Cours de Probabilités et statistiques
Licence 2 d'Informatique
Chapitre 1: Statistique descriptive

Yoann Dabrowski

Université Lyon 1

7 septembre 2022, Séance 1, 3H

Buts de ce cours

1. **But de la statistique** : déduire quelle est la meilleure observation ou le meilleur calcul à faire sur les résultats d'observations pour obtenir efficacement une information cherchée.
2. C'est donc un domaine à l'intersection des probabilités (le modèle mathématique du hasard) et de l'optimisation.

Buts de ce cours :

1. Apprendre les bases de **statistique descriptive** : résumer et illustrer graphiquement des grands ensembles de données.
2. Utiliser le **Python (pandas,scipy,matplotlib)** pour illustrer les notions et résultats du cours.
3. Introduire le **modèle probabiliste**, faire des calculs simples, appréhender la notion d'indépendance. (Illustrer la simulation de ce modèle en Python)
4. Réaliser des **Inférences statistiques** : Estimer la précision d'une mesure (dans un intervalle). Tester sur les données des hypothèses (comparaison de paramètre, indépendance)

Plan du cours

1. Chapitre 1 : **Statistique descriptive** (cours : 4H30, TD : 6H)
2. Chapitre 2 : **Le modèle probabiliste : les Espaces de probabilités** (cours : 2H30)
3. Chapitre 3 : **Probabilités conditionnelles, Indépendance** (cours : 2H00, TD : 3H)
4. Chapitre 4 : **Variables aléatoires discrètes** (cours : 3H, TD : 4H30)
5. Chapitre 5 : **Rappels et Compléments d'Intégration sur $\mathbb{R}$** (cours : 1H30, TD : 3H)
6. Chapitre 6 : **Variables aléatoires continues** (cours : 1H30, TD : 4H30)
7. Chapitre 7 : **Théorèmes limites, application aux intervalles de confiance** (cours :3H, TD : 4H30)
8. Chapitre 8 : **Introduction à la Statistique inférentielle : tests d'hypothèse** (cours : 3H, TD : 4H30)

Évaluation

Organisation

Plan de cette séance : début du chapitre 1

- Présentation du cours.

Chap 1 Statistique descriptive.

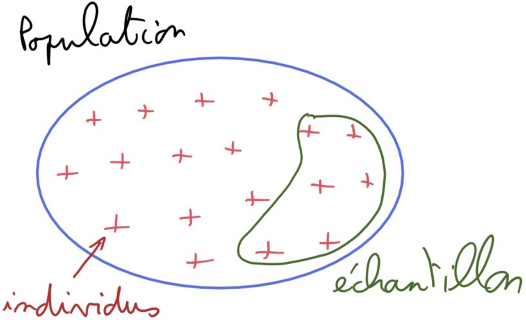
- 1) Vocabulaire statistique et Types de variables statistiques

  2) Représentations graphiques (pour une variable)
- 2.1) Notions d'effectif et de fréquence.
- 2.2) Diagrammes circulaires ou en tuyau d'orgue (pour les variables qualitatives)
- 2.3) Diagrammes cumulatifs (pour les variables ordinales)
- 2.4) Diagrammes en bâton (pour les variables quantitatives discrètes)
- 2.5) Histogrammes (pour les variables quantitatives continues)

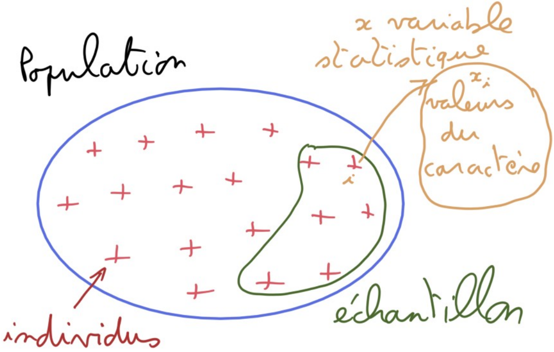
  3) Résumés numériques (pour une variable quantitative)
- 3.1) Moyenne et variance
- 3.2) Médiane et Quartiles
- 3.3) Diagrammes à moustache.

1) Vocabulaire statistique

1. Une observation statistique est réalisée sur une *population*. C'est un ensemble dont les éléments, appelés *individus*, ont des caractères que l'on observe.
2. La partie observée de la population est *l'échantillon (statistique)* (*sample*)



1) Vocabulaire statistique



1. Les caractères observés sont appelés variables statistiques. Ils ne seront donc connus que pour les individus de l'échantillon.
2. Les données statistiques sont les valeurs des variables statistiques pour chaque individu.

1) Exemple de données statistiques

Voici le (début du) fichier texte rassemblant une sélection des résultats recueillis dans une étude sur l'alimentation d'un échantillon de personnes âgées résidant à Bordeaux (Gironde, France), interrogées en 2000 dans le cadre d'une enquête nutritionnelle. L'échantillon est constitué de 226 sujets.

sexe	*the*	*cafe*	*taille*	*poids*	*age*	*viande*	*poisson*	*matgras*
F	0	0	151	58	72	4	3	6
F	1	1	162	60	68	5	2	4
F	0	4	162	75	78	3	1	4
F	0	0	154	45	91	0	4	2
F	2	1	154	50	65	5	3	2
F	2	0	159	66	82	4	2	3
F	2	0	160	66	74	3	3	6
H	0	3	166	80	78	5	0	4
				...				

1) Exemple de données statistiques

Les données sont extraits du livre *Le logiciel R : maîtriser le langage effectuer des analyses statistiques* de Pierre Lafaye de Micheaux, Rémy Drouilhet et Benoit Liquet aux éditions Springer (que nous recommandons) et disponible à la page :

`http://www.biostatisticien.eu/springeR/jeuxDonnees4.html`

Elles sont fournies avec une information (expliquant le codage de l'information recueillie, extrait du même livre)

1) Exemple de codage des données statistiques

Description	Unité ou Codage	Variable
Sexe	F=Femme; H=Homme	sexe
Consommation journalière de thé	Nombre de tasses	the
Consommation journalière de café	Nombre de tasses	cafe
Taille	cm	taille
Poids	kg	poids
Age le jour de l'entretien	Années	age
Consommation de viande	0=Jamais 1=Moins d'une fois par semaine 2=Une fois par semaine 3=2/3 fois par semaine 4=4/6 fois par semaine 5=Tous les jours	viande
Consommation de poisson	Idem	poisson
Matière grasse préférentiellement utilisée pour la cuisson	1=Beurre 2=Margarine 3=Huile d'arachide 4=Huile de tournesol 5=Huile d'olive 6=Mélange d'huile (type Isio4) 7=Huile de colza 8=Graisse de canard ou d'oie	matgras

sexe	*the*	*cafe*	*taille*	*poids*	*age*	*viande*	*poisson*	*matgras*
F	0	0	151	58	72	4	3	6

...

1)Types de variables statistiques : a) qualitatives

Les variables qualitatives (*categorical variables*) sont les variables non numériques. On en distingue 2 types :

1. Les variables qualitatives nominales prennent valeur dans un ensemble fini SANS notion d'ordre naturel sur ses éléments (type `categorical` en python avec option “`ordered=False`").

   Ex : le sexe ou la consommation de matière grasse

2. Les variables qualitatives ordinales prennent valeur dans un ensemble fini AVEC une notion d'ordre naturel mais sans qu'une valeur numérique précise puisse être attribuée à chaque classe (On peut attribuer une telle valeur, mais elle est en parti arbitraire). (type `categorical` en python avec option “ordered=True")

   Ex : la consommation de viande ou de poisson (manger 2 à 3 fois de la viande par semaine est plus que jamais, d'où l'ordre)

1)Types de variables statistiques : b) quantitatives

Les variables quantitatives (*numerical variables*) sont les variables numériques (à valeurs réelles). On en distingue 2 types :

1. Les variables quantitatives discrètes prennent valeur dans un ensemble fini ou dénombrable (c'est à dire en bijection avec les entiers) (type `int64` si elles sont entières). Elles seront représentables par des variables aléatoires discrètes.

   Ex : le nombre de tasse de café ou de thé bus par jour.

2. Les variables quantitatives continues prennent valeur dans les réels, même si ils sont recueillis avec une certaine précision (qui les ramène à des variables discrètes avec un grand nombre de valeurs possibles). (type `float64` en python). Elles seront représentables par des variables aléatoires continues.

   Ex : la taille, l'âge ou le poids.

