Statistique et Informatique (LI323)

Nicolas Baskiotis - Hugues Richard

Université Pierre et Marie Curie (UPMC)
Laboratoire d'Informatique de Paris 6 (LIP6)
(Supports de cours : N. Usunier)

Cours 1 :
Probabilités sur des ensembles discrets et dénombrements

Plan

Description de l'UE

Objectifs du cours

- Présenter les outils de base de:
  - la théorie des probabilités,
  - la statistique,
- donner des exemples de leur application en informatique,
- manipuler quelques algorithmes issus de ces domaines
  $\longrightarrow$ mini-projets.

Organisation

- Calcul des probabilités (Nicolas Baskiotis – cours 1 à 6) :
  - exposé simple sur la théorie des probabilités,
  - quelques applications en informatique.
- L'inférence statistique (Hugues Richard – cours 7 à 11) :
  - recueil et analyse des données,
  - estimation, tests et validation.

Description de l'UE (2)

Informations pratiques

- Site Web : `http://www-connex.lip6.fr/~baskiotisn/`
- Organisation en *mini-projets*:
  - TD/TME 3-5 : projet apprentissage,
  - TD/TME 6-8 : projet bioinformatique,
  - TD/TME 9-11 : projet réseaux.

Evaluation

- Les trois mini-projets des TMEs 3 à 11 sont notés,
- les mini-projets comptent dans la note finale *dans tous les cas*,
- un examen à la fin (pas de partiel).

Evaluation

Calcul de la note finale

- Note de Contrôle Continu : sur 40,
- note d'Ecrit : sur 60,
- note finale (sur 20): $\max\left(\frac{Ecrit}{3}, \frac{CC+Ecrit}{5}\right)$.
- Calcul de la note de CC:

  moyenne (rapportée sur 40) des notes de mini-projets
- Calcul de la note d'Ecrit:

  0.5*(moyenne des projets+examen), rapportée sur 60.

Exemple

- un étudiant a eu 12, 13, et 14 aux mini-projets, et 11 à l'examen. Alors:
- note de CC : $1/3 * (12 + 13 + 14) = 13$ (donc $26/40$),
- note Ecrit : $1/2 * (13 + 11) = 12$ (donc $36/60$),
- note finale : $\max(\frac{36}{3}, \frac{36+26}{5}) = 12{,}4$.

Définitions

Probabilités

- La théorie des probabilités : domaine des mathématiques qui étudie les 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.

Statistique

- La statistique : domaine des mathématiques dans lequel on étudie la collecte, l'analyse, l'interprétation de données,
- en particulier, des données stockées dans les bases de données, sur le Web, ...

Plan

Algorithmique et structures de données

```java
import java.util.Arrays;

public class MyClass {
    Hashtable<String, Integer> myHashTable;
    ...

    public double[] sortedElements(double[] myArray){
        Arrays.sort(myArray);
        return myArray;
    }
}
```

- `public static void sort(double[] a)` : algorithme *tri rapide*
  $\longrightarrow$ meilleure performance "en moyenne" que les autres tris ;
  "en moyenne" $\approx$ les valeurs dans le tableau initial sont aléatoires.

- `Hashtable<String, V>` : utilise `int String.hashCode()`
  $\longrightarrow$ propriété souhaitée de `hashCode` : donner des valeurs différentes aux différentes `String` stockées dans la table de hachage.
  $\longrightarrow$ nécessite un modèle (probabiliste) des chaînes de caractères qui seront stockées.

Algorithmique et structures de données

```java
import java.util.Arrays;

public class MyClass {
    Hashtable<String, Integer> myHashTable;
    ...

    public double[] sortedElements(double[] myArray){
        Arrays.sort(myArray);
        return myArray;
    }
}
```

- `public static void sort(double[] a)` : algorithme *tri rapide*
  $\longrightarrow$ meilleure performance “en moyenne” que les autres tris ;
  “en moyenne” $\approx$ les valeurs dans le tableau initial sont aléatoires.

