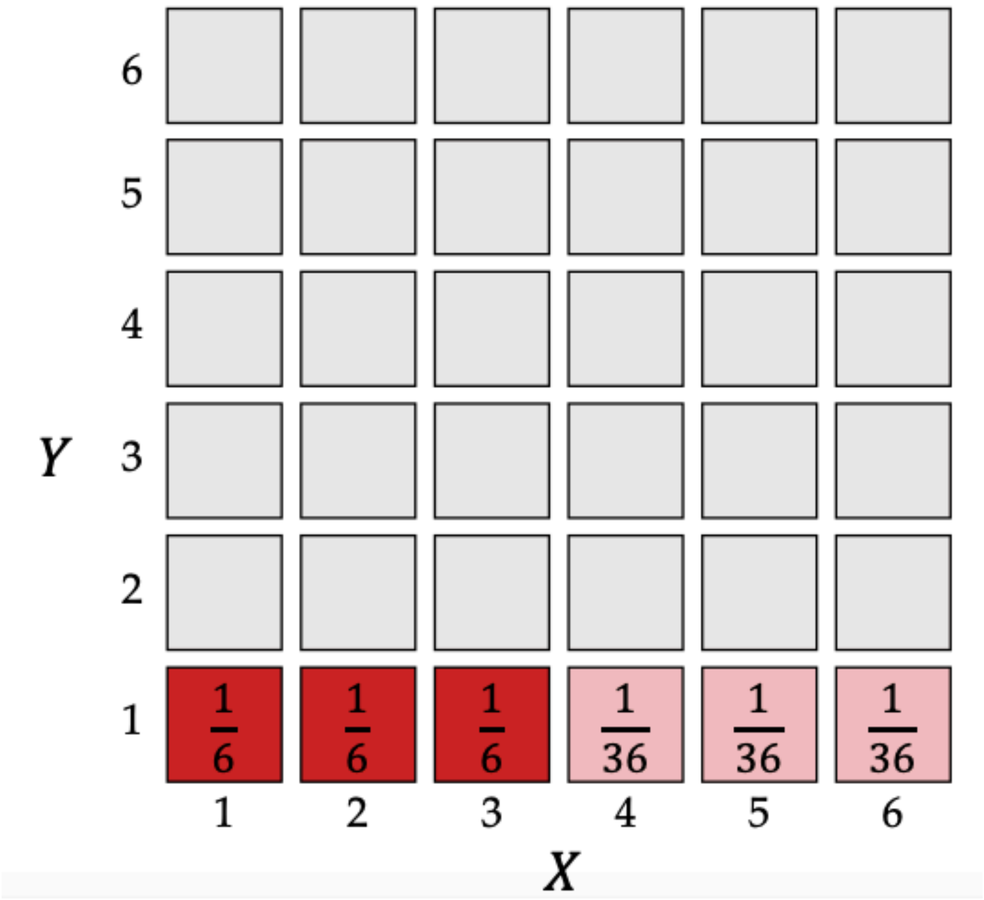
Y \ X	1	2	3	4	5	6
6	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

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



example by Fredrik Johansson

Conditional distributions



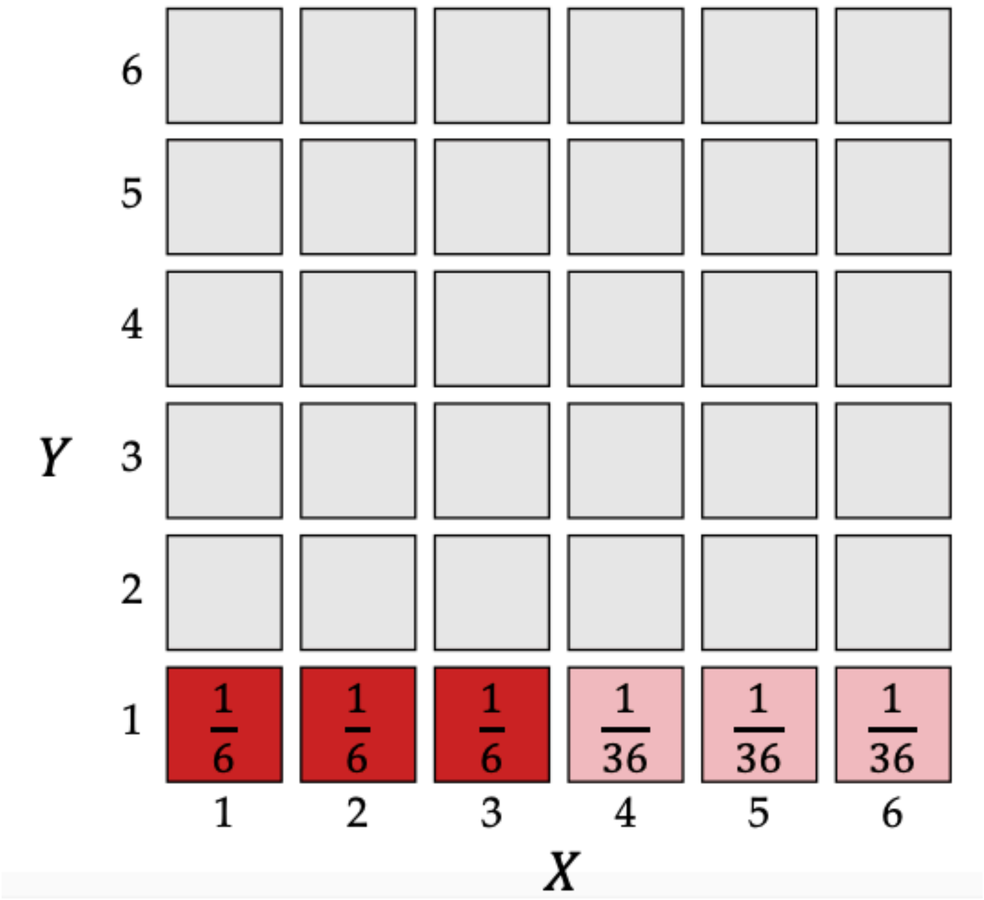
$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

