Information and Communication Theory

Lecture 0
Review of Discrete Probability Theory

Mário A. T. Figueiredo

DEEC, Instituto Superior Técnico,
University of Lisbon, **Portugal**

2023

Information and Communication Theory

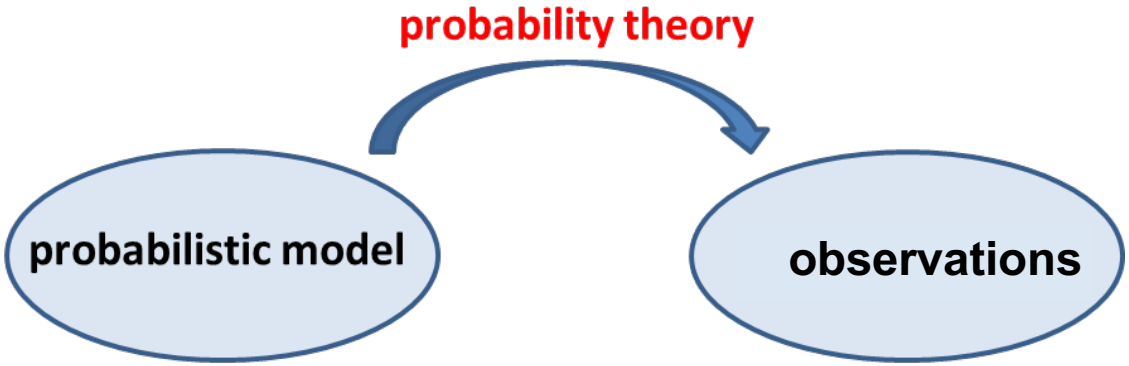
Lecture 0
Review of Discrete Probability Theory

Mário A. T. Figueiredo

DEEC, Instituto Superior Técnico,
University of Lisbon, **Portugal**

2023

Probability theory



- Probability theory has its roots in games of chance

- Great names of science: Bayes, Bernoulli(s), Boltzman, Cardano, Cauchy, Fermat, Huygens, Kolmogorov, Laplace, Pascal, Poisson, ...
- Tool to handle uncertainty, information, knowledge, observations, ...
- ...thus also learning, decision making, inference, science,...

What is probability?

Example: $\mathbb{P}(\text{randomly drawn card is } \clubsuit) = 13/52$.

Example: $\mathbb{P}(\text{getting 1 in throwing a fair die}) = 1/6$.

- Classical definition: $\mathbb{P}(A) = \dfrac{N_A}{N}$

  ...with N mutually exclusive equally likely outcomes,
  N_A of which result in the occurrence of A. *Laplace, 1814*

- Frequentist definition: $\mathbb{P}(A) = \lim\limits_{N\to\infty} \dfrac{N_A}{N}$

  ...relative frequency of occurrence of A in infinite number of trials.

- Subjective probability: $\mathbb{P}(A)$ is a degree of belief. *de Finetti, 1930s*

  .

  ...gives meaning to $\mathbb{P}(\text{“it will rain today”})$, or
  $\mathbb{P}(\text{“Patient A has disease } x\text{”})$

MENU / aeon

The concept of probability is not as simple as you think

Nevin Climenhaga

A summary of some interpretations of probability

	Classical	Frequentist	Subjective	Propensity
Main hypothesis	Principle of indifference	Frequency of occurrence	Degree of belief	Degree of causal connection
Conceptual basis	Hypothetical symmetry	Past data and reference class	Knowledge and intuition	Present state of system
Conceptual approach	Conjectural	Empirical	Subjective	Metaphysical
Single case possible	Yes	No	Yes	Yes
Precise	Yes	No	No	Yes
Problems	Ambiguity in principle of indifference	Circular definition	Reference class problem	Disputed concept

"The mathematics of probability can be developed on an entirely axiomatic basis, independent of any interpretation." (wikipedia)

Key concepts: Sample space and events

- Sample space Ω = set of possible outcomes of a random experiment.