Dans l'exemple ci-dessus, chaque ligne correspond à un individu différent et les données sont incorporées en Python sous la forme d'un tableau individus×variables dits `data.frame`.

sexe	*the*	*cafe*	*taille*	*poids*	*age*	*viande*	*poisson*	*matgras*
F	0	0	151	58	72	4	3	6
F	1	1	162	60	68	5	2	4
F	0	4	162	75	78	3	1	4
F	0	0	154	45	91	0	4	2
F	2	1	154	50	65	5	3	2
F	2	0	159	66	82	4	2	3
F	2	0	160	66	74	3	3	6
H	0	3	166	80	78	5	0	4

...

Dans l'exemple ci-dessus, chaque ligne correspond à un individu différent et les données sont incorporées en Python sous la forme d'un tableau individus×variables dits `data.frame`.

Vous verrez en TP les vecteurs et matrices dont toutes les données doivent être du même type par exemple le plus souvent numérique ou entière. Dans un `data.frame`, les données peuvent être de types différents, comme le texte (type `character`) "H" ou "F" spécifiant le sexe dans l'exemple précédent.

```
import pandas as pan #import du package pour les
    données, on choisit l'abréviation pan
url="http://math.univ-lyon1.fr/homes-www/dabrowski/
    enseignement/StatInfo/nutriageSimplifie.csv"
df=pan.read_csv(url,sep="\t")#on charge les données
print(df.iloc[2,4])#affiche pour le troisième
    individus (car python commence par indicer à 0) le
     poids (la cinquième colonne, colonne 4 en python)
     ici 75
```

Typage des données statistiques en Python : Python ne peut pas deviner l'encodage des données

```
f.dtypes
#Réponse
sexe object
the int64
cafe int64
taille int64
poids int64
age int64
viande int64
poisson int64
matgras int64
```

Typage des données statistiques en Python : exemple simple de création de variables nominales

```
Listecol=["bleu", "bleu","vert","bleu", "bleu","marron
    ","marron"]
col=pan.Series(Listecol,dtype='category')
print(col)#
yeux=pan.DataFrame({"col":Listecol},dtype='category')
print(yeux.col)
#Réponse:
0 bleu
1 bleu
2 vert
3 bleu
4 bleu
5 marron
6 marron
Name: col, dtype: category
Categories (3, object): [bleu, marron, vert]
```

Typage des données statistiques en Python : variables nominales

```
df['sexe']=df['sexe'].astype('category')
print(df['sexe'])
#Réponse:
0 F
1 F
2 F
3 F
4 F
      ..
221 F
222 F
223 H
224 F
225 F
Name: sexe, Length: 226, dtype: category
Categories (2, object): [F, H]
```

Changer le nom des catégories d'une variables nominales

```
df['sexe'].cat.categories=["Femme","Homme"]
print(df['sexe'])
#Réponse:
0  Femme
1  Femme
2  Femme
3  Femme
4  Femme
        ...
221  Femme
222  Femme
223  Homme
224  Femme
225  Femme
Name: sexe, Length: 226, dtype: category
Categories (2, object): [Femme, Homme]
```

Typage des données statistiques : variables ordinales

```
df['viande']=df['viande'].astype('category')
freq=["jamais","<1/sem.","1/sem.","2-3/sem.","4-6/sem.
    ","1/jour"] #on crée une liste avec les noms de
    niveaux
freq_type=pan.CategoricalDtype(categories=freq,ordered
    =True)
df['viande'].cat.categories=freq #on change les noms
df['viande']=df['viande'].astype(freq_type)#on passe
    au type ordonné
#Réponse:
0 4-6/sem.
1 1/jour
        ...
225 2-3/sem.
Name: viande, Length: 226, dtype: category
Categories (6, object): [jamais < <1/sem. < 1/sem. <
    2-3/sem. < 4-6/sem.< 1/jour]
```

Typage des données statistiques : autres variables

```
df['poisson']=df['poisson'].astype('category')
df['poisson'].cat.categories=freq
df['poisson']=df['poisson'].astype(freq_type)

df['taille']=df['taille'].astype('float64')
df['poids']=df['poids'].astype('float64')
df['age']=df['age'].astype('float64')


df.dtypes
#sexe category
#the int64, cafe int64
#taille float64, poids float64,age float64
#viande category,poisson category,matgras category
```

Plan de cette séance : début du chapitre 1

- Présentation du cours.

Chap 1 Statistique descriptive.

- 1) Vocabulaire statistique et Types de variables statistiques

  2) Représentations graphiques (pour une variable)
- 2.1) Notions d'effectif et de fréquence.
- 2.2) Diagrammes circulaires ou en tuyau d'orgue (pour les variables qualitatives)
- 2.3) Diagrammes cumulatifs (pour les variables ordinales)
- 2.4) Diagrammes en bâton (pour les variables quantitatives discrètes)
- 2.5) Histogrammes (pour les variables quantitatives continues)

  3) Résumés numériques (pour une variable quantitative)
- 3.1) Moyenne et variance
- 3.2) Médiane et Quartiles
- 3.3) Diagrammes à moustache.

Notions d'effectif et de fréquence (tous les types de variable)

L'échantillon est constitué de 226 sujets. Soit I l'ensemble des individus (correspondant aux lignes du tableaux). On note $(x_i)_{i \in I}$ la liste des valeurs prises par une variable statistique x (les valeurs d'une colonne).

sexe	*the*	*cafe*	*taille*	*poids*	*age*	*viande*	*poisson*	*matgras*
F	0	0	151	58	72	4	3	6
F	1	1	162	60	68	5	2	4
F	0	4	162	75	78	3	1	4
F	0	0	154	45	91	0	4	2
F	2	1	154	50	65	5	3	2
F	2	0	159	66	82	4	2	3
F	2	0	160	66	74	3	3	6
H	0	3	166	80	78	5	0	4
				...				

On peut toujours compter le nombre d'individus prenant une valeur. Ici, on voit 7 femmes sur les 8 premiers individus.

Notions d'effectif (pour tous les types de variable)

Soit I l'ensemble des individus. On note $(x_i)_{i \in I}$ la liste des valeurs prises par une variable statistique x.

Définition

L'effectif total de l'échantillon est le cardinal de I

$$N = Card(I).$$

L'effectif (partiel) du caractère c (anglais : *absolute frequency*, ATTENTION au faux ami et à l'anglicisme, voir fréquence en français au transparent suivant) est le cardinal suivant :

$$N_c = Card\{i \in I : x_i = c\}.$$

Les modes de la variable statistique x sont les caractères c (souvent le caractère) d'effectif maximum.

```
df["sexe"].size# Calcule l'effectif total de l'
    échantillon sous-jacent à la variable. Attention:
    length(nutriage) est le nombre de caractères, ici
    9.
#Réponse: [1] 226.

TCsexe=df["sexe"].value_counts()# Crée un tableaux des
     comptes des effectifs de chaque caractère

print(TCsexe)

#Réponse:
#Femme 141
#Homme 85
#Name: sexe, dtype: int64
```

Femme est le mode de la variable statistique sexe.

Notions de table de contingence : Ex. en Python

Une table de contingence (d'effectifs) résume les effectifs de paires de caractères. Ici est le calcul en Python.

```
table=pan.crosstab(df["sexe"],df["matgras"])
print(table)
```

	matgras							
sexe	*beurre*	*margarine*	*arachide*	*tournesol*	*olive*	*Melange*	*colza*	*canard*
Femme	5	17	32	47	20	18	1	1
Homme	10	10	16	21	20	5	0	3

Par exemple, 5 femmes de l'échantillon cuisinaient au beurre.

Notion de fréquence (pour tous les types de variable)

Soit $N = Card(I)$ l'effectif total de l'échantillon I.

Définition

La fréquence du caractère c (anglais : *relative frequency*) est le rapport de l'effectif N_c du caractère c sur l'effectif total

$$f(c) = \frac{N_c}{N} = \frac{Card\{i \in I : x_i = c\}}{N}.$$

Attention : pour les variables ordinales et les variables quantitatives, on verra une notion DISTINCTE de fréquence cumulée qu'il ne faudra pas confondre...

Notions de fréquence et de table de contingence de fréquences : Ex. en Python

```
tableFreq=pan.crosstab(df["sexe"],df["matgras"],
    normalize=True,margins=True)
print(tableFreq)
```

	matgras								
sexe	*beurre*	*margarine*	*arachide*	*tournesol*	*olive*	*Melange*	*colza*	*canard*	*Total*
Femme	0.02212	0.07522	0.14159	0.20796	0.08849	0.07964	0.00442	0.00442	0.62389
Homme	0.04424	0.04424	0.07079	0.09292	0.08849	0.02212	0.00000	0.01327	0.37610
Total	0.06637	0.11946	0.21238	0.30088	0.17699	0.10176	0.00442	0.01769	1.00000

Par exemple, 7,079% des personnes interrogées étaient des hommes cuisant à l'huile d'arachide. L'huile de Tournesol est le mode de la variable `matgras` sur cet échantillon.