$$P(X = x \mid Y = 1) = \frac{P(X = x, Y = 1)}{P(Y = 1)} \qquad P(Y = 1) = \frac{21}{36}$$

X 1 2 3 4 5 6

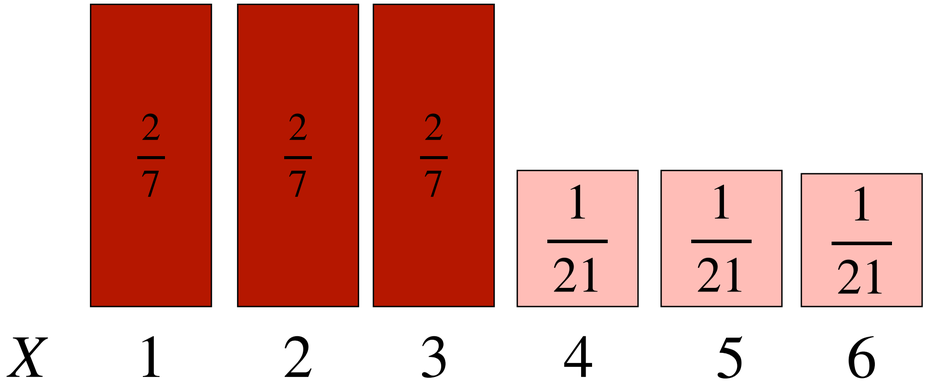
example by Fredrik Johansson

Conditional distributions



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

$$P(X = x \mid Y = 1) = \frac{P(X = x, Y = 1)}{P(Y = 1)} \qquad P(Y = 1) = \frac{21}{36}$$



example by Fredrik Johansson

Conditional distributions

Y \ X	1	2	3	4	5	6
6	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

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

$$P(X = x \mid Y = 2)?$$

Quiz in Canvas

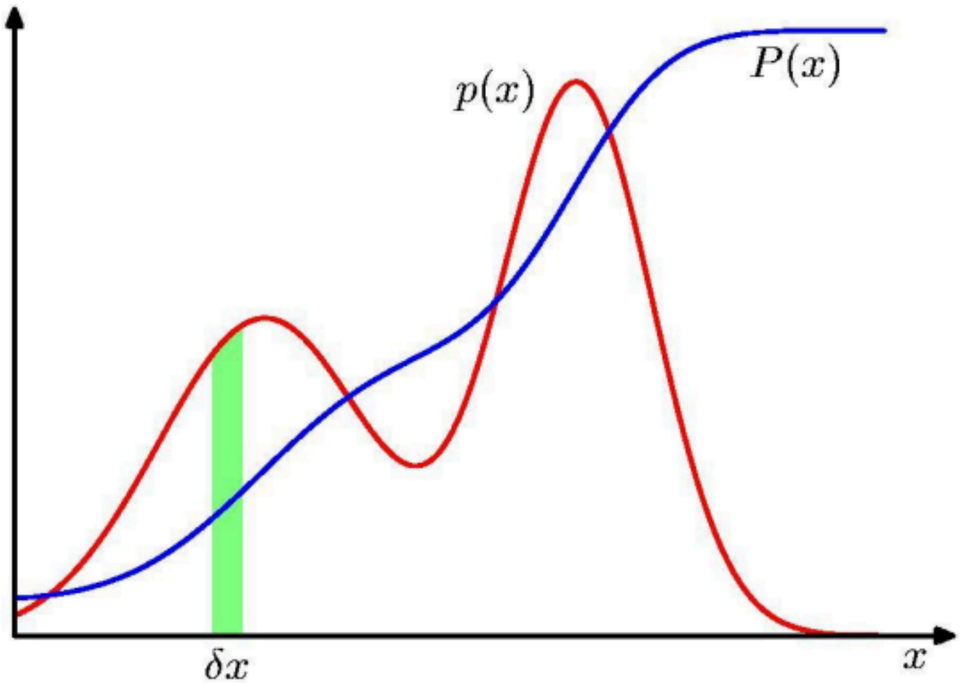
example by Fredrik Johansson

Continuous random variables

- Probability of $x \in \mathbb{R}$ falling in the interval $(x, x + dx)$ is $p(x)dx$

