**UGA – L2/STA401**

# TP5 : Les théorèmes fondamentaux

**Exercice pour l'extraction de données :**

1. Charger le jeu de données cardiaque.csv et l'affecter à `cardiaque`. En extraire l'échantillon des pressions systoliques chez les patients ayant un BMI supérieur ou égal à 23 et en donner les résumés numériques usuels : taille, moyenne, écart-type empirique corrigé (noté $s'$ dans le cours) et quartiles.

2. Charger les jeux de données her.txt, donneesSerie4.csv et diamantsPurs.csv et les affecter aux data.frame nommés respectivement `her`, `serie` et `diamants`. Quelles sont les dimensions de chaque data.frame ?

**Faire un script avec R Markdown :**
Télécharger sur Chamilo le modèle TP5.Rmd et exécuter les tronçons les uns après les autres.

**Objectifs :** *Comprendre la loi des grands nombres et le théorème central limite à l'aide de données simulées.*

## 1 Loi des grands nombres

Soit $X_1, ..., X_n$ une suite de variables aléatoires indépendantes de même loi et telles que $E(X_1) < +\infty$ alors

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow E(X_1) \qquad \text{lorsque} \qquad n \longrightarrow +\infty$$

Lorsque la convergence de $\bar{X}_n$ a lieu presque surement (on accepte qu'elle ne se produise pas sur un ensemble de probabilité nulle) on parle de *la loi* **forte** *des grands nombres* (c'est celle qu'illustrent les exercices de cette section).

**Exercice 1 : Modèle de Bernoulli**

Dans cet exercice on considère un échantillon i.i.d. de $X$ pour $X$ de loi de Bernoulli $\mathcal{B}(p)$ avec $p \in ]0, 1[$ et de taille $n$.

Avec R la loi de Bernoulli de paramètre $p$ s'obtient comme un cas particulier de la loi binômiale : $\mathcal{B}(1, p)$. Sa probabilité en $k$ ($k = 0$ ou $k = 1$) sera calculée avec `dbinom(k,1,p)`, sa FdR en $x$ avec `pbinom(x,1,p)`, son quantile d'ordre $\alpha$ avec `qbinom(`$\alpha$`,1,p)`. Pour finir si on veut tirer $n$ réalisations indépendantes de $X$ (c'est à dire générer un échantillon de taille $n$ de $X$) on utilise `rbinom(n,1,p)`.

1. Créer les objets `p` et `n` auquels seront affectées les valeurs 0.45 pour $p$ et 100 pour $n$. Créer `x` auquel seront affectés $n$ tirages indépendants de $X$.

2. Calculer les moyennes empiriques des $k$ premières valeurs de l'échantillon tiré pour $k = 1, ...., n$ et les affecter à un vecteur qui sera nommé `suitemoy`. On pourra utiliser la fonction `cumsum()` pour éviter l'écriture d'une boucle `for`.

3. Représenter la suite des moyennes obtenues avec `plot()` et y ajouter la droite horizontale qui passe par l'ordonnée $p$. Observer et conclure (on pourra réexécuter plusieurs fois le tirage de $X$ et le graphique ci-dessus pour apprécier le caractère aléatoire de la série calculée mais avec une convergence vers $p$ dans tous les cas).

**Exercice 2 : Modèle Normal**

Dans cet exercice on considère un échantillon i.i.d. de $X$ pour $X$ de loi normale $\mathcal{N}(\mu, \sigma^2)$ et de taille $n$.

Refaire les questions précédentes en utilisant la loi normale au lieu de celle de Bernoulli avec $\mu = -2$ et $\sigma = 5$ (On imposera les limites $[-10, 6]$ sur l'axe de ordonnées). Conclure. Que se passe t-il si on augmente $\sigma$ à 10 ? si on le diminue à 2 ?