Plan de cette séance : début du chapitre 1

- Présentation du cours.

Chap 1 Statistique descriptive.

- 1) Vocabulaire statistique et Types de variables statistiques

2) Représentations graphiques (pour une variable)

- 2.1) Notions d'effectif et de fréquence.
- 2.2) Diagrammes circulaires ou en tuyau d'orgue (pour les variables qualitatives)
- 2.3) Diagrammes cumulatifs (pour les variables ordinales)
- 2.4) Diagrammes en bâton (pour les variables quantitatives discrètes)
- 2.5) Histogrammes (pour les variables quantitatives continues)

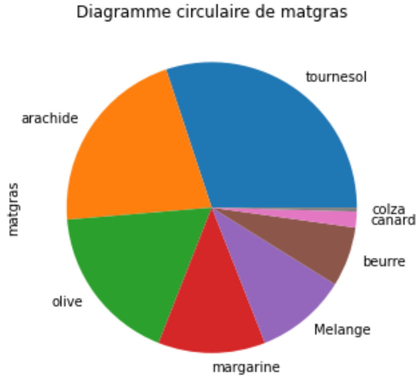
3) Résumés numériques (pour une variable quantitative)

- 3.1) Moyenne et variance
- 3.2) Médiane et Quartiles
- 3.3) Diagrammes à moustache.

Diagrammes circulaires

Pour une une variable nominale, on ne peut représenter que les fréquences $f(c)$ des caractères c. Toutes les représentations, donne une aire du dessin proportionnelle à la fréquence.

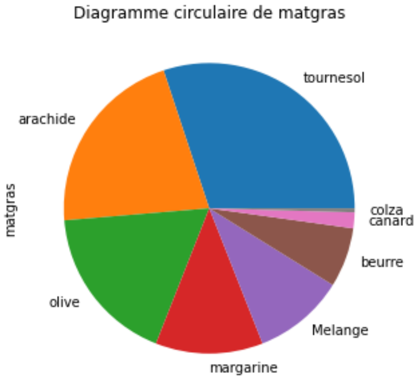
Un diagramme circulaire représente avec un angle $2\pi f(c)$ radian (soit $360 f(c)$ degrés) une colonne du tableau de fréquence $f(c)$.



Un diagramme circulaire représente avec un angle $2\pi f(c)$ radians (soit $360 f(c)$ degrés) une colonne du tableau de fréquence $f(c)$.

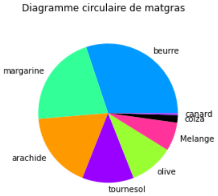
	matgras								
sexe	*beurre*	*margarine*	*arachide*	*tournesol*	*olive*	*Melange*	*colza*	*canard*	*Total*
Femme	0.02212	0.07522	0.14159	0.20796	0.08849	0.07964	0.00442	0.00442	0.62389
Homme	0.04424	0.04424	0.07079	0.09292	0.08849	0.02212	0.00000	0.01327	0.37610
Total	0.06637	0.11946	0.21238	0.30088	0.17699	0.10176	0.00442	0.01769	1.00000

Ex : Le tournesol est représenté par un secteur d'angle : $360 \times 0.30088 = 111,2$ degrés

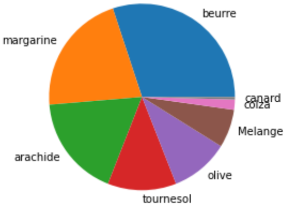


Diagrammes circulaires en Python

```
col=["#0099FF", "#33FF99","#FF9900", "#9900FF", "#99
    FF33","#FF3399","#000000", "#9933FF"]
#On choisit les couleurs.
df["matgras"].value_counts().plot.pie(y='matgras',
    colors=col,figsize=(5,5),title='Diagramme
    circulaire de matgras')
#Attention: pie prend en argument la table de
    contingence df["matgras"].value_counts() et PAS le
     vecteur de données: df["matgras"].
```



```
import matplotlib.pyplot as plt #pour les dessins
#Autre solution 1 avec plt.pie (non recommandée)
plot=plt.pie(df["matgras"].value_counts(), labels=df['
    matgras'].cat.categories)
plt.savefig('/home/Enseignement/Cours/StatInfo/Python/
    DiagCirc.png')#export d'un fichier
plt.show(plot)
```



```
#méthode diagramme pie avec matplotlit et options
fig1, ax1 = plt.subplots()
col=["#0099FF", "#33FF99","#FF9900", "#9900FF", "#99
    FF33","#FF3399","#000000", "#9933FF"]
ax1.pie(df["matgras"].value_counts(), labels=df['
    matgras'].cat.categories,colors=col)
fig1.suptitle('Diagramme circulaire de matgras')
plt.show(ax1)
```

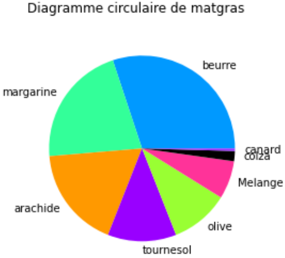
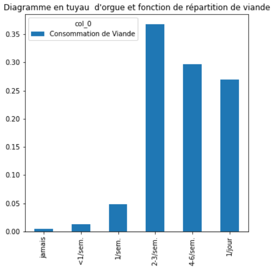


Diagramme en tuyau d'orgue en Python

```
#Diagramme en Tuyau d'orgue (ou barre ) avec pandas
tableViande=pan.crosstab(index = df["viande"],columns=
    "Consommation de Viande",normalize=True)
plotViande=tableViande.plot.bar(figsize=(6,6),title="
    Diagramme en tuyau d'orgue et fonction de
    répartition de viande")
plt.show(plotViande)
```



L'ordonnée est la fréquence $f(c)$.

Plan de cette séance : début du chapitre 1

- Présentation du cours.

Chap 1 Statistique descriptive.

- 1) Vocabulaire statistique et Types de variables statistiques

  2) Représentations graphiques (pour une variable)
- 2.1) Notions d'effectif et de fréquence.
- 2.2) Diagrammes circulaires ou en tuyau d'orgue (pour les variables qualitatives)
- 2.3) Diagrammes cumulatifs (pour les variables ordinales)
- 2.4) Diagrammes en bâton et fonction de répartition empiriques (pour les variables quantitatives discrètes)
- 2.5) Histogrammes (pour les variables quantitatives continues)

  3) Résumés numériques (pour une variable quantitative)
- 3.1) Moyenne et variance
- 3.2) Médiane et Quartiles
- 3.3) Diagrammes à moustache.

Notion de fréquence cumulée.

Soit $N = Card(I)$ l'effectif total de l'échantillon I. On suppose que $(x_i)_{i \in I}$ est une variable qualitative ordinale, ou une variable quantitative.

Définition

La fréquence cumulée du caractère c (anglais : *cumulative frequency*) ou fonction de répartition empirique (anglais : ecdf *empirical cumulative distribution function*) est la fonction

$$F(c) = \frac{Card\{i \in I : x_i \leqslant c\}}{N}.$$

Elle se calcule en fonction de la fréquence f par la formule :

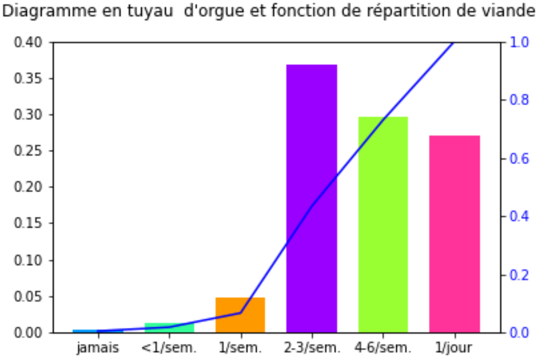
$$F(c) = \sum_{d \leqslant c} f(d).$$

Diagramme en tuyau d'orgue avec fréquence cumulée

```
import numpy as np #librairie pour les calculs
tableViande=pan.crosstab(index = df["viande"],columns=
    "freq",normalize=True)
l=len(tableViande);x = np.arange(l);w=0.7

fig, ax = plt.subplots();ax.set_ylim(0,0.4)
ax.bar(x,np.reshape(tableViande.values,l), width=w,
    color=col)
ax2=ax.twinx()
ax2.tick_params(axis='y', labelcolor='b')
ax2.set_ylim(0,1)
ax2.plot(x,tableViande.cumsum(),color='b')
fig.suptitle("Diagramme en tuyau d'orgue et fonction
    de répartition de viande")
ax.set_xticks(x)
ax.set_xticklabels(tableViande.index)
ax.legend(frameon=False)
```

Diagramme en tuyau d'orgue avec fréquence cumulée



Plan de cette séance : début du chapitre 1

- Présentation du cours.

