Statistiques

Table des matières

Statistiques descriptives unidimensionnelle (Charly)

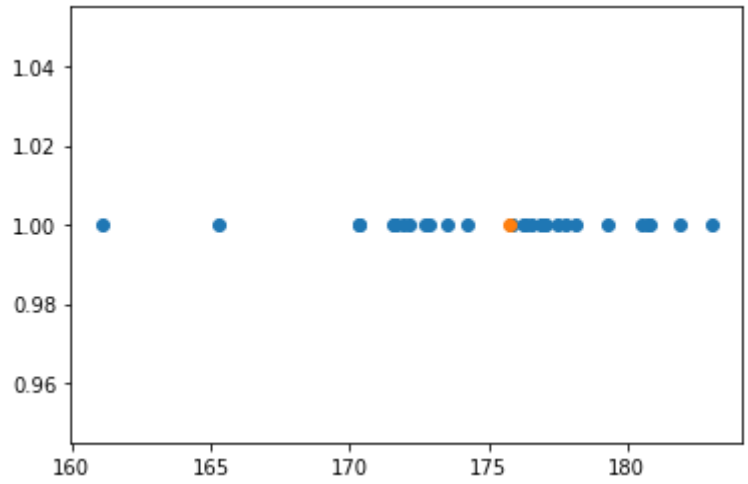
Paramètres de position

Moyenne

La moyenne est un paramètre visant à résumer toutes les données en une seule valeur.

- Symbole : $\bar{x}_A$ ou $\bar{x}$
- Formule :
  - Non-pondérée : $\bar{x}_A = \frac{1}{n}\sum_{i=0}^{n} x_i$
  - Pondérée : $W = \frac{1}{\sum_{i=0}^{n} w_i}\sum_{i=0}^{n} w_i x_i$

Une bonne représentation visuelle de la moyenne est celle d'un point qui minimise la distance cumulative avec toutes les données:



On peut vouloir apporter une importance non uniforme aux données. Pour se faire, nous allons utiliser des poids qui vont multiplier chacune des valeurs de données.

Quantile

Le quantile est un paramètre de position qui permet de connaître la valeur maximum des $\frac{N}{quantile}$ premiers x

- Symbole : $q_{quantile}$
- Formule :

$$x_i où\ i \in \mathbb{N} \wedge i > \frac{n}{quantile} \wedge min(i)$$

Quartile

Un quartile est un quantile de valeur spécifique. Il existe trois quartiles Q1, Q2 et Q3 qui représentent respectivement 25% 50% et 75% des données.

- Symboles :
  - $Q1$
  - $Q2$
  - $Q3$

Médiane

La médiane est le Q2. Elle indique la valeur centrale des données. Elle est souvent utilisée de pair avec la moyenne arithmétique car elle n'est pas sensible aux extrêmes.

Si la médiane est supérieure à la moyenne, alors il y a de petites valeurs dans l'extrême et inversement lorsque la médiane est inférieure.

- Symbole : $Q2$ ou Me

Paramètres de dispersion

Dispersion statistique : la dispersion statistique mesure la variabilité des valeurs d'une série statistique par rapport à une valeur centrale (moyenne). Elle permet de déterminer la dispersion et l'écart entre les différentes valeurs.

Paramètres de dispersion absolue

Étendue

L'étendue est la différence entre la valeur minimale et maximale.

L'étendue de X est $X_{max} - X_{min}$

Écart absolu moyen

Ce paramètre est la moyenne arithmétique de la valeur absolue des écarts à la moyenne. C'est donc la "distance moyenne à la moyenne".
$\frac{1}{n} \sum_{i=1}^{n} |x_i - \overline{x}|$

Variance

La variance est une mesure de la dispersion des valeurs d'un échantillon ou d'une distribution de probabilité.

- Symbole : V
- Formule : $\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \overline{x}^2$

Écart-type

L'écart-type est une mesure de dispersion par rapport à la moyenne.

- Symbole : σ
- Formule : $\sigma = \sqrt{V}$

La règle des 68-95-99,7 stipule que :

- ~68% des valeurs se situent dans un intervalle de 1 écart-type autour de la moyenne.
- ~95% des valeurs se situent dans un intervalle de 2 écarts-types autour de la moyenne.
- ~99,7% des valeurs se situent dans un intervalle de 3 écarts-types autour de la moyenne.

