II – Vecteurs aléatoires

Ce second TP contient trois parties indépendantes permettant de manipuler des vecteurs aléatoires.

A – Lancer de fléchettes

On cherche à générer des points uniformément répartis dans un disque de rayon 1, *i.e.* simuler des valeurs d'un couple (X, Y) de loi $\mathcal{U}(\mathcal{D})$ où

$$\mathcal{D} = \{\, (x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leqslant 1 \,\}.$$

Il est facile de simuler une distribution uniforme sur tout le carré $\mathcal{C} = [-1, 1] \times [-1, 1]$:

```
n = 5000;
x = 2 * rand(n,1) - 1;
y = 2 * rand(n,1) - 1;
plot(x, y, ".", "markersize", 4)
axis("equal")
```

On pourrait choisir de ne garder que les points tombant à l'intérieur du disque $\mathcal{D}$, mais il est un peu dommage de « gaspiller » ainsi des nombres pseudo-aléatoires (d'autant plus qu'avec cette méthode, on ne sait pas précisément à l'avance combien de points on récupère dans $\mathcal{D}$).

Une meilleure idée semble être de passer en coordonnées polaires : le disque est alors décrit par le « rectangle » $(r, \theta) \in [0, 1] \times [0, 2\pi]$. Nous allons donc simuler deux variables aléatoires uniformes,

$$R \sim \mathcal{U}([0, 1]) \quad \text{et} \quad \Theta \sim \mathcal{U}([0, 2\pi]).$$

```
r = rand(n,1);
theta = 2 * pi * rand(n,1);
x = r .* cos(theta);
y = r .* sin(theta);
plot(x, y, ".", "markersize", 4)
axis("equal")
```

On obtient bien des points dans le disque… Mais la densité n'est pas uniforme, elle est plus élevée au centre. L'explication vient de l'expression de l'élément d'aire en coordonnées polaires

$$\mathrm{d}x\,\mathrm{d}y = r\,\mathrm{d}r\,\mathrm{d}\theta :$$

les points uniformément répartis dans $[0, 1] \times [0, 2\pi]$ sont étalés dans le plan (x, y) sur des surfaces d'aire plus grande lorsque r est grand que lorsque r est petit, on observe donc une raréfaction des points dans le disque à mesure que l'on s'approche du bord.

On va tenter de corriger ce biais vers le centre en introduisant un autre paramètre dans nos équations :

$$\begin{cases} X = R^\alpha \cos\Theta \\ Y = R^\alpha \sin\Theta. \end{cases}$$

1) Expérimenter avec différentes valeurs de $\alpha \in [0,1]$ jusqu'à obtenir une distribution qui semble uniforme, et observer à chaque fois les distributions marginales `hist(x,30)` et `hist(y,30)`.

2) Pour la valeur de α trouvée ci-dessus, estimer numériquement l'espérance de $\sqrt{X^2+Y^2}$.

Vous savez donc maintenant à quelle distance du centre se trouvent, en moyenne, des points uniformément répartis dans un disque (NB : ce n'est **pas** la moitié du rayon).

(Sauriez-vous *prouver* tout cela ? Suffit de faire un changement de variables dans une intégrale double . . .)

B – Simulation de variables normales, pt. 2

En modifiant légèrement les formules de la partie précédente, on obtient des distributions assez différentes. Avec toujours R et Θ comme ci-dessus, posons cette fois

$$\begin{cases} X = \sqrt{-\ln R}\,\cos\Theta, \\ Y = \sqrt{-\ln R}\,\sin\Theta. \end{cases}$$

On peut montrer que l'on obtient ainsi un couple de variables (exactement) normales indépendantes.

1) Générer des valeurs du couple (X,Y) ci-dessus, et observer les distributions conjointe (avec un `plot`) et marginales (avec des histogrammes).

En estimant numériquement les paramètres μ et σ^2 des lois normales obtenues, superposer les densités à vos histogrammes comme au TP1.

On en déduit une méthode simple pour obtenir générer des nombres aléatoires normalement distribués appelée méthode de Box-Muller. Quels sont ses avantages et inconvénients par rapport à celle présentée à la fin du TP1 ?

2) Vérifier que les variables X et Y sont décorrélées :

```
corr(x,y)                          % ou encore :

cov(x,y) / std(x) / std(y)         % ou encore, encore :

([x x.^0]\y)(1) * std(x) / std(y)  % que fait-on ici ?
```

Est-ce suffisant pour se convaincre que X et Y sont indépendantes ?

B – Moyennes échantillonnales

Dans cette dernière partie, nous allons observer les tendances asymptotiques pour $n \to \infty$ des moyennes

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

d'une suite de variables aléatoires indépendantes, identiquement distribuées.

Pour bien voir ce qui se passe, choisissons une loi très asymétrique (et pourquoi pas discrète) :

$$X_i \sim \mathcal{G}\left(\frac{1}{20}\right)$$

modélisant, par exemple, le nombre de lancers d'un D20 équilibré à effectuer avant d'obtenir un 13 pour la première fois.

Attention : la loi géométrique fournie par MATLAB compte le nombre d'échecs avant le premier succès, elle souffre donc d'un décalage de 1 par rapport à la définition « standard ».

Simulons donc une suite de valeurs $x_1, x_2, \ldots$ et observons l'évolution de la moyenne échantillonnale au fil de celle-ci :

```
n = 2022;
p = 1/20;
x = geornd(p,n,1) + 1;

xbar = zeros(n,1);
sum = 0;

for i=1:n
  sum = sum + x(i);
  xbar(i) = sum/i;
end

clf
line([1,n],[1/p,1/p],"color","red")
hold on
plot(xbar)
hold off
```

On observe bien une convergence vers l'espérance de la loi géométrique tel que prédit par la loi des grands nombres, mais observez la nature un peu particulière de celle-ci : fluctuations très importantes au départ, qui s'atténuent à la longue mais restent toujours présentes, et globalement une convergence plutôt lente (en gros en $1/\sqrt{n}$).

1) Afficher sur le même graphe les résultats provenant de plusieurs séries de données et observer la variabilité des résultats.

Décrire cette variabilité, c'est précisément décrire la loi des variables aléatoires $\overline{X}_n$. Observons celles-ci en générant, pour un n fixé, un grand nombre d'observations.

```
n = 1      % taille de l'échantillon
m = 10000  % nombre de répétitions

xbars = zeros(m,1);

for i=1:m
  x = geornd(p,n,1) + 1;
  xbars(i) = mean(x);
end

hist(xbars,30)
```

2) Augmenter graduellement la taille n des échantillons et observer comment la distribution des moyennes se resserre et se « normalise ».