Chap 1 Statistique descriptive.

- 1) Vocabulaire statistique et Types de variables statistiques

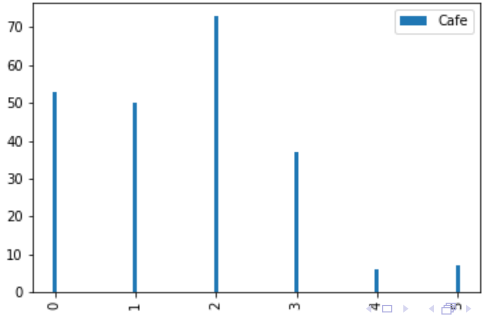
  2) Représentations graphiques (pour une variable)
- 2.1) Notions d'effectif et de fréquence.
- 2.2) Diagrammes circulaires ou en tuyau d'orgue (pour les variables qualitatives)
- 2.3) Diagrammes cumulatifs (pour les variables ordinales)
- 2.4) Diagrammes en bâton et fonction de répartition empirique (pour les variables quantitatives discrètes)
- 2.5) Histogrammes (pour les variables quantitatives continues)

  3) Résumés numériques (pour une variable quantitative)
- 3.1) Moyenne et variance
- 3.2) Médiane et Quartiles
- 3.3) Diagrammes à moustache.

Diagrammes en bâton (pour les variables discrètes)

Pour une variable discrète, la valeur a une position d'abscisse précise, ce n'est pas un nom de classe comme pour une variable ordinale. On représente donc des Diagrammes en bâton avec des barres verticales étroites au lieu des tuyaux d'orgues.

```
tC=pan.crosstab(index =df["cafe"],columns="coffe")
tableCafe=pan.DataFrame({'Cafe': np.reshape(tC.values,
    newshape=len(tC))})#Table en D.F. avec 1 Col. Cafe
axes=tableCafe.plot.bar(width=0.05)
```



Fonction de répartition empirique (pour les variables discrètes)

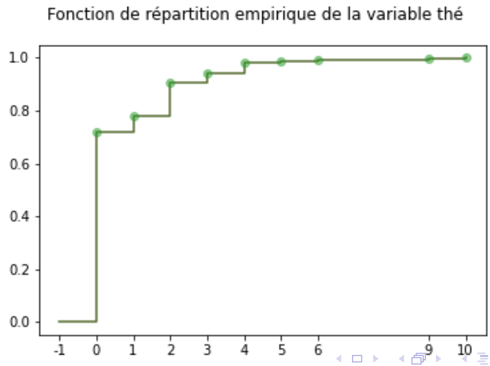
Pour une variable discrète, on peut tracer la **fréquence cumulée**, et elle change à des valeurs exactes, d'où la forme en escalier.

```
tableThe=pan.crosstab(index = df["the"],columns="freq"
    ,normalize=True).cumsum()
xThe = np.array([-1,*(tableThe.index)])#abcisses aux
    valeurs et -1
figThe, axThe = plt.subplots()#crée le graphique
val=np.reshape(tableThe.values,len(tableThe))
axThe.step(xThe,np.array([0,*val]), where='post',color
    ="#556b2f")#trace le diagramme en escalier en
    ajoutant un point de départ à valeur 0
axThe.plot(tableThe.index, val, 'C2o', alpha=0.5)#
    trace les points de sauts
axThe.set_xticks(xThe)
figThe.suptitle("Fonction de répartition empirique de
    la variable thé")
```

Fonction de répartition empirique (pour les variables discrètes)

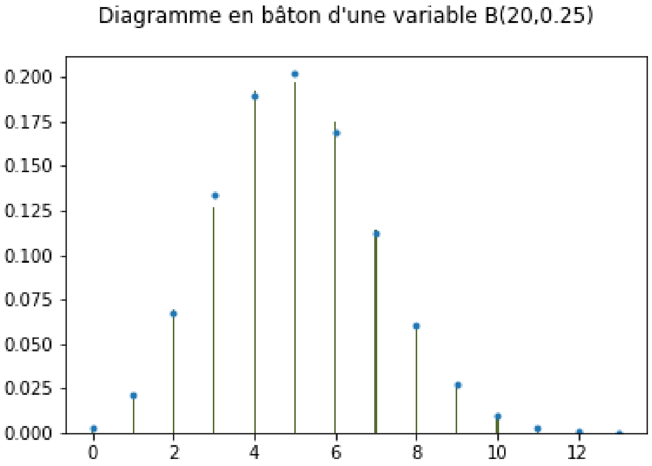
Pour une variable discrète, on peut tracer la **fréquence cumulée**, et elle change à des valeurs exactes, d'où la forme en escalier.

```
axThe.step(xThe,np.array([0,*val]), where='post',color
    ="#556b2f")#trace le diagramme en escalier en
    ajoutant un point de départ à valeur 0
axThe.plot(tableThe.index, val, 'C2o', alpha=0.5)#
    trace l
```



Simulations des variables discrètes

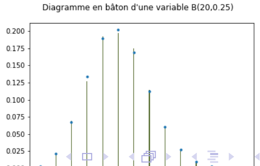
Grâce à la simulation de l'objet principal du modèle probabiliste, les variables aléatoires, on obtient des échantillons statistiques qui ont presque la répartition théorique du modèle probabiliste. On illustre ici avec la loi binomiale (déjà vue au lycée, que l'on reverra bientôt).



Simulations en Python des variables discrètes binomiales

```
import scipy.stats as st
SampleBinom=st.binom.rvs(20,0.25,size=10000)
tableBinom=pan.crosstab(index = SampleBinom,columns="
    freq",normalize=True)
l=len(tableBinom);xBinom = np.arange(l)
BinomTheorique=st.binom.pmf(xBinom,20,0.25)#
    probability mass function

figBinom, axBinom = plt.subplots()
axBinom.bar(xBinom,np.reshape(tableBinom.values,l),
    width=0.05,color="#556b2f")
axBinom.plot(xBinom,BinomTheorique, '.')
figBinom.suptitle("Diagramme en bâton d'une variable B
    (20,0.25)")
plt.show(axBinom)
```



Plan de cette séance : début du chapitre 1

- Présentation du cours.

Chap 1 Statistique descriptive.

- 1) Vocabulaire statistique et Types de variables statistiques

  2) Représentations graphiques (pour une variable)
- 2.1) Notions d'effectif et de fréquence.
- 2.2) Diagrammes circulaires ou en tuyau d'orgue (pour les variables qualitatives)
- 2.3) Diagrammes cumulatifs (pour les variables ordinales)
- 2.4) Diagrammes en bâton et fonction de répartition empirique (pour les variables quantitatives discrètes)
- 2.5) Histogrammes (pour les variables quantitatives continues)

  3) Résumés numériques (pour une variable quantitative)
- 3.1) Moyenne et variance
- 3.2) Médiane et Quartiles
- 3.3) Diagrammes à moustache.

2.5) Histogrammes (pour les variables quantitatives continues)

Pour une variable continue, la valeur est réelle, et chaque valeur est généralement prise peu de fois, c'est donc l'effectif ou la fréquence dans un intervalle qu'il est pertinent de représenter.

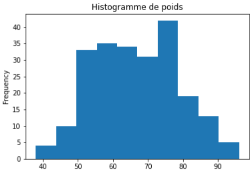
1. Soit $[a, b]$ l'intervalle de valeur de la variable (par exemple a le min et b le max des observations)
2. Soit une subdivision $a = a_0 < \ldots < a_N = b$ donnant N intervalles $I_n = [a_n, a_{n+1}[$ et le dernier $I_{N-1} = [a_{N-1}, a_N]$.
3. On choisit souvent les intervalles de sorte que l'on ait au moins un effectif de 5. On préfère si possible des découpages de pas égaux, mais ce n'est pas toujours adapté (à la condition d'effectif).
4. On calcule alors la hauteur h_n du rectangle de base $[a_n, a_{n+1}[$ pour avoir comme surface la fréquence de la classe. Si N l'effectif total alors la hauteur voulue est

$$h_n = \frac{Card\{i \in I : a_n \leqslant x_i < a_{n+1}\}}{N(a_{n+1} - a_n)}.$$

2.5) Histogrammes (pour les variables quantitatives continues)

Bilan : on trace des rectangles de base $[a_n, a_{n+1}[$ et de hauteur h_n.
Ex : pour la variable age. On utilise la commande `hist` en Python. Par défaut, les segments sont égaux, ici, choisit par R de longueur 5. Attention, par défaut, l'ordonnée est graduée en effectif (avec label anglais *frequency*). L'effectif de la classe [90,95] est trop petit, donc cela suggère de changer les classes pour améliorer l'histogramme.

