Thème : Probabilité, statistique et informatique
Titre : Enquête menée auprès de femmes au BENIN
Auteur : BRO FRÉDÉRIC - TARI KEVIN
Objectifs :

- ▷ Programmer en PYTHON et utiliser notamment le module PANDAS pour étudier une étude statistique.
- ▷ Utiliser une représentation graphique pertinente pour conjecturer un phénomène statistique singulier (Data visualisation)
- ▷ Modéliser le calcul d'une loi de probabilité.

USAID est un organisme public américain qui développe un programme nommé DHS, visant à collecter des données relatives à la démographie et la santé dans les pays en développement.
Une enquête a été menée au **Bénin** auprès de femmes âgées de 15 à 49 ans.
Le fichier `Benin.csv` recense pour chaque femme interrogée :

— l'année de naissance déclarée par celle-ci
— le numéro du mois de naissance déclaré par celle-ci
— l'année durant laquelle celle-ci a été interrogée
— le numéro du mois durant lequel elle a été interrogée
— l'âge déclaré par celle-ci

Objectif :

Étudier la distribution des âges déclarés.

Partie A : Représentation de la distribution des ages

1. **a.** Ouvrir un notebook et recopier les instructions suivantes permettant de charger les modules nécessaires :

```
In [1]: import pandas as pa
        import pylab as pl
        from math import *
        T=pa.read_csv('Benin.csv')
        T.head()
```

T est une « table de données » *(dataframe)* liée au fichier `Benin.csv`.

b. Pour obtenir un résumé statistique de la colonne **age**, exécuter l'instruction :

```
In [2]: T['age'].describe()
```

Compléter alors le tableau ci-joint :

$\overline{x}$	σ	Q_1	m_e	Q_3

c. Combien de femmes ont été interrogées ?

2. **a.** Pour obtenir la table **S** associée à la répartition des âges des femmes interrogées, exécuter :

```
In [4]: S=T['age'].value_counts()
```

b. Exécuter l'instruction `S.sort_index()`. Que représente-elle ?

c. Représenter le diagramme en barres associé à `S.sort_index()`.

Avec Python

Soit **H** une colonne de nombres.
`H.plot(kind='bar')` permet de représenter le diagramme en barres associé.

d. Quelle remarque peut-on faire ?

Objectif :

On choisit au hasard une femme d'âge compris dans [20 ; 50[.
Calculons la probabilité, notée p, que la personne choisie ait son âge égal à un multiple de 5.

Partie B : En théorie

On se limite à la tranche d'âge [20 ; 50[.

1. Compléter la première ligne du tableau suivant :

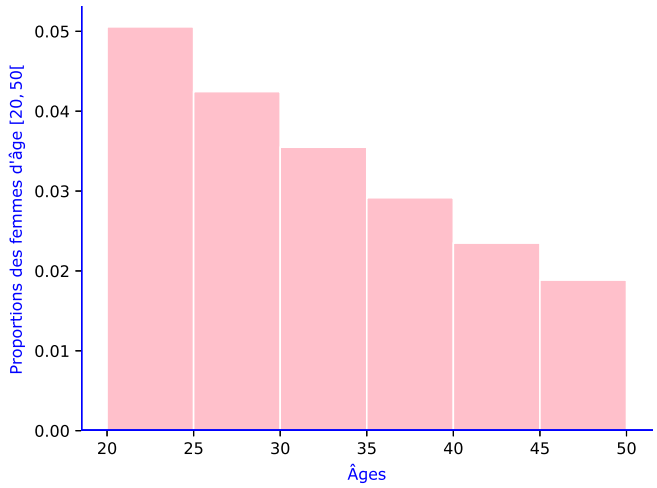
Tranche d'âge	[20 ; 25[	[25 ; 30[	[30 ; 35[	[35 ; 40[	[40 ; 45[	[45 ; 50[
Effectif	318 487	267 423	223 555	183 715	147 920	118 757
Fréquences						

a. Créer la liste `Pop` associée à ces effectifs.

b. Créer la variable `N` qui correspond au nombre total de femmes *(du Bénin)* appartenant à la tranche d'âge [20 ; 50[.

c. En déduire la liste `Freq` associée aux fréquences puis compléter la fin du tableau.

Voici la répartition de l'âge des femmes dans cette tranche :



d. Quelle est l'aire totale de cet histogramme ?

e. On choisit au hasard une femme dans la tranche d'âge [20 ; 50[.

i. Quelle est la probabilité que son âge soit compris dans [20 ; 25[?

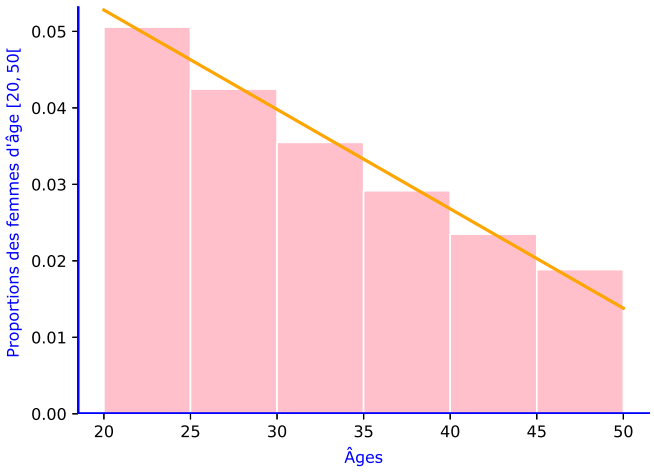
ii. Est-il possible de calculer la probabilité que son âge soit de 20 ans ?

