Master Program in *Data Science and Business Informatics*

Statistics for Data Science

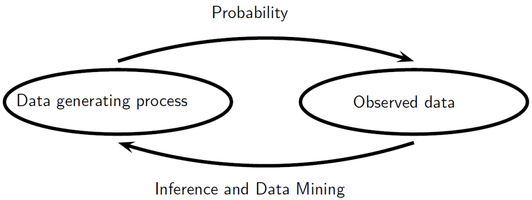
Lesson 01 - Probabilities and independence

Salvatore Ruggieri

Department of Computer Science
University of Pisa, Italy
salvatore.ruggieri@unipi.it

Why Statistics in Data Science?

We need grounded means for reasoning about data generated from real world with some degree of randomness.



What will you learn?

- Probability: properties of data generated by a known/assumed randomness model
- Statistics: properties of a randomness model that could have generated given data
- The R programming language

Sample spaces and events

- An **experiment** is a measurement of a random process
- The **outcome** of a measurement takes values in some set Ω, called the **sample space**.

  Examples:
  - Tossing a coin: $\Omega = \{\text{H, T}\}$ *[Finite]*
  - Month of birthdays $\Omega = \{\text{Jan}, \ldots, \text{Dec}\}$ *[Finite]*
  - Population of a city $\Omega = \mathbb{N} = \{0, 1, 2, \ldots, \}$ *[Countably infinite]*
  - Length of a street $\Omega = \mathbb{R}^+ = (0, \infty)$ *[Uncountably infinite]*
  - Tossing a coin twice: $\Omega = \{\text{H, T}\} \times \{\text{H, T}\} = \{(\text{H, H}),(\text{H, T}),(\text{T, H}),(\text{T, T})\}$
  - Testing for Covid-19 (univariate): $\Omega = \{+, -\}$
  - Testing for Covid-19 (multivariate): $\Omega = \{\text{f, m}\} \times \mathbb{N} \times \{+, -\}$, e..g, $(\text{f}, 25, -) \in \Omega$

  Look at **seeing-theory.brown.edu**
- An **event** is some subset of $A \subseteq \Omega$ of possible outcomes of an experiment.
  - $L = \{$ Jan, March, May, July, August, October, December $\}$ *a long month with 31 days*
- We say that an event A **occurs** if the outcome of the experiment belongs to the set A.
  - If the outcome is Jan then L occurs

Probability functions on finite sample space

A **probability function** is a mapping from events to **real numbers** that satisfies certain axioms. *Intuition: how likely is an event to occur.*

DEFINITION. A *probability function* P on a finite sample space Ω assigns to each event A in Ω a number $\mathrm{P}(A)$ in [0,1] such that
(i) $\mathrm{P}(\Omega) = 1$, and
(ii) $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$ if A and B are disjoint.
The number $\mathrm{P}(A)$ is called the probability that A occurs.

- Fact: $P(\{a_1, \ldots, a_n\}) = P(\{a_1\}) + \ldots + P(\{a_n\})$ *[Generalized additivity]*
  - Assigning probability to a singleton is enough
- Examples:
  - $P(\{\mathrm{H}\}) = P(\{\mathrm{T}\}) = {}^1/_2$
  - $P(\{\mathrm{Jan}\}) = {}^{31}/_{365}$, $P(\{\mathrm{Feb}\}) = {}^{28}/_{365}, \ldots P(\{\mathrm{Dec}\}) = {}^{31}/_{365}$
  - $P(L) = {}^7/_{12}$ or ${}^{31 \cdot 7}/_{365}$?
- $P(\{a\})$ often abbreviated as $P(a)$, e.g., $P(\mathrm{Jan})$ instead of $P(\{\mathrm{Jan}\})$

Properties of probability functions

- $P(A^c) = 1 - P(A)$
- $P(\emptyset) = 0$ *[Impossible event]*
- $A \subseteq B \Rightarrow P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ *[Inclusion-exclusion principle]*
- $P(A \cup B) = P(A) + P(B \setminus A)$
- probability that at least one coin toss over two lands head?

