Statistical Methods for Data Science

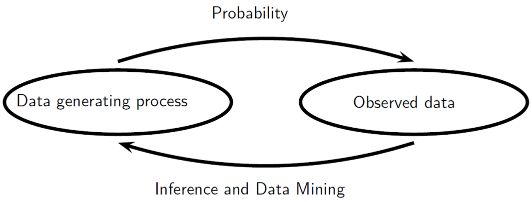
Lesson 01 - Introduction

Salvatore Ruggieri

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

Why Statistics

We need grounded means for reasoning about data science mechanisms.



What will I learn?

- Probability: properties of data generated according to a known randomness model
- Statistics: properties of a randomness model that could have generated given data
- Simulation and R

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: what is Ω?

  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 lies in the set A.
  - If the outcome is Jan then L occurs

Probability functions

A **probability distribution** 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]*
- 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 ${}^{31\cdot 7}/_{365}$?

Properties of probability functions

- Assigning probability is **NOT** an easy task.
  - **Frequentist** interpretation: probability measures a "*proportion of outcomes*".
  - **Bayesian** (or epistemological) interpretation: probability measures a "*degree of belief*".

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

Products of sample spaces

An experiment made of multiple sub-experiments

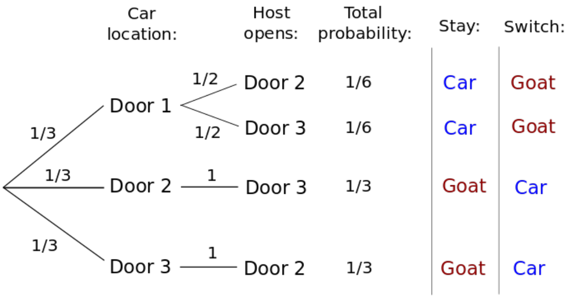
- Eg., $\Omega = \{\text{ H, T }\} \times \{\text{ H, T }\} = \{(H,H),(H,T),(T,H),(T,T)\}$
- $P((H,H)) = {}^1\!/_4$

In general:

- $\Omega = \Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) \mid \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$
- $P((a_1, a_2)) = {}^1\!/_{|\Omega_1|} \cdot {}^1\!/_{|\Omega_2|}$ *[Uniform function over independent experiments]*

The Monty Hall problem

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



Exercise at home: generalize to n doors where host opens $n-2$ doors with goats.

A (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.

- 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'?

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

- **Intuition:** probability of an event in the restricted sample space $\Omega \cap L$

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

Conditional probability

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

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

provided $P(C) > 0$.

Properties:

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

THE MULTIPLICATION RULE. For any events A and C:

$$P(A \cap C) = P(A \mid C) \cdot P(C).$$

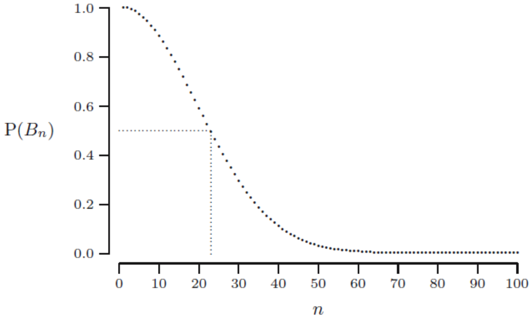
More generally, the **Chain Rule**:

$$P(A_1 \cap A_2 \cap A_3 \ldots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_2, A_1) \cdot \ldots P(A_n|A_{n-1}, \ldots, A_1)$$

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:

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

- **Intuition:** case-based reasoning

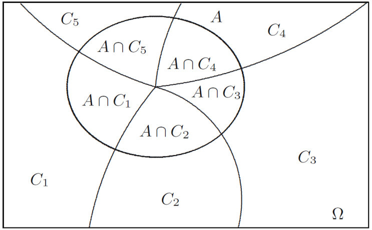


Fig. 3.2. The law of total probability (illustration for $m = 5$).