- `Hashtable<String, V>` : utilise `int String.hashCode()`
  $\longrightarrow$ propriété souhaitée de `hashCode` : donner des valeurs différentes aux différentes `String` stockées dans la table de hachage.
  $\longrightarrow$ nécessite un modèle (probabiliste) des chaînes de caractères qui seront stockées.

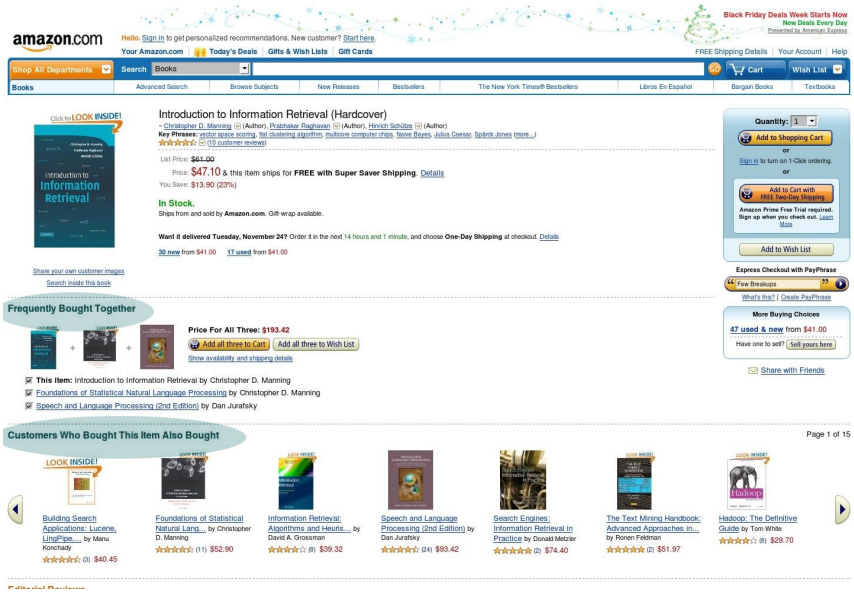
Algorithmique et structures de données

```java
import java.util.Arrays;

public class MyClass {
    Hashtable<String, Integer> myHashTable;
    ...

    public double[] sortedElements(double[] myArray){
        Arrays.sort(myArray);
        return myArray;
    }
}
```

- `public static void sort(double[] a)` : algorithme *tri rapide*
  $\longrightarrow$ meilleure performance "en moyenne" que les autres tris ;
  "en moyenne" $\approx$ les valeurs dans le tableau initial sont aléatoires.

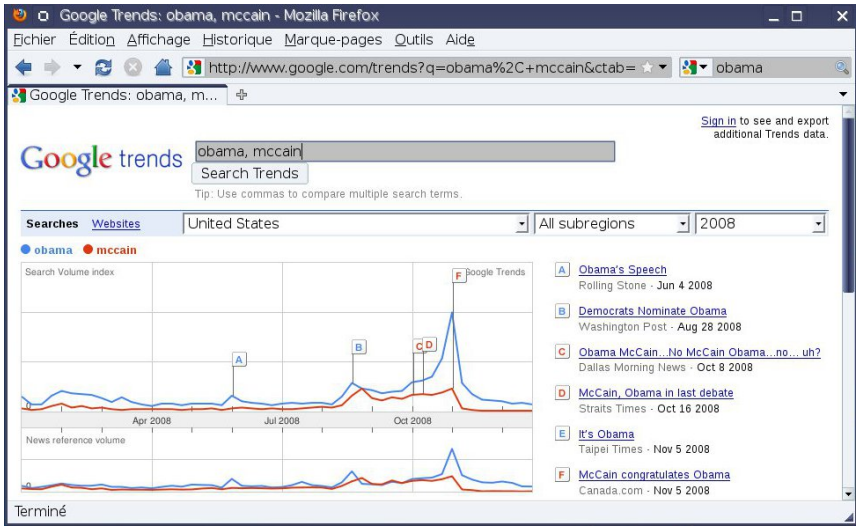
- `Hashtable<String, V>` : utilise `int String.hashCode()`
  $\longrightarrow$ propriété souhaitée de `hashCode` : donner des valeurs différentes aux différentes `String` stockées dans la table de hachage.
  $\longrightarrow$ nécessite un modèle (probabiliste) des chaînes de caractères qui seront stockées.