Defining probability functions

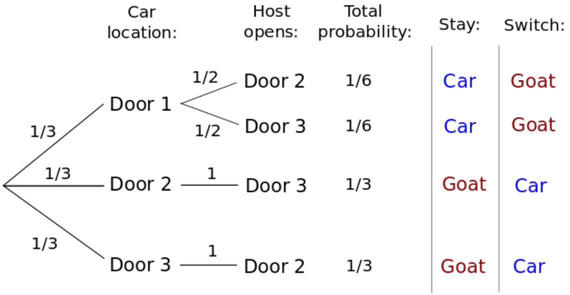
Assigning probability is **NOT** an easy task: a prob. function can be an approximation of reality

- **Frequentist** interpretation: probability measures a "*proportion of outcomes*".
  - A fair coin lands on heads 50% of times
  - $P(A) = {}^{|A|}/_{\Omega}$ *[Counting]*
  - $P(\{\text{ at least one H in two coin tosses}\}) = |\{(H,H),(H,T),(T,H)\}|/4 = {}^{3}/_{4}$
- **Bayesian** (or epistemological) interpretation: probability measures a "*degree of belief*".
  - Iliad and Odissey were composed by the same person at 90%

Counting: The Monty Hall problem

https://math.andyou.com/tools/montyhallsimulator/montysim.htm
(See also Exercise 2.14 of textbook **[T]**)

Tree-based sequential description of probability function



Probability functions on countably infinite sample space

DEFINITION. A *probability function* on an infinite (or finite) sample space Ω assigns to each event A in Ω a number $\mathrm{P}(A)$ in $[0,1]$ such that

(i) $\mathrm{P}(\Omega) = 1$, and

(ii) $\mathrm{P}(A_1 \cup A_2 \cup A_3 \cup \cdots) = \mathrm{P}(A_1) + \mathrm{P}(A_2) + \mathrm{P}(A_3) + \cdots$ if $A_1, A_2, A_3, \ldots$ are disjoint events.

- (ii) is called countable additivity. It is equivalent to σ-additivity: for $A_1 \subseteq A_2 \subseteq \ldots$

$$P(\lim_{n\to\infty} A_i) = \lim_{n\to\infty} P(A_i)$$

- Example
  - Experiment: we toss a coin repeatedly until H turns up.
  - Outcome: the number of tosses needed.
  - $\Omega = \{1, 2, \ldots\} = \mathbb{N}^+$
  - Suppose: $P(H) = p$. Then: $P(n) = (1-p)^{n-1}p$
  - Is it a probability function? $P(\Omega) = \ldots$

Conditional probability

- Long months and months with 'r'
  - $L = \{$ Jan, Mar, May, July, Aug, Oct, Dec $\}$ *a long month with 31 days*
  - $R = \{$ Jan, Feb, Mar, Apr, Sep, Oct, Nov, Dec $\}$ *a month with 'r'*
  - $P(L) = {}^{7}\!/_{12} \quad P(R) = {}^{8}\!/_{12}$
- Anna is born in a long month. What is the probability she is born in a month with 'r'?

$$P(R|L) = \frac{P(L \cap R)}{P(L)} = \frac{P(\{\text{Jan, Mar, Oct, Dec}\})}{P(L)} = \frac{4/12}{7/12} = \frac{4}{7}$$

- **Intuition:** probability of an event in the restricted sample space $\Omega \cap L$
  - *a-priori* probability $P(R) = {}^{8}\!/_{12}$
  - *a-posteriori* probability $P(R|L) = {}^{4}\!/_{7} < {}^{8}\!/_{12}$
- **Example (classification):** probab. of Covid given gender=f and age$\geq$ 60: $P(C|G \cap A)$
  - $\Omega = \{\text{f, m}\} \times \mathbb{N} \times \{+, -\}$
  - $C = \{(\_, \_, +) \in \Omega\} \quad G = \{(f, \_, \_) \in \Omega\} \quad A = \{(\_, a, \_) \in \Omega \mid a \geq 60\}$
  - naming triples with features (gender, age, covid): $P(\text{covid}= + | \text{gender=f, age} \geq 60)$

Another example at **seeing-theory.brown.edu**

Conditional probability

DEFINITION. The *conditional probability of A given C* is given by:

$$\mathrm{P}(A \mid C) = \frac{\mathrm{P}(A \cap C)}{\mathrm{P}(C)},$$

provided $\mathrm{P}(C) > 0$.

