# ADVANCED INFORMATION, COMPUTATION, COMMUNICATION II

Prof. M. Gastpar

Slides by Prof. M. Gastpar and Prof. em. B. Rimoldi

EPFL

Spring Semester 2025— *Slides Version 1.0*

# Outline

AICC-I

- **Computation**
- Algorithms
- Discrete Structures

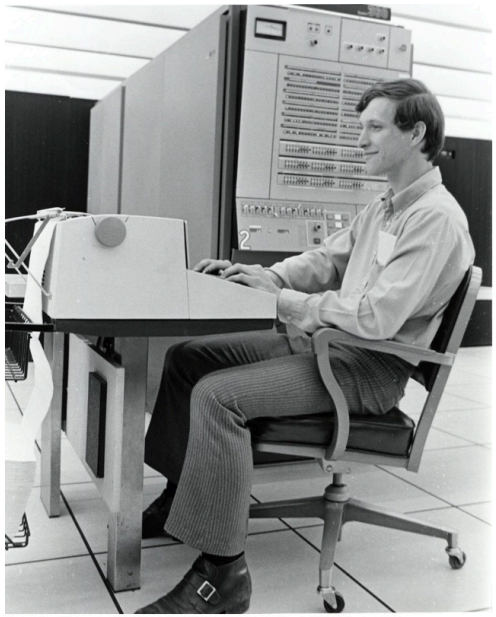
AICC-I

- **Computation**
- Algorithms
- Discrete Structures

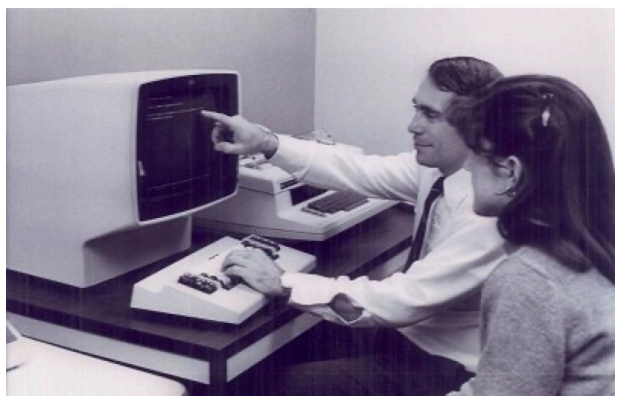
But to have interesting computations, we need data!

AICC-I

- **Computation**
- Algorithms
- Discrete Structures

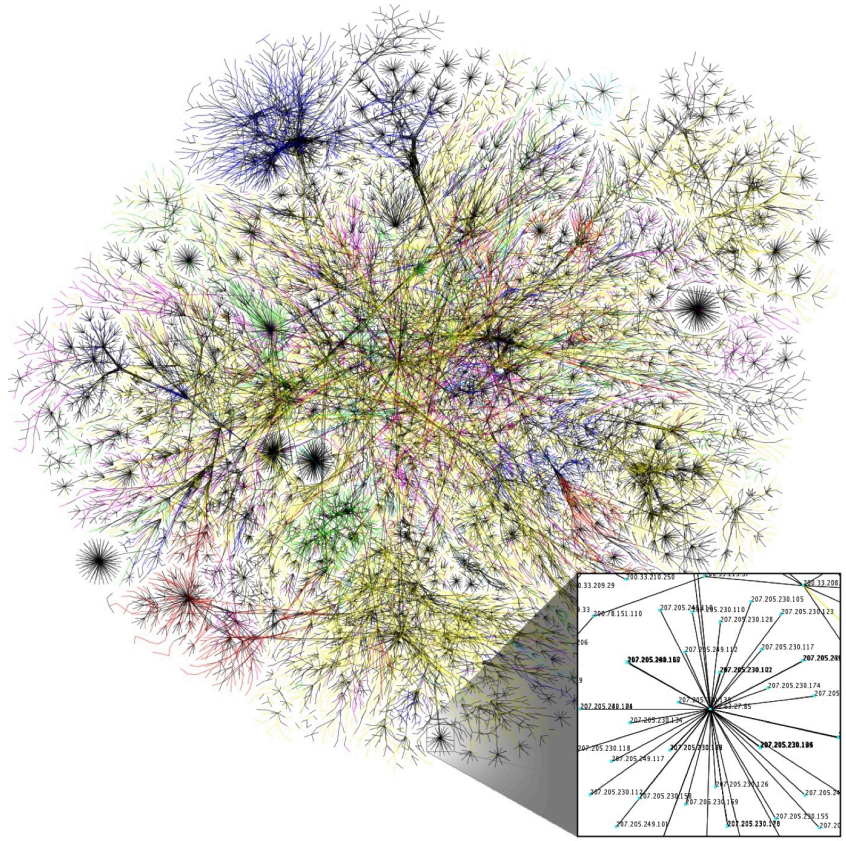


ca. 1980



AICC-I

- **Computation**
- Algorithms
- Discrete Structures



From The Opte Project

AICC-I

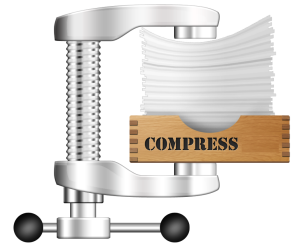
- **Computation**
- Algorithms
- Discrete Structures

AICC-II

- **Communication**
- Information and Data Science
- Cryptography, Secrecy, Privacy

# In this course: three main topics

- **Source Coding:** It is about **compressing information.**




- **Cryptography:** It is about **protecting** the information against undesirable **human** activities: how to provide message **integrity** and **confidentiality**.

- **Channel Coding:** It is about **protecting** the information from **natural** damages.

All three pertain to information **storage/communication**.

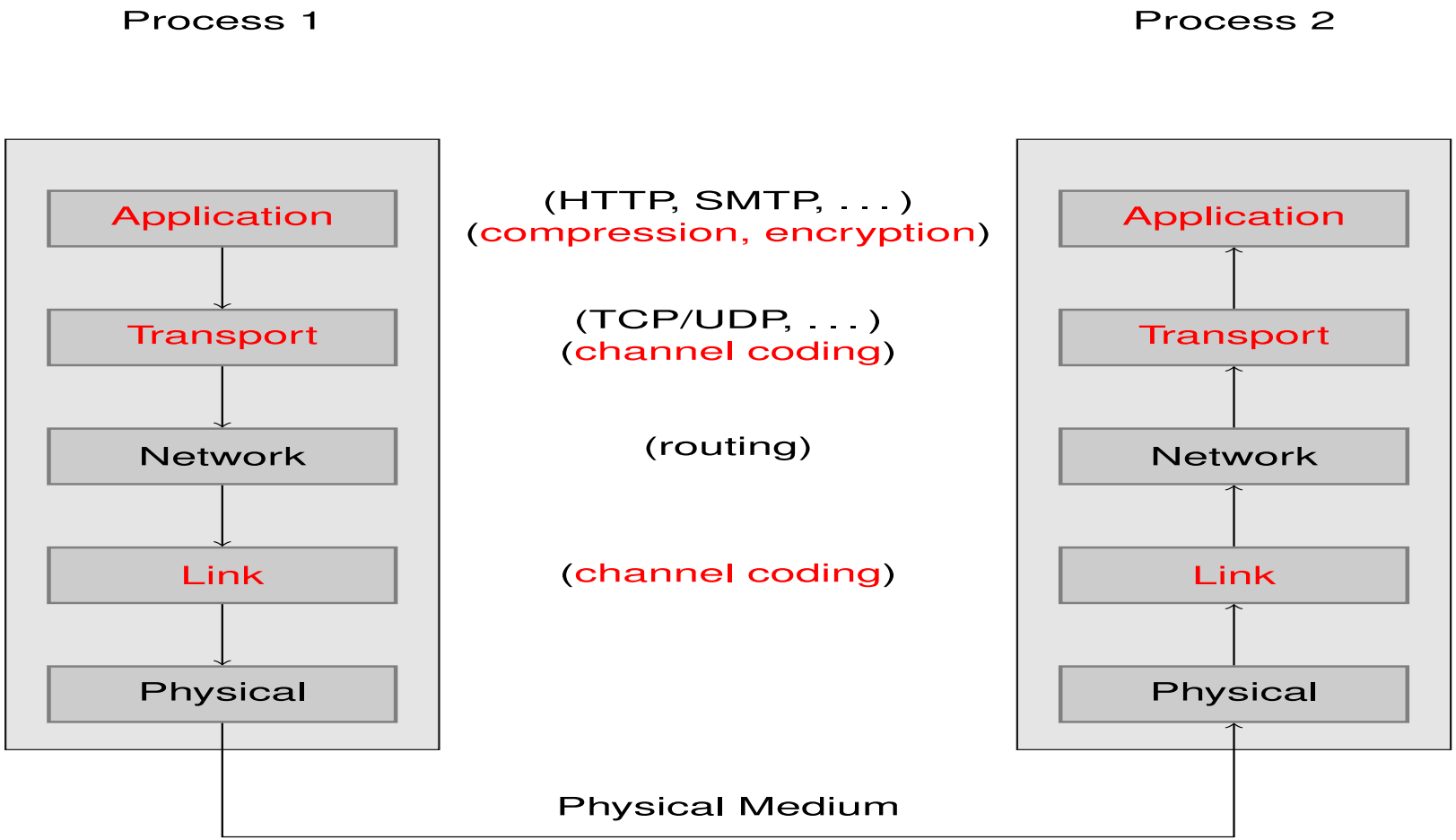
# WE STUDY: SOURCE CODING, CRYPTOGRAPHY, CHANNEL CODING

Why these topics?

