# Travaux Pratiques 11 : statistiques et histogrammes

**Un peu de vocabulaire**

— Une étude statistique consiste à exploiter des informations sur une **population** (= ensemble d'individus).
La population sera notée $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$ et $n = card(\Omega)$ est l'**effectif de la population**.

— Lorsque la population est trop grande, on peut être amené à restreindre l'étude à un **échantillon** (liste finie d'individus de la population sur lesquels on effectue des observations).

— Une étude porte sur un **caractère** de la population. Le caractère peut être **qualitatif** (ex. couleur des yeux) ou **quantitatif** (ex. le poids). Nous restreindrons l'étude aux caractères quantitatifs : le caractère est alors une variable aléatoire sur $\Omega$.

— L'objectif est d'avoir un bon aperçu de la répartition du caractère : il s'agit donc de trouver des grandeurs caractéristiques pertinentes.

**Modalités et fréquences :**

— On appelle série statistique l'énumération des valeurs prises $X(\omega_1), ..., X(\omega_n)$ par le caractère $X$ dans notre population. En enlevant les éventuelles répétitions, obtient $X(\Omega) = \{x_1, ..., x_p\}$ : ces valeurs prises distinctes sont appelées modalités et seront toujours rangées par ordre croissant.

— L'effectif de la modalité $x_i$ est le nombre $n_i$ d'individus qui prennent cette valeur.
Pour présenter la série statistique, on préférera la présentation synthétique sous la forme d'un tableau à deux lignes : $(x_i, n_i)_{1 \leq i \leq p}$, plutôt qu'une liste de valeurs répétées.

— La fréquence de la modalité $x_i$ est le réel $f_i = \frac{n_i}{n}$, qui représente donc la proportion d'individus dans la population qui prennent la valeur $x_i$. (rappel : $n$ est l'effectif de la population).

— On peut également définir l'effectif cumulé associé à la modalité $x_i$ : nombre d'individus prenant une valeur inférieure ou égale à $x_i$. La fréquence cumulée associée à $x_i$ est alors la proportion d'individus prenant une valeur inférieure ou égale à $x_i$.

***Exemple :***

Soit la série statistique suivante : $2, 11, 7, 2, 15, 4, 5, 5, 5, 13, 5, 15, 7, 7, 8, 10, 10, 10, 11, 13, 7, 2, 15, 15$.

Compléter le tableau suivant :

| modalités | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| effectifs | | | | | | | | | |
| fréquences | | | | | | | | | |
| effectifs cumulés | | | | | | | | | |
| fréquences cumulées | | | | | | | | | |

Faire alors le diagramme en bâtons des effectifs de cette série statistique, puis celui des fréquences cumulées.

**Moyenne et écart-type :**

— La moyenne d'une série statistique non rangée est :

$$\bar{x} = \frac{\text{somme de toutes les valeurs prises}}{n} = \frac{\sum_{i=1}^{n} X(\omega_i)}{n} \text{ (où } n \text{ est l'effectif de la population)}$$

Une fois la série statistique rangée par modalités, cette moyenne s'écrit :

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + ... + n_p x_p}{n} = \frac{1}{n} \sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} f_i x_i.$$

*Retour exemple :* Ecrire de 2 manières différentes la moyenne de la série statistique précédente.
*On pourra ne pas faire les calculs jusqu'au bout ....*

— La variance d'une série statistique non rangée est le nombre $v = \frac{1}{n} \sum_{i=1}^{n} (X(\omega_i) - \bar{x})^2$.

Si la série statistique a été regroupée par modalités : $v = \frac{1}{n} \sum_{i=1}^{p} n_i (x_i - \bar{x})^2$.

L'écart-type est le réel $\sigma = \sqrt{v}$.

Ces données mesurent la **dispersion de la série statistique** autour de sa moyenne.

**Médiane et quantiles :**

— La médiane d'une série statistique est un réel qui sépare la série en deux séries de même effectif. Un tel nombre existe toujours (il suffit d'ordonner la série statistique des énumérations et de la couper en 2), mais peut ne pas être unique : par convention, quand plusieurs réels sont possibles on prend le milieu.

*Retour exemple :* Déterminer la médiane de la série statistique.

— Plus généralement, on peut définir des quartiles :

Le premier quartile est la plus petite valeur $Q_1$ de la série telle qu'au moins 25% des valeurs de la série soient inférieures ou égales à $Q_1$.
Le troisième quartile est la plus petite valeur $Q_3$ de la série telle qu'au moins 75% des valeurs de la série soient inférieures ou égales à $Q_3$.

