Statistique et Informatique (LU3IN005)

2020-2021

Nicolas Baskiotis - Pierre-Henri Wuillemin
prenom.nom@lip6.fr

Sorbonne Université
Laboratoire d'Informatique de Paris 6 (LIP6)

Cours 1 : Probabilités sur des ensembles discrets et dénombrements

Description de l'UE

Objectifs du cours

- Introduction aux domaines :
  - de la théorie des probabilités,
  - de la statistique,
- donner des exemples de leurs applications (en informatique),
- pratiquer les concepts introduits sur des exemples $\longrightarrow$ mini-projets.

Organisation

- Calcul des probabilités (Nicolas Baskiotis – cours 1 à 5) :
  - introduction aux probabilités, conditionnement, marginalisation
  - probabilités discrètes, continues
  - Loi des grands nombres et applications
  - études de différentes lois
- L'inférence statistique (Pierre-Henri Wuillemin – cours 6 à 11) :
  - recueil et analyse des données,
  - estimation, tests et validation,
  - Processus séquentiels.

Description de l'UE (2)

Informations pratiques

- Site Web : Moodle
- Organisation en *mini-projets* (en python) :
  - TME 1-4 : projet 1,
  - TME 5-7 : projet 2,
  - TME 8-11 : projet 3 (+ révisions).

Evaluation

- Les trois mini-projets sont notés
- les mini-projets comptent dans la note finale *dans tous les cas* !
- un partiel et un examen.

Note finale

- Partiel : 33$\%$
- Examen : 33$\%$
- Projets : 33$\%$

Plan

De quoi parle ce cours . . .

- Qu'est ce que la chance ? le hasard ? le aléas ?
- Comment mesurer le hasard ?
- Peut on l'utiliser ? Comment ?
- Comment modéliser des phénomènes aléatoires ?
- Comment les étudier ?
- Comment les caractériser ?
- Que peut-on prédire ?
- . . .

Définitions

Probabilités

- domaine des mathématiques qui étudie des phénomènes *aléatoires*,
- fournit des outils pour étudier les *expériences aléatoires* :
  des expériences qui, répétées dans les mêmes conditions, ne donnent pas nécessairement le même résultat,
- modélise à l'aide de ces outils les processus aléatoires pour en étudier le fonctionnement et les résultats.

Exemple : modéliser un lancé de dé

Statistique

- domaine des mathématiques qui étudie la collecte, l'analyse, et l'interprétation de données
- permet d'établir des protocoles expérimentaux et d'analyser les résultats
- permet d'inférer des conclusions sur les processus aléatoires.

Exemple : à partir d'un certain nombre de lancés, le dé est-il biaisé ?

Une (très) petite histoire des probabilités et statistiques

- $-XVI^e$ siècle : préhistoire, (Cardan 1501-1576, Galilée 1564-1642)
- XVI^e-$XVII^e$: la découverte du domaine
  - Fermat (161x-1665), Pascal (1623-1662)
  - Huyghens (1629-1695)
- $XVIII^e - XIX^e$: developpement et premières applications scientifiques
  - Montmort (1678-1719), de Moivre (1667-1754)
  - la dynastie Bernoulli : Jacob (1657-1705), Jean (1667-1748), Daniel (1700-1782), Nicolas (1687-1759)
  - Bayes (1700-1761)
  - Buffon (1707-1788), Simpson (1710-1761), D'alembert (1717-1783)
  - Lagrange (1736-1813), Laplace (1749-1827), Poisson (1781-1840)
- $XIX^e - XX^e$: théorie de la mesure, axiomatisation, applications multiples
  - Tchebychev (1821-1894), Emile Borel (1871-1956), Johann Radon (1887-1956), Paul Lévy (1886-1971), Andreï Kolmogorov (1903-1987)
  - Gibbs (1839-1903), Boltzmann (1844-1906), Poincaré (1854-1912) , Pearson (1857-1936), Markov (1856-1922)
- XX^e- : vraie reconnaissance du domaine, consolidations théoriques et développement de multiples applications.

