DM 2 - TISD - Master 1 Pro
A me remettre dans mon casier (RdC Bat M2) pour le 3/11

Tran Viet Chi, `chi.tran@math.univ-lille1.fr`, bureau 316 (bâtiment M3).

Exercice 1 (Moyenne et variance empiriques : programmation sous R)
Nous nous proposons d'illustrer par des simulations quelques faits simples sur la moyenne et la variance empiriques.

1. Simuler un échantillon de $n = 1000$ variables aléatoires *iid* $X_1, \cdots X_n$ de loi normale $\mathcal{N}(m = 2, \sigma^2 = 121)$.
1.1. Avec **R**, tracer l'histogramme des observations, superposer la densité approchée et la densité théorique. Dessiner un *boxplot* (ou *boîte à moustaches*) des observations.
1.2. Avec **R**, calculer la moyenne empirique et la variance empirique :

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2.$$

1.3. Montrer que les deux définitions de la variance empirique données ci-dessus sont équivalentes.
1.4. Montrer que $\bar{X}_n$ et S_n^2 sont fortement convergentes ?
1.5. Rappeler le Théorème Central Limite. En utilisant la δ-méthode rappelée ci-dessous, donner un théorème de normalité asymptotique pour S_n^2.

δ-méthode : *Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de vecteurs aléatoires de dimension d satisfaisant la convergence en loi suivante :*

$$\sqrt{n}\big(X_n - a\big) \to \mathcal{N}(0, \Sigma)$$

où Σ est une matrice carrée de dimension d symétrique positive, et soit f une fonction de classe $\mathcal{C}^1$ de $\mathbb{R}^d$ dans $\mathbb{R}^p$. Alors :

$$\sqrt{n}\big(f(X_n) - f(a)\big) \to \mathcal{N}(0, {}^t A\Sigma A)$$

où A est la Hessienne de f en a, de dimension $d \times p$, définie par :

$$\forall (i,j) \in [\![1,p]\!] \times [\![1,d]\!]\,,\ A_{ij} = \frac{\partial f_i}{\partial x_j}(a).$$

1.6. Quelles sont les lois de $\bar{X}_n$ et S_n^2 pour n fixé ? Est-ce que $\bar{X}_n$ et S_n^2 sont indépendantes ?

2. Réaliser le programme suivant avec **R** :
- simuler $N = 300$ échantillons de $n = 1000$ variables aléatoires *iid* $X_1, \cdots X_n$ suivant la loi $\mathcal{N}(m = 2, \sigma^2 = 121)$.
- pour l'échantillon $i \in [\![1, N]\!]$, calculer la moyenne empirique $\bar{X}_n^{(i)}$, la variance $S_n^{2,(i)}$, la statistique $\zeta_n^{(i)} = \sqrt{n}(\bar{X}_n^{(i)} - 2)/\sigma$ et la statistique $\xi_n^{(i)} = \sqrt{n}(\bar{X}_n^{(i)} - 2)/S_n^{(i)}$. Nous avons donc N réalisations *iid* de ces variables aléatoires correspondant à des tirages d'échantillons différents. Ces valeurs seront conservées dans un tableau de 4 colonnes et $N = 300$ lignes.

2.2. Tracer avec **R** l'histogramme des $\bar{X}_n^{(1)}, \cdots \bar{X}_n^{(N)}$. Superposer la densité approchée et la densité théorique obtenue à la question 1.6. Faire de même pour les $S_n^{2,(1)}, \cdots S_n^{2,(N)}$, pour les $\zeta_n^{(1)}, \cdots \zeta_n^{(N)}$. Pour trouver la loi théorique des $\xi_n^{(i)}$, on montrera qu'il s'agit d'une loi de Student à $n-1$ degrés de liberté en utilisant que :

Si $X \rightsquigarrow \mathcal{N}(0,1)$, si $Y \rightsquigarrow \chi^2(n-1)$ et si X et Y sont indépendantes, alors $X/\sqrt{Y/(n-1)}$ suit une loi de Student à $n-1$ degrés de liberté.

2.3. Vérifier avec **R**, par des *QQ-plots*, l'adéquation des données $(\bar{X}_n^{(i)})_{i\in[\![1,N]\!]}$, $(S_n^{2,(i)})_{i\in[\![1,N]\!]}$, $(\zeta_n^{(i)})_{i\in[\![1,N]\!]}$ et $(\xi_n^{(i)})_{i\in[\![1,N]\!]}$ avec les lois théoriques trouvées à la question précédente.

3. On réalise maintenant le programme suivant (toujours avec **R**) :
- Pour $n = 100\,k$ avec $k \in [\![1, 1000]\!]$, on simule un échantillon de n variables aléatoires indépendantes de loi exponentielle $\mathcal{E}(\lambda = 2)$.
- Pour chacun de ces échantillons, on calcule $\bar{X}_n$ et $\xi_n = \sqrt{n}(\bar{X}_n - m)/\sigma$, où m et σ sont l'espérance et l'écart-type de $\mathcal{E}(\lambda = 2)$, à préciser.
- On trace ces quantités en fonction de n.

Commenter les convergences obtenues graphiquement. Tout graphique ou argument supplémentaire sera le bienvenu pour appuyer votre propos. ■

Exercice 2 (Répartition salariale sur des données groupées, d'après le partiel de 2007-2008)
Dans une entreprise, les salaires sont les suivants :

Classe de salaire	Salaires mensuels	Nombre de salariés
1	$[500, 1500[$	50
2	$[1500, 2500[$	125
3	$[2500, 5500[$	25

1. Pour chaque classe $i \in \{1, 2, 3\}$ de salaires (notée $[x_{i-1}, x_i[$ et d'effectif n_i), calculer la fréquence empirique f_i, l'amplitude a_i, le centre $c_i = (x_{i-1} + x_i)/2$, la fréquence empirique cumulée F_i (proportion des salaires inférieurs à x_i), la masse salariale approchée $n_i c_i$ et la masse salariale cumulée approchée m_i (approximation de la somme de tous les salaires inférieurs à x_i). On présentera les résultats dans un tableau.

2. Tracer avec **R** l'histogramme de la variable "salaire". Quelle règle faut-il respecter ?

3. Dessiner avec **R** la fonction de répartition. Comme la variable de salaire est quantitative continue, on choisira ici la version continue de la fonction de répartition empirique, obtenue par interpolation linéaire des points (x_i, F_i).

4. Quelle est l'équation de la portion de droite représentant la fonction de répartition empirique sur l'intervalle de salaires $[1500, 2500[$? En déduire la médiane.

5. On s'intéresse à la répartition des salaires sur ces données agrégées.
5.1. Tracer la courbe de Lorenz avec **R** (on pourra utiliser la commande `segments`). Pour des données groupées, cette courbe est obtenue en reliant les points

$$\left(F_i, \frac{\sum_{j\leq i} n_j c_j}{\sum_{j=1}^{3} n_j c_j} \right).$$

5.2. Calculer l'indice de Gini.
5.3. Commenter.