## TD 1 - Analyse de données avec R

Dans ce TD, nous ferons nos premiers pas avec R et Rstudio. Pour quelques explications sur comment installer tout ça :
http://mescal.imag.fr/membres/arnaud.legrand/teaching/2013/RICM4_PS.php

# 1 Régulation des naissances

**Exercice 1. Premiers pas** Une politique de régulation des naissances préconise que chaque couple en âge de procréer n'ait qu'un seul enfant. Toutefois, en raison d'une préférence significative de la population pour les enfants de sexe masculin, les autorités accordent généralement aux parents d'une fille un second et dernier essai. L'exercice vise à déterminer si cette politique a une quelconque influence sur le taux de masculinité de la population pour cette nouvelle génération, en supposant qu'une naissance ait une probabilité 1/2 de donner un garçon.

1. Modéliser ce problème.
2. Programmer en R la génération de la descendance de $n$ couples ($n \in \{100, 1000\}$ par exemple).
3. Calculer le taux de masculinité observé et conclure.

Peut-on expliquer ces résultats ?

**Exercice 2. Une infinité de filles** On généralise le problème en laissant les couples avoir plusieurs chances d'essayer d'enfanter un garçon, sans limite particulière.

1. Programmer cette nouvelle politique en R.
2. Analyser les données de la nouvelle génération et conclure.

Expliquer les résultats.

# 2 Analyse statistique

L'objectif de cette partie est de représenter et d'analyser des données "statistiques". On aborde les sujets suivants :

a) Analyse et comparaison de fréquences sur des variables nominales.

b) Comparaison de variables "métriques" (longueur de mots, de fichiers).

c) Indices de tendance centrale et de dispersion.

Trois outils seront utilisés : les diagrammes en bâton (variables qualitatives), les histogrammes (variables quantitatives), les indices de tendance centrale (mode, médiane, moyenne) et les écarts à l'indice de tendance centrale (quantiles et écart-type).

Le support pour les exercices sera l'analyse de textes : la fréquence des voyelles et la longueur des mots.

**Textes Français et Anglais de Candide**

*Version française :*

> Il y avait en Westfalie, dans le chateau de M. le baron de Thunder-ten-tronckh, un jeune garcon a qui la nature avait donne les moeurs les plus douces. Sa physionomie annoncait son ame. Il avait un jugement assez droit, avec l'esprit le plus simple ; c'est, je crois, pour cette raison qu'on le nommait Candide. Les anciens domestiques de la maison soupconnaient qu'il etait fils de la soeur de Monsieur le baron et d'un bon et honnete gentilhomme du voisinage, que cette demoiselle ne voulut jamais epouser parce qu'il n'avait pu prouver que soixante et onze quartiers, et que le reste de son arbre genealogique avait ete perdu par l'injure du temps.

*Version anglaise :*

> There was in Westfalia, in the castle of Mr. the baron Thunder-ten-tronckh, a young boy to whom the nature had given the softest manners. His aspect announced his heart. He had the rather right judgement, with the simplest spirit ; it is, I believe, for this reason that he was named Candide. The former servants of the house suspected that he was a son of the sister of Monsieur le baron and a good and honest neighbor gentleman, whom the young lady never wanted to marry because he had been able to prove only seventy one quarters, and that the remainder of his genealogic tree was lost by the insult of time.

**Exercice 1. Proportions de voyelles** Dans le texte du premier paragraphe de Candide (en Français), nous avons observé la population des voyelles.

| voyelle | a | e | i | o | u | y |
|---|---|---|---|---|---|---|
| effectifs | 44 | 81 | 37 | 35 | 35 | 2 |

**1)** Calculer les fréquences et choisir une présentation graphique des données.
Peut-on comparer le résultat avec les fréquences de voyelle de langue française :

| voyelle | a | e | i | o | u | y |
|---|---|---|---|---|---|---|
| fréquences | 0.1795 | 0.3739 | 0.1677 | 0.1227 | 0.1489 | 0.0073 |

**2)** Dans la traduction anglaise du même texte on trouve les effectifs suivants :

| voyelle | a | e | i | o | u | y |
|---|---|---|---|---|---|---|
| effectifs | 40 | 69 | 25 | 34 | 12 | 8 |

Représenter graphiquement les données. Comparer le résultat avec les fréquences de voyelles en Français. Peut-on comparer les fréquences avec celle de la langue anglaise ?

| voyelle | a | e | i | o | u | y |
|---|---|---|---|---|---|---|
| fréquences | 0.196 | 0.315 | 0.176 | 0.193 | 0.07 | 0.05 |

Quels conclusions pourrait-on faire ?

**Exercice 2. Longueur des mots** On étudie la variable statistique $X$ qui représente la longueur des mots dans le même texte en Français.

1. Construire le tableau d'effectifs et calculer les fréquences et fréquences cumulées.
2. Représenter le diagramme de fréquences et le graphe de la fonction de répartition empirique.
3. Déterminer et interpréter le (ou les) mode ainsi que la médiane de cette distribution. Calculer sa moyenne.
4. Calculer la variance et l'écart type. Appliquer l'inégalité de Tchebychev et commenter.
5. Déterminer l'intervalle interquartile de la distribution.

**Exercice 3. Taille des fichiers Internet** On considère la répartition d'un échantillon de fichiers disponibles sur Internet selon leur taille d'une population de taille $11852$. [1]

| taille | effectifs |
|---|---|
| $\leq$ 1Ko | 4336 |
| de 1 à 2Ko | 3001 |
| de 2 à 3Ko | 1770 |
| de 3 à 4Ko | 922 |
| de 4 à 5Ko | 618 |
| de 5 à 6Ko | 353 |
| de 6 à 7Ko | 298 |
| de 7 à 8Ko | 231 |
| de 8 à 9Ko | 174 |
| de 9 à 10Ko | 149 |

**1)** Calculer les fréquences et fréquences cumulées et représenter l'histogramme des fréquences et la courbe des fréquences cumulées.

**2)** Calculer la moyenne, la variance et l'écart type, la médiane, l'intervalle interquartile.

**3)** Calculer l'intervalle $[c_5, c_{95}]$ et donner son interprétation.

1. Le fichier de longueur $b$ est dans la classe "de $a$ à $b$" , mais pas le fichier de longueur $a$.