Paramètres de dispersion relative

Les paramètres de dispersion relative permettent de comparer deux distributions à ordre de grandeur ou d'unité de mesure différente.

Le coefficient interquartile relatif

Un indicateur de dispersion moins sensible aux extremes comparé à l'ecart type.

$$\frac{C.I.R.(X) = Q_3 - Q_1}{Q_2}$$

L'ecart absolue moyen

Un indicateur de dispersion similaire à la variance mais avec un calcul plus simple.

$$E.A.M(X) = \frac{1}{n} \sum_{i=1}^{n} |x_i - \overline{x}|$$

L'écart moyen relatif

Cette indicateur est similaire a l'ecart moyen relatif mais il est relatif a un point central souvent la moyenne ca donne une mesure relative qui peut etre comparé avec d'autres donnés

$$E.M.R.(X) = \frac{E.A.M.(X)}{\bar{x}}$$

Mise en application sur notre jeu de données

Masse (kg) \ Taille (cm)	160-165	165-170	170-175	175-180	180-185	185-190	Cumulés
45-50	1						1
50-55			4	1			5
55-60		1	1	1			3
60-65			3	2			5
65-70			1	3	1		5
70-75			2	4	1		7
75-80				1	2		3
80-85							0
85-90							0
90-95				1			1
95-100					1		1
Cumulés	1	1	11	13	5	0	31

Indicateur	Taille (cm)	Masse (kg)
Min	162,5	47,5
Max	182,5	97,5
Moyenne	175,7258065	66,53225806
Médiane	177,5	67,5
1er quartile	172,5	57,5
3e quartile	177,5	72,5
Variance	31,36666667	5,763636364
Ecart-type	5,600595206	2,400757456

Représentation

Il existe de nombreux types de diagramme permettant de représenter des données statistiques. Certaines réprsentations sont plus adaptées que d'autres, en fonction des types de données à analyser. Nous avons décidé de présenter et critiquer quelques diagrammes qui nous semblent pertinents à utiliser dans notre cas de figure.

Diagramme en bâtons

Description

Pour construire un diagramme en bâtons, il nous faut placer les points correspondant à nos valeurs sur un graphique. Pour chaque point, nous traçons un trait vertical le reliant à l'axe des abscisses.

Représentation

Pour le diagramme suivant, nous avons utilisé cette suite de points générée de manière aléatoire :

x	y
0.87009423	0.59815531
0.28730304	0.15536838
0.11426581	0.27674868
0.70423289	0.72387041
0.3701205	0.98997792