*Retour exemple :* Donner les valeurs de $Q_1$ et $Q_3$.

**Syntaxes Python pour ces grandeurs caractéristiques** :
Importer la bibliothèque `numpy` et pour les graphiques la bibliothèque `matplotlib.pyplot`.
Les séries statistiques seront introduites à l'aide de la syntaxe `np.array()`.
Pour les caractéristiques citées ci-dessus, les syntaxes sont : `np.mean, np.var, np.std, np.median ...`
On pourra également avoir besoin des syntaxes `np.min, np.max` pour l'étendue de la série statistique et `np.cumsum` pour les effectifs (ou les fréquences) cumulés.

**Exercice 1:**

Soit la série statistique des notes suivantes $S = [15, 10, 12, 8, 10, 18, 12, 8, 8, 15, 10, 8, 6, 18, 12, 8, 12]$

1. Introduire dans python le vecteur S contenant la série statistique des notes.
2. Calculer avec python la moyenne, la variance, l'écart-type ainsi que la médiane de la série.
3. Ordonner à la main la série en présentant le tableau avec deux lignes : valeurs prises et effectifs.
4. Introduire dans python le vecteur X des valeurs prises et le vecteur E des effectifs.
   Faire déterminer à python le vecteur F des fréquences.
5. Faire tracer le diagramme en bâtons des effectifs dans python : la syntaxe est `plt.bar(X,E)`. Ne pas oublier de finir par `plt.show()`.
6. Variante : afin, de ne pas avoir à introduire $E$ et $X$, on peut utiliser la syntaxe python d'un histogramme qui utilisera $S$ (càd la série statistique initiale, non triée etc.)

```
b=np.arange(5.5,19.5)
plt.hist(S,bins=b)
plt.show()# inutile avec certains éditeurs
```

   Que fait la commande plt.hist () ? Python commence par compter le nombre d'occurences de l'échantillon L qui se retrouvent dans chaque intervalle de bins càd le nombre d'occurences dans l'intervalle [5.5,6.5], puis le nombre d'occurences dans l'intervalle [6.5,7.5],... jusqu'à l'intervalle [17.5,18.5].
   Comme notre échantillon ne contient que des valeurs entières entre 6 et 18, le nombre d'occurences dans l'intervalle [5.5,6.5] correspond bien au nombre de 6 obtenus, etc. Puis python dessine des barres dont la hauteur correspond à ce nombre d'occurences.
   On peut également spécifier la couleur des barres via l'argument `color`, ou la largeur des barres via l'argument `rwidth`.
   *Remarque :* pour faire afficher les intervalles de `bins` ainsi que le nombre d'occurences qui tombent dans chaque intervalle, il faut parfois (selon les éditeurs) rajouter la syntaxe `print(plt.hist(....))` dans le script.
   L'histogramme pourra également s'avérer intéressant quand les valeurs des caractères seront des réels et qu'il faudra les regrouper par petits intervalles (cf variables à densité en 2e année)
7. Pour pouvoir comparer des graphiques issus de populations de tailles variables, on choisira de mettre en ordonnées les fréquences ds valeurs prises plutôt que les effectifs : on parle alors de diagramme ou d'histogramme **renormalisé**.
   Pour le diagramme en bâtons, la syntaxe devient `plt.bar(X,F)` et pour l'histogramme
   `plt.hist(S,bins=b, density=True)`
   Faire afficher dans python le diagramme et l'histogramme renormalisé.

**Lois et histogrammes :**

On a vu dans le TP 10 qu'une valeur approchée de la probabilité d'un événement $A$ pouvait être obtenue en calculant la fréquence d'apparition de cet événement, lors de la réalisation de $n$ expériences, avec $n$ suffisamment grand.

Imaginer que l'on cherche maintenant à **connaitre la loi** d'une variable aléatoire $X$.

Connaître sa loi, c'est connaître toutes ses probabilités d'apparition.

En combinant le résultat rappelé ci-dessus et la notion de série statistique, on aboutit au **protocole** suivant :

| |
|---|
| On réalise une série statistique (ou échantillon) en simulant $n$ fois notre variable aléatoire $X$. On regarde alors les effectifs puis les fréquences de chaque valeur prise(= modalité). Si on représente ces fréquences en fonction des valeurs prises (diagramme en bâtons ou histogramme), on peut ainsi obtenir une approximation de la loi théorique de $X$. |

Dans les exercices suivants, on va valider informatiquement ce protocole, et voir que choisir $n = 10000$ reste un bon compromis entre précision et temps de calcul. On l'appliquera alors pour voir que la loi Binomiale peut être approchée par la loi de Poisson dans certains cas ...