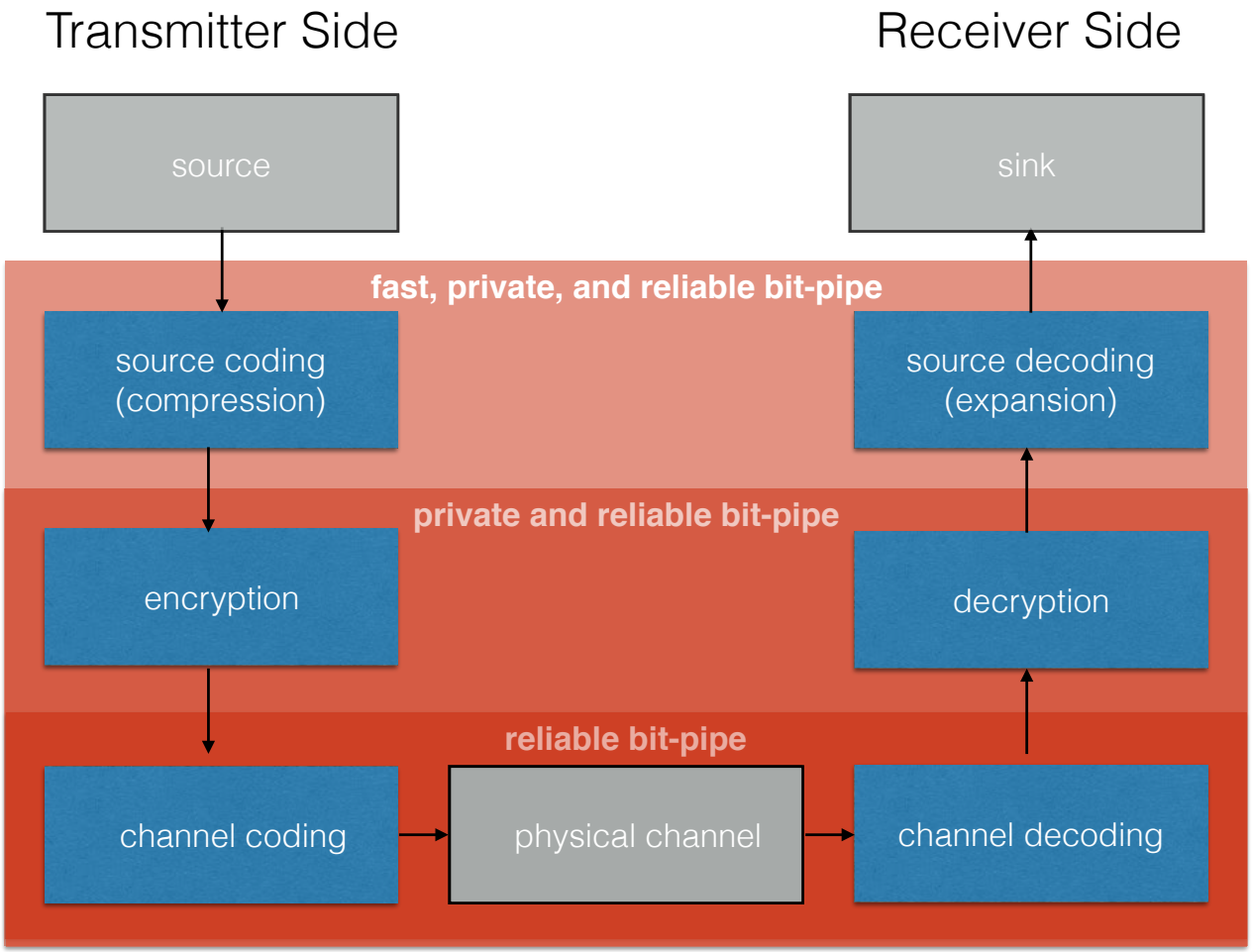
- important building blocks of communication systems
- non-evident topics and the results are often surprising
- intimately related to fundamental concepts (probability theory, linear algebra, number theory)
- have a common root: the notion of entropy
- require/promote rigorous thinking

# Digital Communication: The "Big Picture"

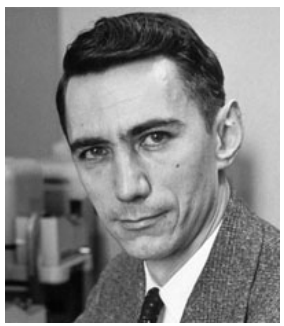
# Communication over the Internet



# POINT-TO-POINT COMMUNICATION SYSTEM

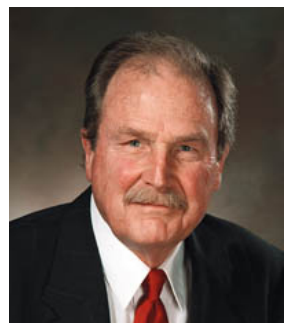


We will rely on **discrete probability theory** and on the work of various people including:



Shannon

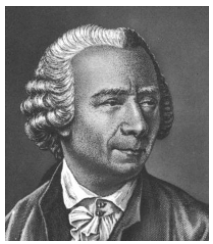
Fano



Huffman

# Second Topic: Cryptography

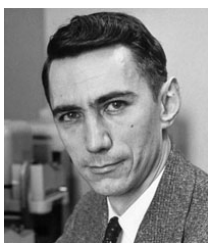
We will rely on **number theory**



Euler

Fermat

as well as on **group theory** and on the work of various people including:



Shannon

Clifford Cocks

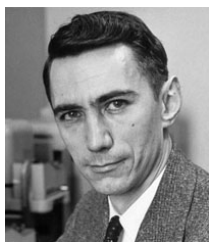
Rivest, Shamir, Adleman

# THIRD TOPIC: CHANNEL CODING

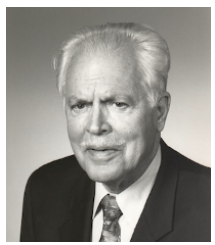
We will rely on **finite fields**

Galois

as well as on **linear algebra** and on the work of various people including:



Shannon



Reed

Solomon

# Course Organization

# Outline

# Teaching Crew

- Professor:
  Michael Gastpar
- Senior Teaching Assistants:
  Adrien Vandenbroucque, Millen Kanabar, Yunzhen Yao
- Student TAs:

| | | |
|---|---|---|
| Roxanne Chevalley | Ait Lalim Adrien | Mehdi Zoghlami |
| Michaël Brasey | Yuki Crivelli | Valerio de Santis |
| Théo Hollender | Gersende Kerjan | Simon Lefort |
| Mattia Metzler | Emmanuel Omont | Laura Paraboschi |
| | Anthony Tamberg | |

# SCHEDULE

- Tuesdays 15:15 - 17:00
  Lecture
  RLC E1 240

- Wednesdays 13:15 - 15:00
  Lecture
  RLC E1 240

- Wednesdays 15:15 - 17:00
  Exercises
  Various rooms, see Moodle

# Grading Formula

- **90%** Final exam during exam period.

  *Note:* No documents or electronic devices allowed during the exam.
- **10%** Quizzes (on-line on Moodle).
  - There will be 6 Quizzes. Only the best 5 count.
  - The Quiz questions are very similar to the final exam questions in style and difficulty.
  - On the Quizzes, you can update your answer as many times as you want before the deadline.
  - However, once the deadline is passed, you can no longer change your answers.

- There is also a weekly homework set:
  - The Quizzes are highly correlated with the homework.
  - If you did not do the homework, you **should not expect to be able to do the Quizzes!**
  - We do not grade the homework.