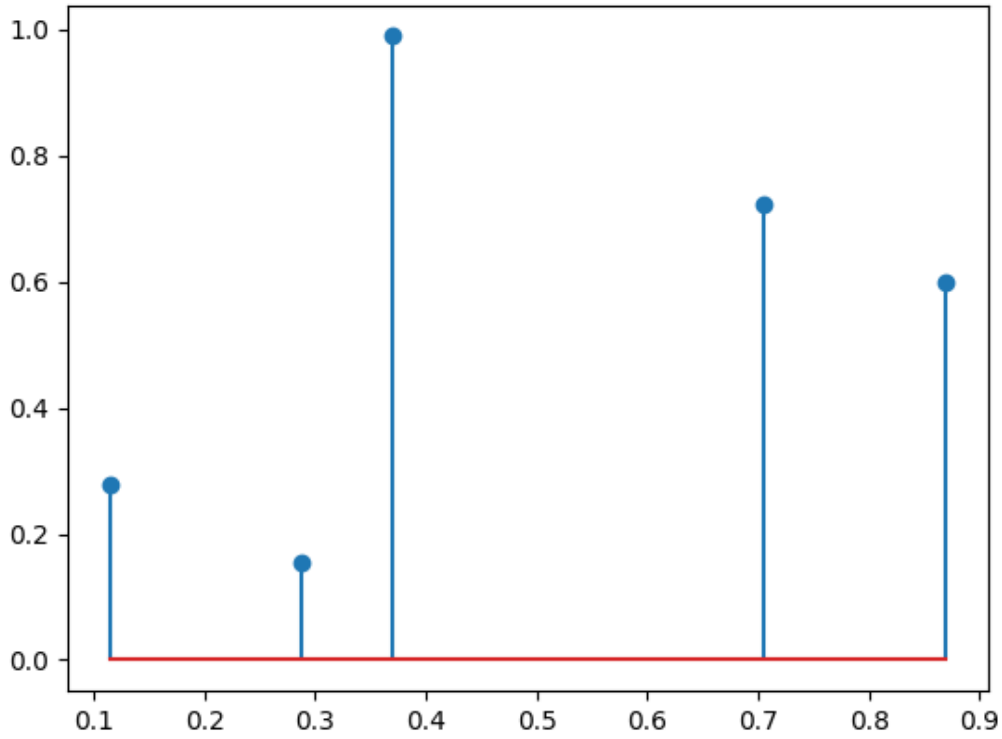


Diagramme cumulatif

Description

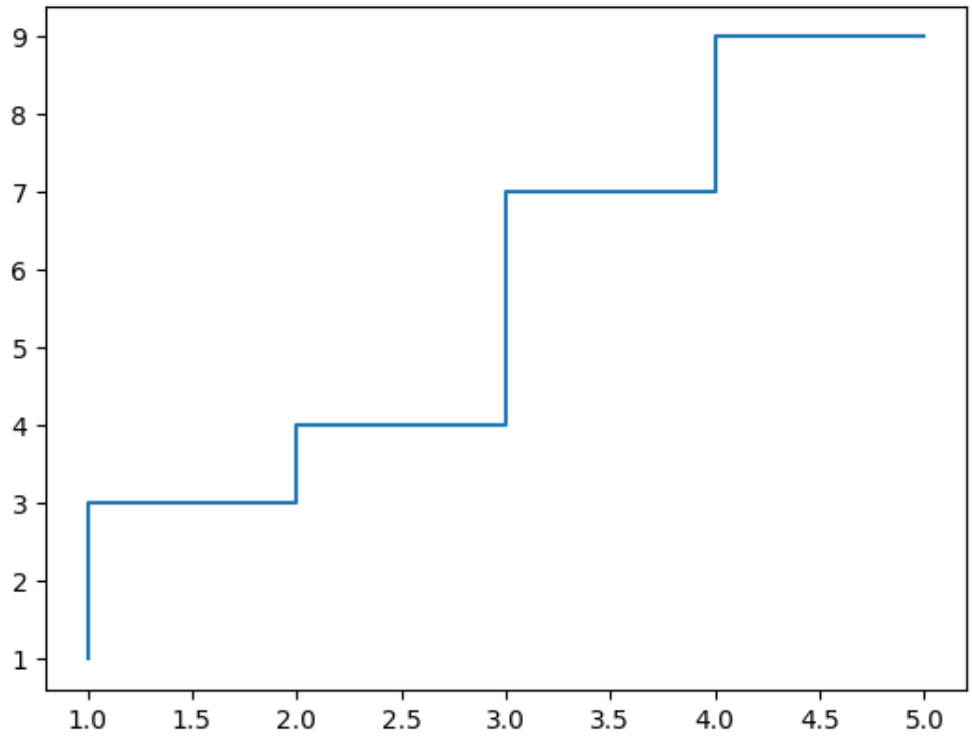
Ce diagramme permet de représenter graphiquement la distribution des effectifs. Il se base donc sur l'effectif cumulé. On peut exprimer l'effectif cumulé d'une valeur comme étant la somme de l'effectif de cette valeur additionné aux effectifs qui lui sont inférieurs.

Représentation

Pour cette partie, nous avons utilisé la suite de valeurs générée de manière aléatoire décrite ci-dessous. Par exemple, avec cette dernière, si l'on souhaite obtenir l'effectif cumulé de la 3ème valeur (v3), nous pouvons faire $1 + 2 + 1$ soit 4. Si nous préférons celle de la 5ème, $v3 + 3 + 2$ soit 9.

x	y
1	1
2	2
3	1
4	3
5	2

La représentation graphique de l'effectif cumulé de la suite ci-dessus est la suivante :



Histogramme

Description

L'histogramme représente la répartition empirique d'une variable aléatoire en la représentant avec des colonnes. Ces dernières portent le nom de **classes**.

Dans un histogramme, les classes sont représentées par un rectangle. La hauteur d'un rectangle représente la quantité de valeurs de cette classe.