  Examples:
  - Tossing two coins: $\Omega = \{HH, TH, HT, TT\}$
  - Roulette: $\Omega = \{1, 2, ..., 36\}$
  - Draw a card from a shuffled deck: $\Omega = \{A\clubsuit, 2\clubsuit, ..., Q\diamondsuit, K\diamondsuit\}$.

- An event A is a subset of Ω: $A \subseteq \Omega$ (also written $A \in 2^{\Omega}$).

  Examples:
  - "exactly one H in 2-coin toss": $A = \{TH, HT\}$.
  - "odd number in the roulette": $B = \{1, 3, ..., 35\}$.
  - "drawn a $\heartsuit$ card": $C = \{A\heartsuit, 2\heartsuit, ..., K\heartsuit\}$

Key concepts: Sample space and events

- Sample space $\Omega =$ set of possible outcomes of a random experiment.

  (More delicate) examples:
  - Distance travelled by tossed die: $\Omega = \mathbb{R}_+$
  - Location of the next rain drop on a given square tile: $\Omega = \mathbb{R}^2$

- Properly handling the continuous case requires deeper concepts:
  - Sigma algebras, Borel sets, measurable functions, ...

...mathematically **heavier** stuff, not covered here

Kolmogorov's Axioms for Probability

- Probability is a function that maps events A into the interval $[0, 1]$.

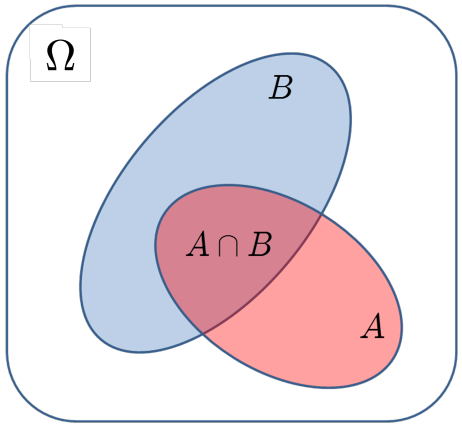
  Kolmogorov's axioms (1933) for probability

  - For any A, $\mathbb{P}(A) \geq 0$
  - $\mathbb{P}(\Omega) = 1$
  - If A_1, A_2 ... $\subseteq \Omega$ are disjoint events, then $\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$

- From these axioms, many results can be derived.

  Examples:

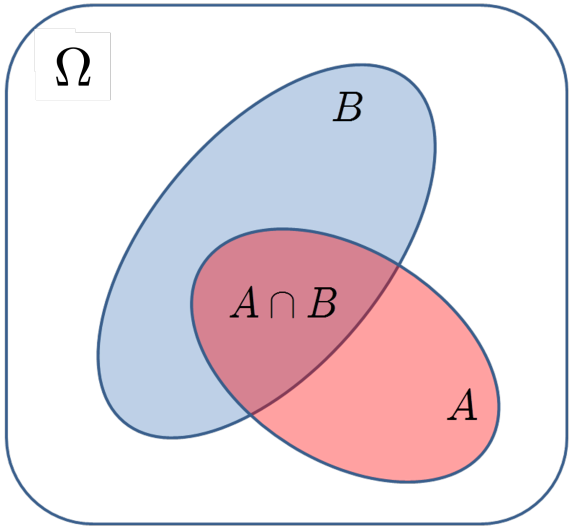
  - $\mathbb{P}(\emptyset) = 0$
  - $C \subset D \Rightarrow \mathbb{P}(C) \leq \mathbb{P}(D)$
  - $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
  - $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ (union bound)



Conditional Probability and Independence

- If $\mathbb{P}(B) > 0$, $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ (conditional prob. of A, given B)

- ...satisfies all of Kolmogorov's axioms:



- For any $A \subseteq \Omega$, $\mathbb{P}(A|B) \geq 0$
- $\mathbb{P}(\Omega|B) = 1$
- If $A_1, A_2, \ldots \subseteq \Omega$ are disjoint,

$$\mathbb{P}\Big(\bigcup_i A_i \Big| B\Big) = \sum_i \mathbb{P}(A_i|B)$$

- Independence: A, B are independent ($A \perp\!\!\!\perp B$):