# How To Be Efficient And Do Well In This Course

Before class (stay ahead):

- browse through the slides to know what to expect
- review the background material as needed

After class:

- read the notes: they are the reference
- do the review questions

Before the exercise session:

- are you up-to-date with the theory?
- solve what you can ahead of time and finish during the exercise session
- write down YOUR own solution

# Course Web Site

- moodle.epfl.ch > Informatique (IN) > Bachelor > COM-102 Advanced information, computation, communication II
  (Password protected if not registered to AICC-II)

- There you'll find:
  - Lecture slides
  - Link to videos
  - Homework assignments
  - Solutions
  - Quizzes
  - Forums (news and questions/answers)

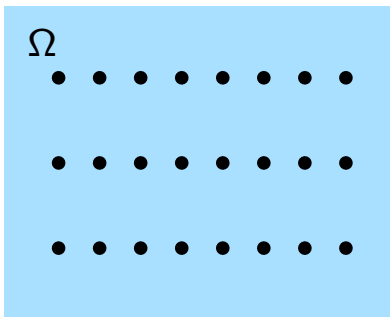
# Outline

# Review of Discrete Probability:
# (Book Chapter 0)

# Outline

# Initial case: Finite $\Omega$ with Equally Likely Outcomes

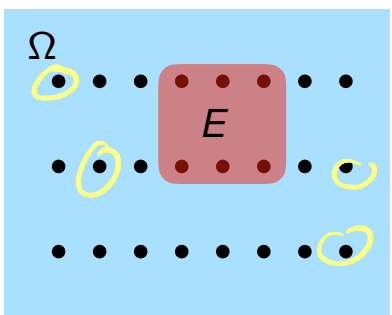
**Sample space $\Omega$:** set of all possible outcomes.




$$\Omega = \{\omega_1, \ldots, \omega_n\}$$

**Event $E$:** subset of $\Omega$. Since the outcomes are equally likely,

$$p(E) = \frac{|E|}{|\Omega|}.$$




$$p(E) = \tfrac{6}{24}$$

## EXAMPLE : TOSSING A FAIR COIN.

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

$$p(H) = 1/2$$

$$p(T) = 1/2$$

EXAMPLE: ROLL A FAIR DIE.

$\Omega = \{$  $\}$

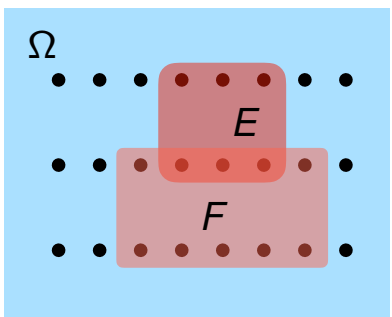
$$p(\omega) = 1/6 \quad , \forall \omega \in \Omega$$

# Conditional Probability

ALL $\omega \in \Omega$ ARE EQUALLY LIKELY.

The **conditional probability** $p(E|F)$ is the probability that $E$ occurs, given that $F$ has occurred (hence assuming that $|F| \neq \emptyset$):

$$p(E|F) = \frac{|E \cap F|}{|F|}.$$




$$p(E|F) = \tfrac{3}{10}$$

We may think of $F$ as a new sample space.

TOSS TWO FAIR COINS.

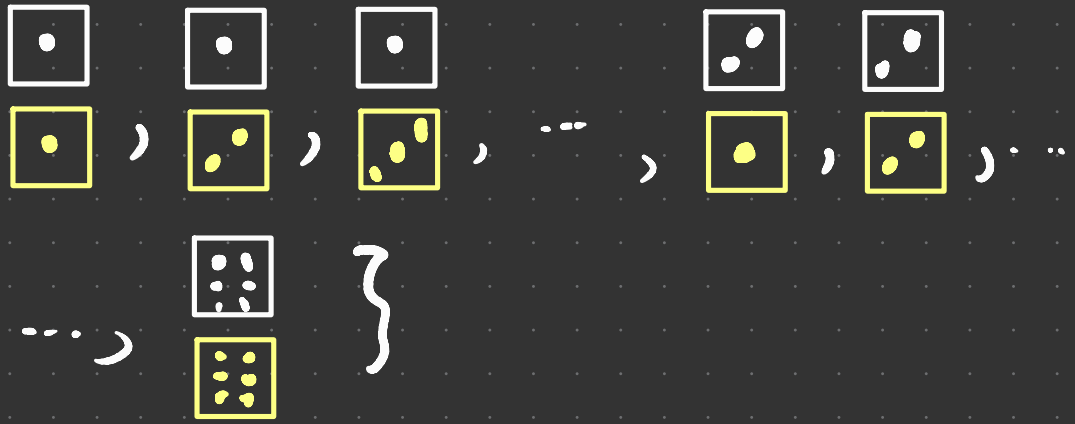
- ONE COMES UP HEADS.
- WHAT IS THE PROBABILITY THAT THE OTHER COIN IS ALSO HEADS?

$\Omega = \{HH, HT, TH, TT\}$

$F = \{HH, HT, TH\}$

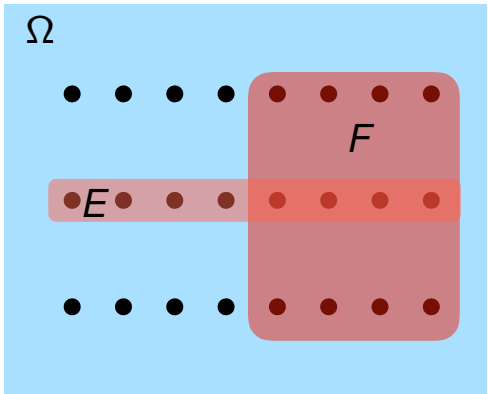
$E = \{HH\}$

EXAMPLE: ROLL TWO FAIR DICE:

$\Omega = \{$  $\}$

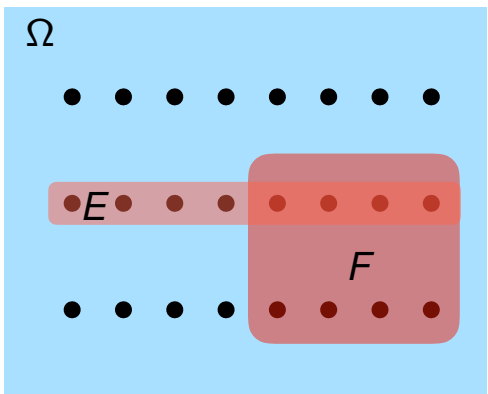
# Independent Events

Events $E$ and $F$ are called **independent** if $p(E|F) = p(E)$.




$$p(E|F) = \frac{1}{3} = p(E)$$

$E$ and $F$ are independent

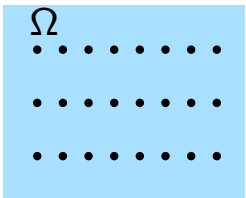



$$p(E|F) = \frac{1}{2} \neq p(E) = \frac{1}{3}$$

$E$ and $F$ are NOT independent

# General Case: Finite $\Omega$, arbitrary $p(\omega)$

**Sample space $\Omega$:** set of all possible outcomes.




$$\Omega = \{\omega_1, \ldots, \omega_n\}$$

**Probability distribution (probability mass function) $p$:**
A function $p : \Omega \to [0, 1]$ such that

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

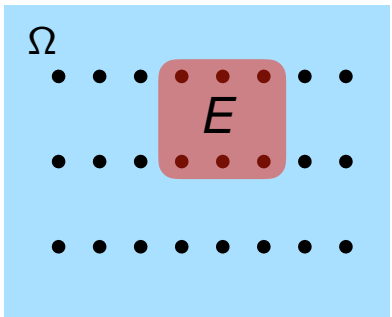
## EXAMPLE : TOSSING A BIASED COIN.

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

$$P(H) = 2/3$$

$$P(T) = 1/3$$

**Event $E$:** a subset of $\Omega$.



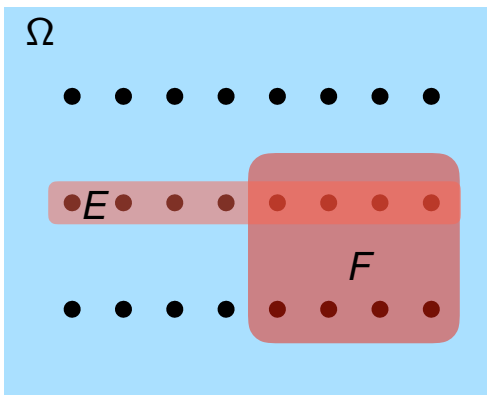

The domain of the probability mass function $p$ is extended to the power set of $\Omega$:

$$p(E) = \sum_{\omega \in E} p(\omega).$$

The general form for the conditional probability is

$$p(E|F) = \frac{p(E \cap F)}{p(F)}$$

for $F$ such that $p(F) \neq 0$.

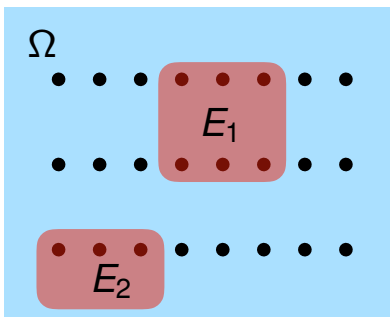



**Independent Events.** Exactly as before, events $E$ and $F$ are called independent if $p(E|F) = p(E)$. Equivalently, $E$ and $F$ are independent if $p(E \cap F) = p(E)p(F)$.