Si l'on dispose d'une suite de valeurs, il existe plusieurs formules pour choisir le nombre de classes. Entre autres, il y a celle de Herbert Sturges qui définit que pour un nombre de valeurs N on suggère un nombre de K classes avec $K = 1 + \log_2(N) \approx 1 + \frac{10}{3}\log_{10}(N)$.

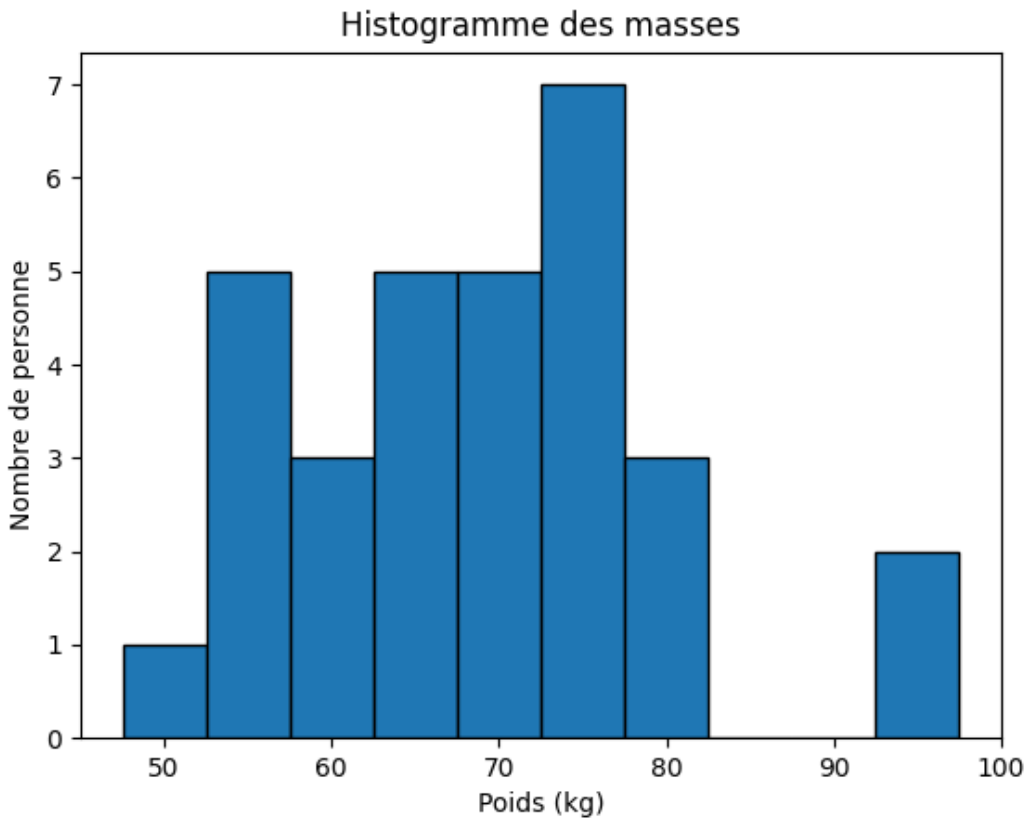
On parle de l'amplitude d'un histogramme comme étant la plage sur laquelle est définie l'histogramme. Elle se calcule avec la formule suivante : $A = Vmax - Vmin$.

Par conséquent, la largeur d'une classe peut être calculée comme suit : $W = A/K$.

L'histogramme est un moyen rapide et efficace d'analyser la répartition d'une certaine valeur.

Représentation

Pour le diagramme ci-dessous, nous avons utilisé les données relatives au poids des personnes de la classe ainsi que la bibliothèque matplotlib de python.



Dans notre cas, nous n'avons pas eu besoin de calculer le nombre de classes ainsi que leur largeur car les données prélevées l'ont été sur des plages de 5 en 5 ([45kg;50kg[,[50kg;55kg[, ...).