Fouille de données



Systèmes de recommandation :
Les clients qui ont acheté ...
ont aussi acheté ...

Fondés sur des analyses statistiques des achats/recherches des différents produits

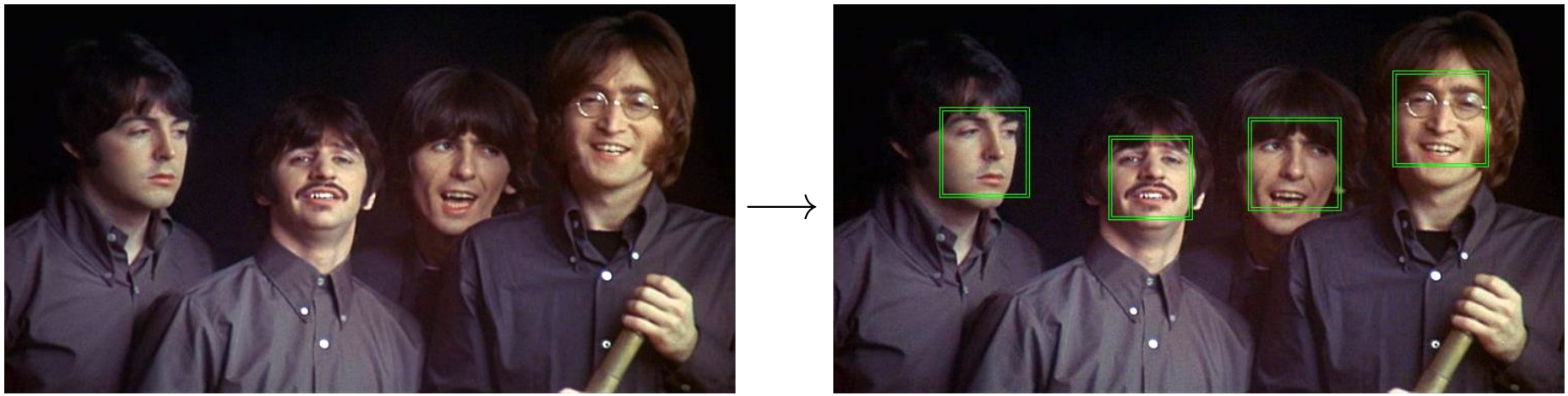


Google Trends : analyse des requêtes effectuées par les utilisateurs de Google.

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

Analyse prédictive/apprentissage automatique

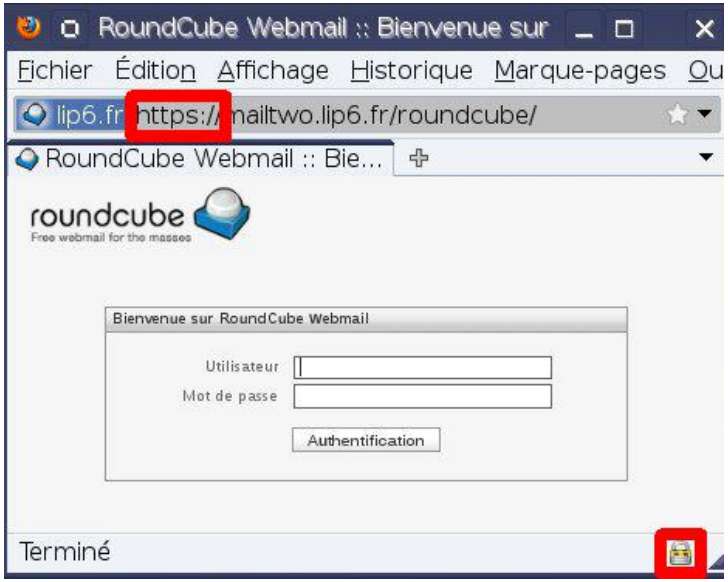
Exemple : détection de visages (`http://www.idiap.ch/onlinefacedetector/`)



