TP6 : Généralités sur la fonction `rand` et simulations de v.a.r. uniformes

Pré-requis : je vous invite à consulter les chapitres de cours correspondants sur ma page (support informatique). Pour ce TP, on pourra en particulier se reporter à la section « Tracés d'histogrammes » du CH 8.

▶ Dans votre dossier `Info_2a`, créez le dossier `TP_6`.

I. Avant-propos

I.1. Sur la fonction `rand`

La fonction `rand` est un générateur de nombres **pseudo-aléatoires**.
Il vérifie notamment les propriétés suivantes :

× si $X(\omega)$ désigne le résultat de l'appel `rand()`, alors X est une v.a.r. de loi uniforme sur $[0,1]$.

× si $X_1(\omega),\dots,X_N(\omega)$ désignent les résultats de N appels successifs de `rand()`, alors les variables aléatoires $X_1,\dots,X_N$ sont mutuellement indépendantes.

En mathématiques, on travaille souvent avec des variables aléatoires, mais en informatique, on simule **une réalisation** de ces variables aléatoires (ce qui correspond à la notion d'observation en statistique).

I.2. Sur la loi (faible) des grands nombres pour une v.a.r. discrète

La loi (faible) des grands nombres[(1)] est un résultat clé en statistique inférentielle.
Illustrons brièvement ce résultat. On considère une population $\mathscr{P}$.

- Les membres de $\mathscr{P}$ appartiennent à l'une des catégories suivantes :

  1. enfant, ***2.*** adolescent, ***3.*** adulte.

- On considère l'expérience aléatoire consistant à tirer au sort une personne de la population.
  On note alors X la v.a.r. donnant le numéro de la catégorie de la personne choisie.
  Ainsi, $\mathbb{P}([X=3])$ donne la proportion d'adultes dans la population.
- Afin d'obtenir des informations sur la population, on va effectuer des tirages successifs d'un individu.
  Ces tirages se font avec remise ce qui permet d'en assurer l'**indépendance**. On note alors X_i la v.a.r. qui donne la catégorie de la $i^{\text{ème}}$ personne choisie.
  Évidemment, X_i suit la même loi que X (on a notamment $\mathbb{P}([X_i=3])=\mathbb{P}([X=3])$).
  On définit ce qu'on appelle un **échantillon** $(X_1,\dots,X_N)$ de la v.a.r. X.

Le théorème stipule que l'on peut approcher $\mathbb{E}(X)$ à l'aide d'un échantillon $(X_1,\dots,X_N)$.
Pour ce faire, on observe la valeur $(x_1,\dots,x_N)$ de cet échantillon.

Lorsque N est grand, la moyenne des valeurs observées $\dfrac{1}{N}\sum_{k=1}^{N} x_k$ devient proche de $\mathbb{E}(X)$.

(1) Ce résultat sera vu dans le chapitre « Convergences et approximations ».

Remarque

- On ne développe pas ici le formalisme de la loi (faible) des grands nombres. L'idée est plutôt de comprendre qu'il existe un théorème qui correspond à l'intuition mathématique suivante.

> Afin d'approcher la moyenne probabiliste $\mathbb{E}(X)$ d'une variable aléatoire X, il suffit de prendre la moyenne arithmétique d'un **grand nombre** d'observations de X.

- Par « grand nombre », il faut évidemment comprendre qu'il y a une notion de limite sous-jacente. Plus le nombre d'observations est grand, plus on se rapproche du résultat théorique $\mathbb{E}(X)$. On obtiendrait la valeur exacte de $\mathbb{E}(X)$ si l'on savait faire une infinité d'observations.

▶ Revenons à notre exemple. On considère maintenant la v.a.r. Z suivante.

$$\begin{array}{rccl} Z & : & \Omega & \to & \mathbb{R} \\ & & \omega & \mapsto & \begin{cases} 1 & \text{si } X(\omega) = 3 \\ 0 & \text{si } X(\omega) \neq 3 \end{cases} \end{array}$$

Démontrer que Z est une v.a.r. discrète et déterminer son espérance.

- $Z(\Omega) = \{0, 1\}$. Ainsi, Z est une v.a.r. finie (et donc discrète).
- Par définition de l'espérance, on a alors :

$$\begin{aligned} \mathbb{E}(Z) &= \sum_{z \in Z(\Omega)} z\, \mathbb{P}([Z = z]) \\ &= 1 \times \mathbb{P}([Z = 1]) + 0 \times \cancel{\mathbb{P}([Z = 0])} \\ &= \mathbb{P}([Z = 1]) \;=\; \mathbb{P}([X = 3]) \end{aligned}$$

▶ En déduire une méthode permettant d'obtenir une valeur approchée de la quantité $\mathbb{P}([X = 3])$.