Exemples d'application : en informatique fondamentale

- Algorithmique : *tri rapide*
  $\longrightarrow$ meilleure performance “en moyenne” que les autres tris ;
  “en moyenne” $\approx$ les valeurs dans le tableau initial sont aléatoires.

- Calcul d'une coupe minimale dans un graphe : algorithme de Karger
- structure de données : *Table de Hashage*
  $\longrightarrow$ propriété souhaitée de `hashCode` : donner des valeurs différentes aux différentes chaînes de caractères stockées dans la table de hachage.
  $\longrightarrow$ nécessite un modèle (probabiliste) des chaînes de caractères qui seront stockées.
- Compression de données : codage d'Huffman

Cryptographie et cryptanalyse

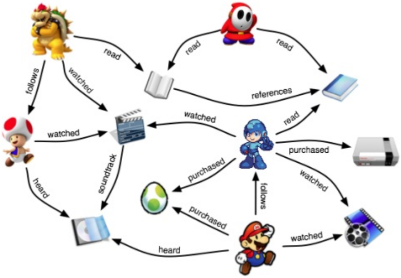
Enigma : machine de cryptage allemande pendant la Seconde Guerre mondiale.

Le décryptage des messages par les alliés a été facilité par un mauvais algorithme de génération de *permutations* aléatoires.

La sécurité des communications sur Internet est gérée par des algorithmes de cryptographie.

Les algorithmes de cryptographie utilisent des générateurs de nombres aléatoires.

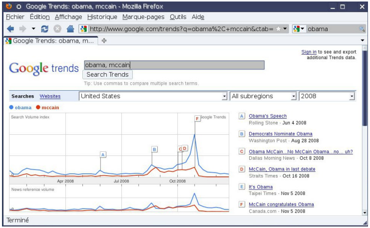
Réciproquement : les cryptanalystes cherchent les *régularités* (déviations par rapport à l'aléatoire) dans les textes cryptés.



Systèmes de reco. (Netflix, Amazon, . . .) :
Les clients qui ont acheté/vu ...
ont aussi acheté/vu ...

Analyses statistiques des achats/recherches des différents produits

Ciblage publicitaire

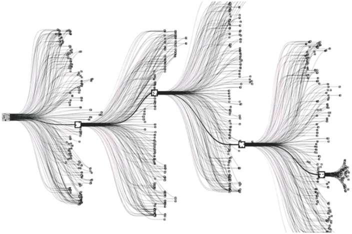


Google Trends : analyse des requêtes effectuées par les utilisateurs de Google.

Applications possibles : suivi des intérêts dans une population, détection des épidémies, ...

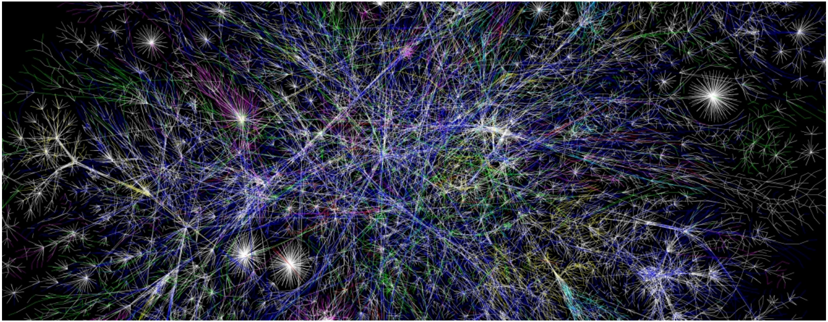
IA et jeux

- Exploration efficiente des possibilités dans une combinatoire élevée,
- Modélisation de l'adversaire (Poker par exemple)
- Matchmaking