Autres exemples :

- traduction automatique,
- reconnaissance de la parole, ...

Cryptographie et cryptanalyse



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.

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.

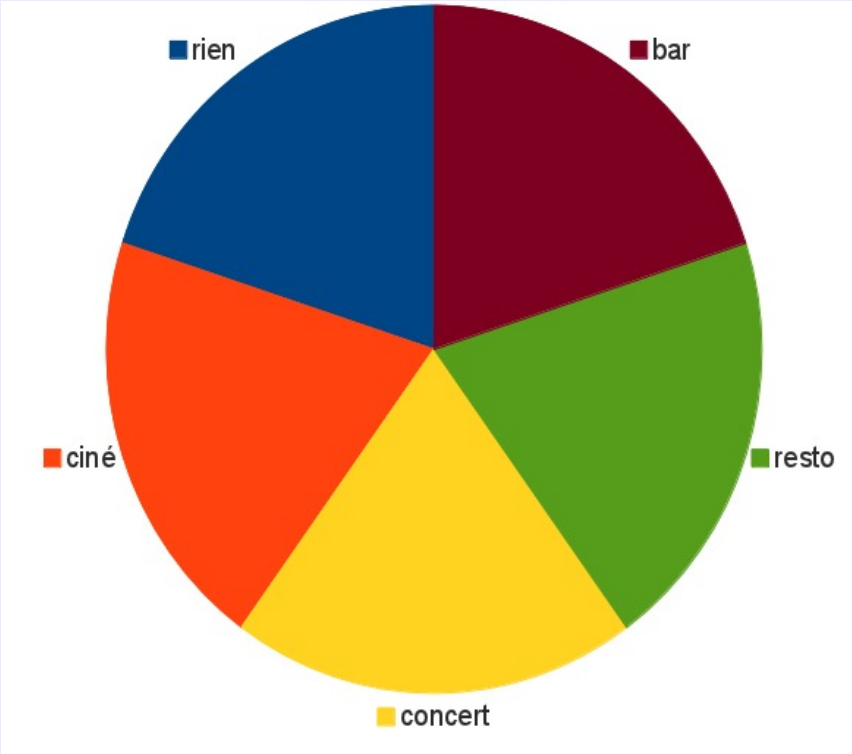
Et bien d'autres...