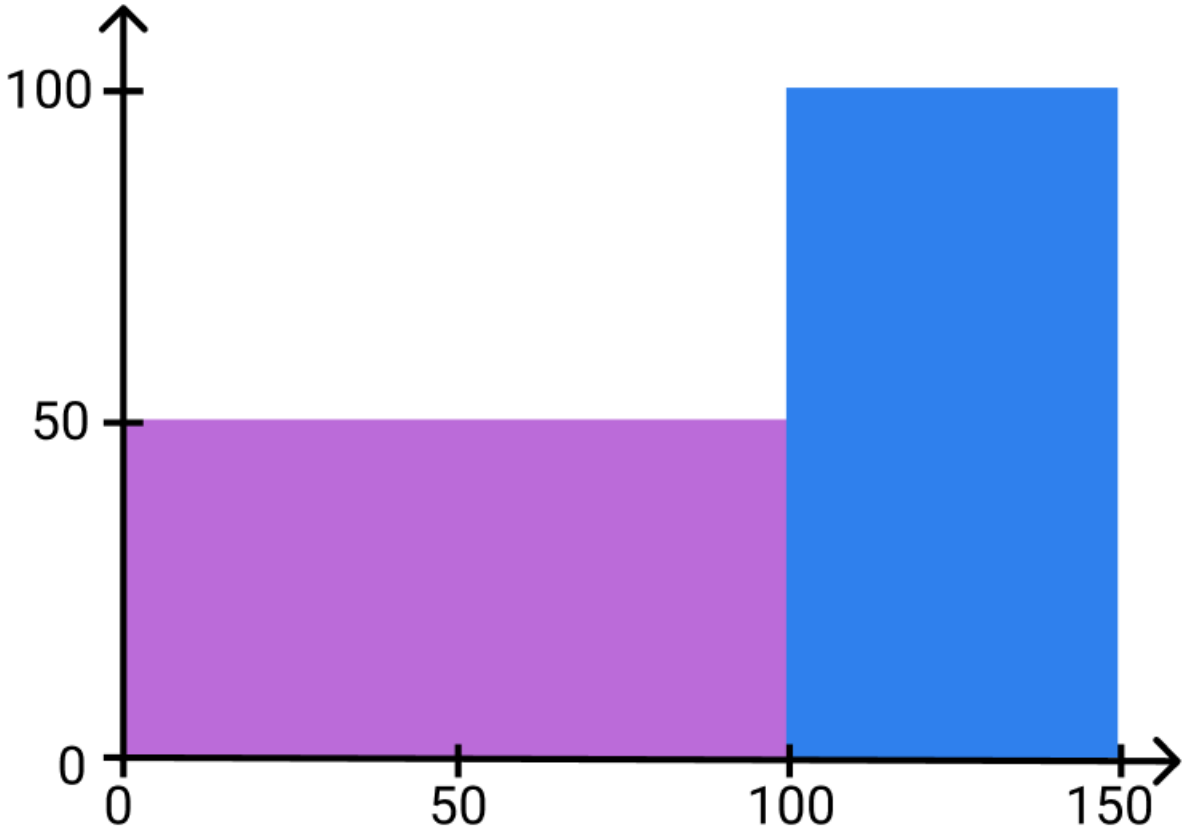
Histogramme à largeur variable

Il est possible de choisir des largeurs variables pour chaque classe de l'histogramme afin d'affiner la représentation. On peut choisir de prendre des intervalles plus larges là où la population est moins dense afin de réduire le bruit (données sont homogénéisées dans l'intervalle) et des intervalles plus étroits là où la population est la plus importante.

La difficulté est qu'on ne peut plus se contenter d'avoir la hauteur de chaque barre proportionnelle à l'effectif de leur classe car c'est l'aire qui doit être proportionelle à l'effectif et on la hauteur. Par exemple si on a deux classes :

Classe :	**0-100**	**100-50**
Effectif :	100	100

On aurait pu être tenté de donner une hauteur égale aux deux barres de l'histogramme car leur effectif est le même. Cependant on remarque que l'intervalle 0-100 est deux fois plus grand que l'intervalle 100-50 donc pour que les aires soient égales il faut que la première barre soit deux fois moins haute que la seconde :



Ici un bean, une case délimité par une graduation verticale et une graduation horzontale correspond à un effectif de 50.