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B).$$

Conditional Probability and Independence

- If $\mathbb{P}(B) > 0$, $\quad \mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- Events A, B are independent $(A \perp\!\!\!\perp B) \Leftrightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.
- Relationship with conditional probabilities:

$$A \perp\!\!\!\perp B \Leftrightarrow \mathbb{P}(A|B) = \mathbb{P}(A)$$

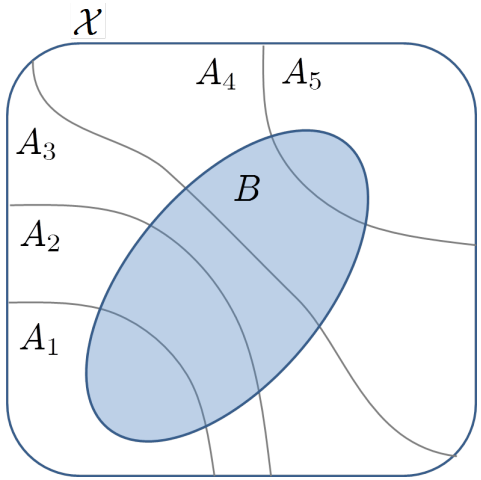
- Example: $\Omega =$ "52 cards", $A = \{4\heartsuit, 4\clubsuit, 4\diamondsuit, 4\spadesuit\}$, and $B = \{A\heartsuit, 2\heartsuit, ..., K\heartsuit\}$; then, $\mathbb{P}(A) = 1/13$, $\mathbb{P}(B) = 1/4$

$$\begin{aligned} \mathbb{P}(A \cap B) &= \mathbb{P}(\{4\heartsuit\}) = \frac{1}{52} \\ \mathbb{P}(A)\,\mathbb{P}(B) &= \frac{1}{13}\frac{1}{4} = \frac{1}{52} \\ \mathbb{P}(A|B) &= \mathbb{P}(\text{“4”}|\text{“}\heartsuit\text{”}) = \frac{1}{13} = \mathbb{P}(A) \end{aligned}$$

Bayes Theorem

- Law of total probability: if $A_1, ..., A_n$ are a partition of Ω

$$\begin{aligned}\mathbb{P}(B) &= \sum_i \mathbb{P}(B|A_i)\mathbb{P}(A_i) \\ &= \sum_i \mathbb{P}(B \cap A_i)\end{aligned}$$



- Bayes' theorem: if $\{A_1, ..., A_n\}$ is a partition of Ω

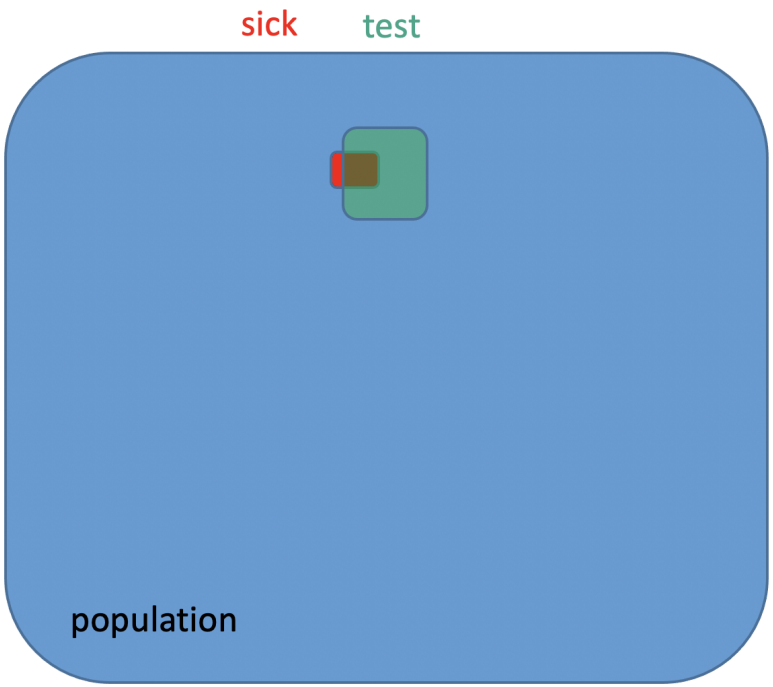
$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B \cap A_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\ \mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)}$$

