FICHE TP 1

Ce TP a pour but d'illustrer le lien entre la loi trinomiale (qui généralise la loi binomiale au cas d'une variable discrète ternaire et non plus binaire) et la loi normale bivariée (qui généralise la loi normale usuelle), soit le théorème central limite bidimensionnel…

Exercice 1 (rappels de 1A). Une v.a. Z est de loi normale $N(\mu, \sigma^2)$ si elle admet la densité de probabilité :

$$\mathbf{f_Z(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(z-\mu)^2}{2\sigma^2} \right)}$$

Une v.a. S (ou N) est de **loi binomiale B(n, p) de paramètres n et p** si

$$\mathbf{P(S = k) = \binom{n}{k} p^k (1-p)^{n-k}} \qquad \text{pour } k = 0, 1, \ldots, n.$$

En pratique, S = « nombre de succès lors de n épreuves de Bernoulli B(p) indépendantes » :

$$S = \mathbf{1}_{X_1 = 1} + \ldots + \mathbf{1}_{X_n = 1} \text{ avec } X_1, \ldots, X_n \text{ i.i.d. de loi de Bernoulli B(p)}$$

On notera encore par la suite $S = N_1$ pour signifier que S est le nombre aléatoire N_1 de fois où 1 apparaît dans une suite de Bernoulli (cf. la suite $X_1, \ldots, X_n$ est une suite de 0 ou 1).

Lien avec le Théorème Central Limite (TCL) :

- $P(S = k) \underset{\text{n grand}}{\approx} P(Z \in [k - 0.5, k + 0.5])$ où Z de loi $N(np, np(1-p))$
- Si n est grand, la variable centrée-réduite $\sqrt{\frac{n}{p(1-p)}} \times \left(\frac{S}{n} - p \right) = \frac{S - np}{\sqrt{np(1-p)}}$ est à peu près de loi $N(0, 1) \Rightarrow \frac{S}{n} \approx p + \sqrt{\frac{p(1-p)}{n}} \times N(0, 1)$ si n grand.

A l'aide du code R fourni sur Campus, vérifier la validité de ces résultats.

Exercice 2. On considère maintenant une variable X à K = 3 modalités codées 0, 1 ou 2.
On note $p_1 = P(X = 1)$ et $p_2 = P(X = 2)$ de sorte que $p_0 = P(X = 0) = 1 - p_1 - p_2$.

On cherche à estimer les proportions p_1 et p_2 à partir de n copies indépendantes $X_1, \ldots, X_n$ de la v.a. X. On calcule pour cela

$N_1 = \mathbf{1}_{X_1 = 1} + \ldots + \mathbf{1}_{X_n = 1}$ le nombre d'occurrence de la modalité n°1
$N_2 = \mathbf{1}_{X_1 = 2} + \ldots + \mathbf{1}_{X_n = 2}$ le nombre d'occurrence de la modalité n°2

sachant que $N_0 = n - N_1 - N_2$.

On s'intéresse à la loi limite du vecteur bidimensionnel $\mathbf{Z} = \left(\frac{N_1 - np_1}{\sqrt{np_1(1-p_1)}}, \frac{N_2 - np_2}{\sqrt{np_2(1-p_2)}} \right)$ lorsque n tend vers $+\infty$ sachant que ses marges $Z_1 = \frac{N_1 - np_1}{\sqrt{np_1(1-p_1)}}$ et Z_2 sont de loi limite la loi normale centrée-réduite (cf. exercice 1).

1. Pour $0 \le k_1 \le n$, $0 \le k_2 \le n$ et $k_1 + k_2 \le n$, on a

$$\mathbf{P(N_1 = k_1 \text{ et } N_2 = k_2) = \frac{n!}{k_1!k_2!(n - k_1 - k_2)!}\, p_1^{k_1} p_2^{k_2} (1 - p_1 - p_2)^{n - k_1 - k_2}}$$

On parle de **loi trinomiale de paramètres n, p_1 et p_2** qui est une simple extension de la loi binomiale (cf. exercice 1). C'est un cas particulier de la loi multinomiale où la variable X discrète sous-jacente est à K modalités avec K = 3.

Sous R, c'est la fonction **rmultinom()** qui permet de simuler une telle loi de probabilité. En complétant le code fourni, chercher à comprendre ce que renvoie cette fonction en jouant sur N le nombre de simulations indépendantes.

Pour $p_0 = 0.1$, $p_1 = 0.5$, $p_2 = 0.4$, n = 500 et N = 100, obtenir un jeu de réalisations indépendantes de taille N du vecteur aléatoire $Z = \left(\frac{N_1 - np_1}{\sqrt{np_1(1 - p_1)}}, \frac{N_2 - np_2}{\sqrt{np_2(1 - p_2)}}\right)$ sous la forme d'une matrice de taille 2×N notée Z.

2. Vérifier que les composantes Z_1 et Z_2 sont à peu près de loi normale N(0, 1) si n est grand (utiliser **qqnorm()** et **qqline()**).

3. Représenter le nuage de points associé au couple de variables (Z_1, Z_2). Que peut-on dire de la corrélation ρ entre Z_1 et Z_2 ? Estimer cette corrélation et comparer avec la valeur théorique

$$\rho = -\frac{p_1 p_2}{\sqrt{p_1(1 - p_1)}\sqrt{p_2(1 - p_2)}}$$

En prenant p_0 proche de 0, observer que la corrélation devient en module maximale égale à 1 ($\rho \to -1$). Expliquer ce phénomène.

4. On considère une Analyse dite en Composantes Principales du nuage de points précédent (matrice de données Z). Pour cela, on considère le nouveau repère orthonormé formé par les deux vecteurs de $\mathbb{R}^2$ suivants :

$$u_1 = \frac{\sqrt{2}}{2}(1, 1) \; ; \; u_2 = \frac{\sqrt{2}}{2}(-1, 1)$$

Calculer les nouvelles composantes C_1 et C_2 des N points dans ce nouveau repère et représenter le plan associé. Que peut-on dire des nouvelles variables C_1 et C_2 associées ?

5. Vérifier que C_1 et C_2 correspondent à des variables normalement distribuées. Estimer leurs variances respectives et comparer aux deux valeurs $1+\rho$ et $1-\rho$.

6. Le TCL multidimensionnel affirme que la loi limite cherchée est la loi du vecteur aléatoire

$$\sqrt{1+\rho}\, \varepsilon_1 u_1 + \sqrt{1-\rho}\, \varepsilon_2 u_2$$

où ε_1, ε_2 sont indépendantes de même loi N(0, 1).

Cela est-il conforme à vos observations ? Représenter la densité de probabilité associée à cette loi limite.