- On applique une nouvelle fois la méthode fournie par la loi (faible) des grands nombres. On considère donc $(Z_1, \ldots, Z_N)$ un N-échantillon de la v.a.r. Z avec N grand. On réalise alors une observation $(z_1, \ldots, z_N)$ de cet échantillon. La loi (faible) des grands nombres permet de conclure que :

$$\frac{1}{N} \sum_{k=1}^{N} z_k \;\simeq\; \mathbb{E}(Z) = \mathbb{P}([X = 3])$$

- En pratique, cela veut dire qu'on fait N tirages indépendants dans la population. Pour chaque tirage, on incrémente un compteur de 1 si l'individu est de la catégorie 3. Dans le cas contraire, on ne met pas à jour ce compteur. On divise alors ce résultat par le nombre N d'individus choisis en tout. Autrement dit, on détermine la proportion (on parle aussi de fréquence) d'individus de catégorie 3 de ce N-tirage.

De manière plus générale, lorsqu'on effectue un N-tirage, on note l'effectif (*i.e.* le nombre d'individus) de chaque catégorie. La loi faible permet d'affirmer que si N est grand, le diagramme des fréquences des observations $\left(\dfrac{\text{effectif de chaque catégorie}}{\text{nombre d'observations}}\right)$ fournit un diagramme approché de la distribution de probabilité $(\mathbb{P}([X = 1]), \mathbb{P}([X = 2]), \mathbb{P}([X = 3]))$.

I.3. Sur la loi (faible) des grands nombres pour une v.a.r. à densité

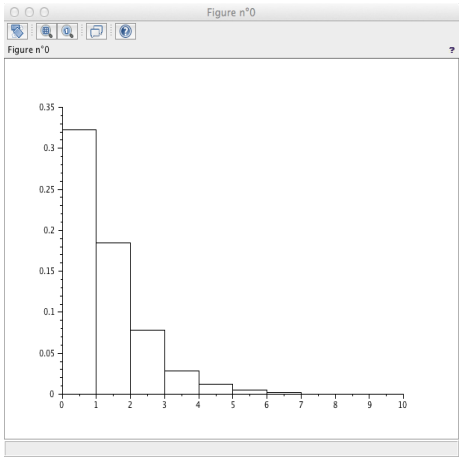
Dans l'épreuve EDHEC 2017, on s'intéressait à la loi de Gumbel et à la manière de la simuler informatiquement. Par définition, on dit qu'une v.a.r. X suit une loi de Gumbel si sa fonction de répartition F_X est donnée par :

$$\forall x \in \mathbb{R},\ F_X(x) = \mathrm{e}^{-\mathrm{e}^{-x}}$$

On démontrait dans cet énoncé que si on dispose d'une v.a.r. V telle que $V \hookrightarrow \mathcal{E}(1)$ alors la v.a.r. $W = -\ln(V)$ suit une loi de Gumbel.

On considérait alors le script et l'histogramme résultat suivant :

```
1  V = grand(1,10000,'exp',1)
2  W = -log(V)
3  s = linspace(0,10,11)
4  histplot(s,W)
```



L'énoncé donnait l'explication suivante :

Ce script simule 10000 *variables indépendantes, regroupe les valeurs renvoyées en 10 classes qui sont les intervalles* $[0,1]$, $]1,2]$, $]2,3]$, …, $]9,10]$ *et trace l'histogramme correspondant (la largeur de chaque rectangle est égale à* 1 *et leur hauteur est proportionnelle à l'effectif de chaque classe).*

► Que représente chaque barre de l'histogramme ?
Détailler chaque ligne de la simulation en faisant le lien avec la loi (faible) des grands nombres.

- La ligne 1 permet d'obtenir des valeurs $(v_1, \ldots, v_{10000})$ qui correspondent à l'observation d'un 10000-échantillon $(V_1, \ldots, V_{10000})$ de la v.a.r. V qui suit la loi $\mathcal{E}(1)$.
(les V_i sont indépendantes et ont même loi que V)
- La ligne 2 permet d'obtenir des valeurs $(w_1, \ldots, w_{10000})$ qui correspondent à l'observation d'un 10000-échantillon $(W_1, \ldots, W_{10000})$ de la v.a.r. W qui suit la loi de Gumbel.
(les W_i sont indépendantes et ont même loi que W)
- Les lignes 3 et 4 permettent d'obtenir un histogramme des fréquences : on considère 10 classes et on compte la fréquence de chaque classe $\left(\dfrac{\text{effectif de la classe}}{\text{nombre d'observations}}\right)$.
- Si on souhaite faire un lien plus formel avec la loi (faible) des grands nombres, on peut, pour la classe 3, introduire la v.a.r. Z suivante.