On peut donc calculer l'effectif réel ainsi :

- Pour la première barre : effectif = 50 * 2 (*car l'intervalle est sur 2 graduations*).
- Pour la seconde : effectif = 100 * 1

Vidéo explicative YouTube - Les Bons Profs :

https://www.youtube.com/watch?v=IWyaMBV76EE

Diagramme en boîte / box-plot

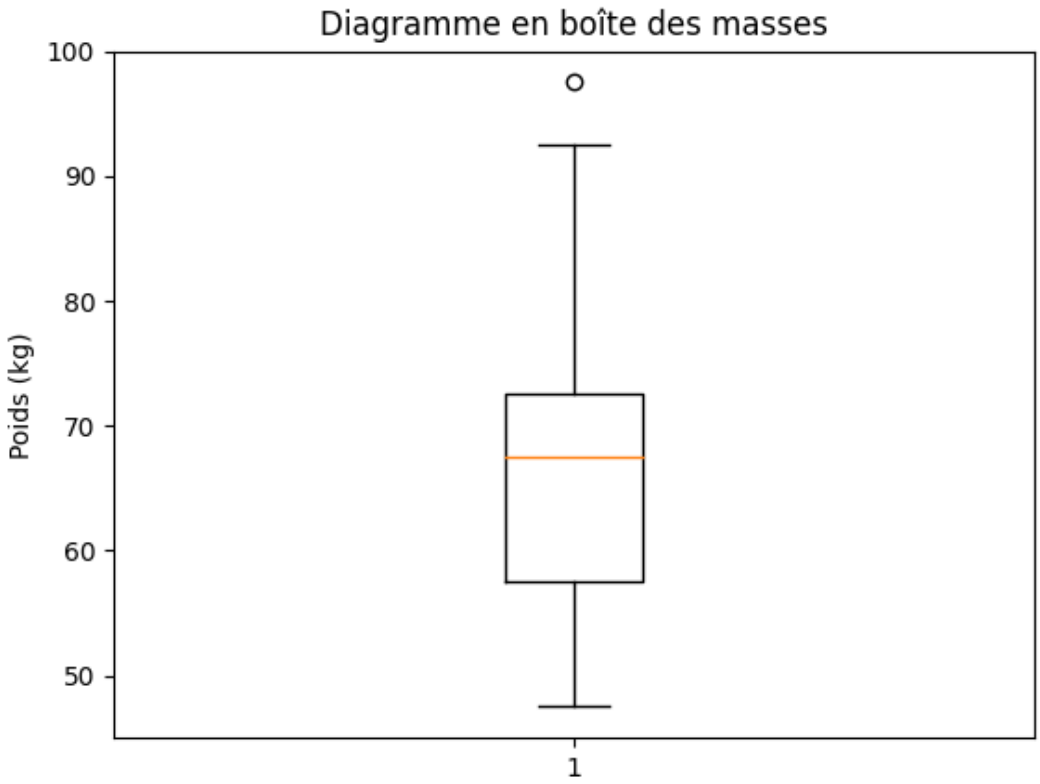
Description

Le diagramme en boîte, connu aussi sous le nom de boîte à moustaches ou boîte de Tukey, permet de visualiser l'essentiel d'une série statistique.

On y retrouve la médiane, les quartiles, la valeur minimale et maximale.

Représentation

Pour le diagramme ci-dessous, nous avons utilisé les données relatives au poids des personnes de la classe ainsi que la bibliothèque matplotlib de python.



Sur ce diagramme, nous pouvons aisément identifier la médiane, (barre orange dans la boite), la valeur minimale (trait du bas), et la valeur maximale (trait du haut). La boîte intégre 25% des données inférieures et supérieures à son milieu. Ainsi la boite intègre 50% des valeurs totales.

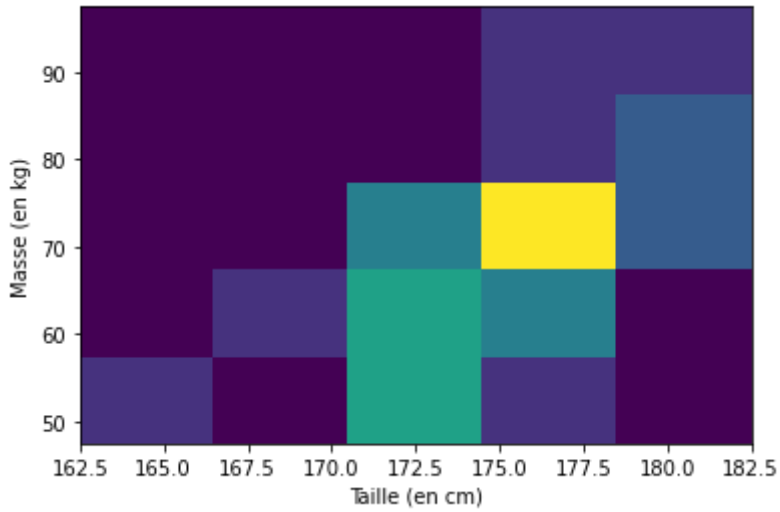
Nous pouvons également remarquer que la valeur `97,5` est considérée comme aberrante car elle n'est pas à l'intérieur de la moustache. Ce cas de figure sous entend qu'il est parfois nécessaire d'en vérifier l'intégrité, afin de statuer de sa véracité. De ce fait, nous éliminons un grand nombre de valeurs aberrantes. À terme, nous nous sommes aperçus que c'était bien le cas ici.