**Exercice 2:** *Loi Uniforme discrète sur ⟦1,6⟧*

1. Dans la console, créer un vecteur $L$ contenant 100 simulations de la loi $\mathcal{U}(\llbracket 1,6 \rrbracket)$, à l'aide de la syntaxe `rd.randint()` et le visualiser.
2. Quelle est la moyenne de cet échantillon ? Et la variance ? A comparer avec l'espérance et la variance théorique de la loi $\mathcal{U}(\llbracket 1,6 \rrbracket)$.
3. Taper puis exécuter le code suivant dans l'éditeur :
```
import numpy as np
import numpy.random as rd
import matplotlib.pyplot as plt
L=rd.randint(1,7,100)
b=np.arange(0.5,7.5)
plt.hist(L,range=(1,6),bins=b,color='r',rwidth=0.5)
plt.show()# inutile avec certains éditeurs
```
4. Rajouter l'argument `density=True` dans `plt.hist()` afin d'afficher l'histogramme renormalisé : les hauteurs des bâtons ne sont plus les nombres d'occurences mais les fréquences d'apparitions.
   La renormalisation est nécessaire pour pouvoir comparer des histogrammes dont les échantillons n'ont pas même taille ou pour les comparer avec des histogrammes "théoriques" c'est-à-dire des diagrammes en bâtons.
5. Refaire l'histogramme plusieurs fois (bien sûr en changeant L). Puis le refaire plusieurs fois en prenant 10 000 simulations. Que constatez-vous ?
6. Enfin, refaire un dernier histogramme à partir de 10 000 simulations et le comparer au diagramme en bâtons de la loi théorique en rajoutant la syntaxe `plt.bar(np.arange(1,7), 1/6*np.ones(6), color='b', width=0.3)`. Que constatez-vous ?

**Regroupez vous par 3**, pour l'exercice suivant : chacun choisit une loi différente et montre le graphique aux autres.

**Exercice 3:** *Loi Binomiale* $\mathcal{B}(10, 0.3)$, *Loi Géométrique* $\mathcal{G}(1/3)$, *Loi de Poisson* $\mathcal{P}(5)$

1. Créer un vecteur $L$ de 10000 simulations (= échantillon) de la loi choisie.
2. Quelle est la moyenne et la variance de cet échantillon ? Comparer avec les valeurs théoriques attendues.
3. Quelles sont les valeurs extrêmes de votre échantillon ? A comparer avec l'ensemble de valeurs prises théoriquement.
4. Faire alors l'histogramme de votre échantillon.
5. Mémoriser l'allure obtenue en répondant aux questions suivantes : forme de cloche ? symétrique ? centrée autour de quelle valeur ? espérance ? nombre de valeurs prises en pratique ?
6. Pour la loi géométrique uniquement, on va comparer le diagramme en bâtons de la loi théorique avec l'histogramme obtenu à partir de 10 000 simulations (pour les autres, aidez votre voisin à comprendre ce qui suit !).
   Commencer par renormaliser votre histogramme puis réaliser qu'il suffit de se positionner sur $\llbracket 1, 20 \rrbracket$ car au-delà, les probabilités sont quasi-nulles. Il reste à rajouter le script suivant dans votre fichier et à conclure :
```
x=np.arange(1,21)
y=(1/3)*(2/3)**(x-1)
plt.bar(x,y,color='r',width=0.3)
```

**Exercice 4:** *Approximation de la loi* $\mathcal{B}(n, \frac{\lambda}{n})$ *par la loi* $\mathcal{P}(\lambda)$ *lorsque* $n \to +\infty$

1. Créer un vecteur `xB` contenant 10 000 simulations de la loi $\mathcal{B}(30, \frac{5}{30})$, et un vecteur `xP` contenant 10 000 simulations de la loi $\mathcal{P}(5)$.
2. Représenter les deux histogrammes associés à ces simulations, puis commenter. (Penser à changer de couleur et d'épaisseur afin de gagner en lisibilité).
3. *bonus :* Soit $k \in \mathbb{N}$ fixé. Montrer par le calcul que $P(X_n = k) \underset{n\to+\infty}{\longrightarrow} P(Y = k)$ où $X_n \hookrightarrow \mathcal{B}(n, \frac{\lambda}{n})$ et $Y \hookrightarrow \mathcal{P}(\lambda)$.