Bayesian inference

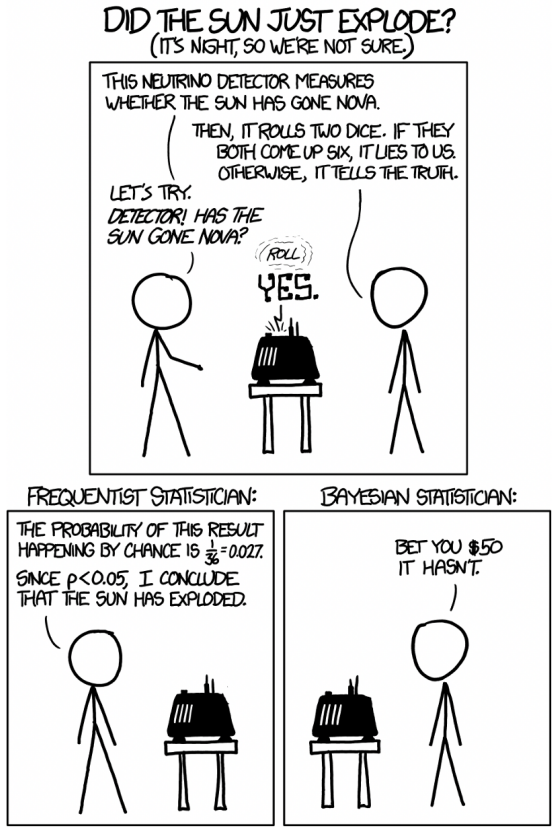
$$P(\text{sick} \mid \text{test}) = \frac{P(\text{test}, \text{sick})}{P(\text{test})} = \frac{\overbrace{P(\text{test} \mid \text{sick})}^{\text{sensitivity}} \overbrace{P(\text{sick})}^{\text{prevalence}}}{P(\text{test} \mid \text{sick}) P(\text{sick}) + \underbrace{P(\text{test} \mid \text{not sick})}_{\text{false positive}} \underbrace{P(\text{not sick})}_{(1 - \text{prevalence})}}$$

Bayes (1763)

Laplace (1812)

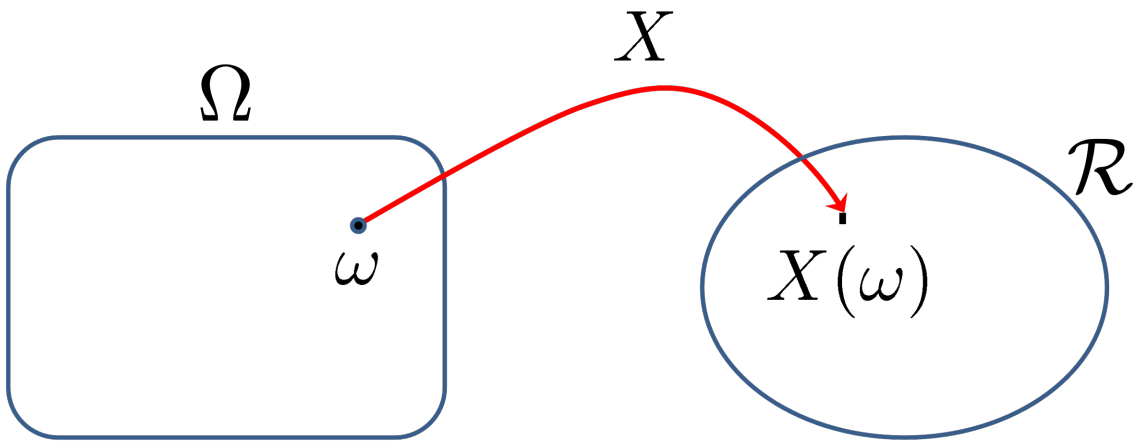


$\frac{\text{■}}{\text{■}}$ = prevalence = P(sick)