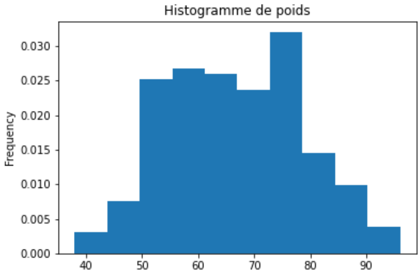
```
plotPoids=df["poids"].plot.hist(title='Histogramme de
    poids')
plt.show(plotPoids)
```



2.5) Histogrammes (pour les variables quantitatives continues)

ATTENTION, contrairement aux autres fonctions `pie`, `plot`, la fonction `hist` prend les données (ici df["poids"]) comme entrée et non la table (car elle va calculer elle-même une table correspondant à la subdivision choisie).

```
plotPoids=df["poids"].plot.hist(title='Histogramme de
    poids',density=True)
```



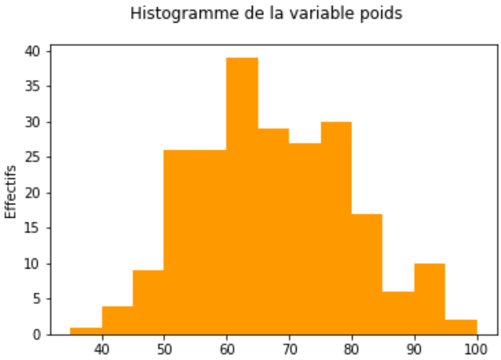
2.5) Histogrammes (pour les variables quantitatives continues)

Une version améliorée est la suivante.
Ici, on prend un pas de 5 kg sur [35,100].

```
figPoids, axPoids = plt.subplots()
bins=35+np.arange(14)*5
axPoids.hist(df["poids"],bins, color=col[2],histtype='
    bar')
axPoids.legend(prop={'size': 13})
axPoids.set_ylabel('Effectifs')
axPoids.legend(frameon=False)
figPoids.suptitle("Histogramme de la variable poids")
```

2.5) Histogrammes (pour les variables quantitatives continues)

```
plt.savefig('/home/Enseignement/Cours/StatInfo/Python/
    HistoPoids3.png')
plt.show(axPoids)#facultatif dans les dernières
    versions de Spyder
```



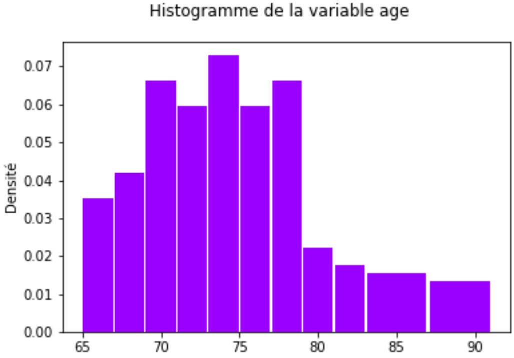
2.5) Histogramme pour Age (pour les variables quantitatives continues)

Une version améliorée pour Age est la suivante.
Ici, on prend un pas de 2 ans au début puis un intervalle [87,91] à la fin. Remarque :Pour les intervalles inégaux, Il faut absolument l'option density=True sinon, c'est totalement faux.

```
figAge, axAge = plt.subplots()
binsAge=[*(65+np.arange(10)*2),87,91]#* pour extraire
    la liste du np.array
axAge.hist(df["age"],binsAge, color=col[3],histtype='
    bar',rwidth=0.95,density=True)#rwidth pour les
    espaces
axAge.legend(prop={'size': 11})
axAge.set_ylabel('Densité')
axAge.legend(frameon=False)
figAge.suptitle("Histogramme de la variable age")
plt.savefig('/home/Enseignement/Cours/StatInfo/Python/
    HistoAge.png');plt.show(axAge)
```

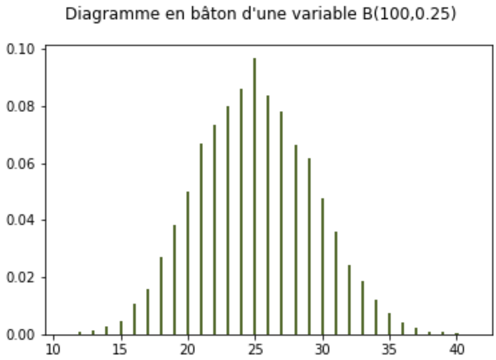
2.5) Histogramme pour Age (pour les variables quantitatives continues)

Une version améliorée pour Age est la suivante.
Ici, on prend un pas de 2 ans au début puis un intervalle [87,91] à la fin. Remarque :Pour les intervalles inégaux, Il faut absolument l'option `density=True` sinon, c'est totalement faux.



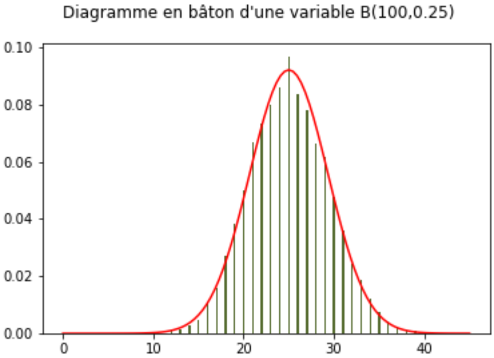
2.5) Histogrammes :Illustration des variables aléatoires continues

La notion de variable aléatoire sera le modèle théorique des variables statistiques. Un diagramme en bâton d'une simulation d'un variable binômiale $B(100, 0.25)$ qui prend des valeurs dans $[\![0, 100]\!]$ illustre que l'on pourrait vouloir considérer l'échantillon obtenu comme presque une variable continue.



2.5) Histogrammes :Illustration des variables aléatoires continues

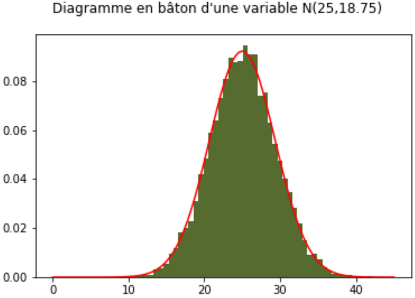
On verra (dans un résultat fondamental du cours : le Théorème central limite) que la variable binomiale $\mathcal{B}(100, 0.25)$ est approchée par une loi prenant toutes les valeurs réelles (et pas seulement les entiers), l'exemple le plus important de loi continue, une loi normale $\mathcal{N}(25, 18.75)$.



2.5) Histogrammes :Illustration des variables aléatoires continues

Voici une simulation de la loi approximative $\mathcal{N}(25, 18.75)$, et de la valeur théorique (en rouge). La fameuse loi normale à la fameuse forme de courbe en cloche.

```
SNorm=st.norm.rvs(25,np.sqrt(18.75),size=10000)
figBinom, axBinom = plt.subplots()
axBinom.hist(SNorm,density=True,bins=50)
axBinom.plot(xNorm,NormT, color="red")
figBinom.suptitle("Diagramme en bâton d'une variable N
    (25,18.75)")
```



Plan de cette séance : début du chapitre 1

- Présentation du cours.

Chap 1 Statistique descriptive.

- 1) Vocabulaire statistique et Types de variables statistiques
  2) Représentations graphiques (pour une variable)
- 2.1) Notions d'effectif et de fréquence.
- 2.2) Diagrammes circulaires ou en tuyau d'orgue (pour les variables qualitatives)
- 2.3) Diagrammes cumulatifs (pour les variables ordinales)
- 2.4) Diagrammes en bâton et fonction de répartition empirique (pour les variables quantitatives discrètes)
- 2.5) Histogrammes (pour les variables quantitatives continues)

  3) Résumés numériques (pour une variable quantitative)
- 3.1) Moyenne et variance
- 3.2) Médiane et Quartiles
- 3.3) Diagrammes à moustache.

3.1) Moyenne et variance

Les résumés numériques permettent de donner une première idée quantitative d'une variable quantitative $x \in \mathbb{R}^n$. Le plus simple est la moyenne.