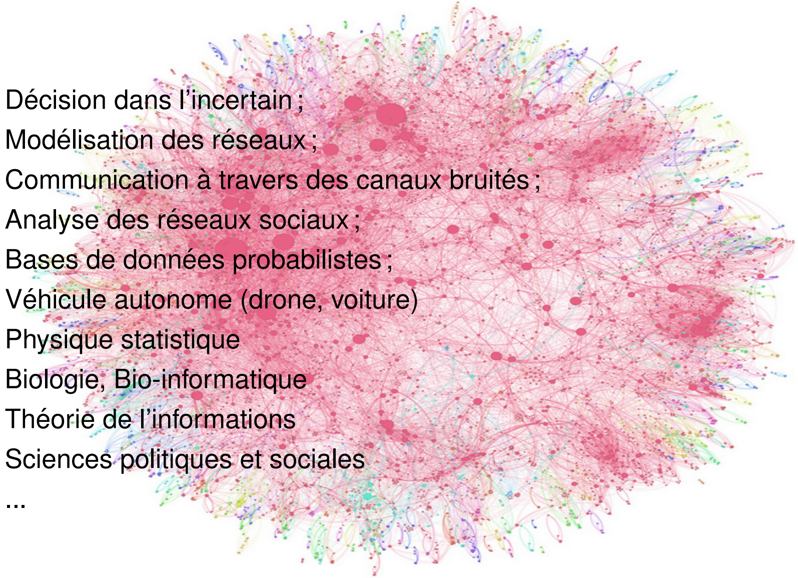


Machine Learning (Apprentissage statistique)

- Apprentissage bayésien
- Réseaux de neurones
- Applications en
  - Classification (image, texte, . . .)
  - traduction automatique,
  - génération automatique (musique, textes)
  - Moteur de recherche
  - Interface cerveau-machine (BCI)
  - Recommendation



Et bien d'autres...



- Décision dans l'incertain ;
- Modélisation des réseaux ;
- Communication à travers des canaux bruités ;
- Analyse des réseaux sociaux ;
- Bases de données probabilistes ;
- Véhicule autonome (drone, voiture)
- Physique statistique
- Biologie, Bio-informatique
- Théorie de l'informations
- Sciences politiques et sociales
- ...

Plan

Une probabilité ?

Trois sachets, un croissant . . .

- Pourquoi a-t-on une chance sur 3 de trouver le croissant ?
- Est-ce toujours le cas ?
- Que veut dire *chance* dans ce contexte ?

Une probabilité ?

Trois sachets, un croissant . . .

- Pourquoi a-t-on une chance sur 3 de trouver le croissant ?
- Est-ce toujours le cas ?
- Que veut dire *chance* dans ce contexte ?

La probabilité d'un événement

- c'est la fréquence d'apparition de l'événement, le nombre de fois où il apparaît rapporté au nombre d'expériences.
- Notions (intuitives) :
  - d'expérience : un cadre bien défini, avec des conditions initiales et un ensemble de résultats déterminés
  - d'événement : un résultat de l'expérience
  - de répétition : l'expérience peut être reproduite dans les mêmes conditions !

Un peu plus compliqué

30 croissants, 30 pains au chocolat, 20 pains aux raisins, 10 pains au lait, 10 chaussons

Qu'est ce qui est équiprobable ?
Quelle est la probabilité :

- d'un pain au chocolat ?
- si 5 croissants ont été tirés avant ?
- qu'un pain soit tiré ?

Plusieurs tirages

Qu'elle est la probabilité :

- de tirer aucun croissants au bout de deux tirages ? au bout de trois ?
- de tirer au moins un croissant ?

Et le café ?

Une machine à café a un seul bouton

Elle peut faire un café court ou long, avec du sucre et/ou du lait.

- Quelle est la probabilité d'avoir un café court sucré ?

Et le café ?

Une machine à café a un seul bouton

Elle peut faire un café court ou long, avec du sucre et/ou du lait.