**Disjoint Events:**

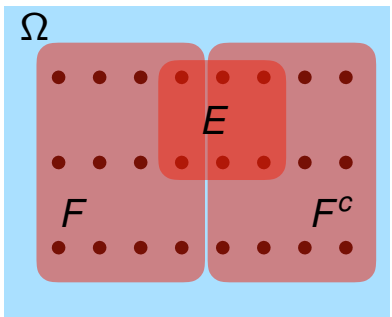
If $E_1$ and $E_2$ are disjoint events then

$$p(E_1 \cup E_2) = p(E_1) + p(E_2).$$



**Law of Total Probability:**

For any $F \subseteq \Omega$ and its complement $F^c$,



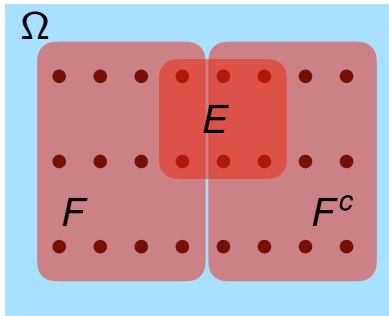

$$p(E) = p(E|F)p(F) + p(E|F^c)p(F^c).$$

More generally, if $\Omega$ is the union of disjoint events $F_1, F_2, \ldots, F_n$,

$$p(E) = p(E|F_1)p(F_1) + p(E|F_2)p(F_2) + \cdots + p(E|F_n)p(F_n).$$

(Divide and conquer.)

**Proof:** We prove the law of total probability for $\Omega = F \cup F^c$. (The general case follows straightforwardly.)




$$p(E) = p(( \underbrace{E \cap F) \cup (E \cap F^c)}_{\text{union of disjoint sets}} )$$

$$= p(E \cap F) + p(E \cap F^c)$$

$$= \frac{p(E \cap F)}{p(F)} p(F) + \frac{p(E \cap F^c)}{p(F^c)} p(F^c)$$

$$= p(E|F)p(F) + p(E|F^c)p(F^c).$$

□

## EXERCISE

**Example of Total Probability:** Two factories supply light bulbs.

- Factory $F_1$'s bulbs work for over 5000 hours in 99% of cases;
- Factory $F_2$'s bulbs work for over 5000 hours in 95% of cases.
- It is known that factory $F_1$ supplies 60% of the total bulbs.

What is the chance that a bulb chosen at random works for longer than 5000 hours?

## Solution

***Answer:***

*$\Omega$ is the space of all bulbs.*

*(Optional: to picture the partitioning of $\Omega$ into subsets, you may want to imagine each bulb being labeled by the factory's name and the number of hours that it works.)*

*Let $E \subseteq \Omega$ be the set that consists of all bulbs that work for longer than 5000 hours and let $F_i \subseteq \Omega$ be the set of bulbs from factory $i = 1, 2$.*

- $p(E|F_1) = .99$
- $p(E|F_2) = .95$
- $p(F_1) = .6$

$$p(E) = p(E|F_1)p(F_1) + p(E|F_2)p(F_2) = \frac{99}{100} \times \frac{6}{10} + \frac{95}{100} \times \frac{4}{10} = \frac{974}{1000}.$$

Sometimes we are given $p(E)$, $p(F)$ and $p(E|F)$, and we need $p(F|E)$.

In this case we use **Bayes' Rule:**

$$p(F|E) = \frac{p(E|F)p(F)}{p(E)}.$$

**Proof:** We use the definition of conditional probability to write $p(E \cap F)$ two ways and solve for $p(F|E)$:

$$p(F|E)p(E) = p(E \cap F) = p(E|F)p(F).$$

□

**Example:**

Let $\Omega$ be a population of drivers (e.g. of Switzerland, on New Year's eve).

Let $A$ be the event that a driver has an accident.

Let $D$ be the event that a driver is drunk.

From observations, the police knows $p(A)$, $p(D)$ as well as $p(D|A)$.

$p(A|D)$ cannot be easily obtained from observations. Yet, knowing it might discourage a drunk person to drive.

Let us be concrete (numbers are made up):

$$p(A) = 10^{-6},$$
$$p(D) = 0.1,$$
$$p(D|A) = 0.8.$$

Now

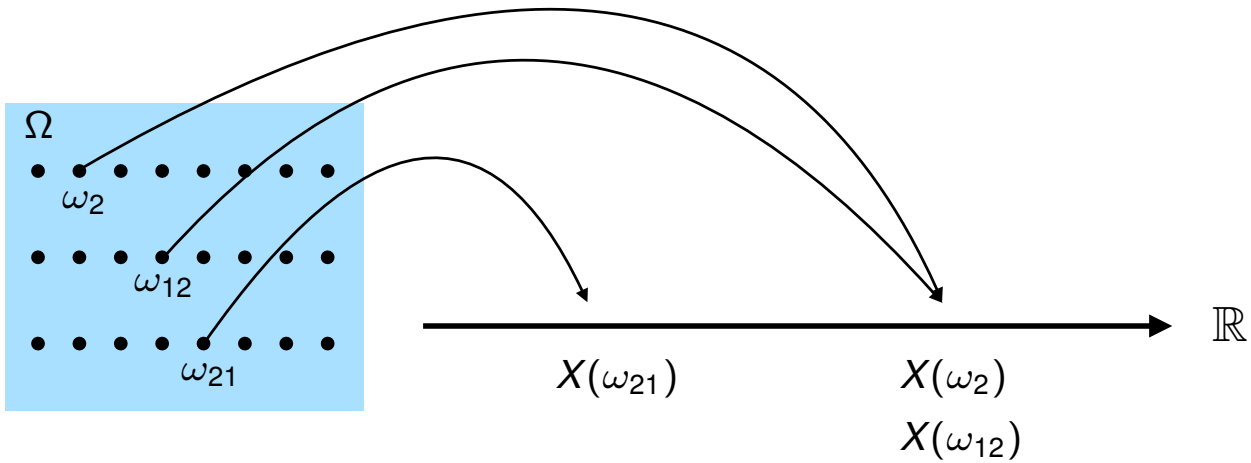
$$p(A|D) = \frac{p(D|A)p(A)}{p(D)} = \frac{0.8 \times 10^{-6}}{0.1} = 8 \times 10^{-6}.$$

We can also compute

$$p(A|D^c) = \frac{p(D^c|A)p(A)}{p(D^c)} = \frac{(1 - 0.8) \times 10^{-6}}{(1 - 0.1)} = \frac{2}{9} \times 10^{-6}.$$

Notice that, in this case, $\frac{p(A|D)}{p(A|D^c)} = 36$.

**Random Variable** $X$: A function $X : \Omega \to \mathbb{R}$.



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

$$X(\omega) = \begin{cases} 1, & \omega = H \\ -1, & \omega = T \end{cases}$$

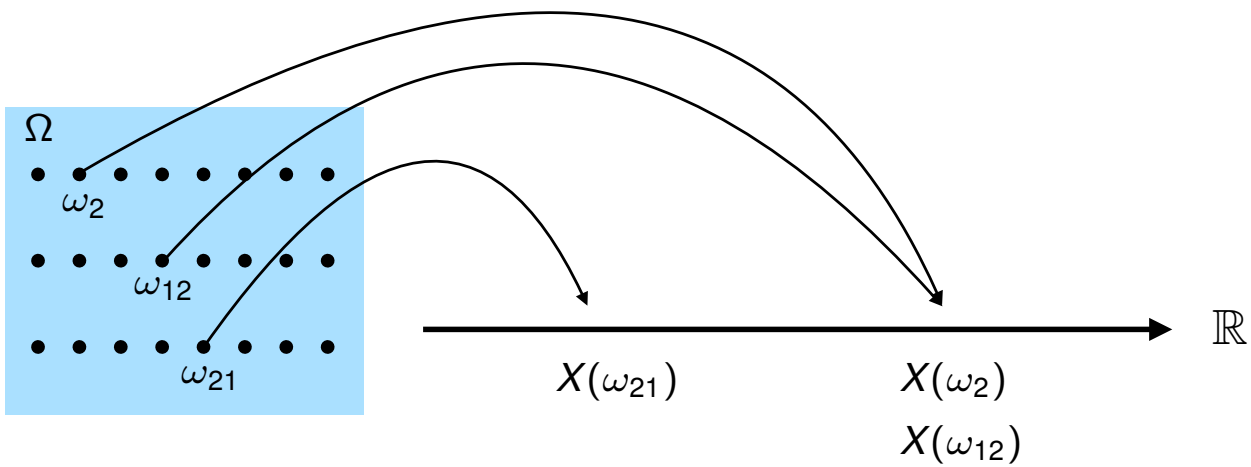
$$Y(\omega) = \begin{cases} 1, & \omega = H \\ 0, & \omega = T \end{cases}$$

**Probability distribution** $p_X$: $p_X(x)$ is the probability that $X = x$, i.e. the probability of the event

$$E_x = \{\omega \in \Omega : X(\omega) = x\}.$$

Hence,

$$p_X(x) = \sum_{\omega \in E_x} p(\omega).$$



## Example (Lucky Dice)

You roll a dice.

If the outcome is 6, you receive 10 CHF. Otherwise, you pay 1 CHF.