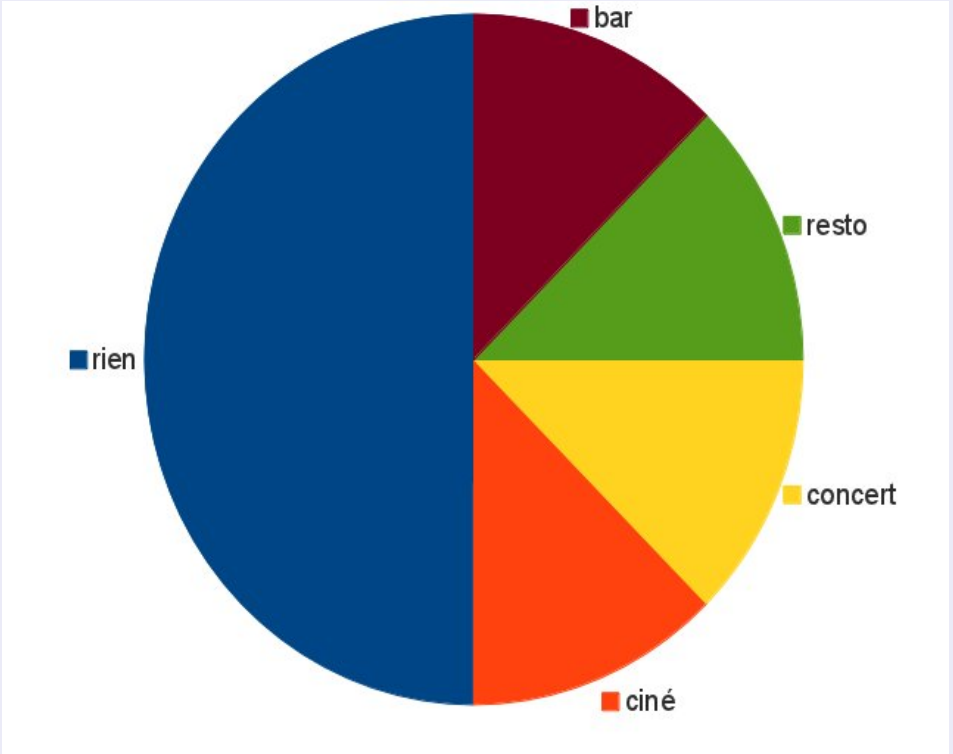
- Décision dans l'incertain ;
- Modélisation des réseaux ;
- Communication à travers des canaux bruités ;
- Analyse des réseaux sociaux ;
- Bases de données probabilistes ;
- ...

Plan

Les probabilités ... intuitives ?

Roue de la fortune : un ticket pour ...





laquelle choisir ?

Probabilités sur les ensembles discrets

Traduction mathématique

- une case de la roue $\equiv$ un **événement élémentaire**, noté ω;
- un ensemble de cases de la roue $\equiv$ un **évenement**, noté E;
  (par exemple les cases gagnantes)
- l'ensemble des cases de la roue $\equiv$ **l'univers**, noté Ω.

Que veut-on ? la solution en 3 axiomes

- une mesure de la "chance" qu'un événement se réalise
  $\Rightarrow$ **notion de probabilité** : une fonction P dans $[0, 1]$
  pas de chance négative : *axiome de positivité*
- au moins une case sort à chaque tirage
  $\Rightarrow P(\Omega) = 1$: *axiome de certitude*
- la probabilité d'un événement est proportionnelle à l'aire qu'il occupe
  si $A \cap B = \emptyset$, $P(A) + P(B) = P(A \cup B) \Rightarrow$ *axiome d'additivité*

Probabilités sur les ensembles discrets

Événements

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

Exemple : lancer simultané de deux dés

- L'univers Ω est :

$$\Omega = \{(1,1),(1,2),(1,3),...(5,6),(6,6)\},$$

- $E = \{(1,6),(2,6),(3,6),(4,6),(5,6),(6,6)\}$ représente l'événement ?

Probabilités sur des ensembles discrets (2)

Mesure de probabilité

Soit $\mathcal{P}(\Omega)$ l'ensemble des sous-ensembles de Ω. Une mesure de probabilité sur Ω est une fonction $P : \mathcal{P}(\Omega) \to [0, 1]$ vérifiant:

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

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 $P(E)$
  $\to$ le résultat appartient à l'ensemble E avec une fréquence $P(E)$.