Définition

La moyenne empirique (anglais *mean*) d'un échantillon $x = (x_1, ..., x_n)$ est la grandeur :

$$m(x) \equiv \overline{x} = \frac{x_1 + ... + x_n}{n}$$

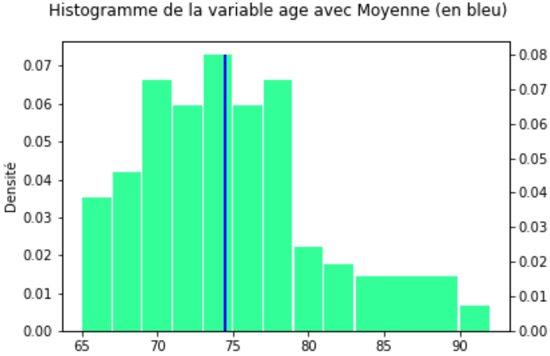
Ex : Soit $x = (1, 2, 4, 93)$ on obtient $n = 4$ et :

$$m(x) \equiv \overline{x} = \frac{1 + 2 + 4 + 93}{4} = 25.$$

La moyenne donne une idée du milieu de la distribution.
Ex : Soit $x = (a, \cdots, a)$:

$$m(x) \equiv \overline{x} = a.$$

```
figAge, axAge = plt.subplots()
axAge.hist(df["age"],[*(65+np.arange(10)*2),90,92],
    color=col[2],rwidth=0.95,density=True)
ax2=axAge.twinx()
ax2.bar(np.mean(df["age"]),0.08,color='blue',width=.2)
axAge.set_ylabel('Densité')
figAge.suptitle("Histogramme de la variable age avec
    Moyenne (en bleu)")
np.mean(df["age"])
#Réponse: 74.477876106
```



3.1) Propriétés de la Moyenne : linéarité

Proposition

La moyenne empirique est linéaire : Soient $x = (x_1, \dots, x_n), y = (y_1, \dots, y_n)$ *des échantillons (de même longueur) et* $\lambda \in \mathbb{R}$

$$\overline{\lambda x + y} = \lambda \overline{x} + \overline{y}.$$

Démonstration.

C'est juste la linéarité des sommes

$$\begin{aligned}
\overline{\lambda x + y} &= \frac{1}{n}\Big(\sum_{k=1}^{n} \lambda x_k + y_k\Big) \\
&= \frac{1}{n}\Big(\lambda \sum_{k=1}^{n} x_k + \sum_{k=1}^{n} y_k\Big) \\
&= \lambda \overline{x} + \overline{y}
\end{aligned}$$

□

3.1) Propriétés de la Moyenne : monotonie

Proposition

La moyenne empirique préserve l'ordre Soient $x = (x_1, ..., x_n), y = (y_1, ..., y_n)$ des échantillons (de même longueur) si $x_i \leqslant y_i$ pour tout i, alors :

$$\overline{x} \leqslant \overline{y}.$$

En particulier, une moyenne d'un échantillon positif $x \in [0, +\infty[^n$ est positive : $\overline{x} \geqslant 0$.

3.1) Moyenne et variance

Variances et écart type donnent une idée de la largeur de la distribution autour de la moyenne

Définition

La variance empirique (anglais *variance*) est la moyenne des carrés des écarts à la moyenne :

$$V(x) = m((x - \overline{x})^2) = \frac{1}{n} \sum_{k=1}^{n} (x_k - \overline{x})^2 = \frac{(x_1 - \overline{x})^2 + \dots + (x_n - \overline{x})^2}{n}.$$

La variance (empirique) non biaisée est la variante suivante :

$$var(x) = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2.$$

Sa racine carré $\sigma(x) = \sqrt{var(x)}$ s'appelle l'écart type empirique (non biaisé) (anglais *standard deviation*).

La variance non biaisée est meilleure en pratique (surtout si n petit).

3.1) Variance en Python

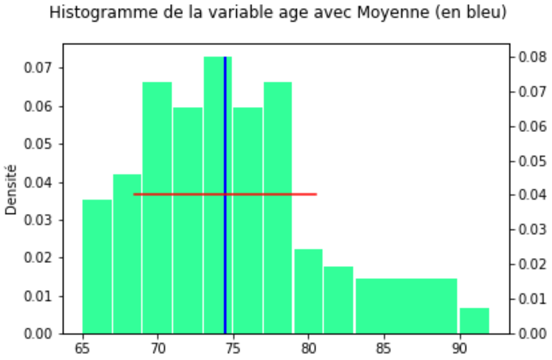
La commande `var` calcule par défaut la Variance non biaisée $var(x)$.
Il faut définir une fonction pour obtenir la variance empirique.

```
np.var(df["age"])#Variance empirique de age : 35.90438
np.var(df["age"],ddof=1)#Variance non biaisée de age :
     36.06395 (toujours plus grande)
np.std(df["age"],ddof=1)#écart type (non biaisé) de
    age: 6.005327
```

3.1) Variance en Python

On ajoute avec R un intervalle $[\overline{x}-\delta\sigma(x), \overline{x}+\delta\sigma(x)]$ (avec $\delta = 1$). On verra que ce type d'intervalle contient une "grande proportion" de l'échantillon (pour un bon choix de δ).

```
ax2.bar(np.mean(df["age"]),0.08,color='blue',width=.2)
ax2.plot([np.mean(df["age"])-np.std(df["age"]),np.mean
    (df["age"])+np.std(df["age"])],[0.04,0.04],color='
    red')
```



3.1) Propriétés de la Variance : Formule alternative

La définition $V(x) = m[(x - \overline{x})^2]$ n'est pas toujours pratique, on a une formule alternative :

Proposition

Soit $x = (x_1, \dots, x_n)$ *un échantillon et* $x^2 = (x_1^2, \dots, x_n^2)$.

$$V(x) = m(x^2) - (m(x))^2 = \overline{x^2} - (\overline{x})^2.$$

Attention, cette formule ne fonctionne pas avec la variance non-biaisée $var(x) = \frac{n}{n-1} V(x)$!

3.1) Propriétés de la Variance : Preuve de la Formule alternative

La définition $V(x) = m[(x - \overline{x})^2]$ n'est pas toujours pratique, on a une formule alternative $V(x) = m(x^2) - (m(x))^2$

Démonstration.

Dans la définition $\overline{x} = (\overline{x}, \cdots, \overline{x})$ désigne le vecteur constant. Le carré correspond au produit d'échantillon terme à terme $xy = (x_1 y_1, \cdots, x_n y_n)$. On développe le carré

$$(x - \overline{x})^2 = x^2 - 2x\overline{x} + \overline{x}^2.$$

Puis on applique la linéarité de la moyenne :

$$\begin{aligned} m[(x - \overline{x}))^2] &= m(x^2) - 2m(x)\overline{x} + \overline{x}^2 m(1) \\ &= m(x^2) - 2m(x)^2 + m(x)^2 \\ &= m(x^2) - m(x)^2. \end{aligned}$$

car $m(1) = 1$. $\square$

3.1) Propriétés de la Variance : homogénéité

Proposition

Soient $x = (x_1, \dots, x_n)$ *un échantillon et* $\lambda \in \mathbb{R}$

$$V(\lambda x) = \lambda^2 V(x).$$

$$var(\lambda x) = \lambda^2 var(x).$$

$$\sigma(\lambda x) = |\lambda| \sigma(x).$$

Attention, la variance et l'écart type ne sont PAS LINÉAIRES. On verra la relation qui remplace avec l'étude du cas de 2 variables (la semaine prochaine).

Démonstration.

Les deux autres relations suivent de la première :

$$var(\lambda x) = \frac{n}{n-1} V(\lambda x) = \frac{n}{n-1} \lambda^2 V(x) = \lambda^2 var(x).$$

$$\sigma(\lambda x) = \sqrt{\lambda^2 var(x)} = |\lambda| \sigma(x).$$

Proposition

Soient $x = (x_1, ..., x_n)$ *un échantillon et* $\lambda \in \mathbb{R}$

$$V(\lambda x) = \lambda^2 V(x).$$

Démonstration.

Remarquez que $(\lambda x)^2 = \lambda^2 x^2, m(\lambda x) = \lambda m(x)$. On déduit :

$$\begin{aligned} V(\lambda x) &= m((\lambda x)^2) - m(\lambda x)^2 \\ &= m(\lambda^2 x^2) - \lambda^2 m(x)^2 \\ &= \lambda^2 m(x^2) - \lambda^2 m(x)^2 \end{aligned}$$

par linéarité de la moyenne. En factorisant par λ^2, on conclut :

$$V(\lambda x) = \lambda^2 \left(m(x^2) - m(x)^2 \right) = \lambda^2 V(x).$$

$\square$

3.1) Propriétés de la Variance : Cas d'annulation

Proposition

Soient $x = (x_1, ..., x_n)$ un échantillon $V(x) = 0$ si et seulement si il existe $c \in \mathbb{R}$ tel que $x = (c, \cdots, c)$.

Démonstration.

$\Leftarrow$ On a déjà vu que si $x = (c, \cdots, c)$, $\overline{x} = c$, donc $x - \overline{x} = 0$. On conclut

$$V(x) = m((x - \overline{x})^2) = m(0) = 0.$$

$\Rightarrow$ On a $nV(x) = \sum_{k=1}^{n}(x_k - \overline{x})^2$ est une somme de termes positifs, qui est donc nulle si et seulement si tous les termes sont nuls, soit pour tout k

$$x_k = \overline{x}.$$

Il suffit donc de prendre $c = \overline{x}$. $\square$

Plan de cette séance : début du chapitre 1

- Présentation du cours.