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

$$\text{For each } \omega, p(\omega) = 1/6.$$

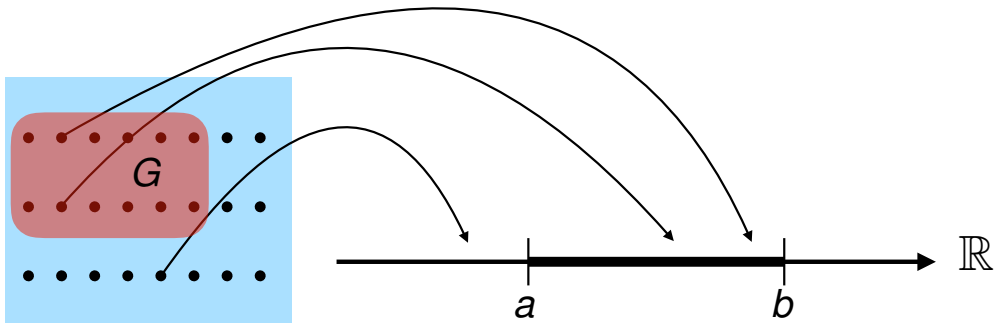
Then, define:

$$X(\omega) = \begin{cases} 10, & \omega = 6 \\ -1, & \omega \in \{1, 2, 3, 4, 5\}. \end{cases}$$

Hence, we have

$$p_X(x) = \begin{cases} \frac{1}{6}, & x = 10 \\ \frac{5}{6}, & x = -1. \end{cases}$$

How about the probability that $X \in [a, b]$?

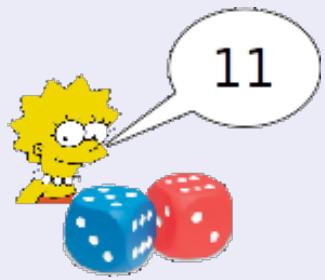



We can compute it two ways:

- using $p$: $\sum_{\omega \in G} p(\omega)$.
- using $p_X$: $\sum_{x \in [a,b]} p_X(x)$.

## Exercise (Lisa rolls two dice)

- Lisa rolls two dice and announces the sum $L$ written as a two digit number.
- The alphabet of $L = L_1 L_2$ is $\{02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12\}$.
  - The alphabet of $L_1$ is $\{0, 1\}$.
  - The alphabet of $L_2$ is $\{0, 1, \ldots, 9\}$.
- Determine $p_L$.



## Solution

$\Omega = \Big\{(i,j) : i,j \in \{1,\ldots,6\}\Big\}$.

$L : \Omega \to \mathbb{R}$ defined by $L\big((i,j)\big) = i + j$ written as a two-digit number.

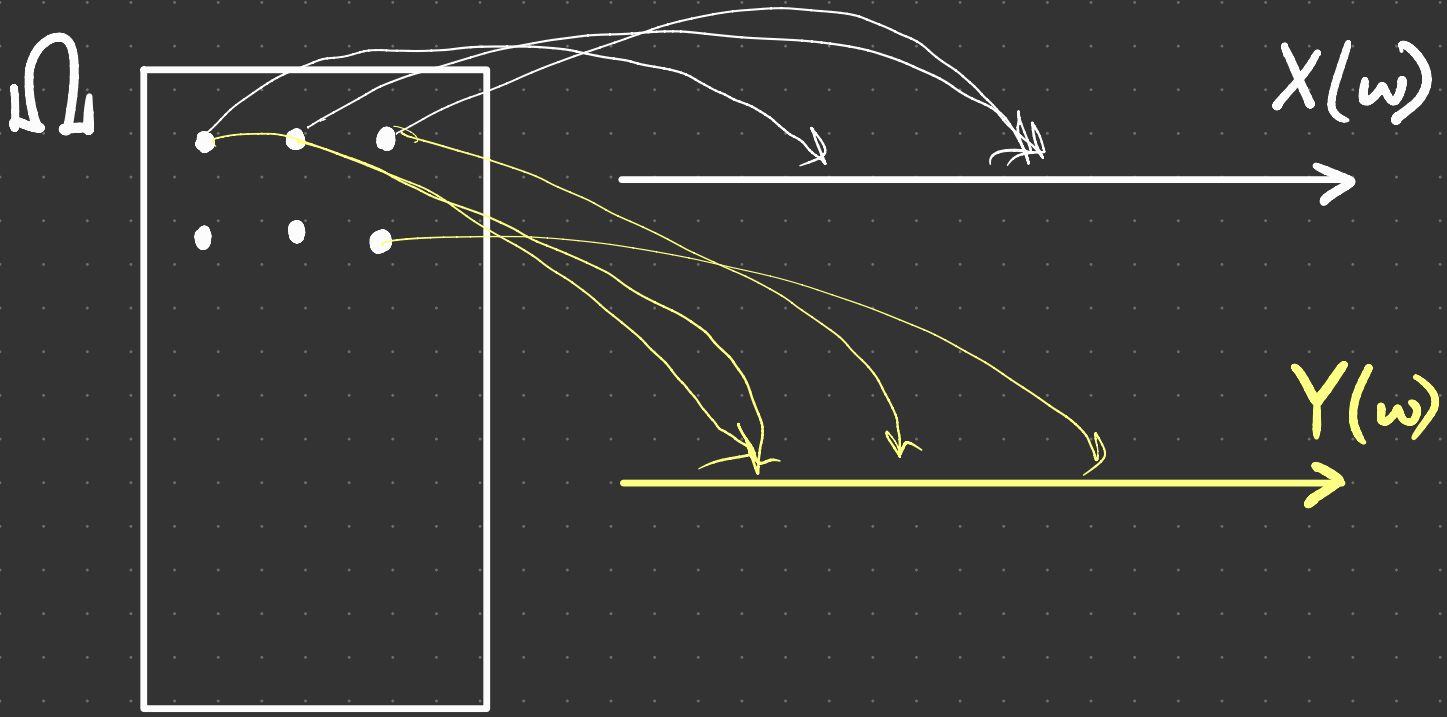
$L = 02$ iff $\omega = (1,1)$, which has probability $\frac{1}{36}$.

$L = 03$ iff $\omega \in \{(1,2)\} \cup \{(2,1)\}$. The events $\{(2,1)\}$ and $\{(1,2)\}$ are disjoint, with probability $\frac{1}{36}$ each. Hence $L = 03$ with probability $\frac{2}{36}$.

Etc.

| $L$ | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $p_L$ | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |

# TWO RANDOM VARIABLES



**Two Random Variables**:

Let $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ be two random variables.

The probability of the event $E_{(x,y)} = \{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\}$ is

$$p_{X,Y}(x, y) = \sum_{\omega \in E_{(x,y)}} p(\omega).$$

JOINT DISTRIBUTION OF X & Y.

We can compute $p_X$ from $p_{X,Y}$:

$$p_X(x) = \sum_y p_{X,Y}(x, y).$$

$p_X$ is called **marginal distribution** (of $p_{X,Y}(x, y)$ with respect to $x$).

$p_Y$ can be computed similarly.

# TWO RANDOM VARIABLES

SNOW IN { ZERMATT $X \in \{0, 1\}$ ; NENDAZ $Y \in \{0, 1\}$

| Y \ X | 0 | 1 |
|---|---|---|
| 0 | 0.4 | 0.2 |
| 1 | 0.3 | 0.1 |

$P(Y)$

$0.6 = p(Y=0)$

$0.4 = p(Y=1)$

$p(X) = (0.7 \quad 0.3)$

## EXERCISE

Determine the probability $p_{L_1}$, knowing $p_L$, where $L = L_1 L_2$.

| $L$ | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $p_L$ | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |

| $L_1$ | 0 | 1 |
|---|---|---|
| $p_{L_1}$ | | |

## SOLUTION

| $L$ | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $p_L$ | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |

We marginalize:

$$p_{L_1}(1) = \sum_x p_{L_1,L_2}(1,x) = \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{6}{36} = \frac{1}{6}.$$

Hence

| $L_1$ | 0 | 1 |
|---|---|---|
| $p_{L_1}$ | $\frac{5}{6}$ | $\frac{1}{6}$ |

The **Expected Value** $\mathbb{E}[X]$ of a random variable $X : \Omega \to \mathbb{R}$ is

$$\begin{aligned}\mathbb{E}[X] &= \sum_{\omega} X(\omega)p(\omega) \\ &= \sum_{x} xp_X(x).\end{aligned}$$

To see that these two expressions are indeed equal, we reorganize the sum:

$$\sum_{\omega} X(\omega)p(\omega) = \sum_{x} \sum_{\omega:X(\omega)=x} X(\omega)p(\omega) = \sum_{x} x \sum_{\omega:X(\omega)=x} p(\omega) = \sum_{x} xp_X(x).$$

## Exercise Lucky Dice

You roll a dice.

If the outcome is 6, you receive 10 CHF. Otherwise, you pay 1 CHF.

What is your expected gain or loss?