Statistiques descriptives bidimensionnelles

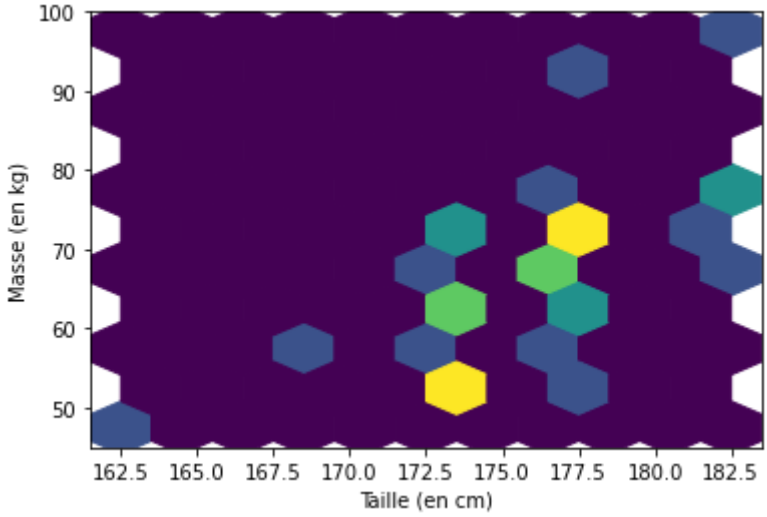
Histogrammes 2D

Les histogrammes 2D permettent de visualiser la répartition des données et utilisent des couleurs pour en indiquer la concentration.

Comme les histogrammes en une seule dimension, les histogrammes 2D sont dépendants de la taille des divisions. Plus les divisions sont grandes, moins on a d'informations sur la distribution. À l'inverse, plus les divisions sont petites, plus elles sont sensibles au bruit. Il est aussi important de sélectionner correctement les couleurs utilisées pour représenter la concentration. En effet, les couleurs ne sont pas perçues linéairement.



Un des problèmes principaux de l'histogramme 2D classique concerne sa répartition en corbeilles. Parfois, deux points éloignés se retrouvent dans la même corbeille, alors que d'autres points, plus proches, sont séparés. Une autre représentation similaire permet d'atténuer ce problème en utilisant des hexagones.



Loi jointe

Définition : Soit (X,Y) un vecteur aléatoire réel. On appelle loi conjointe de (X,Y) la probabilité définie sur R2 par :
$P_{(X,Y)}(I \times J) = \mathbb{P}(X \in I \; et \; Y \in J)$

Les lois de probabilité de X et Y sont alors appelées lois marginales de (X, Y).

En particulier, lorsque X et Y sont à valeurs finies, la loi conjointe de (X, Y) est l'ensemble des $P((X = x_i) \cap (Y = y_j))$

Loi unidirectionnelle ou marginale

La loi marginale d'une variable aléatoire à plusieurs dimensions est la loi de probabilité d'une de ses composantes.

Il est toujours possible, à l'aide de la loi conjointe, de retrouver les lois marginales :

si $X(\Omega) = \{x_1, \ldots, x_p\}$ et $Y(\Omega) = \{y_1, \ldots, y_p\}$,

on a : $P(X = x_i) = \sum_{j=1}^{q} P(X = x_i, Y = y_j)$

Si les composantes sont indépendantes, on a la propriété suivante :

$$P(X = x_i, Y = y_j) = P(X = x_i) \times P(Y = y_j)$$

sinon, dans le cas contraire, on a:

$$P(X = x_i, Y = y_j) = P(X = x_i) \times P(Y = y_j | X = x_i)$$

Loi conditionnelle

Une distribution conditionnelle est une distribution sattistique où on peut déduire VAR à partir d'une autre.