Chap 1 Statistique descriptive.

- 1) Vocabulaire statistique et Types de variables statistiques
  2) Représentations graphiques (pour une variable)
- 2.1) Notions d'effectif et de fréquence.
- 2.2) Diagrammes circulaires ou en tuyau d'orgue (pour les variables qualitatives)
- 2.3) Diagrammes cumulatifs (pour les variables ordinales)
- 2.4) Diagrammes en bâton et fonction de répartition empirique (pour les variables quantitatives discrètes)
- 2.5) Histogrammes (pour les variables quantitatives continues)

  3) Résumés numériques (pour une variable quantitative)
- 3.1) Moyenne et variance
- 3.2) Médiane et Quartiles
- 3.3) Diagrammes à moustache.

Il existe d'autres indicateurs du milieu de la distribution que la moyenne et de la largeur de la distribution que l'écart type. Le premier est la médiane.

Définition

La médiane d'un échantillon $(x_1, ..., x_n)$ est obtenu de deux façons différentes selon la parité de n. On commence par regarder le réarrangement croissant de l'échantillon $(x_1^*, ..., x_n^*)$, c'est à dire une permutation $(x_1^*, ..., x_n^*) = (x_{\sigma(1)}, ..., x_{\sigma(n)})$ de la suite $(x_1, ..., x_n)$ (c'est à dire $\sigma : [\![1, n]\!] \to [\![1, n]\!]$ bijection) telle que $x_1^* \leqslant x_2^* \leqslant \cdots \leqslant x_n^*$. La médiane est définit :

1. Si $n = 2l + 1$ par $mediane(x) = x_{l+1}^* = x_{\frac{n+1}{2}}^*$.
2. Si $n = 2l$ par
$$mediane(x) = \frac{x_l^* + x_{l+1}^*}{2}.$$

3.2) Exemple de Médiane

Exemple

Dans l'échantillon $x = (1, 93, 4, 2)$, on remet dans l'ordre l'échantillon et on obtient

$$x^* = (1, 2, 4, 93).$$

On a $n = 4 = 2l$ pair, donc on regarde $x_l^* = 2, x_{l+1}^* = 4$ et on obtient la médiane

$$mediane(x) = \frac{x_l^* + x_{l+1}^*}{2} = \frac{2+4}{2} = 3.$$

La moyenne est $\frac{1+2+4+93}{4} = 25$. On voit que la moyenne prend plus en compte les valeurs extrêmes (ici 93) que la médiane même si elles sont de faible probabilité (ici 0.25). La médiane est une notion de milieu de la distribution moins sensible aux valeurs extrêmes.

Attention, il y a différentes définitions pour la notion de quartile, on choisit la plus simple, souvent utilisée théoriquement. Les quartiles partagent la distribution en 4 parties comptant environ un quart des individus de la distribution.

Définition

Le fractile d'ordre $\beta \in]0, 1[$ d'une variable statistique x est la plus petite valeur x_β telle que la fréquence cumulée (fonction de répartition empirique) F_x vérifie $F_x(x_\beta) \geqslant \beta$ soit

$$x_\beta = \inf\{t : F_x(t) \geqslant \beta\}.$$

Les quartiles sont $Q_1(x) = x_{1/4}, Q_2(x) = x_{1/2}, Q_3(x) = x_{3/4}$.

En pratique, si $(x_1^*, \ldots, x_n^*)$ est le réarrangement croissant de x :

$$Q_1(x) = x^*_{\lceil n/4 \rceil}, Q_2(x) = x^*_{\lceil n/2 \rceil}, Q_3(x) = x^*_{\lceil 3n/4 \rceil},$$

avec $\lceil x \rceil$ l'unique entier (partie entière supérieure) telle que $\lceil x \rceil - 1 < x \leqslant \lceil x \rceil$.

Exemple

On reprend l'échantillon $x = (1, 93, 4, 2)$, on remet dans l'ordre l'échantillon et on obtient $x^* = (1, 2, 4, 93)$. On a $F_x(1) = 1/4, F_x(2) = 1/2, F_x(4) = 3/4$ donc

$$Q_1(x) = 1, Q_2(x) = 2, Q_3(x) = 4$$

sont les 3 quartiles. Notez que $Q_2(x) \neq mediane(x)$ avec cette définition, c'est encore une autre variante du milieu de la distribution.

```
np.quantile([1,2,3,4,5,6,7,8],[0,0.25,0.5,0.75,1],axis
    =0,interpolation='lower'))#ATTENTION, sans l'
    option, interpolation='lower', Python calcule
    selon une autre définition, non vue en cours (
    parmi 5 choix de variantes). Il renvoie aussi le
    minimum et le maximum de l'échantillon.
```

0%	25%	50%	75%	100%
1	2	4	6	8

Plan de cette séance : début du chapitre 1

- Présentation du cours.

Chap 1 Statistique descriptive.

- 1) Vocabulaire statistique et Types de variables statistiques
  2) Représentations graphiques (pour une variable)
- 2.1) Notions d'effectif et de fréquence.
- 2.2) Diagrammes circulaires ou en tuyau d'orgue (pour les variables qualitatives)
- 2.3) Diagrammes cumulatifs (pour les variables ordinales)
- 2.4) Diagrammes en bâton et fonction de répartition empirique (pour les variables quantitatives discrètes)
- 2.5) Histogrammes (pour les variables quantitatives continues)

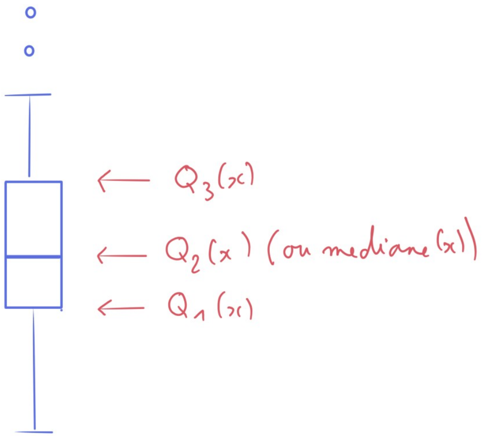
  3) Résumés numériques (pour une variable quantitative)
- 3.1) Moyenne et variance
- 3.2) Médiane et Quartiles
- 3.3) Diagrammes à moustache.

3.3) Définition d'un Diagramme à moustache

Un diagramme à Moustache ressemble à cela :

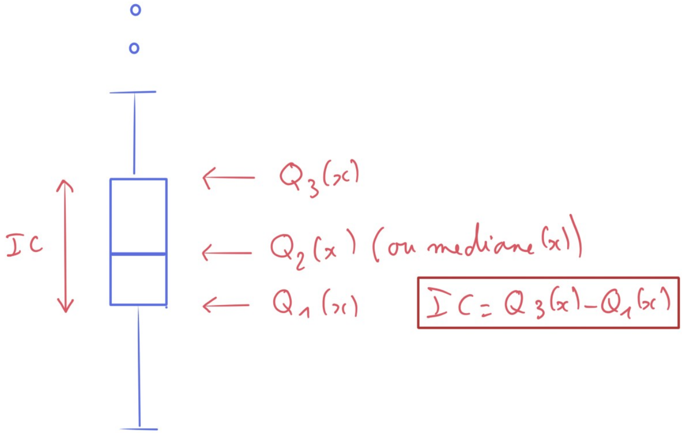
3.3) Définition d'un Diagramme à moustache

On place d'abord la boîte en calculant les quartiles :



3.3) Définition d'un Diagramme à moustache

Avant de calculer la moustache, on a besoin de calculer la distance interquartile IC