Probabilités sur des ensembles discrets (3)

Propriétés

- $P(\varnothing) = 0$, ($\varnothing$ est l'événement impossible)
- $P(\bar{E}) = 1 - P(E)$ ($\bar{E}$: complémentaire de E dans Ω),
- $E \subset F \Rightarrow P(F) = P(F \backslash E) + P(E) \Rightarrow P(E) \leq P(F)$
  ($F \backslash E$: ensemble des éléments de F qui ne sont pas dans E),
- $P(E \cup F) = P(E) + P(F) - P(E \cap F)$
- $P(\bigcup_i E_i) \leq \sum_i P(E_i)$

Fonction de masse

On note p la fonction de masse de probabilité associée à P:

$$\forall \omega \in \Omega, p(\omega) = P(\{\omega\})$$

Alors, pour tout événement E:

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

Mesures de Probabilités uniformes

Probabilité uniforme

- Considérons un ensemble *fini* Ω. La probabilité uniforme sur Ω est définie par la fonction de masse :

$$p(\omega) = \frac{1}{card(\Omega)}.$$

- De façon équivalente, la loi uniforme est définie de la façon suivante :

$$\text{pour tout événement } E,\ P(E) = \frac{card(E)}{card(\Omega)}.$$

Exemple : lancer simultané de deux dés

Soit E l'événement

`La somme des deux chiffres est inférieure ou égale à` 5,

alors

$$P(E) = \frac{10}{36}.$$

En effet : $E = \{(1,1),(1,2),(1,3),(1,4),(2,1),(2,2),(2,3),(3,1),(3,2),(4,1)\}$.

Problème du Prince de Toscane

Exemple

Pourquoi en lançant trois dés, obtient-on plus souvent un total de 10 points qu'un total de 9 points, alors qu'il y a 6 façons d'obtenir ces deux totaux?

9 pts	10 pts
$6+2+1$	$6+3+1$
$5+2+2$	$6+2+2$
$5+3+1$	$5+4+1$
$4+3+2$	$5+3+2$
$4+4+1$	$4+4+2$
$3+3+3$	$4+3+3$

Problème du Prince de Toscane

Exemple

Pourquoi en lançant trois dés, obtient-on plus souvent un total de 10 points qu'un total de 9 points, alors qu'il y a 6 façons d'obtenir ces deux totaux?

9 pts	10 pts
$6+2+1$	$6+3+1$
$5+2+2$	$6+2+2$
$5+3+1$	$5+4+1$
$4+3+2$	$5+3+2$
$4+4+1$	$4+4+2$
$3+3+3$	$4+3+3$

L'univers est $\Omega = \{(i,j,k) \mid i \in \{1,..,6\}, j \in \{1,..,6\}, k \in \{1,..,6\}\}$,

On a $card(\Omega) = 6^3 = 216$ et, $\forall (i,j,k) \in \Omega, P((i,j,k)) = \frac{1}{216}$.

Problème du Prince de Toscane

Exemple

Pourquoi en lançant trois dés, obtient-on plus souvent un total de 10 points qu'un total de 9 points, alors qu'il y a 6 façons d'obtenir ces deux totaux?

9 pts	10 pts
$6+2+1$	$6+3+1$
$5+2+2$	$6+2+2$
$5+3+1$	$5+4+1$
$4+3+2$	$5+3+2$
$4+4+1$	$4+4+2$
$3+3+3$	$4+3+3$

On considère les événements suivants, qui prennent en compte uniquement les chiffres affichés sur les dés:

$$\text{pour } i \geq j \geq k, \Omega_{i,j,k} = \{(i,j,k),(j,i,k),(j,k,i),(k,j,i),(k,i,j),(i,k,j)\}$$

Il y a 6 événements $\Omega_{i,j,k}$ qui donnent une somme à 10, et 6 qui donnent une somme à 9.

Problème du Prince de Toscane

Exemple

Pourquoi en lançant trois dés, obtient-on plus souvent un total de 10 points qu'un total de 9 points, alors qu'il y a 6 façons d'obtenir ces deux totaux?