Random Variables

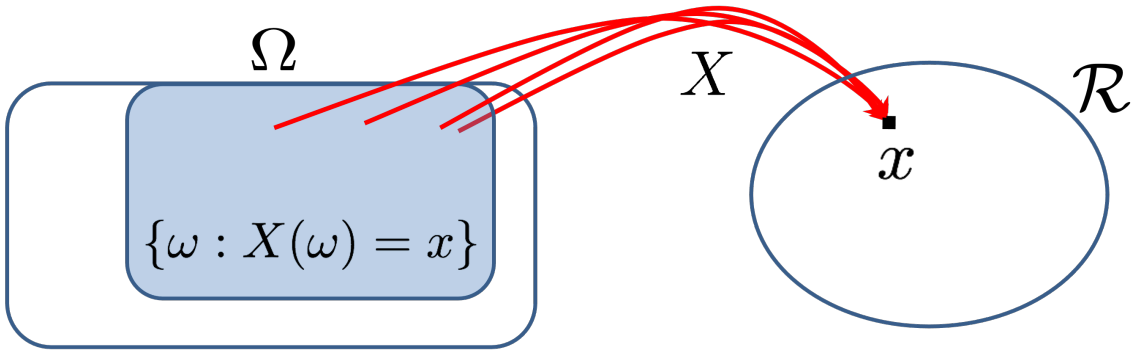
- A (real) random variable (RV) is a function: $X : \Omega \to \mathcal{R}$



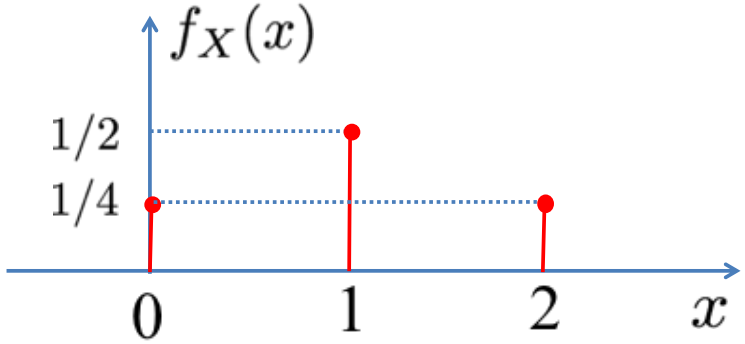
  - Discrete RV: $\mathcal{R}$ is countable (*e.g.*, $\mathbb{N}$ or $\{0, 1\}$ or $\{\text{yes}, \text{no}, \text{maybe}\}$)
  - Continuous RV: range of X is uncountable (*e.g.*, $\mathbb{R}$ or $[0,\ 1]$)
  - Example: number of heads in tossing two coins,
    $\Omega = \{HH, HT, TH, TT\}$,
    $X(HH) = 2$, $X(HT) = X(TH) = 1$, $X(TT) = 0$.
    Range of $X = \{0, 1, 2\}$.
  - Example: distance traveled by a tossed coin; range of $X = \mathbb{R}_+$.

Discrete Random Variables

- Probability mass function: $f_X(x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\})$



- Example: number of heads in tossing 2 coins; $\mathcal{R} = \{0, 1, 2\}$.



- $\sum_{x \in \mathcal{R}} f_X(x) = 1$ (can you show why?)

Important Discrete Random Variables

- Uniform: $X \in \{x_1, ..., x_K\}$, pmf $f_X(x_i) = 1/K$.