- $p(x)$ is the **probability density**

$$p(x \in (a, b)) = \int_a^b p(x)dx$$



https://www.microsoft.com/en-us/research/wp-content/uploads/2016/05/prml-slides-1.pdf

Continuous random variables

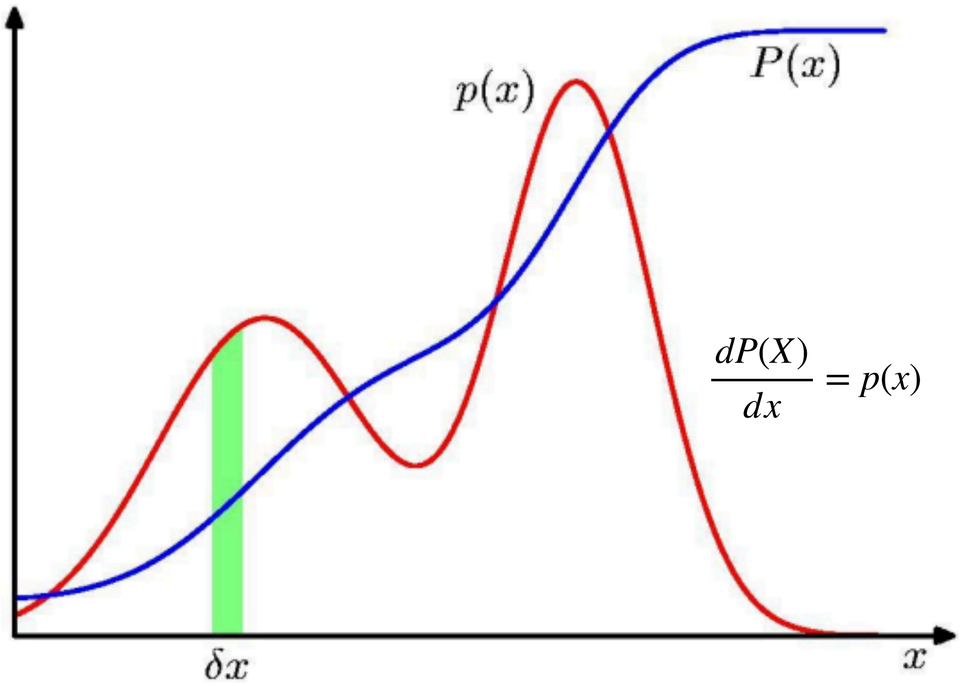
- Probability of $x \in \mathbb{R}$ falling in the interval $(x, x + dx)$ is $p(x)dx$

- $p(x)$ is the **probability density**

$$p(x \in (a, b)) = \int_a^b p(x)dx$$

- **Cumulative distribution P(X)**

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



https://www.microsoft.com/en-us/research/wp-content/uploads/2016/05/prml-slides-1.pdf

Bayes theorem

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

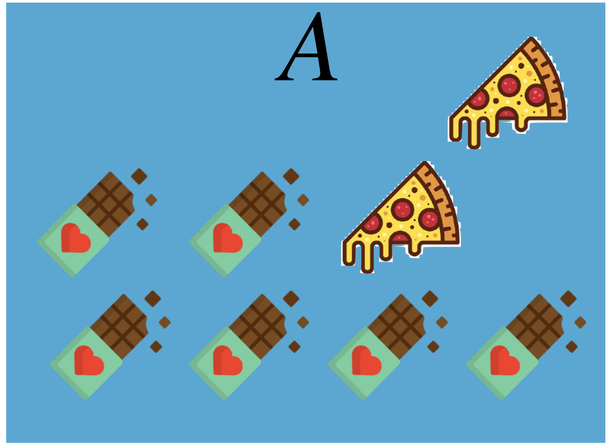
Bayes theorem

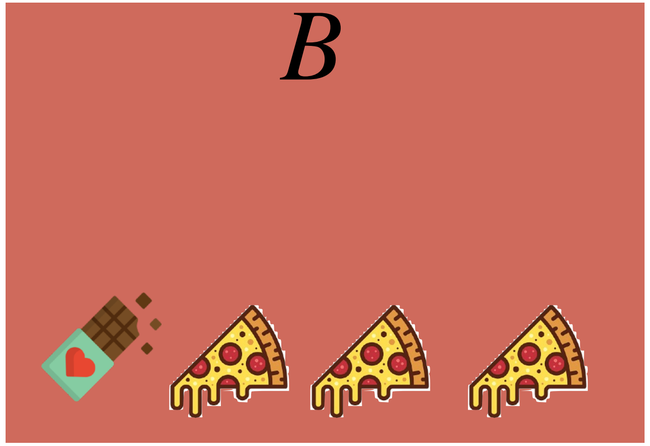
$$P(Y|X) = \frac{P(X,Y)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)}$$