$$\begin{array}{rcccl} Z & : & \Omega & \to & \mathbb{R} \\ & & \omega & \mapsto & \begin{cases} 1 & \text{si } W(\omega) \in]2,3] \\ 0 & \text{sinon} \end{cases} \end{array}$$

C'est une v.a.r. finie qui a pour espérance :

$$\mathbb{E}(Z) = 1 \times \mathbb{P}([2 < W \leqslant 3]) + \cancel{0 \times \mathbb{P}([2 < W \leqslant 3])} = \mathbb{P}([2 < W \leqslant 3]) = F_W(3) - F_W(2)$$

La loi (faible) des grands nombre stipule que : $\dfrac{1}{N}\sum_{i=1}^{N} z_i \simeq \mathbb{E}(Z) = F_W(3) - F_W(2)$ pour N suffisamment grand. Cette valeur est approchée par **l'aire** de la barre entre 2 et 3.

Par défaut, un histogramme est normalisé : en sommant les aires de chaque barre de l'histogramme, on obtient 1. C'est donc bien l'aire de chaque barre qui est importante et pas la hauteur !

On considère maintenant le programme suivant.

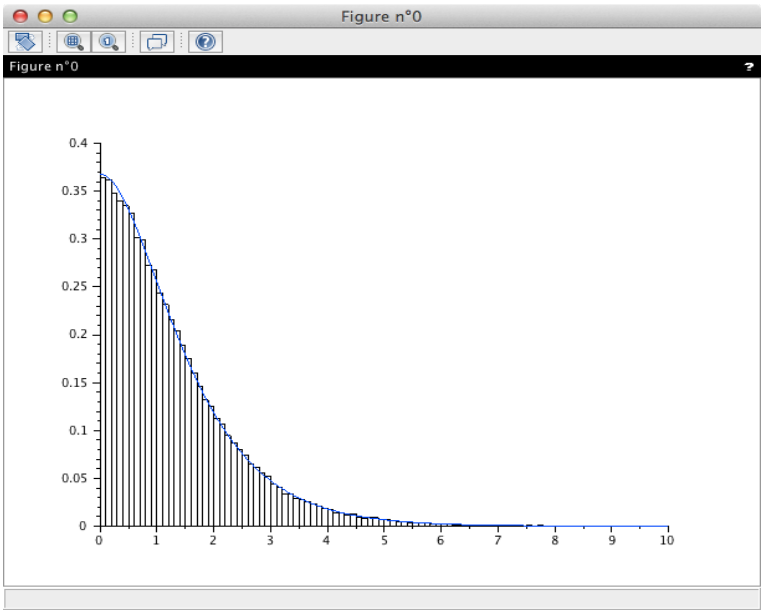
```
1 V = grand(1,100000,'exp',1) // plus d'observations
2 W = -log(V)
3 s = 0:0.1:10 // des classes plus petites
4 histplot(s,W)
5 plot(s, exp(-exp(-s)) .* exp(-s))
```

▶ Qu'observe-t-on ? Expliquer le résultat obtenu.

- En plus de la représentation d'un histogramme des fréquences, ce programme permet d'obtenir la représentation graphique sur $[0, 10[$ de la fonction $x \mapsto \mathrm{e}^{-x}\ \mathrm{e}^{-\mathrm{e}^{-x}}$ qui est une densité de la v.a.r. W sur $[0, +\infty[$.
- L'histogramme a été modifié. On considère ici 100 classes toutes de largeur $0,1$.
  Ainsi, la barre de la $31^{\text{ème}}$ est une approximation de la probabilité :

$$\mathbb{P}([3 < W \leqslant 3.1]) \ = \ \int_3^{3.1} f_W(t)\ dt$$

  On est donc en train d'approcher une aire sous la courbe de f_W sur $[3, 3.1]$ par un rectangle de largeur 0.1. Il est logique que ce rectangle soit de hauteur approximative $f_W(3)$: on retrouve en fait le schéma associé à la méthode des rectangles.



II. Simulation de v.a.r. suivant une loi uniforme continue

II.1. Simulation de v.a.r. suivant une loi uniforme sur $[0, 1]$

- Évaluer `help rand` et prendre connaissance des informations notamment celles présentes dans la section `Générer des nombres aléatoires`.
- Évaluer `rand(2,6)`. Qu'obtient-on ?