Properties:

- $P(A|C) \neq P(C|A)$, in general
- $P(\Omega|C) = 1$
- if $A \cap B = \emptyset$ then $P(A \cup B|C) = P(A|C) + P(B|C)$

$P(\cdot|C)$ is a probability function

THE MULTIPLICATION RULE. For any events A and C:

$$\mathrm{P}(A \cap C) = \mathrm{P}(A \mid C) \cdot \mathrm{P}(C)\,.$$

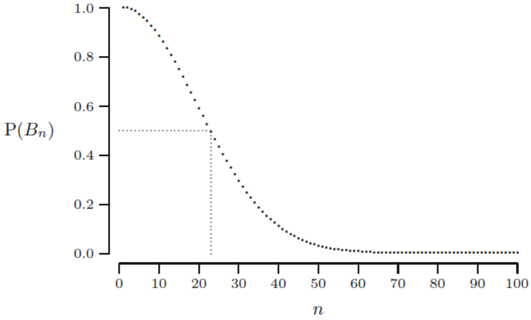
More generally, the **Chain Rule**:

$$P(A_1 \cap A_2 \cap A_3 \ldots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot \ldots \cdot P(A_n| \cap_{i=1}^{n-1} A_i)$$

Example: no coincident birthdays

- $B_n = \{n \text{ different birthdays}\}$
- For $n = 1$, $P(B_1) = 1$
- For $n > 1$,

$$
\begin{aligned}
P(B_n) &= P(B_{n-1}) \cdot P(\{\text{the } n\text{-th person's birthday differs from the other } n-1\} | B_{n-1}) \\
&= P(B_{n-1}) \cdot (1 - \frac{n-1}{365}) = \ldots = \prod_{i=1}^{n-1} (1 - \frac{i}{365})
\end{aligned}
$$



Example: case-based reasoning

Factory 1's light bulbs work for over 5000 hours in 99% of cases.
Factory 2's bulbs work for over 5000 hours in 95% of cases.
Factory 1 supplies 60% of the total bulbs on the market and Factory 2 supplies 40% of it.
What is the chance that a purchased bulb will work for longer than 5000 hours?

- $A = \{\text{bulbs working for longer than 5000 hours}\}$
- $C = \{\text{bulbs made by Factory 1}\}$, hence $C^c = \{\text{bulbs made by Factory 2}\}$
- Since $A = (A \cap C) \cup (A \cap C^c)$ with $(A \cap C)$ and $(A \cap C^c)$ disjoint:

$$P(A) = P(A \cap C) + P(A \cap C^c)$$

- and then by the multiplication rule:

$$P(A) = P(A|C) \cdot P(C) + P(A|C^c) \cdot P(C^c)$$

Answer: $P(A) = 0.99 \cdot 0.6 + 0.95 \cdot 0.4 = 0.974$

The law of total probability

THE LAW OF TOTAL PROBABILITY. Suppose $C_1, C_2, \ldots, C_m$ are disjoint events such that $C_1 \cup C_2 \cup \cdots \cup C_m = \Omega$. The probability of an arbitrary event A can be expressed as:

$$\mathrm{P}(A) = \mathrm{P}(A \mid C_1)\mathrm{P}(C_1) + \mathrm{P}(A \mid C_2)\mathrm{P}(C_2) + \cdots + \mathrm{P}(A \mid C_m)\mathrm{P}(C_m).$$

- **Intuition:** case-based reasoning

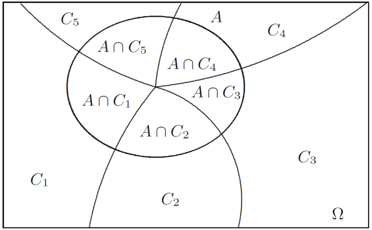


Fig. 3.2. The law of total probability (illustration for $m = 5$).

Exercise: Prisoners and guard dilemma

3 prisoners, 2 of which will be released.