In ML we often call:

- $P(Y)$ is the **prior probability of Y=y**
- $P(X)$ is the **evidence of X=x**
- $P(X|Y)$ is the **likelihood of X=x given Y=y**
- $P(Y|X)$ is the **posterior of Y=y given X=x**

Applying Bayes' theorem to pizza and chocolate





Quiz in Canvas

Independent random variables

- **Definition:** Two random variables X and Y are **independent** iff:

$$\forall x, y : P(X = x, Y = y) = P(X = x)P(Y = y)$$

- We then write $X \perp\!\!\!\perp Y$, otherwise we write $X \not\perp\!\!\!\perp Y$ **(dependent)**

Independent random variables

- **Definition:** Two random variables X and Y are **independent** iff:

$$\forall x, y : P(X = x, Y = y) = P(X = x)P(Y = y)$$

- We then write $X \perp\!\!\!\perp Y$, otherwise we write $X \not\perp\!\!\!\perp Y$ **(dependent)**

- This is equivalent to $P(X = x \mid Y = y) = P(X = x)$ (and vice versa for Y)

- Intuitively, this means that knowing the value of Y **will not tell us anything** about the distribution of X.

Independent random variables - example

- Consider throwing an **unbiased coin** two times:

- Knowing that $X_1 = H$ will not change $P(X_2 = H \mid X_1 = H) = P(X_2 = H) = 0.5$

Independent random variables - example

- Consider throwing an **unbiased coin** two times:

- Knowing that $X_1 = H$ will not change $P(X_2 = H \mid X_1 = H) = P(X_2 = H) = 0.5$

- On the other hand, in the previous dice example, the two dice throws were dependent on each other, i.e. $X \not\perp\!\!\!\perp Y$

Y \ X	1	2	3	4	5	6
6	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

Conditional independence - definition

- X is independent of Y **conditioned/given** Z iff

$$\forall x, y, z : P(X = x \,|\, Y = y, Z = z) = P(X = x \,|\, Z = z)$$

- We then write $X \perp\!\!\!\perp Y \,|\, Z$

Conditional independence - definition

- X is independent of Y **conditioned/given** Z iff

$$\forall x, y, z : P(X = x \mid Y = y, Z = z) = P(X = x \mid Z = z) \qquad (\text{for } P(Z = z) > 0)$$

- We then write $X \perp\!\!\!\perp Y \mid Z$, otherwise $X \not\perp\!\!\!\perp Y \mid Z$

- Intuitively this means that **Y does not add any information** to predict X that isn't already offered by Z

- Z can be a **set of variables,** e.g. $X \perp\!\!\!\perp Y \mid Z_1, Z_2$

Conditional Independence - Example 1

- Alice and Bob each independently throw a coin.
- They try to guess what the other person has thrown.
- Is that possible by just looking at their own coin?

slide by Patrick Forré

Conditional Independence - Example 1

- Alice and Bob each independently throw a coin.
- They try to guess what the other person has thrown.
- Is that possible by just looking at their own coin?

$A \perp\!\!\!\perp B$

slide by Patrick Forré

Conditional Independence - Example 1

- Alice and Bob each independently throw a coin.
- They try to guess what the other person has thrown.
- Is that possible by just looking at their own coin?

$A \perp\!\!\!\perp B$

- Eve just happens to see both coin outcomes and tells Bob: “I don’t tell you what Alice has, the outcomes are different anyways.”
- What about now?

slide by Patrick Forré

Conditional Independence - Example 1

- Alice and Bob each independently throw a coin.
- They try to guess what the other person has thrown.
- Is that possible by just looking at their own coin?

$A \perp\!\!\!\perp B$

- Eve just happens to see both coin outcomes and tells Bob: “I don’t tell you what Alice has, the outcomes are different anyways.”
- What about now?

$A \not\perp\!\!\!\perp B \mid E$

slide by Patrick Forré

Conditional Independence - Example 2

- Alice, Bob and Casper each throw a coin.
- Alice tries to guess the number of ‘tails’ T.
- Alice has a ‘head’, what can she tell?

slide by Patrick Forré

Conditional Independence - Example 2

- Alice, Bob and Casper each throw a coin.
- Alice tries to guess the number of ‘tails’ T.
- Alice has a ‘head’, what can she tell?

A $\not\perp$ T

slide by Patrick Forré

Conditional Independence - Example 2

- Alice, Bob and Casper each throw a coin.
- Alice tries to guess the number of ‘tails’ T.
- Alice has a ‘head’, what can she tell?