- Quelle est la probabilité d'avoir un café court sucré ?

$\Rightarrow$ Telle quelle, la question a autant de sens que "quel est l'âge du client" ...

Et le café ?

Une machine à café a un seul bouton

Elle peut faire un café court ou long, avec du sucre et/ou du lait.

- Quelle est la probabilité d'avoir un café court sucré ?

$\Rightarrow$ Telle quelle, la question a autant de sens que "quel est l'âge du client" ...

De quoi a-t-on besoin pour répondre à la question ?

Et le café ?

Une machine à café a un seul bouton

Elle peut faire un café court ou long, avec du sucre et/ou du lait.

- Quelle est la probabilité d'avoir un café court sucré ?

⇒ Telle quelle, la question a autant de sens que "quel est l'âge du client" . . .

De quoi a-t-on besoin pour répondre à la question ?

- Idéalement, la probabilité de chaque événement *(court, lait, sucré)*, *(court, lait, non sucré)*, *(long, lait, sucré)*, *(long, lait, non sucré)*, *(court,non lait, sucré)*, *(court, non lait, non sucré)*, *(long, non lait, sucré)*, *(long, non lait, non sucré)*
- Peut-on tout calculer ?

⇒ Oui, ce sont tous les événements , tous les résultats attendus

- Exemple : probabilité de *court lait* = *court, lait, sucré* + *court, lait, non sucré*

Et le café ?

Une machine à café a un seul bouton

Elle peut faire un café court ou long, avec du sucre et/ou du lait.

- Quelle est la probabilité d'avoir un café court sucré ?

$\Rightarrow$ Telle quelle, la question a autant de sens que "quel est l'âge du client" . . .

De quoi a-t-on besoin pour répondre à la question ?

- Probabilités des événement *lait*, *sucré* suffisent ?

Et le café ?

Une machine à café a un seul bouton

Elle peut faire un café court ou long, avec du sucre et/ou du lait.

- Quelle est la probabilité d'avoir un café court sucré ?

$\Rightarrow$ Telle quelle, la question a autant de sens que "quel est l'âge du client" ...

De quoi a-t-on besoin pour répondre à la question ?

- Probabilités des événement *lait*, *sucré* suffisent ?

$\Rightarrow$ non, on ne sait pas court ou long ...

- Et avec *court* en plus ?

Et le café ?

Une machine à café a un seul bouton

Elle peut faire un café court ou long, avec du sucre et/ou du lait.

- Quelle est la probabilité d'avoir un café court sucré ?

$\Rightarrow$ Telle quelle, la question a autant de sens que "quel est l'âge du client" ...

De quoi a-t-on besoin pour répondre à la question ?

- Probabilités des événement *lait*, *sucré* suffisent ?

$\Rightarrow$ non, on ne sait pas court ou long ...

- Et avec *court* en plus ?

$\Rightarrow$ oui, *court* et *long* sont complémentaires !

- probabilité de *court* = 1 - *long*

Plan

Probabilités sur les ensembles discrets

Pour modéliser une expérience aléatoire :

Nous avons besoin des notions de :

- *événement élémentaire* : un résultat simple non composé de l'expérience
- *événement* : un résultat simple ou composé de plusieurs événements élémentaires de l'expérience
- *univers* : l'ensemble de tous les résultats possibles

Formalisation

- Soit Ω, un ensemble dénombrable, appelé univers,
  - Ω représente l'ensemble de tous les résultats possibles d'une expérience aléatoire
- un élément $\omega \in \Omega$ est *un événement élémentaire*,
- un sous-ensemble *E* de Ω est un *événement*.