$A1 = \{$ Prisoner 1 is released $\}$, $A2 = \{$ Prisoner 2 is released $\}$, $A3 = \{$ Prisoner 3 is released $\}$

$$P(A1^c \cap A2 \cap A3) = P(A1 \cap A2^c \cap A3) = P(A1 \cap A2 \cap A3^c) = 1/3$$

You are Prisoner 1:

- at your's present state of knowledge, the probability of being released is $2/3$
  - $P(A1) = P(A1 \cap A2^c \cap A3) + P(A1 \cap A2 \cap A3^c) = 2/3$
- if you ask a friendly guard to tell you who is the prisoner other than yourself that will be released, your probability of being released will become $1/2$
  - $P(A1|A2) = P(A1 \cap A2)/P(A2) = (1/3)/(2/3) = 1/2$

What is wrong with this line of reasoning?

- $P(A1) = P(A1|A2)P(A2) + P(A1|A3)P(A3) - P(A1|A2 \cap A3)P(A2 \cap A3) = 1/2 \cdot 2/3 + 1/2 \cdot 2/3 - 0 \cdot 1/3 = 2/3$

Independence of events

Intuition: whether one event provides any information about another.

Independence

An event A is independent of B, if $P(B) = 0$ or

$$P(A|B) = P(A)$$

- For $P(R|L) = {}^4/_7 \neq {}^8/_{12} = PR(R)$ - knowing Anna was born in a long month change the probability she was born in a month with 'r'!
- Tossing 2 coins:
  - A_1 is "H on toss 1" and A_2 is "H on toss 2"
  - $P(A_1) = P(A_2) = {}^1/_2$
  - $P(A_2|A_1) = P(A_2 \cap A_1)/P(A_1) = {}^1/_4/{}^1/_2 = {}^1/_2 = P(A_1)$
- Physical and stochastic independence
- Properties:
  - A independent of B iff $P(A \cap B) = P(A) \cdot P(B)$
  - A independent of B iff B independent of A *[Symmetry]*
  - A independent of B iff A^c independent of B

Conditional independence of events

Intuition: whether one event provides any information about another given a third event occurred. Technically, consider $P(\cdot|C)$ in independence.

Conditional independence

An event A is conditionally independent of B given C such that $P(C) > 0$, if $P(B|C) = 0$ or

$$P(A|B \cap C) = P(A|C)$$

- Properties:
  - A conditionally independent of B iff $P(A \cap B|C) = P(A|C) \cdot P(B|C)$
  - A conditionally independent of B iff B conditionally independent of A *[Symmetry]*
- **Exercise at home.** Prove or disprove:
  - If A is independent of B then A is conditionally independent of B given C

Independence of two or more events

INDEPENDENCE OF TWO OR MORE EVENTS. Events A_1, A_2, ..., A_m are called independent if

$$P(A_1 \cap A_2 \cap \cdots \cap A_m) = P(A_1)\,P(A_2)\cdots P(A_m)$$

and this statement *also* holds when any number of the events A_1, ..., A_m are replaced by their complements throughout the formula.

Alternative definition

Events $A_1, A_2, \ldots, A_m$ are called independent if for every $J \subseteq \{1, \ldots, m\}$:

$$P(\bigcap_{i \in J} A_i) = \prod_{i \in J} P(A_i)$$

- **Exercise at home**: show the two definitions are equivalent

Independence of two or more events

Alternative definition

Events $A_1, A_2, \ldots, A_m$ are called independent if for every $J \subseteq \{1, \ldots, m\}$:

$$P(\bigcap_{i \in J} A_i) = \prod_{i \in J} P(A_i)$$

- It is **stronger** than **pairwise independence**

$$P(A_i \cap A_j) = P(A_i) \cdot P(A_j) \text{ for } i \neq j \in \{1, \ldots, m\}$$

- Example: what is the probability of at least one head in the first 10 tosses of a coin?
  $A_i = \{\text{head in } i\text{-th toss}\}$

$$P(\bigcup_{i=1}^{10} A_i) = 1 - P(\bigcap_{i=1}^{10} A_i^c) = 1 - \prod_{i=1}^{10} P(A_i^c) = 1 - \prod_{i=1}^{10} (1 - P(A_i))$$