  Example: a fair roulette $X \in \{1, ..., 36\}$, with $f_X(x) = 1/36$

  Example: a fair die $X \in \{1, ..., 6\}$, with $f_X(x) = 1/6$

- Bernoulli RV: $X \in \{0, 1\}$, pmf $f_X(x) = \begin{cases} p & \Leftarrow \quad x = 1 \\ 1-p & \Leftarrow \quad x = 0 \end{cases}$

  Compact form: $f_X(x) = p^x (1-p)^{1-x}$.

  Example: a coin toss; heads $= 0$, tails $= 1$

  fair, if $p = 1/2$; unfair, if $p \neq 1/2$

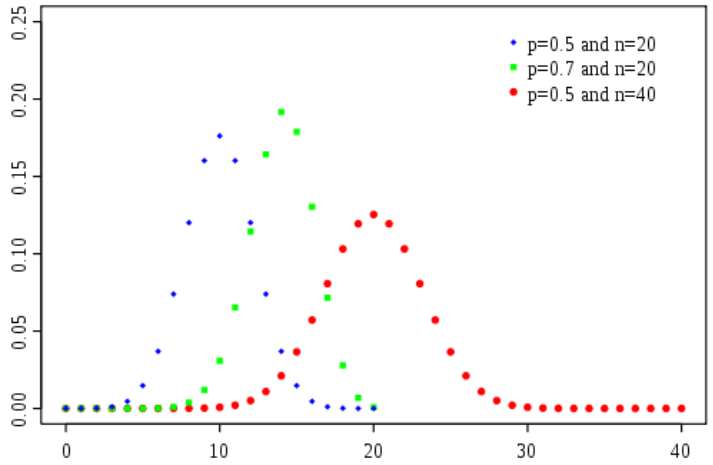
Important Discrete Random Variables

- Binomial RV: $X \in \{0, 1, ..., n\}$ (sum of n Bernoulli RVs)

$$f_X(x) = \text{Binomial}(x; n, p) = \binom{n}{x} p^x \, (1-p)^{(n-x)}$$

Binomial coefficients
("n choose x"):

$$\binom{n}{x} = \frac{n!}{(n-x)!\, x!}$$



Example: number of heads in n coin tosses.

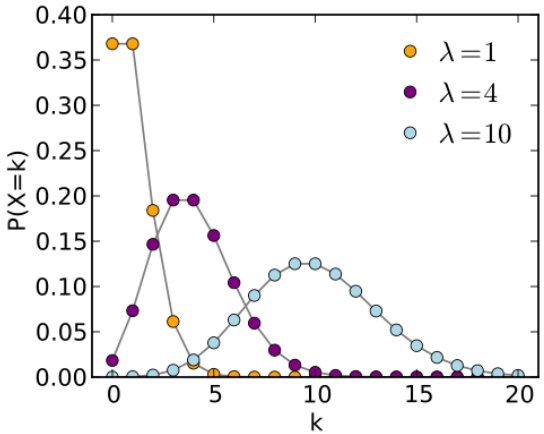
Other Important Discrete Random Variables

- Geometric(p): $X \in \mathbb{N}$, pmf $f_X(x) = p(1-p)^{x-1}$.

  Example: number of coin tosses until first heads.

- Poisson(λ):

$$X \in \mathbb{N} \cup \{0\},$$
$$\text{pmf } f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$



"...probability of the number of independent occurrences in a fixed (time/space) interval, if these occurrences have known average rate"

Examples: number of rain drops per second on a given area, number of calls per hour in a call center, number of tweets per day by DT, ...