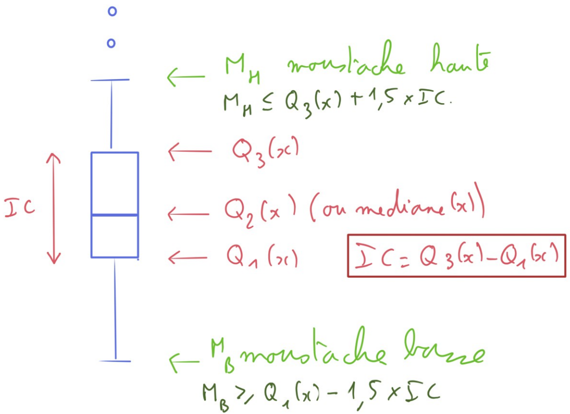


3.3) Définition d'un Diagramme à moustache

La moustache haute est

$$M_H = M_H(x) = \sup\{x_i : x_i \leqslant Q_3(x) + 1.5IC\}$$

c'est la plus grande valeur de l'échantillon qui ne dépasse pas $Q_3(x) + 1.5IC$

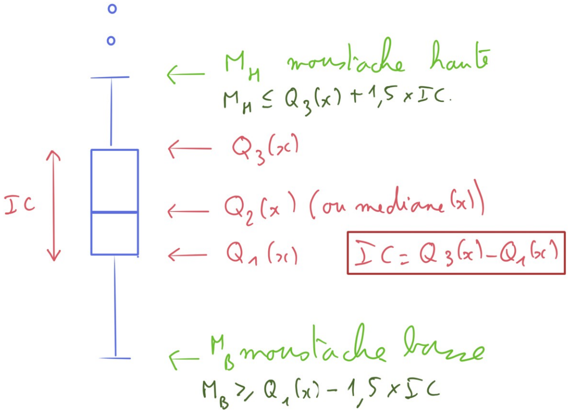


3.3) Définition d'un Diagramme à moustache

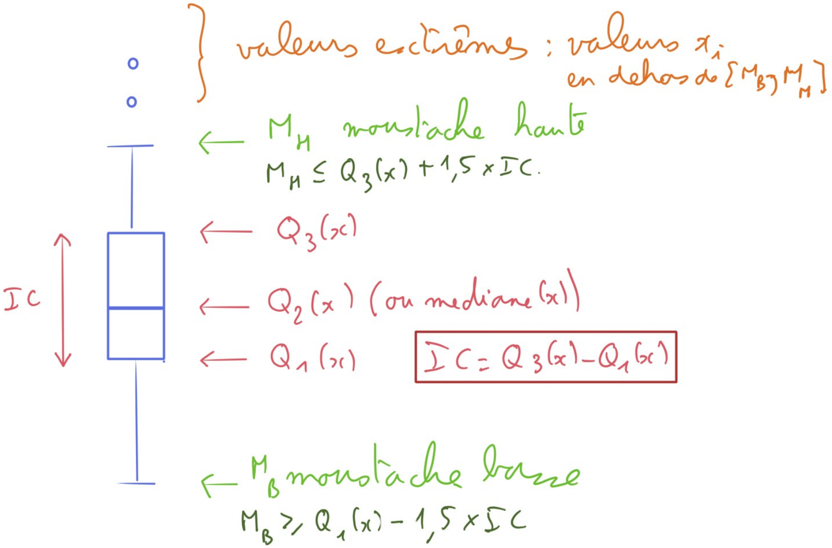
De même, la moustache basse est

$$M_B = M_B(x) = \inf\{x_i : x_i \geqslant Q_1(x) - 1.5IC\}$$

c'est la plus petite valeur de l'échantillon au dessus de $Q_1(x) - 1.5IC$.



Enfin, les valeurs en dehors de la moustache $[M_B, M_H]$, sont dites "valeurs extrêmes" et sont tracès par des ronds.



On prend $x = (1, 2, 2, 2, 4, 5, 5.5, 10)$
On calcule $n = 8$. $x = x^*$ est déjà ordonné.

$$Q_1(x) = x^*_{8/4} = 2, Q2(x) = x^*_{8/2} = 2, Q_3(x) = x^*_{3*8/4} = 5$$
$$IC = Q_3(x) - Q_1(x) = 5 - 2 = 3$$

Calcul des moustâches :

$$Q_1(x) - 1.5IC = 2 - 4.5 = -2.5 < Min(x) = 1 \Rightarrow M_B(x) = Min(x) = 1$$

$$Q3(x) + 1.5IC = 5 + 4.5 = 9.5 \in [5.5, 10] \Rightarrow M_H(x) = 5.5$$

$10 > M_H(x)$ est la seule valeur extrême.

Dessin

3.3) Ex de Diagrammes à moustache

On prend $x = (1, 2, 2, 2, 4, 5, 5.5, 10)$
On calcule $n = 8$. x est déjà ordonné.

$$Q_1(x) = x^*_{8/4} = 2, Q2(x) = x^*_{8/2} = 2, Q_3(x) = x^*_{3*8/4} = 5$$

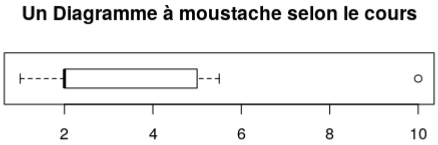
$$IC = Q_3(x) - Q_1(x) = 5 - 2 = 3$$

Calcul des moustâches :

$$Q_1(x) - 1.5IC = 2 - 4.5 = -2.5 < Min(x) = 1 \Rightarrow M_B(x) = Min(x) = 1$$

$$Q3(x) + 1.5IC = 5 - 4.5 = 9.5 \in [5.5, 10] \Rightarrow M_H(x) = 5.5$$

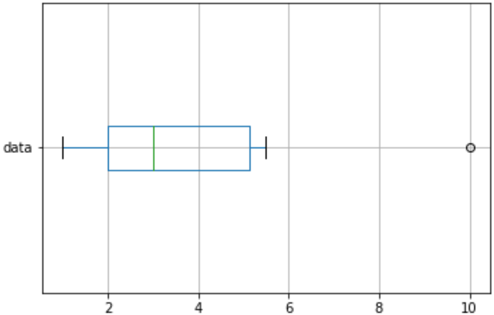
$10 > M_H(x)$ est la seule valeur extrême.



3.3) Diagrammes à moustache en Python : Version par défaut (à utiliser même si différent du cours)

Pour comparaison, voici le résultat avec boxplot.

```
exbox=pan.DataFrame({'data':[1,2,2,2,4,5,5.5,10]})
exbox.boxplot(vert=False)
np.quantile(exbox,[0,0.25,0.5,0.75,1],axis=0,
    interpolation='lower')
# 1.00 2.00 3.00 5.25 10.00
```



3.3) Diagrammes à moustache selon le cours, (FACULTATIF via importation de R)

Par défaut la commande `boxplot` ne calcule pas avec la définition des quantiles du cours (mais celle par défaut pour la fonction quantile). On peut l'obtenir par une variante du package R "qboxplot".

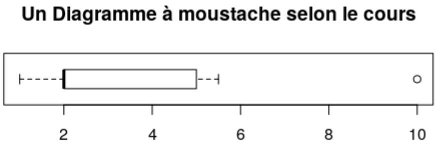
```
pip install rpy2#la première fois, après installation
    de Rbase sur votre ordi
import rpy2
from rpy2.robjects.packages import importr
import rpy2.robjects as ro
from rpy2.robjects import pandas2ri
from rpy2.robjects.conversion import localconverter

base = importr("base");qb=importr("qboxplot")#import
    des librairies R nécessaires
rqboxplot=ro.r['qboxplot']#import des fonctions R
```

3.3) Diagrammes à moustache selon le cours, (FACULTATIF via importation de R)

Par défaut la commande `boxplot` ne calcule pas avec la définition des quantiles du cours (mais celle par défaut pour la fonction quantile). On peut l'obtenir par une variante du package R "qboxplot".

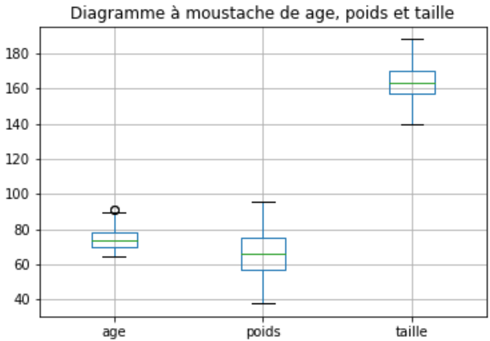
```
with localconverter(ro.default_converter + pandas2ri.
    converter):
    rbox=ro.conversion.py2rpy(exbox)#conversion du
        DataFrame Pandas en un dataFrame de R
rqboxplot(rbox,qtype=1,horizontal=True,main="Un
    Diagramme à moustache selon le cours")#dessin du
    boxplot
```



3.3) Diagrammes à moustache en Python

La commande boxplot suffit à réaliser les Diagrammes à moustache (dans la plupart des cas sans grand changement). Ici, on en trace ceux de l'âge du poids et de la taille dans notre exemple.

```
box=df.boxplot(column=['age','poids','taille'])
box.set_title('Diagramme à moustache de age, poids et
    taille')
plt.show(box)
```



3.3) Diagrammes à moustache en Python

On remarquera ci-dessous les nombreuses valeurs extrêmes pour la variable the.

```
figBox, axesBox = plt.subplots(figsize=(5, 4))
all_data=(df.iloc[:,1:3].values)
bplot=axesBox.boxplot(all_data,vert=True,patch_artist=
    True,labels=['the','cafe'])
axesBox.set_title('Diagramme à moustache de thé et
    café')
plt.show(axesBox)
```