> On obtient une matrice de taille $(2, 6)$ contenant des réels choisis de manière aléatoire uniforme dans $[0, 1]$ (à l'aide du générateur pseudo-aléatoire implémentant `rand`).

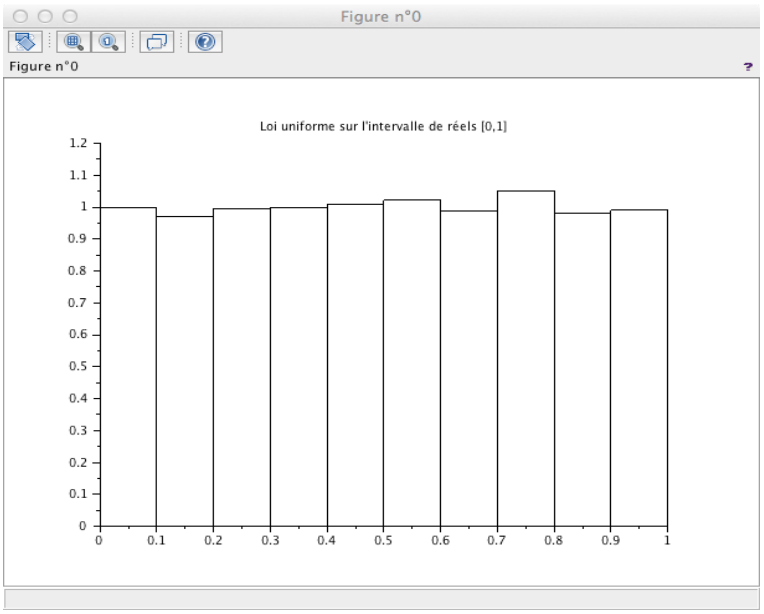
- Comparer ce résultat avec votre voisin. Commenter brièvement.

> Tout le monde dans la classe obtient le même résultat. En effet, la fonction `rand` est un générateur `pseudo-aléatoire`. Partant de la même graine, tout le monde obtient la même suite de nombres. On peut se reporter, pour plus de détails au TP de l'an dernier.

On considère le programme `uniforme_continue.sce` suivant.

```
1 clf()
2 U = rand(1,10000) // correspond à 10000 appels successifs de rand()
3 histplot(0:0.1:1,U)
4 xtitle("Loi uniforme sur [0,1] : distribution approchée")
```

- Recopier ce fichier dans l'éditeur **SciNotes** et l'exécuter (touche `F5`).
  Donner une allure de la figure obtenue et l'expliquer brièvement.



L'histogramme obtenu confirme la distribution uniforme issue de l'appel à la fonction `rand`.

Le nombre 1 affiché en ordonnée est un peu surprenant. Par défaut, l'histogramme est normalisé *i.e.* que l'ordonnée est choisie de telle sorte que la somme des aires des rectangles vaut 1.

- Remplacer la ligne 3 par `histplot(0:0.1:1, U, normalization=%f)`.
  À quoi sert de désactiver l'option `normalization` ?
  Noter le nombre d'éléments du vecteur `U` appartenant à la première classe.

> Cet argument optionnel permet de désactiver la normalisation. Les ordonnées mentionnent alors les effectifs de chaque classe. En l'occurrence, la première classe (ainsi que les autres) comporte ici environ 1000 individus.

▶ L'histogramme précédent contient 10 barres. Repérer le paramètre permettant de définir ces barres et modifiez-le de sorte à afficher 20 barres. Noter la ligne modifiée.

- L'appel `histplot(0:0.05:1, U)` permet d'obtenir 20 classes.
- On peut aussi directement demander `histplot(20, U)`. Dans ce cas, la première classe démarre au nombre le plus petit obtenu et la dernière finit au nombre le plus grand.

II.2. Loi uniforme sur $[a, b]$

On dispose du programme `uniforme_cont_2_7.sce` suivant.

```
clf()
U = 2+(7-2)*rand(1,10000) // correspond à 10000 appels successifs
histplot(2:0.5:7,U)
xtitle("Loi uniforme sur [2,7] : distribution approchée")
```

▶ Copier ce fichier dans l'éditeur **SciNotes** et l'exécuter (touche `F5`).
Expliquer brièvement le résultat obtenu.

- À l'aide de `rand`, on obtient des réels dans $[0, 1]$ (répartis de manière uniforme).
- Ce résultat est transporté dans $[0, 5]$ par la multiplication par `7-2`.
- Puis dans $[2, 7]$ par ajout de `2`.