$A \not\perp T$

- Eve just saw that Alice and Bob have 2 ‘heads’ together.
- Can Alice now tell if there now 2 or 3 ‘tails’ in total?

slide by Patrick Forré

Conditional Independence - Example 2

- Alice, Bob and Casper each throw a coin.
- Alice tries to guess the number of ‘tails’ T.
- Alice has a ‘head’, what can she tell?

$A \not\perp\!\!\!\perp T$

- Eve just saw that Alice and Bob have 2 ‘heads’ together.
- Can Alice now tell if there now 2 or 3 ‘tails’ in total?

$A \perp\!\!\!\perp T \mid E$

slide by Patrick Forré

Expectation/Mean

- The expectation of a function f of a random variable $X \sim p$ is defined as:

$$\mathbb{E}[f] = \mathbb{E}_{x \sim p(X)}[f(x)] = \begin{cases} \sum_x f(x)p(x) & \textbf{Discrete} \\ \int_x f(x)p(x)dx & \textbf{Continuous} \end{cases}$$

Expectation/Mean

- The expectation of a function f of a random variable $X \sim p$ is defined as:

$$\mathbb{E}[f] = \mathbb{E}_{x \sim p(X)}[f(x)] = \begin{cases} \sum_x f(x)p(x) & \textbf{Discrete} \\ \int_x f(x)p(x)dx & \textbf{Continuous} \end{cases}$$

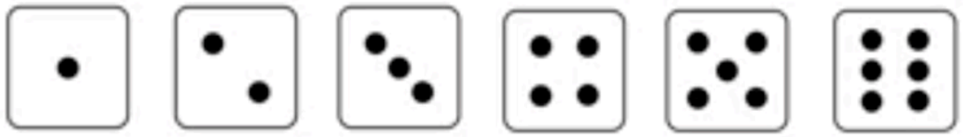
- Conditional expectation:

$$\mathbb{E}[f] = \mathbb{E}_{x \sim p(X|Y=y)}[f(x)] = \begin{cases} \sum_x f(x)p(x \mid Y = y) & \textbf{Discrete} \\ \int_x f(x)p(x \mid Y = y)dx & \textbf{Continuous} \end{cases}$$

Expectation example

- Consider a fair dice and f=identity

$$P(\{1\}) = \frac{1}{6}, \ldots, P(\{6\}) = \frac{1}{6},$$

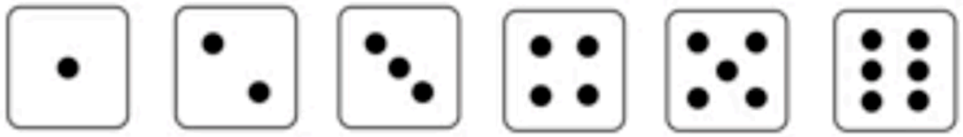


$$\mu_X = \mathbb{E}[x] = \sum_x xp(x) = \quad 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

Expectation example

- Consider a fair dice and f=identity

$$P(\{1\}) = \frac{1}{6}, \ldots, P(\{6\}) = \frac{1}{6},$$



$$\mu_X = \mathbb{E}[x] = \sum_x xp(x) = \quad 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

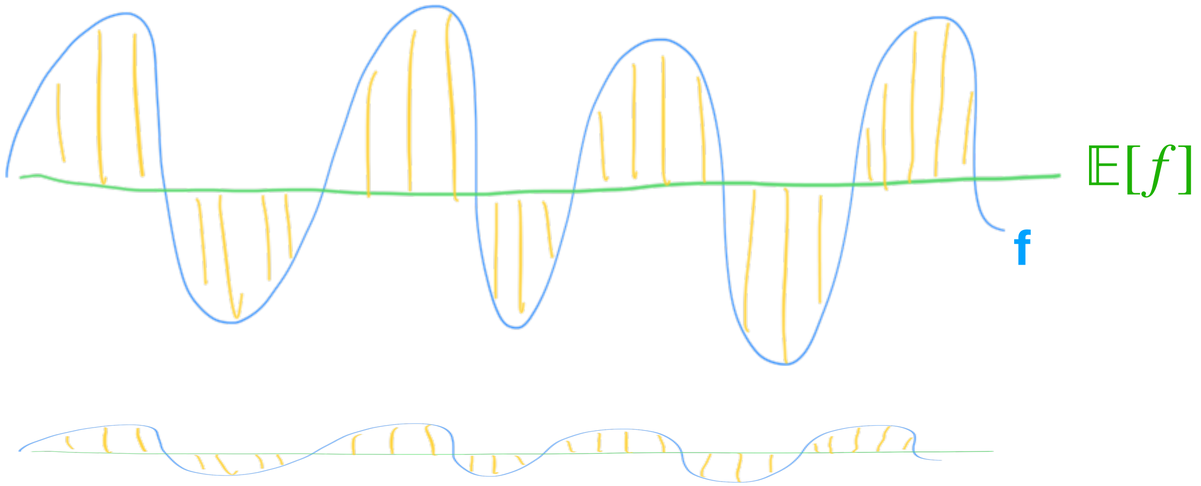
- Consider a fair dice and $f = X^2$

$$\mathbb{E}[f(x)] = \sum_x f(x)p(x) = \quad 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} = 15.17$$