## Solution

*Recall:* $\Omega = \{1, 2, 3, 4, 5, 6\}$ *and for each* $\omega$, $p(\omega) = 1/6$.

$$X(\omega) = \begin{cases} 10, & \omega = 6, \\ -1, & \omega \in \{1, 2, 3, 4, 5\}. \end{cases}$$

*Then,*

$$\mathbb{E}[X] = \sum_{\omega} X(\omega)p(\omega) = \frac{1}{6}(-1) + \frac{1}{6}(-1) + \frac{1}{6}(-1) + \frac{1}{6}(-1) + \frac{1}{6}(-1) + \frac{1}{6} \cdot 10$$

$$= \sum_{x} x p_X(x) = \frac{5}{6}(-1) + \frac{1}{6} \cdot 10$$

Expectation is a linear operation in the following sense:

Let $X_1, X_2, \ldots, X_n$ be random variables and $\alpha_1, \alpha_2, \ldots, \alpha_n$ be scalars. Then

$$\mathbb{E}\Big[\sum_{i=1}^{n} X_i \alpha_i\Big] = \sum_{i=1}^{n} \alpha_i \mathbb{E}[X_i].$$

(See e.g. Rosen.)

# INDEPENDENCE

Recall that two events $E$ and $F$ are independent iff

$$p(E|F) = p(E)$$

or, equivalently, iff

$$p(E \cap F) = p(E)p(F).$$

# DEFINITION:

Two random variables $X$ and $Y$ are independent iff, for all realizations $x$ and $y$,

$$p(\{X = x\} \cap \{Y = y\}) = p(\{X = x\})p(\{Y = y\}),$$

or, more concisely, iff

$$p_{X,Y}(x, y) = p_X(x)p_Y(y).$$

Generalization to $n$ random variables is straightforward: $X_1, \ldots, X_n$ are independent iff

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i).$$

The conditional distribution of $Y$ given $X$ is the function $p_{Y|X}$ defined by

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}.$$

It is defined for all $x$ such that $p_X(x) \neq 0$.

The following statements are equivalent to the statement that $X$ and $Y$ are independent random variables:

- $p_{X,Y} = p_X p_Y$;
- $p_{Y|X}(y|x) = p_Y(y)$ (for all $x$ for which it is defined and for all $y$);
- $p_{Y|X}(y|x)$ is not a function of $x$;
- $p_{X|Y}(x|y) = p_X(x)$ (for all $y$ for which it is defined and for all $x$);
- $p_{X|Y}(x|y)$ is not a function of $y$.

## EXERCISE

Let $L$ be the random variable modeling Lisa's experiment.

Let $L_1$ and $L_2$ be the first and the second digit of $L$, respectively.

Are $L_1$ and $L_2$ independent ?

Hint: Compute $p_L(13)$.

FOR INDEP, WE WOULD NEED

$$p_L(13) = p_{L_1}(1)\, p_{L_2}(3)$$

## SOLUTION

$$p_{L_1}(1) = \frac{1}{6} \text{ (found earlier)}$$

$$p_{L_2}(3) = \frac{2}{36} \text{ (see table below)}$$

$$p_L(13) = 0 \neq p_{L_1}(1)p_{L_2}(3)$$

Hence $L_1$ and $L_2$ are NOT independent.

| $L$ | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $p_L$ | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |

| $L_1$ | 0 | 1 |
|---|---|---|
| $p_{L_1}$ | $\frac{5}{6}$ | $\frac{1}{6}$ |

| $L_2$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| $p_{L_2}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ |

$$\mathbb{E}[X_1 + Y_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2$$

The expected value of a product is NOT always the product of the expected values.

Example: $X = Y \in \{-1, 1\}$, uniformly distributed.

$$\mathbb{E}[XY] = 1$$
$$\mathbb{E}[X]\mathbb{E}[Y] = 0$$

However, if $X$ and $Y$ are independent random variables, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

(See e.g. Rosen.)

# Summary — Probability Review

- Random variable
- Probability distribution
  - Joint distribution of multiple random variables.
  - Marginal distribution.
  - Conditional distribution.
- Independence

# Sources and Entropy
# (Book Chapter 1)

# Outline

# LOGARITHM

IF $b^x = y$ THEN $x = \log_b y$

(1) $\log_b (1) = 0$

(2) $\log_b (xy) = \log_b (x) + \log_b (y)$

(3) $\log_b \left(\frac{1}{x}\right) = -\log_b (x)$

(4) $\log_b (x) = \log_b (a) \log_a (x)$

How do we communicate in the digital world?

We communicate by revealing the value of a sequence of variables that we call **(information) symbols**.

The $i$-th symbol might represent

- the intensity of the $i$th pixel of a black/white digital photo
- your score in your $i$th exam
- the $i$th bit of a binary file
- the $i$th letter of a text
- etc.

In an article that appeared in 1928, Hartley (Bell Labs) wrote: A symbol provides information only if there had been other possibilities for its value, besides that which was revealed.

In modern language, Hartley was saying that the value of a symbol provides information only if the symbol is a (non-constant) **random variable**.

WEATHER

$W \in \{ \text{sun}, \text{rain}, \text{fog}, \text{snow}, \text{hail} \}$

↳ INFORMATION IS

# WEATHER

$W \in \{ \text{sun}, \text{rain}, \text{fog}, \text{snow}, \text{hail} \}$

↳ INFORMATION IS

$(W_{today}, W_{tomorrow}) \in \{ (\text{sun}, \text{sun}), (\text{sun}, \text{rain}), \ldots \}$

↳ INFORMATION IS

In the same article, Hartley gave a tentative answer to the following related question: How much information is carried by a symbol such as $S$?

Hartley's answer:

- Suppose that $S \in \mathcal{A}$ is a symbol that can take on $|\mathcal{A}|$ different values.
- The amount of information conveyed by $n$ such symbols should be $n$ times the information conveyed by one symbol.
- There are $|\mathcal{A}|^n$ possible values for the $n$ symbols.
- This suggests that $\log |\mathcal{A}|^n = n \log |\mathcal{A}|$ is the appropriate measure for information, where we are free to choose the base for the logarithm.

## Example

In a village that has 8 telephones, we can assign a different three-digit binary number, such as 001, to each phone.

Hence it takes 3 bits of information to identify a phone. Mathematically, the phones are represented by a uniformly distributed random variable $S \in \mathcal{A} = \{1, 2, \ldots, 8\}$.

## Example

The world population in 2024 is estimated to be 8.1 billion.

Hence it takes $\log_2(8.1 \times 10^9) = 32.9$ bits of information to identify a person. A person is represented by a uniformly distributed random variable $S \in \mathcal{A} = \{1, 2, \ldots, 8.1 \times 10^9\}$.

The world population in 1970 is estimated to have been 3.7 billion. How many bits did it take back then?

The following example shows that something is not right with Hartley's measure of information.

## Example

Suppose that $S_n \in \{\text{sunny}, \text{rainy}\}$ is the weather prognosis for day $n+1$, revealed on day $n$. Suppose that $S_n = \text{rainy}$ has probability $\frac{5}{365}$.

It seems intuitively obvious that the amount of information provided by $S_n = \text{rainy}$ is much higher than that provided by $S_n = \text{sunny}$.

Hartley's measure assigns $\log_2(2) = 1$ bit of information to both, $S_n = \text{sunny}$ and $S_n = \text{rainy}$.

In an article that appeared in 1948, Shannon fixes the problem by defining the notion of **uncertainty** or **entropy** $H(S)$ associated to a discrete random variable $S$.

### Definition (entropy, uncertainty)

$$H_b(S) := -\sum_{s \in \mathrm{supp}(p_S)} p_S(s) \log_b p_S(s),$$

where $\mathrm{supp}(p_S) = \{s : p_S(s) > 0\}$.

A few comments are in order regarding

$$H_b(S) := -\sum_{s \in \mathrm{supp}(p_S)} p_S(s) \log_b p_S(s) :$$

- The condition $s \in \mathrm{supp}(p)$ is needed because $\log_b p_S(s)$ is not defined if $p_S(s) = 0$.
- To simplify the notation, we declare that $p_S(s) \log p_S(s) = 0$ when $p_S(s) = 0$. This convention allows us to simplify the notation to

$$H_b(S) = -\sum_{s \in \mathcal{A}} p_S(s) \log_b p_S(s).$$

- The choice of the base $b$ determines the unit. Typically $b = 2$. In this case, the unit is the **bit**.

We can think of evaluating

$$H(S) = -\sum_{s \in \mathcal{A}} p_S(s) \log p_S(s)$$

by first computing $-\log p_S(s)$ for each $s \in \mathcal{A}$, and then take the average (excluding zero-probability terms).

Hence we can write

$$H(S) = \mathbb{E}[-\log p_S(S)].$$

## EXAMPLE

When $p_S$ is the uniform distribution over the alphabet $\mathcal{A}$, $p_S(s) = \frac{1}{|\mathcal{A}|}$ and

$$-\log p_S(s) = \log |\mathcal{A}|, \text{ which is constant.}$$

In this case

$$H(S) = \mathbb{E}[\log |\mathcal{A}|] = \log |\mathcal{A}|,$$

which is Hartley's information measure.