▶ En vous inspirant du résultat précédent, écrire un programme qui :

× demande à l'utilisateur d'entrer au clavier deux réels `a` et `b` et un entier `N`,

× affiche l'histogramme des fréquences d'un échantillon de taille `N` de v.a.r. suivant la loi uniforme sur l'intervalle `[a, b]` simulée par la fonction `rand`.
Le titre de la figure devra être `Loi uniforme sur [`va`,`vb`] : distribution approchée` où va et vb désignent les valeurs entrées au clavier par l'utilisateur.

```
a = input("Entrer la borne gauche a : ")
b = input("Entrer la borne droite b : ")
N = input("Entrer la taille de l''échantillon : ")
clf()
U = a + (b-a)*rand(1,N) // correspond à N appels successifs
histplot(a:(b-a)/10:b, U)
chaineIntervalle = "[" + string(a) + "," + string(b) + "]"
xtitle("Loi uniforme sur" + chaineIntervalle + ": distrib approchée")
```

III. Simulation de v.a.r. suivant une loi uniforme discrète

III.1. Loi uniforme sur $[\![n1, n2]\!]$

On considère la fonction suivante.

```
function y = unifDiscrete(n1,n2)
    y = n1 + floor((n2-n1+1) * rand())
endfunction
```

▶ Que réalise cette fonction ?

- À l'aide de `rand`, on obtient un réel dans $[0, 1]$.
- Ce résultat est transporté dans `[0, n2-n1+1]` par la multiplication par `n2-n1+1`.
- Puis dans $[\![$`0, n2-n1`$]\!]$ par application de `floor`.
  (il faut noter que la probabilité d'obtenir 1 *à l'aide de la fonction* `rand` *est nulle)*
- Puis dans $[\![$`n1, n2`$]\!]$ par ajout de `n1`.

▶ Écrire un programme qui :

× demande à l'utilisateur d'entrer au clavier deux entiers `n1` et `n2` et un entier `N`,

× stocke `N` appels successifs à la fonction `unifDiscrete` dans un vecteur ligne `U`,

× affiche le diagramme des fréquences d'un échantillon de taille `N` de v.a.r. suivant la loi uniforme sur l'intervalle $[\![$`n1, n2`$]\!]$ simulée par la fonction `rand`.
Le titre de la figure devra être `Loi uniforme sur [|`$vn1$`,`$vn2$`|] : distribution approchée` où $vn1$ et $vn2$ désignent les valeurs entrées au clavier par l'utilisateur.

```
n1 = input("Entrer la borne gauche n1 : ")
n2 = input("Entrer la borne droite n2 : ")
N = input("Entrer la taille de l''échantillon : ")
clf()
U = zeros(1,N)
for i = 1:N
    U(i) = unifDiscrete(n1, n2)
end
histplot(n2-n1+1, U)
chaineIntervalle = "[|" + string(n1) + "," + string(n2) + "|]"
xtitle("Loi uniforme sur" + chaineIntervalle + ": distrib approchée")
```

Il faut noter que dans l'appel à `hisplot` on spécifie le nombre de barres et pas les intervalles sur lesquels on les construit. Ceci pour éviter que la première classe soit définie par `[n1, n1+1]`, ce qui produirait une barre deux fois plus grandes que les autres.

Il faut réserver l'appel à `hisplot` au travail sur des v.a.r. à densité.

▶ Évaluer dans la console `X = tabul(U)` puis `Y = tabul(U,"i")`. À quoi servent ces commandes ?

L'appel `tabul(U)` renvoie une matrice de deux colonnes.

- La première colonne liste les différentes valeurs présentes dans `U`.
- La deuxième colonne liste l'effectif de chacune de ces valeurs.

L'argument optionnel permet de choisir l'ordre dans lequel sont listés les éléments de la première colonne (`"i"` pour *increasing* et `"d"` pour *decreasing*).

▶ Évaluer alors `clf(); Y = tabul(U,"i"); bar(Y(:,1), Y(:,2))`.
Expliquer le fonctionnement de la fonction `bar` en détaillant le contenu de `Y(:,1)` et `Y(:,2)`.

L'appel `bar(X, Y)` permet de représenter un diagramme en bâtons. Le vecteur `X` représente les abscisses des bâtons et le vecteur `Y` contient la hauteur de chaque bâton.

▶ Commenter l'intérêt de l'option `width` en évaluant `clf(); bar(Y(:,1), Y(:,2), width=0.1)`.

L'option `width` permet de spécifier la largeur de chaque barre.