Variance

- The variance is the expected quadratic distance between f and its expectation $\mathbb{E}[f]$

$$\sigma^2 = \mathrm{Var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f])^2]$$



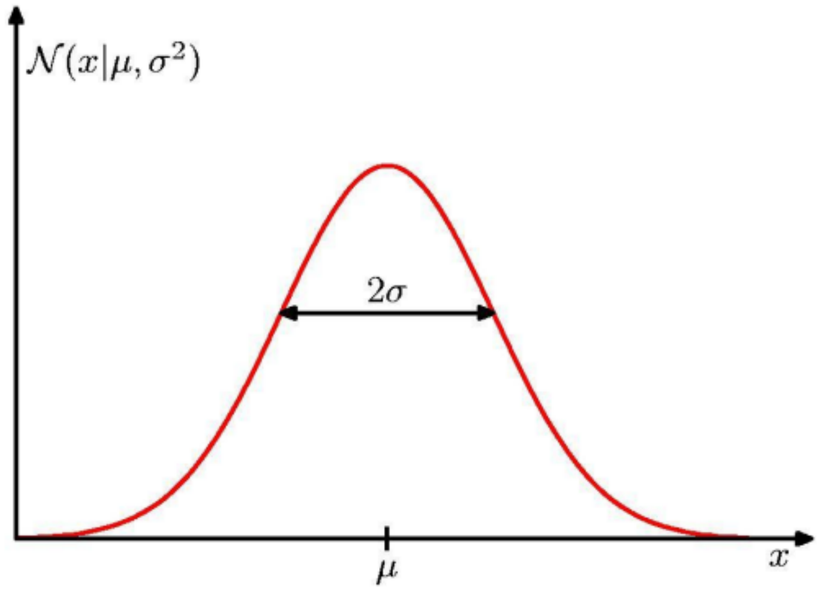
Example: Gaussian distribution

- The Normal (Gaussian) distribution has a squared exponential PDF

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mathbb{E}[x] = \mu$$

$$\mathrm{Var}[x] = \sigma_X^2$$



https://www.microsoft.com/en-us/research/wp-content/uploads/2016/05/prml-slides-1.pdf

Linear combination of Gaussians is a Gaussian

- The Normal (Gaussian) distribution has a squared exponential PDF

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- For two Gaussian variables A and B, their linear combination aA+bB is also a Gaussian with mean and variance:

$$\mathbb{E}[aA + bB] = a\mathbb{E}[A] + b\mathbb{E}[B]$$

$$\mathrm{Var}[aA + bB] = a^2\mathrm{Var}[A] + b^2\mathrm{Var}[B] \quad \text{if} \quad A \perp\!\!\!\perp B$$

Covariance and correlation

- The covariance measures the degree in which two variables vary together **linearly**

$$\sigma_{XY} = Cov[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Covariance and correlation

- The covariance measures the degree in which two variables vary together **linearly**

$$\sigma_{XY} = Cov[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

- This is often normalised to the **correlation coefficient:**

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

σ_X, σ_Y **are standard deviations (**$\sqrt{\mathrm{Var(X)}}, \sqrt{\mathrm{Var(Y)}}$**)**