Expectation of (Real) Random Variables ($\mathcal{R} \subset \mathbb{R}$)

- Expectation: $\mathbb{E}(X) = \sum_{x \in \mathcal{R}} x\, f_X(x)$
- Example: Bernoulli, $f_X(x) = p^x\,(1-p)^{1-x}$, for $x \in \{0,\, 1\}$.

  $$\mathbb{E}(X) = 0\,(1-p) + 1\,p = p.$$

- Example: Binomial, $f_X(x) = \binom{n}{x} p^x\,(1-p)^{n-x}$, for $x \in \{0, ..., n\}$.

  $$\mathbb{E}(X) = n\,p.$$

- Linearity of expectation:

  $$\mathbb{E}(\alpha\, X + \beta\, Y) = \alpha\, \mathbb{E}(X) + \beta\, \mathbb{E}(Y), \qquad \alpha, \beta\, \in \mathbb{R}$$

Expectation of Functions of RVs

- Expectation: for $g : \mathcal{R} \to \mathbb{R}$, $\mathbb{E}(X) = \sum_{x \in \mathcal{R}} g(x)\, f_X(x)$
- Example: variance, $\text{var}(X) = \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)^2\Big) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$
- Example: Bernoulli variance, $\mathbb{E}(X^2) = \mathbb{E}(X) = p$, thus $\text{var}(X) = p(1-p)$.
- Probability as expectation of indicator, $\mathbf{1}_A(x) = \begin{cases} 1 & \Leftarrow \quad x \in A \\ 0 & \Leftarrow \quad x \notin A \end{cases}$

$$\mathbb{P}(X \in A) = \int_A f_X(x)\, dx = \int \mathbf{1}_A(x)\, f_X(x)\, dx = \mathbb{E}(\mathbf{1}_A(X))$$

Two (or More) Random Variables

- Joint pmf of two discrete RVs: $f_{X,Y}(x,y) = \mathbb{P}(X = x \wedge Y = y)$.
- Joint pmf of N discrete RVs:

$$f_{X_1,X_2,...,X_N}(x_1,x_2,...,x_N) = \mathbb{P}(X_1 = x_1 \wedge X_2 = x_2 \wedge \cdots \wedge X_N = x_N)$$

- Marginalization: $f_Y(y) = \sum_x f_{X,Y}(x,y)$,
- Marginalization:

$$\sum_{x_i} f_{X_1,X_2,...,X_N}(x_1,x_2,...,x_N) = f_{X_1,...,X_{i-1},X_{i+1}...,X_N}(x_1,...,x_i,x_{i+1},...,x_N)$$

- Independence:

$$X \perp\!\!\!\perp Y \Leftrightarrow f_{X,Y}(x,y) = f_X(x)\, f_Y(y) \begin{array}{c}\Rightarrow \\ \not\Leftarrow\end{array} \mathbb{E}(X\,Y) = \mathbb{E}(X)\,\mathbb{E}(Y).$$

Conditionals and Bayes' Theorem

- Conditional pmf (discrete RVs):

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

- Bayes' theorem: $f_{X|Y}(x|y) = \dfrac{f_{Y|X}(y|x)\, f_X(x)}{f_Y(y)}$

- Bayesian jargon: $\text{posterior} = \dfrac{\text{likelihood} \times \text{prior}}{\text{evidence}}$

Joint, Marginal, and Conditional Probabilities: An Example

- A pair of binary variables $X, Y \in \{0, 1\}$, with joint pmf:

$f_{X,Y}(x,y)$	$Y=0$	$Y=1$
$X=0$	1/5	2/5
$X=1$	1/10	3/10

- Marginals: $f_X(0) = \frac{1}{5} + \frac{2}{5} = \frac{3}{5}, \quad f_X(1) = \frac{1}{10} + \frac{3}{10} = \frac{4}{10},$

  $f_Y(0) = \frac{1}{5} + \frac{1}{10} = \frac{3}{10}, \quad f_Y(1) = \frac{2}{5} + \frac{3}{10} = \frac{7}{10}.$