9 pts	10 pts
$6+2+1$	$6+3+1$
$5+2+2$	$6+2+2$
$5+3+1$	$5+4+1$
$4+3+2$	$5+3+2$
$4+4+1$	$4+4+2$
$3+3+3$	$4+3+3$

Les événements $\Omega_{i,j,k}$ ne sont pas équiprobables:

- si $i \neq j \neq k \neq i$, alors $P(\Omega_{i,j,k}) = \frac{6}{216}$,
- si $i = j \neq k$, alors $P(\Omega_{i,j,k}) = \frac{3}{216}$ (idem pour $i = k \neq j$ et $k = j \neq i$),
- si $i = j = k$, alors $P(\Omega_{i,j,k}) = \frac{1}{216}$

On a alors $P(\{i+j+k=9\}) = \frac{25}{216}$ et $P(\{i+j+k=10\}) = \frac{27}{216}$

Plan

Mise en jambe

- Combien y'a-t-il de mots de 2 lettres ?
- Combien y'a-t-il de mots de 2 lettres formés d'une voyelle et d'une consonne ?
- Un numéro de téléphone est composé de 5 chiffres, dont le premier est 0, le deuxième compris entre 1 et 5, et les 3 derniers libres. Combien de numéro différents peut-on former ? Combien de numéro avec des chiffres tous différents ?
- On tire 5 cartes dans un jeu de 32 cartes. Combien de résultats possibles ?

Dénombrements

Dénombrement de n-uplets

Soit E un ensemble fini de taille n, et k un entier.

- nombre de k-uplets d'éléments de E: n^k,
- nombre de k-uplets d'éléments distincts:

$$A_n^k = n \times (n-1) \times \ldots \times (n-k+1).$$

A_n^k est appelé le nombre d'arrangements de k parmi n, ou le nombre de k-arrangements de E.

- Cas particulier : nombre de permutations (cas $n = k$):

$$n! = n \times (n-1) \times \ldots \times 1$$

(une permutation est une façon d'ordonner n éléments distincts).

Nombre de sous-ensembles

Nombre de sous-ensembles

- soit E un ensemble fini de cardinal n, Le nombre de sous-ensembles distincts de cardinal k contenus dans E:

$$C_n^k = \frac{n!}{k!(n-k)!}$$

C_n^k s'appelle aussi le *nombre de combinaisons de k parmi n éléments*

- Remarque : Formule du binôme de Newton

$$(x+y)^n = \sum_{k=0}^{n} C_n^k x^{n-k} y^k \Rightarrow card\left(\mathcal{P}(\Omega)\right) = 2^n$$

.

- Remarque 2 : $C_n^k = \frac{A_n^k}{k!}$.

nb k-arrangements = $\Big($ nb combinaisons de k parmi n $\Big)$

$\times \Big($ nb permutations de k éléments $\Big)$.

Rappel : une permutation est une façon d'ordonner les éléments.

Dénombrements : exemples (1)

Exemple

Tirer deux cartes, sans remise, dans un jeu de 52 cartes. L'ensemble de tous les événements élémentaires:

$$\Omega = \{\{a, b\} \mid a \text{ et } b \text{ sont deux cartes différentes du jeu}\}$$

Tous les sous-ensembles sont de cardinal 2 et sont équiprobables :

$$P(\{a, b\}) = \frac{1}{1326}, \forall \{a, b\} \in \Omega$$

Soit E l'événement *au moins une des deux cartes est une dame*

$$P(E) = 1 - \frac{C_{48}^2}{1326} = 0.149$$

Dénombrements : exemples (2)

Exemple : PMU

Un joueur parie toujours sur le même résultat :

- pour le quarté : les chevaux 1, 2, 3 et 4 vont terminer la course en premier (dans cet ordre).
- pour le 2 sur 4 : les chevaux 1 et 2 seront dans les 4 premiers arrivés.

On suppose qu'il y a toujours 15 chevaux dans une course, et que l'ordre d'arrivée des chevaux suit une probabilité uniforme.

Quelle est la probabilité que le joueur gagne au quarté et au 2 sur 4 ?