- $\rho_{X,Y}$ is the slope of the best-fit line after we normalise X and Y

Correlation examples from previous class

- $\rho_{X,Y}$ is the slope of the best-fit line after we normalise X and Y

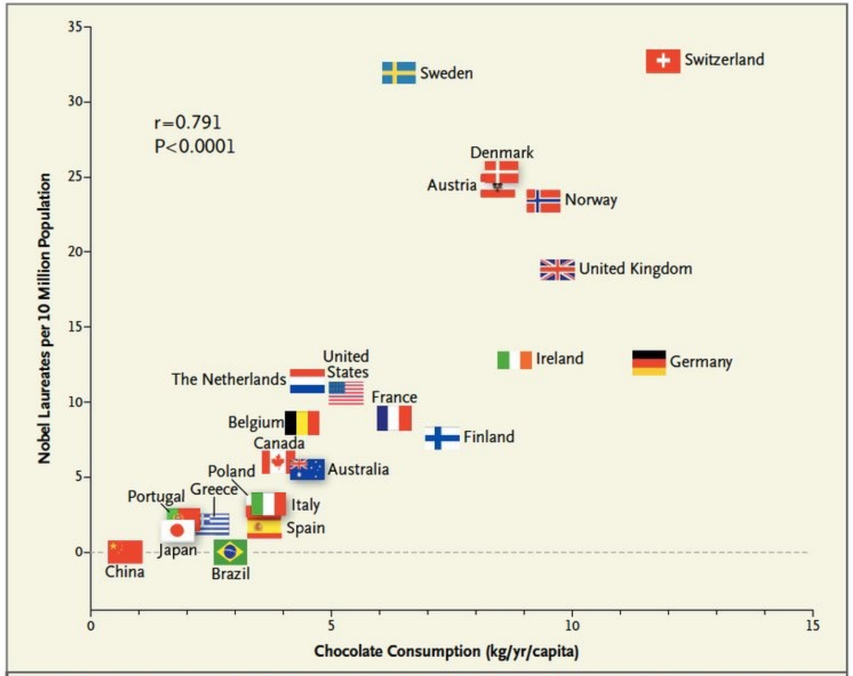
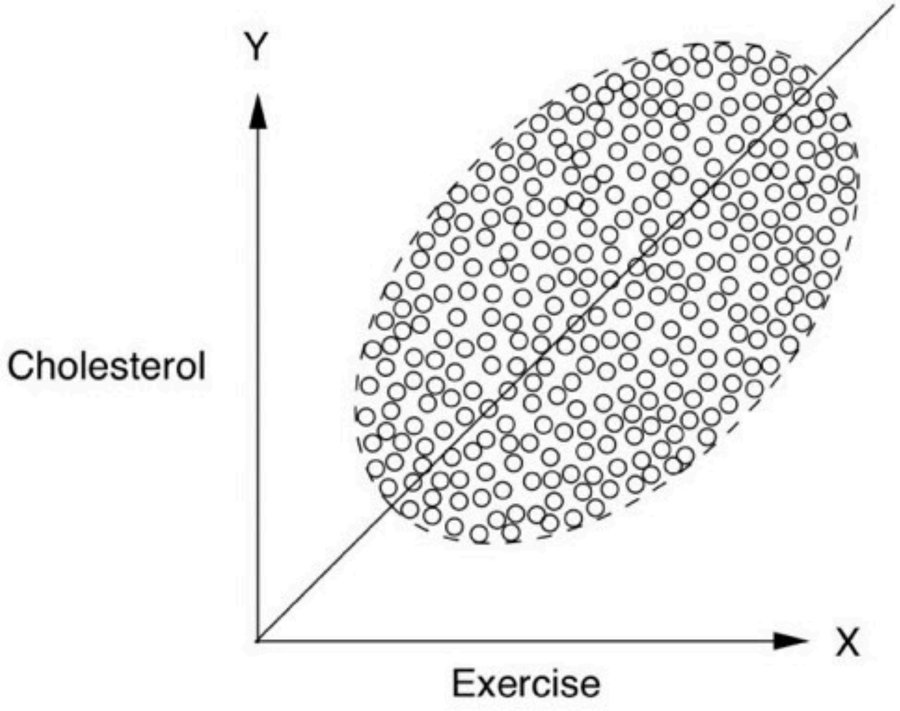


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

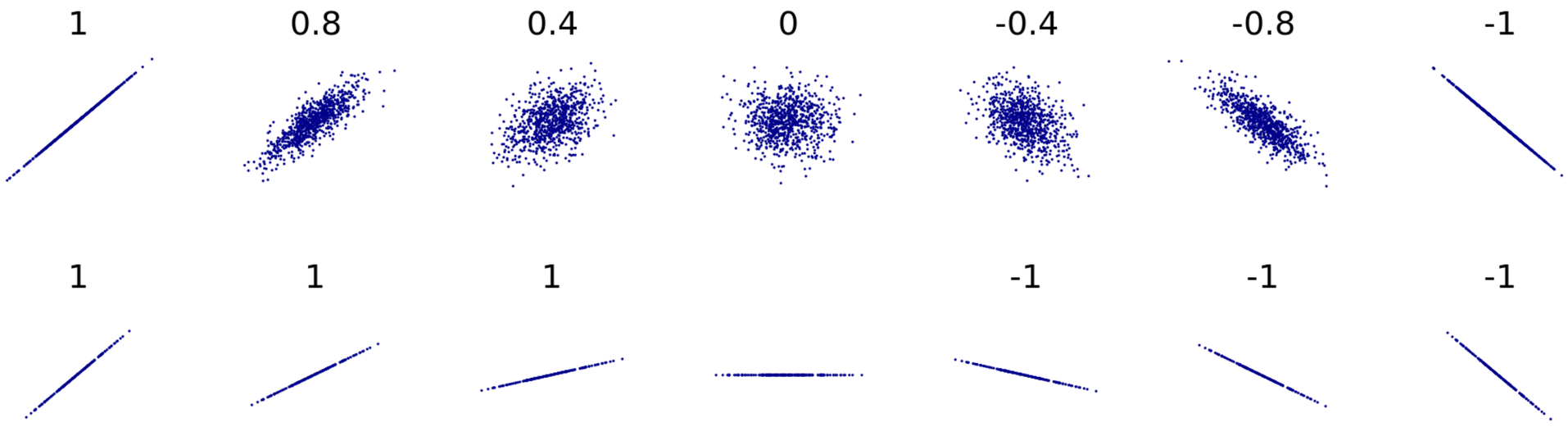


[Messerli, 2012] https://www.nejm.org/doi/full/10.1056/NEJMon1211064

From the Book of Why [Pearl 2018]