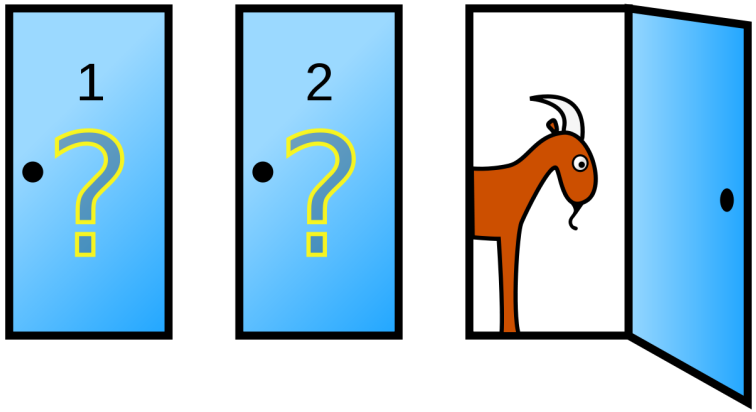
- Conditional probabilities:

$f_{X\|Y}(x\|y)$	$Y=0$	$Y=1$
$X=0$	2/3	4/7
$X=1$	1/3	3/7

$f_{Y\|X}(y\|x)$	$Y=0$	$Y=1$
$X=0$	1/3	2/3
$X=1$	1/4	3/4

- Bayes example: $f_{X|Y}(1|0) = \frac{f_{Y|X}(0|1)\, f_X(1)}{f_Y(0)} = \frac{(1/4)\,(4/10)}{(3/10)} = \frac{1}{3}$

Let's play the Monty Hall Game!



- Three doors: a car and two goats.
- You choose a door; say door 1.
- The host opens a door with a goat.
- The host gives you the chance of keeping the original choice or switch.

What is the best option (higher probability of wining the car)?

Monty Hall Game: Bayes Solution

- Door with car $C \in \{1, 2, 3\}$; pmf $f_C(1) = f_C(2) = f_C(3) = 1/3$.
- You choose door 1
- Door opened with goat $D \in \{2, 3\}$; conditional probabilities

$$
\begin{aligned}
&f_{D|C}(2|1) = \tfrac{1}{2}, \quad f_{D|C}(3|1) = \tfrac{1}{2} && \text{(host can open either door)};\\
&f_{D|C}(2|2) = 0, \quad f_{D|C}(3|2) = 1 && \text{(host can only open door 3)};\\
&f_{D|C}(2|3) = 1, \quad f_{D|C}(3|3) = 0 && \text{(host can only open door 2)};
\end{aligned}
$$

- Posterior probabilities

$$
\begin{aligned}
f_{C|D}(1|2) &= \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + \frac{1}{3}} = \tfrac{1}{3}, \quad f_{C|D}(2|2) = 0, \qquad f_{C|D}(3|2) = \tfrac{2}{3}\\
f_{C|D}(1|3) &= \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + \frac{1}{3}} = \tfrac{1}{3}, \quad f_{C|D}(2|3) = \tfrac{2}{3}, \qquad f_{C|D}(3|3) = 0
\end{aligned}
$$

- It is better to switch!

An Important Multivariate RV: Multinomial

- Multinomial: $X = (X_1, ..., X_K)$, $X_i \in \{0, ..., n\}$, s.t. $\sum_i X_i = n$,

$$f_X(x_1, ..., x_K) = \begin{cases} \binom{n}{x_1\ x_2 \cdots x_K} p_1^{x_1} p_2^{x_2} \cdots p_K^{x_K} & \Leftarrow \sum_i x_i = n \\ 0 & \Leftarrow \sum_i x_i \neq n \end{cases}$$

$$\binom{n}{x_1\ x_2\ \cdots\ x_K} = \frac{n!}{x_1!\, x_2! \cdots x_K!}$$

Parameters: $p_1, ..., p_K \geq 0$, such that $\sum_i p_i = 1$.

- Generalizes the binomial from binary to K-classes.

- Example: tossing n independent fair dice, $p_1 = \cdots = p_6 = 1/6$.
  $x_i =$ number of outcomes with i dots (of course, $\sum_i x_i = n$)

Recommended Reading

- A. Maleki and T. Do, "Review of Probability Theory", Stanford University, 2017 (`https://tinyurl.com/pz7p9g5`)
- L. Wasserman, "All of Statistics: A Concise Course in Statistical Inference", Springer, 2004.