Probabilités sur les ensembles discrets

Pour modéliser une expérience aléatoire :

Nous avons besoin des notions de :

- *événement élémentaire* : un résultat simple non composé de l'expérience
- *événement* : un résultat simple ou composé de plusieurs événements élémentaires de l'expérience
- *univers* : l'ensemble de tous les résultats possibles

Formalisation

- Soit Ω, un ensemble dénombrable, appelé univers,
  - Ω représente l'ensemble de tous les résultats possibles d'une expérience aléatoire
- un élément $\omega \in \Omega$ est *un événement élémentaire*,

$\Rightarrow$ Un croissant, un pain au chocolat, un pain aux raisins, un café court sucré sans lait

- un sous-ensemble *E* de Ω est un *événement*.

$\Rightarrow$ Un pain, quelque chose sans beurre, un café court

Probabilités sur des ensembles discrets (2)

Mesure de probabilité (ou distribution) : caractérise l'aléa

- elle définie une probabilité pour chaque événement, qui correspond à la fréquence d'apparition de l'événement, entre 0 et 1
- la probabilité de l'univers est de 1 : au moins un des événements de l'univers se réalise lors d'une expérience
- la probabilité de deux événements qui ne peuvent arriver en même temps (incompatibles) est la somme de leur probabilité.
- la probabilité qu'aucun événement de l'univers n'arrive est donc de 0.

Elle est entièrement définie par les probas des événements élémentaires.

Définitions

Soit l'univers Ω, ensemble discret.

Une mesure de probabilité sur Ω est une fonction $P : \mathcal{P}(\Omega) \to [0, 1]$ tq :

- $P(\Omega) = 1$ (Ω est l'événement certain),
- pour tout événement E, $P(E) \geq 0$,
- Pour toute suite $(E_i)_{i \in \mathbb{N}}$ d'événements deux à deux disjoints (*incompatibles*, $E_i \cap E_j = \emptyset, i \neq j$) : $P(\bigcup_i E_i) = \sum_i P(E_i)$.

avec Ω ensemble discret et $\mathcal{P}(\Omega)$ l'ensemble des sous-ensembles de Ω.

Remarques

- Ces 3 axiomes définissent le cadre des probabilités discrètes.
- Les événements élémentaires sont forcément incompatibles.
- La définition de la probabilité comme étant la limite de la fréquence du nombre d'apparition de l'événement en répétant à l'infini l'expérience n'est pas considérée comme axiome (pourquoi ?), mais est déduite de ces 3 axiomes.

Probabilités sur des ensembles discrets (3)

Définitions

- Fonction de masse p associée à P : $\forall\omega \in \Omega, p(\omega) = P(\{\omega\})$
  (rappel : ω événement élémentaire !)
- Pour tout événement E :

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

- La probabilité uniforme sur un univers *fini* Ω est définie par la fonction de masse : $p(\omega) = \frac{1}{card(\Omega)}$. Dans ce cas, $\forall E \in \mathcal{P}(\Omega),\ P(E) = \frac{card(E)}{card(\Omega)}$

Interprétation : si on répète (indéfiniment) l'expérience aléatoire

- le résultat de l'expérience sera ω avec une fréquence de $P(\{\omega\})$,
- un événement E se produit avec une fréquence $P(E)$
  $\rightarrow$ le résultat appartient à l'ensemble E avec une fréquence $P(E)$.

Probabilités sur des ensembles discrets (4)

Propriétés

- $P(\varnothing) = 0$, ($\varnothing$ est l'événement impossible)
  ne pas avoir de café ...
- $P(\bar{E}) = 1 - P(E)$ ($\bar{E}$: complémentaire de E dans Ω)
  café court et café long
- $P(E \cup F) = P(E) + P(F) - P(E \cap F)$
  P(*café au lait* ou *café sucré*) = P(*café au lait*)+P(*café sucré*) - P(*café au lait et sucré*)
- $E \subset F \Rightarrow P(F) = P(F \backslash E) + P(E) \Rightarrow P(E) \leq P(F)$
  ($F \backslash E$: ensemble des éléments de F qui ne sont pas dans E),
  La probabilité d'un café au lait sucré est inférieure à celle d'un café au lait
- $P(\bigcup_i E_i) \leq \sum_i P(E_i)$

Comment résoudre un problème de probabilité discrète ?