Hence Shannon's entropy equals Hartley's measure of information if (and only if as we will see) the random variable has uniform distribution.

We will see that Shannon's entropy it is indeed the answer to very practical engineering questions.

## Example (Anne's lock)

A sequence of 4 decimal digits $s_1, s_2, s_3, s_4$ representing the number to open Anne's lock can be seen as the output of a source $S_1, S_2, S_3, S_4$ with $S_i \in \mathcal{A} = \{0, 1, \ldots, 9\}$.

If Anne picks each of the 4 digits at random and independently, then all 4-digit sequences are equiprobable, i.e.,

$$p_{S_1,S_2,S_3,S_4}(s_1, s_2, s_3, s_4) = \frac{1}{10^4} \quad \text{for all 4-digit numbers } s_1 s_2 s_3 s_4.$$

Notation: When no confusion can arise, we write $p(s_1, s_2, s_3, s_4)$ instead of $p_{S_1,S_2,S_3,S_4}(s_1, s_2, s_3, s_4)$.

## Example (Anne's lock, alternative view)

We can also take the view that Anne's lock number is modeled by a single random variable

$$S \in \mathcal{A} = \{0000, 0001, \ldots, 9998, 9999\}$$

$$p(s) = \frac{1}{10^4} \quad \text{for all 4-digit numbers.}$$

Since the distribution is uniform over the alphabet $\mathcal{A}$,

$$H(S) = \log_2 |\mathcal{A}| = \log_2 10^4 \approx 13.3 \text{ bits.}$$

## Example (Entropy of french)

A monkey produces text by selecting letters at random from a French text

$$H = 3.9425 \text{ bits.}$$

As we will see shortly, the maximum entropy of a source with $|\mathcal{A}| = 26$ is $\log_2 26 = 4.7004$ bits.

| Letter | Prob. |
|---|---|
| A | 0.0811 |
| B | 0.0081 |
| C | 0.0338 |
| D | 0.0428 |
| E | 0.1769 |
| F | 0.0113 |
| G | 0.0119 |
| H | 0.0074 |
| I | 0.0724 |
| J | 0.0018 |
| K | 0.0002 |
| L | 0.0599 |
| M | 0.0229 |
| N | 0.0768 |
| O | 0.0520 |
| P | 0.0292 |
| Q | 0.0083 |
| R | 0.0643 |
| S | 0.0887 |
| T | 0.0744 |
| U | 0.0523 |
| V | 0.0128 |
| W | 0.0006 |
| X | 0.0053 |
| Y | 0.0026 |
| Z | 0.0012 |

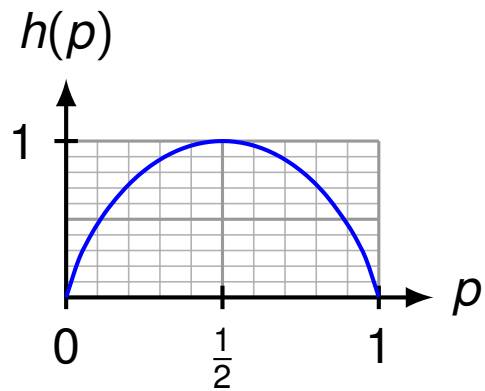
# Binary Entropy Function

An interesting special case is when $|\mathcal{A}| = 2$.

In this case, $p_S$ has only two possible values, say $p$ and $(1-p)$.
The corresponding entropy is $H(S) = h(p)$ where

$$h(p) := -p \log_2 p - (1-p) \log_2 (1-p).$$

$h(p)$ is called the **binary entropy function**.




- For $p = 0$ and for $p = 1$, $h(p) = 0$.
- For $p = \frac{1}{2}$, $h(p) = 1$.
- For $p \in \{0.0001, 0.9999\}$, $H(S) \approx 0.001$.

## EXAMPLE

Let $S_n$ be the above weather forecast.

The probabilities of $S_n$ are $p = \frac{5}{365}$ and $(1-p) = \frac{360}{365}$.

$$H_2(S_n) = h\left(\frac{5}{365}\right) = -\frac{5}{365}\log_2\frac{5}{365} - \frac{360}{365}\log_2\frac{360}{365} \approx 0.072 \text{ bits.}$$

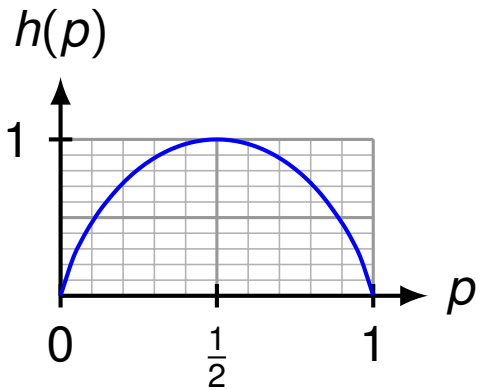
$$= \frac{5}{365}\log_2\frac{365}{5} + \frac{360}{365}\log_2\frac{365}{360}$$

## EXAMPLE

Let $S$ be the answer to the question "Is 8950 Anne's lock number ?".

Now $S \in \mathcal{A} = \{YES, NO\}$ is a binary random variable with $p_S(YES) = \frac{1}{10^4}$.

Hence $H(S) = h\left(\frac{1}{10^4}\right) \approx 0.001$ bits.



# Information-Theory Inequality

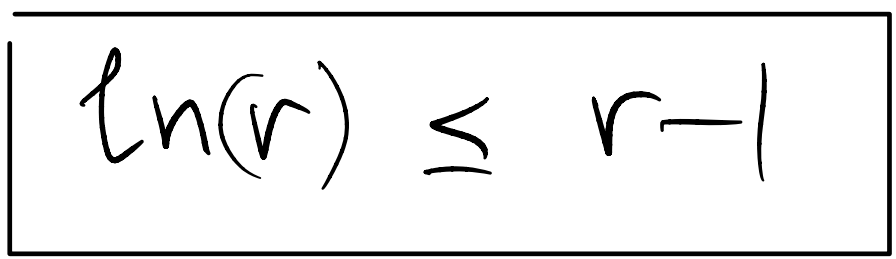
Surprisingly many results in information theory are a direct consequence of the following key inequality.

## Lemma (IT-Inequality)

For a positive real number $r$,

$$\log_b r \leq (r-1)\log_b(e),$$

with equality iff (if and only if) $r = 1$.



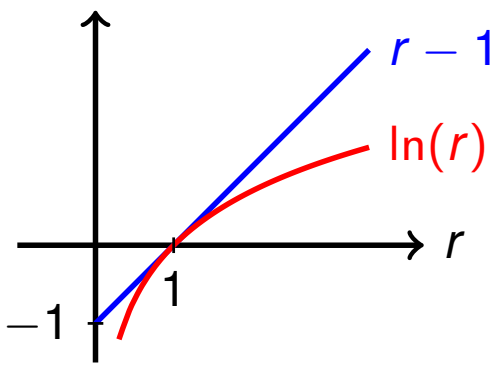
**Proof [IT-Inequality]:** Because $\log_b(r) = \ln(r)\log_b(e)$, it suffices to show that

$$\ln r \leq (r-1),$$

with equality iff $r = 1$.

The inequality is true (see graph below) because:

- the functions $\ln r$ and $r-1$ coincide at $r = 1$,
- the function $r - 1$ has slope 1 throughout,
- $\frac{d}{dr}\ln(r) = \frac{1}{r} < 1$ for $r > 1$,
- $\frac{d}{dr}\ln(r) = \frac{1}{r} > 1$ for $r < 1$.



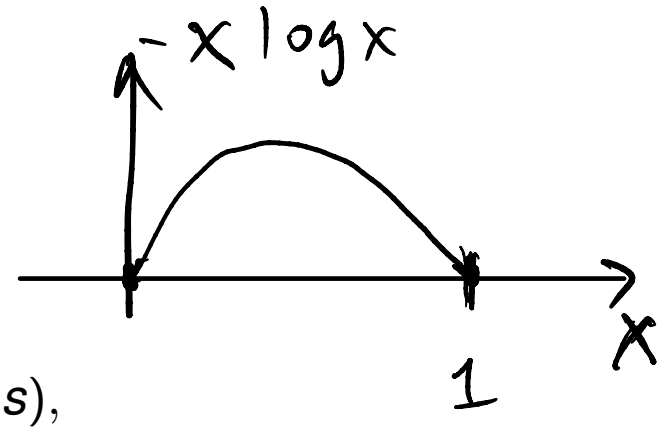
## Theorem (Entropy Bounds)

The entropy of a discrete random variable $S \in \mathcal{A}$ satisfies

$$0 \leq H_b(S) \leq \log_b |\mathcal{A}|,$$

with equality on the left iff $p_S(s) = 1$ for some $s$, and with equality on the right iff $p_S(s) = \frac{1}{|\mathcal{A}|}$ for all $s$.