Modélisation :

— On ajuste au mieux cet histogramme par une droite (D). Elle a pour équation :

$$y = -0,0013x + 0,0788.$$

La zone délimité par l'axe des abscisses, la droite (D) et les droites d'équations $x = 20$ et $x = 50$ a pour aire 1.



— On choisit au hasard une femme d'âge compris dans [20 ; 50[.
On modélise la probabilité que son âge soit égal à A par l'aire du trapèze délimité par l'axe des abscisses, la droite (D) et les droites d'équations $x = A$ et $x = A + 1$.

2. Écrire la fonction nommée **f** de paramètre x *(l'abscisse d'un point de* (D)*)* qui renvoie y *(l'ordonnée d'un point de* (D)*)* .

3. Écrire la fonction nommée **proba** de paramètre x *(l'âge d'une femme)* qui renvoie la probabilité que l'âge de la femme choisie au hasard soit égal à x.

4. Colorier la zone d'aire égale à la probabilité p, que l'âge d'une femme choisie au hasard soit un multiple de 5.

5. Calculer p.

Partie C : Pour l'échantillon

1. Écrire la fonction nommée **multiple_5** de paramètre age *(l'âge d'une femme de l'échantillon)*.
Elle renvoie 1 si age est un multiple de 5 et 0 sinon.

2. On applique cette fonction à chaque élément de **T** contenu dans la colonne **age** et on stocke les résultats dans une colonne nommée **multiple**.
Exécuter le code suivant :

```
In [15]: T['multiple']=T['age'].apply(multiple_5)
```

3. **Fréquence des femmes d'âge supérieur à** 20 **ans e**

a. Quelle requête permet d'obtenir les lignes de **T** associées seulement aux femmes ayant un âge égal à un multiple de 5 et un âge supérieur à 20 ans ?

Avec Python

- `T.query('multiple==1')` est la requête qui crée la table contenant les lignes de **T** qui ont un 1 seulement dans la colonne **multiple**.
- `len(T)` donne le nombre de ligne de **T**.

b. En déduire la fréquence observée des femme d'âge supérieur à 20 ans et égal à un multiple de 5.

On notera f cette fréquence.

c. Vérifier que le nombre de femmes qui sont âgées de plus de 20 ans *(dans ce pannel)* est $n = 19\,160$.

d. Considérons un échantillon de $n = 19\,160$ femmes de plus de 20 ans, interrogées au hasard dans la population.

*Étant donné que la population de femmes au **Bénin** est suffisamment grande par rapport à n, les tirages peuvent être considérés comme des tirages avec remise et indépendants les uns des autres.*

Déterminer l'intervalle de fluctuation de la fréquence observée de femmes ayant un âge de plus de 20 ans et égal à un multiple de 5 *(avec un seuil de confiance de 95%)*.

e. Quelle conclusion peut-on donner quant au panel de ces femmes interrogées ?

Partie D : Croisement de la variable age avec la variable etude

La colonne de **T** nommée **etude** concerne le niveau d'étude de la femme interrogée :

Niveau	0	1	2	3
Signification	pas scolarisée	primaire	secondaire	universitaire

1. Quel l'effectif de chaque niveau dans le panel des femmes interrogées ?

2. Calculer la part que représente chaque niveau par rapport à l'ensemble du panel.

3. **Tableau des effectifs :**

a. Exécuter le code suivant :

```
In [15]: M=pa.crosstab(T['catu'],T['grav'],margins=True)
         M
```

Ceci permet d'obtenir le tableau suivant :

Out[15]:

multiple etude	0	1	All
0	9565	5580	15145
1	2188	933	3121
2	769	294	1063
3	57	12	69
All	12579	6819	19398

b. On observe qu'il y a 5 580 femmes qui ont un âge égal à un multiple de 5 et qui n'ont pas fait d'études.

Quel est le nombre de femmes qui ont été à l'université et qui ont un âge égal à un multiple de 5 ?

4. Tableau des fréquences :

a. Pour obtenir le tableau des fréquences par rapport à l'ensemble des femmes du panel, exécuter le

```
In [23]: M=pa.crosstab(T['etude'],T['multip
                       margins=True,normali
         M.round(4)*100
```

Avec python

`M.round(4)` arrondit chaque élément de `M` avec 4 chiffres après la virgule

b. Exécuter le code suivant et indiquer ce qu'il représente :

```
In [24]: M[1]/M['All']
```

c. Compléter les instructions en pointillé ci-dessous pour substituer chaque colonne de **M** par elle-même divisée par la colonne `'All'` :

```
In [25]: for x in M.columns:
             M[x]=...................
         M.round(4)*100
```

Out[25]:

multiple etude	0	1	All
0			
1			
2			
3			
All			

d. L'annonce d'un âge égal à un multiple de 5 par les femmes est-il dépendant du niveau d'études ?