Testing for Covid-19

A new test for Covid-19 (or Mad-Cow desease, or drug use) has been developed.

- $+ = \{$ people tested positive $\}$ $\quad - = \{$ people tested negative $\} = +^c$
- $C = \{$ people with Covid-19 $\}$ $\quad C^c = \{$ people without Covid-19 $\}$

In lab experiments, people with and without Covid-19 tested

- $P(+|C) = 0.99$ *[Sensitivity/Recall/True Positive Rate]*
- $P(-|C^c) = 0.99$ *[Specificity/True Negative Rate]*

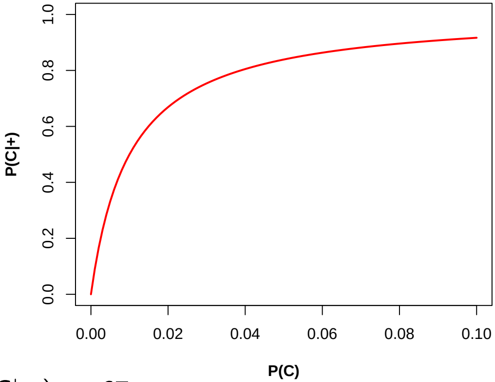
What is the probability I really have Covid-19 given that I tested positive? *[Precision]*

$$P(C|+) = \frac{P(C \cap +)}{P(+)} = \frac{P(+|C) \cdot P(C)}{P(+)} = \frac{P(+|C) \cdot P(C)}{P(+|C) \cdot P(C) + P(+|C^c) \cdot P(C^c)}$$

$$P(C|+) = \frac{0.99 \cdot P(C)}{0.99 \cdot P(C) + 0.01 \cdot (1 - P(C))}$$

Testing for Covid-19

$P(C)$, the probability of having Covid-19, **is unknown**. Let's plot $P(C|+)$ over $P(C)$:



- For $P(C) = 0.02$, $P(C|+) = .67$
- For $P(C) = 0.06$, $P(C|+) = .86$
- For $P(C) = 0.10$, $P(C|+) = .92$

Bayes' Rule

BAYES' RULE. Suppose the events $C_1, C_2, \ldots, C_m$ are disjoint and $C_1 \cup C_2 \cup \cdots \cup C_m = \Omega$. The conditional probability of C_i, given an arbitrary event A, can be expressed as:

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

- It follows from $P(C_i|A) = \frac{P(A|C_i) \cdot P(C_i)}{P(A)}$ and the law of total probability
- Useful when:
  - $P(C_i|A)$ not easy to calculate
  - while $P(A|C_j)$ and $P(C_j)$ are known for $j = 1, \ldots, m$
  - E.g., in classification problems (see Bayesian classifiers from Data Mining)
- $P(C_i)$ is called the *prior* probability
- $P(C_i|A)$ is called the *posterior* probability (after seeing event A)

Independence of events

Intuition: whether one event provides any information about another.

DEFINITION. An event A is called *independent of* B if

$$P(A \mid B) = P(A).$$

- For $P(C) = 0.10$, $P(C|+) = .92$ - knowing test result changes prob. of being infected!
- Tossing 2 coins:
  - A_1 is "H on toss 1" and A_2 is "H on toss 2"
  - $P(A_1) = P(A_2) = 1/2$
  - $P(A_2|A_1) = P(A_2 \cap A_1)/P(A_1) = 1/4/1/2 = 1/2 = P(A_1)$
- 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]*

Independence of two or more events

INDEPENDENCE OF TWO OR MORE EVENTS. Events $A_1, A_2, \ldots, A_m$ 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 $A_1, \ldots, A_m$ are replaced by their complements throughout the formula.

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

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

Independence of two or more events

Alternative definition

Events $A_1, A_2, \dots, A_m$ are called independent if

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

for every $J \subseteq \{1, \dots, m\}$

- **Exercise at home**: show the two definitions are equivalent
- Example: what is the probability of at least one head in the first 10 tosses of a coin? $A_i = \{\text{head in } i\text{-th toss}\}$

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