**Proof of the left inequality:**




Recall that

$$H(S) = \sum_{s \in \mathcal{A}} -p_S(s) \log p_S(s),$$

and observe that

$$-p_S(s) \log p_S(s) = \begin{cases} 0 & \text{if } p_S(s) \in \{0, 1\} \\ > 0 & \text{if } 0 < p_S(s) < 1. \end{cases}$$

Thus, $H(S) \geq 0$, with equality iff $p_S(s) \in \{0, 1\}$ for all $s \in \mathcal{A}$.

$p_S(s) \in \{0, 1\}$ for all $s \in \mathcal{A}$ iff $p_S(s) = 1$ for some $s$. □

To prove the right inequality, we use a trick that often works in inequalities involving entropies:

To prove, say, $A \leq B$, we prove $A - B \leq 0$ by means of the IT-Inequality.

**THM:** $H_b(S) \leq \log_b |A|$

**PROOF:**

$$H_b(S) - \log_b |A| \leq 0$$

$$= -\sum_s p(s) \log_b p(s) - \log_b |A|$$

$$= -\sum_s p(s) \log_b p(s) - \sum_s p(s) \log_b |A|$$

$$= -\sum_s p(s) \left[ \log_b p(s) + \log_b |A| \right]$$

$$= -\sum_s p(s) \log_b \left( p(s) |A| \right)$$

$$= \sum_s p(s) \log_b \frac{1}{p(s)|A|}$$

$$\leq \sum_s p(s) \left( \frac{1}{p(s)|A|} - 1 \right) \log_b(e)$$

$$= \left\{ \sum_s \frac{1}{|A|} - \sum_s p(s) \right\} \log_b(e)$$

$$\leq \{ \leq 1 - 1 \} \log_b(e)$$

$$\leq 0.$$

**Proof of the right inequality:**

$$
\begin{aligned}
H(S) - \log|\mathcal{A}| &= \mathbb{E}\Big[-\log p_S(S)\Big] - \log|\mathcal{A}| \\
&= \mathbb{E}\Big[-\log p_S(S) - \log|\mathcal{A}|\Big] \\
&= \mathbb{E}\left[\log \frac{1}{p_S(s)|\mathcal{A}|}\right] \\
&= \sum_{s\in\mathcal{A}} p_S(s)\left[\log \frac{1}{p_S(s)|\mathcal{A}|}\right] \\
&\overset{\text{(IT-Inequality)}}{\leq} \sum_{s\in\mathcal{A}} p_S(s)\left[\frac{1}{p_S(s)|\mathcal{A}|} - 1\right]\log(e) \\
&= \log(e)\sum_{s\in\mathcal{A}}\left[\frac{1}{|\mathcal{A}|} - p_S(s)\right] \\
&= \log(e)\,(1-1) = 0,
\end{aligned}
$$

with equality iff $p_S(s)|\mathcal{A}| = 1$ for all $s$. □

## Example

Let $S$ be Anne's lock number. Its entropy is maximized if Anne chooses at random over all $10^4$ possibilities.

In this case, and only in this case,

$$H(S) = \log |\mathcal{A}| = \log 10^4 \approx 13.3 \text{ bits.}$$

## EXAMPLE

Let $S$ be Anne's grandmother's lock number. She always picks $S = 0000$. Then

$$H(S) = 0$$

The formula for the entropy of a random variable $S$ extends to any number of random variables. If $X$ and $Y$ are two discrete random variables, with (joint) probability distribution $p_{X,Y}$ then

$$H(X, Y) = \mathbb{E}[-\log p_{X,Y}(X, Y)],$$

which means

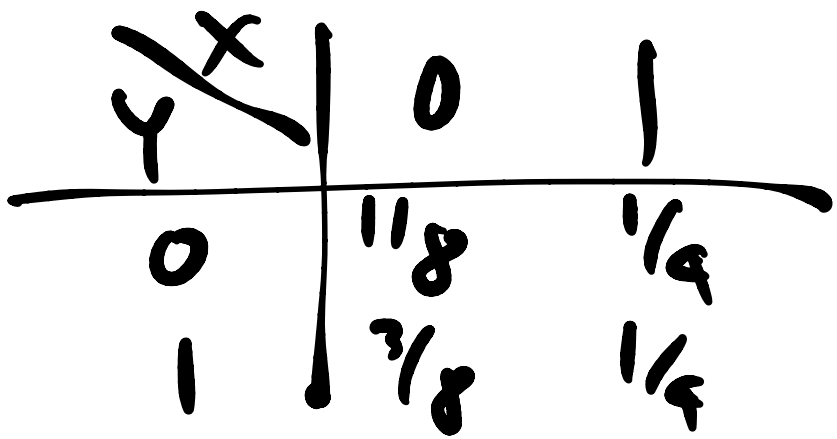
$$H(X, Y) = -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{X,Y}(x, y) \log p_{X,Y}(x, y).$$

## EXAMPLE

Let $p_{X,Y}$ be given by the following table

| $x$ | $y$ | $p_{XY}(x,y)$ |
|---|---|---|
| 0 | 0 | 1/8 |
| 0 | 1 | 3/8 |
| 1 | 0 | 1/4 |
| 1 | 1 | 1/4 |

$$H(X,Y) = -\tfrac{1}{8}\log_2\tfrac{1}{8} - \tfrac{3}{8}\log_2\tfrac{3}{8} - \tfrac{1}{4}\log_2\tfrac{1}{4} - \tfrac{1}{4}\log_2\tfrac{1}{4}.$$



We are mainly interested is sources that emit a large number of random variables. (We want to compress large amounts of data.)

The sequence of random variables can be

- finite, like in $(S_1, \ldots, S_n)$ (random vector)
- infinite, like in $S_1, S_2, \ldots$ (random sequence, random process), also denoted by $\{S_i\}_{i=1}^{\infty}$.
- sometimes it is convenient to consider $\ldots, S_{-1}, S_0, S_1, \ldots$, also denoted by $\{S_i\}$.

A collection of random variables $(S_1, \ldots, S_n)$ is specified by the joint probability distribution $p_{S_1,\ldots,S_n}$. This is all we need to compute the entropy $H(S_1, \ldots, S_n)$.

## Example (Coin-Flip Source)

A sequence of coin flips can be seen as the output of a source $S_1, S_2, \ldots, S_n$ with $S_i \in \mathcal{A} = \{H, T\}$, where $H$ stands for head, and $T$ for tail, $i = 1, \ldots, n$.

If the coin is fair, all sequences are equally likely:

$$p(s_1, s_2, \ldots, s_n) = \prod_i p(s_i) = \frac{1}{2^n} \quad \text{for all } (s_1, s_2, \ldots, s_n) \in \mathcal{A}^n$$

Notation: $\mathcal{A}^n = \underbrace{\mathcal{A} \times \mathcal{A} \times \cdots \times \mathcal{A}}_{n \text{ times}}$ is the $n$-fold cartesian product of $\mathcal{A}$.

The following statement is a corollary to two fundamental results that we will prove next week.

### Theorem (1.4 of textbook)

Let $S_1, \ldots, S_n$ be discrete random variables. Then

$$H(S_1, S_2, \ldots, S_n) \leq H(S_1) + H(S_2) + \cdots + H(S_n),$$

with equality iff $S_1, \ldots, S_n$ are independent.

## EXAMPLE

Let $S_1$ and $S_2$ be the random variables associated to Bart's two dice rolls.

$$H(S_1) = H(S_2) = \log 6 \quad \text{(the two distributions are uniform)}$$
$$H(S_1, S_2) = \log 36 \quad \text{(the distribution of } (S_1, S_2) \text{ is uniform)}$$

We verify that

$$H(S_1, S_2) = \log 36 = \log 6^2 = 2 \log 6 = H(S_1) + H(S_2).$$

## Example (Entropies in Lisa's experiment)

- $H(L_1) = 0.65$ bits
- $H(L_2) = 3.22$ bits
- $H(L_1, L_2) = 3.27$ bits

$H(L_1, L_2) < H(L_1) + H(L_2)$. Hence $L_1$ and $L_2$ are **not** independent.

We will see that entropy is fundamental in all three topics:

- Source coding: To derive the limit to how much a source can be compressed.
- Cryptography: To derive the length of the shortest key for which perfect secrecy is possible.
- Channel coding: To derive the highest rate at which we can communicate reliably across an unreliable communication channel.

Stay tuned!

Ex. "HAT PARTY 1950"

- $n$ men, all have the same hat.

Ex. "HAT PARTY 1950"

- $n$ men, all have the same hat.
- they throw hats in a corner.
- leaving, they randomly take a hat.

PROBLEM:

$\mathbb{E}[$ Number of men who leave with their own hat $] = ?$

Let $R_i =$

Let $R_i = \begin{cases} 1, & \text{if man } i \text{ leaves with his own hat.} \\ 0, & \text{otherwise.} \end{cases}$

$$\mathbb{E}[R_1 + R_2 + \dots + R_n]$$

$$= \mathbb{E}[R_1] + \mathbb{E}[R_2] + \dots + \mathbb{E}[R_n]$$

$$= \frac{1}{n} + \frac{1}{n} + \dots + \frac{1}{n} = 1$$