4 étapes :

1. Bien définir l'univers ! une majorité d'erreurs/paradoxes proviennent d'une confusion sur les événements élémentaires.
2. Déterminer le(s) événement(s) d'intérêt(s).
3. Assigner les probabilités $\Rightarrow$ la plupart du temps, savoir compter ! (plus difficile qu'il n'y paraît, cf dénombrements).
4. Calculer la probabilité des événements d'intérêts (généralement simple, par addition/soustraction des probabilités définies précédemment).

Prochaine partie : les dénombrements, ou comment apprendre à compter

Plan

1. Introduction et exemples d'applications
2. Probabilités discrètes : introduction
3. Probabilités discrètes : axiomatique
4. Dénombrements

Principes élémentaires

Dénombrer les résultats de tuples d'expériences

- Soit 2 expériences telles que il y ait n_1 résultats possibles pour la première expérience et pour chacun de ces résultats n_2 résultats pour la deuxième. Alors il y a en tout $n_1 \times n_2$ résultats possibles.
- Généralisation : Soit r expériences et $(n_1, \ldots, n_r)$ le nombre de résultats possibles pour chacune de ces expériences indépendamment des autres. Alors le nombre de résultats possibles est $n_1 \times n_2 \times \ldots \times n_r$.

Exemples

- Nombre de résultats pour le jet d'un dé suivi du jet d'une pièce ?
- Nombre de plaques d'immatriculation formé de deux lettres, 3 chiffres puis deux lettres ?
- Nombre de configurations au jeu de go (plateau de 19×19) ?
- Combien de questions binaires (oui/non) sont nécessaires pour différencier 10 millions de personnes ?

Principes élémentaires

Dénombrer les résultats de tuples d'expériences

- Soit 2 expériences telles que il y ait n_1 résultats possibles pour la première expérience et pour chacun de ces résultats n_2 résultats pour la deuxième. Alors il y a en tout $n_1 \times n_2$ résultats possibles.
- Généralisation : Soit r expériences et $(n_1, \ldots, n_r)$ le nombre de résultats possibles pour chacune de ces expériences indépendamment des autres. Alors le nombre de résultats possibles est $n_1 \times n_2 \times \ldots \times n_r$.

Exemples

- Nombre de résultats pour le jet d'un dé suivi du jet d'une pièce ?
  6 résultats pour le dé, 2 pour la pièce, donc $6 * 2$.
- Nombre de plaques d'immatriculation formé de deux lettres, 3 chiffres puis deux lettres ?
- Nombre de configurations au jeu de go (plateau de 19×19) ?
- Combien de questions binaires (oui/non) sont nécessaires pour différencier 10 millions de personnes ?

Principes élémentaires

Dénombrer les résultats de tuples d'expériences

- Soit 2 expériences telles que il y ait n_1 résultats possibles pour la première expérience et pour chacun de ces résultats n_2 résultats pour la deuxième. Alors il y a en tout $n_1 \times n_2$ résultats possibles.
- Généralisation : Soit r expériences et $(n_1, \ldots, n_r)$ le nombre de résultats possibles pour chacune de ces expériences indépendamment des autres. Alors le nombre de résultats possibles est $n_1 \times n_2 \times \ldots \times n_r$.

Exemples

- Nombre de résultats pour le jet d'un dé suivi du jet d'une pièce ?
- Nombre de plaques d'immatriculation formé de deux lettres, 3 chiffres puis deux lettres ?
  $26 * 26 * 10 * 10 * 10 * 26 * 26$
- Nombre de configurations au jeu de go (plateau de 19×19) ?
- Combien de questions binaires (oui/non) sont nécessaires pour différencier 10 millions de personnes ?

