# Devoir maison
# Statistiques numériques et analyse de données

Pour le 13 janvier 2015

**Exercice 1**

On cherche une méthode pour repérer les faux billets de 100 francs suisses. Pour cela, on dispose d'un échantillon de 200 billets dont il sait s'ils sont vrais ou faux. Il y en a 100 vrais et 100 faux. Pour chacun de ces billets, on mesure très précisément six dimensions :

- Length
- Left
- Right
- Bottom
- Top
- Diagonal

Chacune de ces dimensions est en *mm*.

On cherche maintenant une règle de décision permettant de différentier les faux billets des vrais. Pour cela, on réalise une analyse en composante principale sur les données.

1. Quelle est la dimension de la matrice des données d'entrée ?
2. On donne les valeurs propres résultant de l'analyse en composante principale :

   ```
   2.9455582 1.2780838 0.8690326 0.4497687 0.2686769 0.1888799
   ```

   (a) L'analyse a-t-elle été effectuée sur des données normées ?

   (b) Dessiner rapidement le scree-plot de cette analyse en composantes principales. Combien de composantes fait garder le critère de Kaiser ? Combien en garderiez-vous ?
3. Le graphe ci-dessous représente les données dans le plan engendré par les deux premières composantes principales. Les faux billets sont représentés par une croix, les vrais par un cercle. Donner une règle de décision simple permettant de déceler un faux billet.
4. On donne ci-dessous les coordonnées des deux premières composantes principales dans la base formée des variables d'origine.

   ```
                    PC1         PC2
   Length    0.006987029 -0.81549497
   Left     -0.467758161 -0.34196711
   Right    -0.486678705 -0.25245860
   Bottom   -0.406758327  0.26622878
   Top      -0.367891118  0.09148667
   Diagonal  0.493458317 -0.27394074
   ```

   (a) Expliquer comment cette matrice donne accès aux corrélations entre les deux premières composantes et les variables d'origine.

   (b) Faire figurer, sur un cercle des corrélations, les variables Left, Right, Bottom et Top.

   (c) Quelles variables vous semblent les plus utiles pour cette analyse ?

   (d) On donne les caractéristiques suivantes à propos d'un billet de banque :

   ```
   Length   1       Left     1
   Right    1       Bottom   2
   Top      2       Diagonal 3
   ```

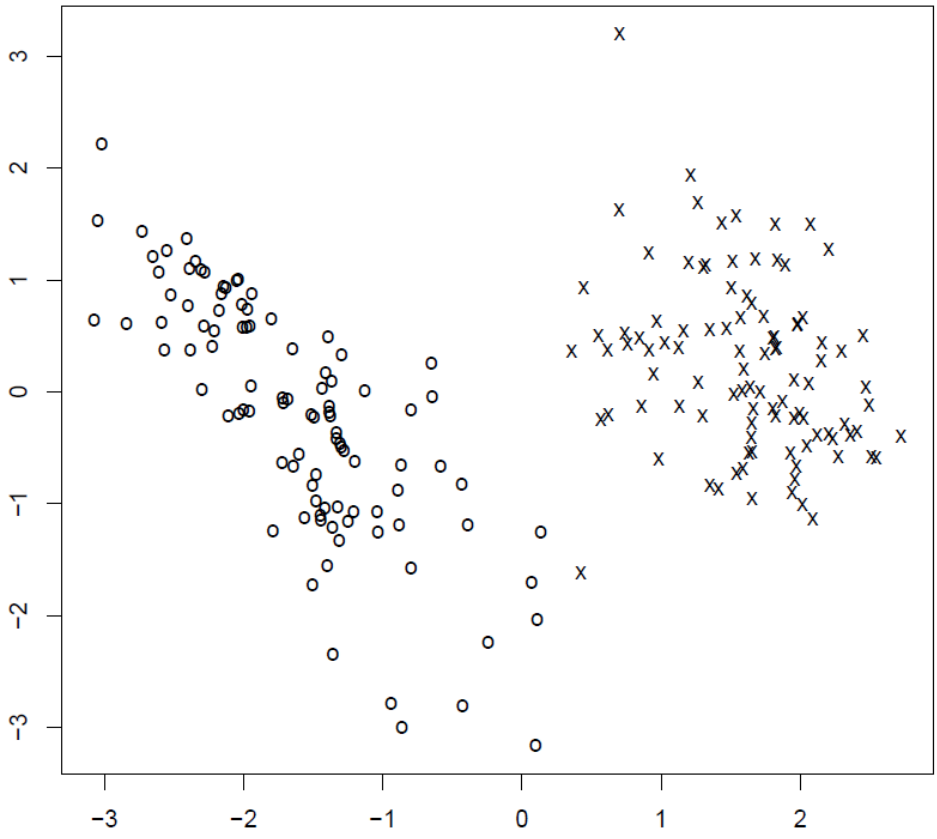
   Ce billet vous semble-t-il vrai ou faux ?

**Exercice 2**

On rappelle qu'une qu'une variable aléatoire $W$ suit une loi géométrique de paramètre $1-p$ si pour tout entier $k$ strictement positif

$$P(W = k) = p^{k-1}(1-p).$$

De plus $E[W] = \frac{1}{1-p}$ et var$(W) = \frac{p}{(1-p)^2}$.




1. Pour un individu, on considère que la probabilité de ne pas avoir d'enfants est $\alpha$ et que, si il a des enfants, le nombre de ses enfants est une variable aléatoire de loi géométrique de paramètre $1-\beta$, $(0<\beta<1)$. On note $X$ le nombre d'enfants.
   (a) Montrer que la loi de $X$ est définie par $P(X=0)=\alpha$ et, pour tout $j \geq 1$,
   $$P(X=j) = (1-\alpha)(1-\beta)\beta^{j-1}$$
   On note $\mathcal{E}(\alpha,\beta)$ cette loi.
   (b) Calculer $E[X]$.
2. On considère maintenant un échantillon $(X_1,\cdots,X_n)$ de loi $\mathcal{E}(\alpha,\beta)$. On note $Y_n$ le nombre d'individus de l'échantillon n'ayant aucun enfant (correspondant à $X_i=0$) et $Z_n$ le nombre total d'enfants :
   $$Z_n = \sum_{i=1}^{n} X_i.$$
   (a) Montrer que $Y_n$ suit une loi binomiale dont on précisera les paramètres. *On rappelle qu'une variable aléatoire suit une loi binomiale si elle s'écrit comme une somme de variables aléatoires de Bernouilli iid.*
   (b) Calculer $E[Y_n]$. En déduire un estimateur sans biais de $\alpha$.
   (c) Calculer les estimateurs du maximum de vraisemblance $\hat{\alpha_n}$ et $\hat{\beta}_n$ pour $\alpha$ et $\beta$.
   (d) Montrer que ces estimateurs sont consistants.
3. On choisit à présent $\hat{\alpha}_n = Y_n/n$ comme estimateur de $\alpha$.
   (a) Montrer que $\hat{\alpha}_n$ est asymptotiquement normal et préciser la variance limite.
   (b) Déterminer un intervalle de confiance asymptotique pour $\alpha$ au niveau 0.95.
4. (a) Montrer que :
   $$\text{var}(\alpha_n) \leq \frac{1}{4n}.$$
   (b) En utilisant l'inégalité de Markov, en déduire un intervalle de confiance non asymptotique de niveau 0.95 pour $\alpha$.