Correlation examples

- $\rho_{X,Y}$ is the slope of the best-fit line after we normalise X and Y



https://en.wikipedia.org/wiki/Correlation#/media/File:Correlation_examples2.svg

Uncorrelated but dependent

- When two variables are **independent**, the correlation coefficient is zero

$$X \perp\!\!\!\perp Y \implies \rho_{X,Y} = 0$$

- On the other hand, this does not mean that the inverse is true

https://en.wikipedia.org/wiki/Correlation#/media/File:Correlation_examples2.svg

Factorizing joint distributions

- A joint distribution can always be factorized in several ways by iterating the **chain rule**

$$P(X, Y) = P(X \mid Y)P(Y)$$

Factorizing joint distributions

- A joint distribution can always be factorized in several ways by iterating the **chain rule**

$$P(X, Y) = P(X \mid Y)P(Y)$$

- For example:
  - $P(X, Y, Z) = P(X \mid Y, Z)P(Y \mid Z)P(Z)$
  - $P(X, Y, Z) = P(Z \mid Y, X)P(Y \mid X)P(X)$
  - $P(X, Y, Z) = P(Y \mid Z, X)P(Z \mid X)P(X)$ ...

Using conditional independences to simplify the factorisation

- A joint distribution can always be factorized in several ways by iterating the **chain rule**

$$P(X, Y) = P(X \mid Y)P(Y)$$

- If we know some conditional independences, we can further simplify:

if $X \perp\!\!\!\perp Y \mid Z$:

$$P(X, Y, Z) = P(X \mid \cancel{Y}, Z)P(Y \mid Z)P(Z)$$

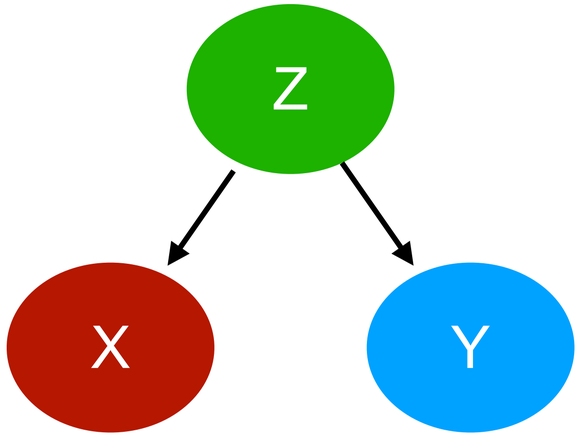
$$P(X, Y, Z) = P(X \mid Z)P(Y \mid Z)P(Z)$$

Next time: Probabilistic graphical models

- We can represent a factorisation of joint probability as a (directed acyclic) **graph**

$$P(X, Y, Z) = P(X|Z)P(Y|Z)P(Z)$$

- Each variable is represented by a **node**, **edges** indicate **dependences**

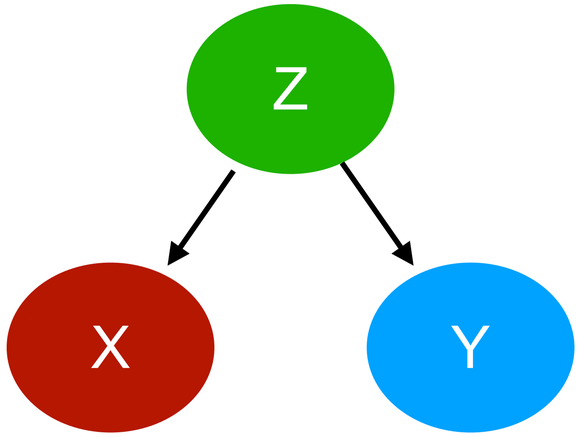


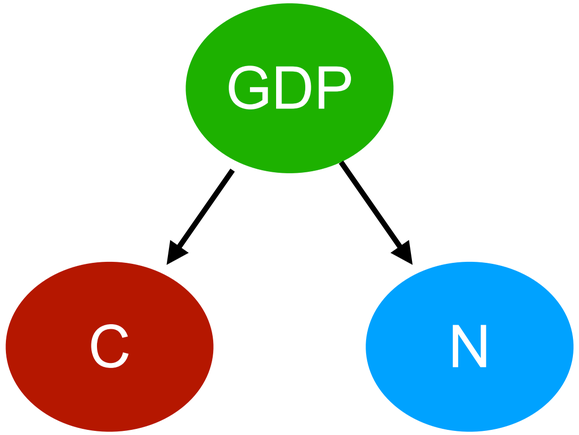
Next time: Probabilistic graphical models

- We can represent a factorisation of joint probability as a (directed acyclic) **graph**

$$P(X, Y, Z) = P(X|Z)P(Y|Z)P(Z)$$

- Each variable is represented by a **node**, **edges** indicate **dependences**





Questions??