On a comme formule :

$$P_X\big(\{x\}|(Y = y)\big) = \frac{P\big((X=x) \cap (Y=y)\big)}{P(Y=y)}$$

Paramètres de position et de dispersion

Moyenne

C'est la moyenne des moyennes de chaque variable.

Moyenne de distance par rapport au point centrale

Une maniere de caracteriser la dispersion d'une loi aleatoire 2D est de quantifier la distence moyenne entre le points de donné et un point centrale comme la moyenne.

$$\frac{1}{n} \sum_{i=1}^{n} ||\mathbf{z}_i - \mathbf{c}||$$

$$où$$

$$c = (\bar{x}, \bar{y})$$

Notions de dépendance (EN COUR)

Espérance

L'espérance mathématique correspond à une moyenne pondérée des résultats d'une expérience aléatoire dans laquelle les facteurs de pondération sont les probabilités d'obtenir chacun des résultats.

Pour X, une variable aléatoire réelle prenant les valeurs $x1, \ldots, x2$, la formule de l'espérance est la suivante:

$$E(X) = \sum_{i=1}^{n} x_i P(X = x_i)$$

Covariance

La covariance de deux variables aléatoires X et Y est définie par la formule suivante:

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

Si X et Y sont indépendante alors $Cov(X,Y) = 0$.

La covariance d'une variable avec elle-même (autocovariance) est tout simplement la variance.
$Cov(X,X) = Var(X)$

Corrélation

La corrélation est une mesure basée sur la covariance. Elle détermine le degré auquel deux variables se déplacent de façon équivalente. Cette valeur est comprise entre -1 et 1. Plus la valeur de la corrélation est proche de 0, moins les deux variables sont liées. Ainsi si la corrélation est égale à 0, on peut dire que les deux variables sont indépendantes. À l'inverse, plus la valeur de la corrélation est proche de 1 ou -1, plus les variables sont liées (et donc plus il est simple de prédire l'une à partir de l'autre). Dans le cas d'une valeur positive on parle de corrélation positive, et de corrélation négative le cas échéant.

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{(Var(X)Var(Y))}}$$

Matrice de variance-covariance

Var(taille)	**Cov(taille;poids)**
Cov(taille;poids)	Var(poids)

Foret d'arbre de decision

Droite de régression

Paramètres de position et de dispersion minime

La droite de régression linéaire est la droite de forme $y = ax + b$,qui minimise l'écart entre la droite et le nuage des points expérimentaux.
Il y a plusieurs
Nous choisissons en général le carré de la différence entre le point théorique et le point expérimental, c'est-à-dire (yi−(axi+b))2. L'écart total est donc :

Échantillonnage

Introduction

En général il n'est pas possible de réaliser une étude sur l'ensemble d'une population car le nombre d'éléments à étudier est trop important. Dans ce cas on a recours à l'échantillonnage.

L'échantillonnage consiste à sélectionner une sous-partie représentative d'un ensemble d'individus ou, de manière générale, d'un groupe d'éléments varié, afin de produire une série d'échantillons à étudier.

Différents types d'échantillonnages

- Échantillonnage aléatoire et simple : le tirage des individus de l'échantillon est aléatoire, c'est-à-dire que chaque individu a la même probabilité d'être choisi, et simple, c'est-à-dire que les choix des différents individus sont réalisés indépendamment les uns des autres.
- Échantillonnage systématique : le premier individu est choisi de manière aléatoire, puis les suivants sont déterminés à intervalle régulier. Par exemple, dans un verger, on choisit au hasard le 7e pommier, puis les 27e, 47e, 67e, etc.
- Échantillonnage stratifié : on subdivise la population en plusieurs parties avant de prendre l'échantillon.
- Échantillonnage par quotas : la composition de l'échantillon doit être représentative de celle de la population selon certains critères jugés particulièrement importants. On utilise cette méthode pour réaliser les sondages d'opinions.

Cas pratique : sondage des élections présidentielles

En France on utilise la méthode des quotas. Elle consiste à interroger un échantillon représentatif de la population, en s'appuyant sur les statistiques de l'Insee.