Principes élémentaires

Dénombrer les résultats de tuples d'expériences

- Soit 2 expériences telles que il y ait n_1 résultats possibles pour la première expérience et pour chacun de ces résultats n_2 résultats pour la deuxième. Alors il y a en tout $n_1 \times n_2$ résultats possibles.
- Généralisation : Soit r expériences et $(n_1, \ldots, n_r)$ le nombre de résultats possibles pour chacune de ces expériences indépendamment des autres. Alors le nombre de résultats possibles est $n_1 \times n_2 \times \ldots \times n_r$.

Exemples

- Nombre de résultats pour le jet d'un dé suivi du jet d'une pièce ?
- Nombre de plaques d'immatriculation formé de deux lettres, 3 chiffres puis deux lettres ?
- Nombre de configurations au jeu de go (plateau de 19×19) ?
  3 états possibles par case, donc $3^{19\times 19}$
- Combien de questions binaires (oui/non) sont nécessaires pour différencier 10 millions de personnes ?

Principes élémentaires

Dénombrer les résultats de tuples d'expériences

- Soit 2 expériences telles que il y ait n_1 résultats possibles pour la première expérience et pour chacun de ces résultats n_2 résultats pour la deuxième. Alors il y a en tout $n_1 \times n_2$ résultats possibles.
- Généralisation : Soit r expériences et $(n_1, \ldots, n_r)$ le nombre de résultats possibles pour chacune de ces expériences indépendamment des autres. Alors le nombre de résultats possibles est $n_1 \times n_2 \times \ldots \times n_r$.

Exemples

- Nombre de résultats pour le jet d'un dé suivi du jet d'une pièce ?
- Nombre de plaques d'immatriculation formé de deux lettres, 3 chiffres puis deux lettres ?
- Nombre de configurations au jeu de go (plateau de 19×19) ?
- Combien de questions binaires (oui/non) sont nécessaires pour differencier 10 millions de personnes ?
  Soit n ce nombre, on veut $2^n = 10^7$, donc $n = 10^7 = 23.2$, donc 24.

Exemples d'application

Lancé simultané de trois dés

- Univers ? Nombre d'événements élémentaires ?
- Probabilité de l'événement E : *La somme des 3 chiffres est inférieure stricte à* 5 ?
- Quelle est la probabilité :
  - d'obtenir aucun 6 ?
  - que $i = j = k$?

Exemples d'application

Lancé simultané de trois dés

- Univers ? Nombre d'événements élémentaires ?
  Evénements élémentaires : $(i, j, k) \in \{1, \ldots 6\}^3$. L'univers (l'ensemble des événements possibles) est :

$$\begin{aligned}\Omega &= \{(i, j, k) \mid i \in \{1, .., 6\}, j \in \{1, .., 6\},\ k \in \{1, .., 6\}\} \\ &= \{(1, 1, 1), (1, 1, 2), ..., (1, 2, 1), (1, 2, 2) ..., (6, 6, 6)\}, \\ card(\Omega) &= 6^3 = 216\end{aligned}$$

  Si on considère chaque dé équilibré, alors chaque événement élémentaire est équiprobable :

$$\forall (i, j, k) \in \Omega, P((i, j, k)) = \frac{1}{216}$$

  $E = \{(1, 1, 1), (2, 2, 2), (3, 3, 3), (4, 4, 4), (5, 5, 5), (6, 6, 6)\}$ représente l'événement : *les 3 dés sont égaux*.
- Probabilité de l'événement E : *La somme des 3 chiffres est inférieure stricte à* 5 ?

Exemples d'application

Lancé simultané de trois dés

- Univers ? Nombre d'événements élémentaires ?
- Probabilité de l'événement E : *La somme des 3 chiffres est inférieure stricte à* 5 ?
  $P(E) = \frac{4}{256}$. En effet : $E = \{(1,1,1),(1,1,2),(1,2,1),(2,1,1)\}$.
- Quelle est la probabilité :
  - d'obtenir aucun 6 ?
  - que $i = j = k$?