# 2 Théorème central limite

Soient $X_1, ..., X_n$ une suite de variables aléatoires indépendantes de même loi et telles que $E(X_1) = \mu < +\infty$ et $V(X_1) = \sigma^2 < +\infty$ et $U$ une variable aléatoire normale centrée réduite (i.e. de loi N(0,1)) alors :

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\mathcal{L}} U \qquad \text{lorsque} \qquad n \longrightarrow +\infty.$$

$\xrightarrow{\mathcal{L}}$ désigne la convergence en loi. Ce théorème signifie que lorsque $n \longrightarrow +\infty$, la fonction de répartition (resp. la densité) de $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ converge en tout point vers la fonction de répartition (resp. la densité) de $U$.

Ce théorème implique également que la fonction de répartition de $\bar{X}_n$ peut-être approchée par celle d'une variable de loi $\mathcal{N}(\mu, \sigma^2/n)$ et que celle de $n\bar{X}_n = \sum_{i=1}^{n} X_i$ peut-être approchée par la FdR d'une variable de loi $\mathcal{N}(n\mu, n\sigma^2)$.

**Exercice 3 : modèle normal**

Dans cet exercice on considère d'abord un échantillon i.i.d. de $X$ pour $X$ de loi normale $\mathcal{N}(\mu, \sigma^2)$ et de taille $n$. Ensuite on considère $N$ échantillons de taille $n$ de $X$.

1. Définir $n = 5$, $\mu = -2$ et $\sigma = 2$ et tirer un échantillon `x` de taille $n$ de $X$. Calculer sa moyenne que l'on affectera à `moyx` et l'afficher. Observer la variation de la moyenne calculée lorsque l'on retire l'échantillon (il suffit pour cela de réexécuter la commande qui définit `x` et celle qui calcule sa moyenne).

2. Refaire la question précédente avec $N = 2$ échantillons de taille $n$ et affecter les deux moyennes calculées à un objet nommé `moyennes`. On pourra utiliser `rbind` pour définir une matrice à $N$ lignes et $n$ colonnes où ranger les données simulées. Faire ensuite le calcul des deux moyennes en une seule commande, grâce à la fonction `rowMeans()`.

3. Tirer à présent $N = 100$ échantillons de taille $n$ et les affecter à une matrice qui sera nommée `Mdata` (on utilisera la fonction `matrix()`).

4. Calculer les moyennes en ligne de `Mdata`. Soit avec `rowMeans()` soit avec `apply(Mdata,MARGIN=1,mean)` et les affecter à `moyennes`. Afficher ce dernier objet dans lequel on trouve les $N$ réalisations de $\bar{X}_n$.

5. Calculer la moyenne empirique et l'écart-type empirique corrigé de `moyennes`.

6. Représenter l'histogramme des $N$ réalisations de $\bar{X}_n$ et y superposer la densité d'une loi normale dont les paramètres seront convenablement choisis. Ajouter également une verticale rouge passant par l'abscisse indiquant la moyenne de l'échantillon représenté ici.

7. Calculer à présent les $N$ réalisations de $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ et les affecter à `moyennescr`. Faire l'histogramme de `moyennescr` où sera superposée la densité de la normale centrée réduite et la verticale passant par la moyenne empirique de l'échantillon.

8. Observer les changements lorsque l'on fait varier $n$ ou $N$ et expliquer les résultats obtenus.

**Exercice 4 : modèle de Bernoulli**

Dans cet exercice on considère d'abord un échantillon i.i.d. de $X$ pour $X$ de loi $\mathcal{B}(1,p)$ et de taille $n$. Ensuite on considère $N$ échantillons de taille $n$ de $X$.

Refaire toutes les questions de l'exercice précédent (on pourra copier coller toutes les instructions du script de l'exercice 3 et remplacer seulement le tirage des observations sous la loi normale par un tirage sous la loi de Bernoulli de paramètre $p$ qu'on aura défini au préalable; ajuster aussi les lignes de commandes notamment pour définir les paramètres des densités à superposer....) Que se passe-t-il pour $n = 5$ et $N = 10000$ ? Le théorème central limite s'applique t-il-dans ce cas ? Expliquer. Choisir à présent $n = 100$. Pour finir, conclure sur les cas limites pour $n \to +\infty$ et $N \to +\infty$.