Table des matières

Introduction et quelques définitions

La Statistique est l'ensemble de méthodes et outils mathématiques visant à collecter, décrire et analyser des données afin d'obtenir de l'information permettant de prendre des décisions malgré la présence d'incertitude.
La statistique est l'outil de confrontation d'une théorie scientifique à l'observation.
La statistique descriptive est l'étude de résultats observés (statistiques) sur des ensembles comportant un grand nombre d'observations dans le but de les condenser pour en avoir une vue synthétique. Ce condensé se fera en perdant un certain nombre d'observations mais en gardant l'allure générale du phénomène étudié.

A. Population, individu, échantillon

- **Population** : collection complète (dans le sens où elle inclut tous les individus à étudier) d'individus sur laquelle porte l'étude
- **Individu** : Chaque élément de la population s'appelle individu ou unité statistique
- **Echantillon** : sous-ensemble d'individus obtenus à partir de la population (méthodes de sondage)
- **Paramètre** : mesure numérique décrivant une caractéristique de la population
- **Une statistique** : mesure numérique décrivant une caractéristique de l'échantillon
- **Variable ou Caractère** : Caractéristique dont la valeur change d'un individu à l'autre dans la population
- **Donnée** : fait numérique ou non porteur d'information

B. Variables

Chaque individu d'une population peut être décris par une ou plusieurs **variables.** Ces variables peuvent être

- **Variable qualitative** : caractéristiques qui prend des valeurs non numériques qui sont appelées **modalités**. Une variable qualitative peut être :
  - **Nominale** : données non numériques qui ne peuvent pas être ordonnées (genre). Une variable est dite **dichotomique** : si elle ne prend que 2 modalités (comme le genre qui prends les modalités : Homme, Femme)
  - **Ordinale** : données non numériques possédant un ordre naturel (avis pédagogiques)
    Exemple : Un questionnaire de satisfaction demande aux consommateurs d'évaluer une prestation en cochant l'une des six catégories suivantes : (a) nulle, (b) médiocre, (c) moyenne, (d) assez bonne, (e) très bonne, (f) excellente. Il s'agit de modalités ordinales puisqu'elles peuvent être hiérarchisées : une prestation excellente est meilleure qu'une prestation bonne, etc.

- **Variable quantitative** : variable numérique qui peut être :
  - **Discrète** : valeurs numériques discrètes, isolées
    Exemple : On a questionné 100 ménages sur le nombre d'ampoules électriques utilisées dans leur domicile. Les données sont regroupées par nombre d'ampoules.
  - **Continue** : valeurs numériques sur un intervalle continu
    Exemple : On a mesuré 20 personnes et les résultats sont (en cm)
- Les données brutes sont souvent collectées dans tableau dit **tableau des données**

Variables / Individus	X_1	X_2	⋯	X_j	⋯	X_p
1	$x_{1,1}$	$x_{1,2}$		$x_{1,j}$		$x_{1,p}$
2	$x_{2,1}$	$x_{2,2}$		$x_{2,j}$		$x_{2,p}$
⋮						
I	$x_{i,1}$	$x_{i,2}$		$x_{i,j}$		$x_{i,p}$
⋮						
N	$x_{n,1}$	$x_{n,2}$		$x_{n,j}$		$x_{n,p}$

- n est nombre d'individus ou **taille** de la population ou de l'échantillon à étudier ;
- p est nombre de variables étudiées ;
- $x_{i,j}$ est la valeur de la j-ème variable mesurée sur le i -ème individu.
- Ce tableau est appelé **série statistique à p dimensions**

Exemple.

Le tableau suivant une partie de données simulées de livraison de pizza. Les données se réfèrent à un restaurant italien qui propose la livraison de pizzas à domicile. Ce restaurant possède trois succursales du restaurant. La livraison de pizza est gérée de manière centralisée : un opérateur reçoit un appel téléphonique et transmet la commande à l'agence la plus proche du l'adresse du client. L'un des cinq chauffeurs livre la commande. L'ensemble de données capture le nombre de pizzas (Pizzas) commandées comme ainsi que l'addition finale (bill). Le propriétaire de l'entreprise a observé une augmentation du nombre de plaintes, principalement parce que les pizzas arrivent trop tard (time) et trop froides (température). L'analyse des données visent à déterminer les facteurs qui influencent le délai de livraison et la température des pizzas

Day	Date	time	operator	branch	driver	temperature	bill	Pizzas
Thursday	01-May-14	23,8	Laura	West	Bruno	64,6	25,9	1
Thursday	01-May-14	36,1	Laura	Centre	Bruno	64,0	45,2	3
Friday	02-May-14	32,5	Laura	East	Salvatore	52,5	41	3
Friday	02-May-14	34,9	Melissa	Centre	Bruno	63,7	57,4	3
Saturday	03-May-14	42,3	Laura	East	Domenico	76,5	23,5	2
Saturday	03-May-14	20,2	Melissa	East	Domenico	71,8	44,6	2
Sunday	04-May-14	31,4	Melissa	West	Mario	64,8	35,3	5
Sunday	04-May-14	37,8	Laura	West	Domenico	66,1	35,9	2
Monday	05-May-14	41,3	Laura	Centre	Bruno	53,9	37	4
Tuesday	06-May-14	44,0	Melissa	West	Bruno	66,5	56,6	6
Tuesday	06-May-14	30,8	Melissa	Centre	Mario	61,3	24,6	2
Tuesday	06-May-14	32,0	Melissa	East	Mario	65,8	46,2	1
Tuesday	06-May-14	43,3	Melissa	Centre	Bruno	50,2	48,6	4
Wednesday	07-May-14	28,9	Melissa	East	Salvatore	67,2	28,6	2
Wednesday	07-May-14	34,1	Laura	East	Salvatore	70,2	50,4	3
Wednesday	07-May-14	40,2	Melissa	Centre	Salvatore	62,8	39,1	4
Wednesday	07-May-14	42,0	Laura	West	Mario	56,9	39,7	3
Thursday	08-May-14	41,3	Laura	East	Bruno	59,7	31,7	3

Friday	09-May-14	37,5	Laura	East	Luigi	59,5	53,7	2
Saturday	10-May-14	43,7	Melissa	West	Mario	58,6	58	4
Saturday	10-May-14	32,3	Laura	West	Luigi	67,3	38,1	5
Saturday	10-May-14	27,6	Laura	West	Salvatore	57,0	26,4	2
Sunday	11-May-14	24,7	Melissa	East	Domenico	74,4	30,1	1
Sunday	11-May-14	39,8	Melissa	West	Mario	59,3	57,1	4
Sunday	11-May-14	40,4	Melissa	East	Bruno	68,6	52,4	3
Sunday	11-May-14	37,1	Laura	West	Bruno	61,5	67,1	4
Sunday	11-May-14	28,6	Laura	East	Luigi	74,1	24,7	3
Monday	12-May-14	34,5	Laura	West	Salvatore	60,7	40,3	4
Monday	12-May-14	31,6	Laura	East	Bruno	54,7	37,6	4
Wednesday	14-May-14	34,5	Laura	East	Mario	66,3	59,1	2
Wednesday	14-May-14	28,9	Laura	West	Mario	68,1	21	1
Wednesday	14-May-14	39,7	Laura	East	Mario	63,8	31,1	1
Thursday	15-May-14	28,4	Laura	Centre	Mario	65,8	44,1	4
Thursday	15-May-14	28,9	Melissa	West	Mario	51,5	33,5	4
Friday	16-May-14	32,7	Melissa	East	Luigi	66,4	51,3	2
Friday	16-May-14	32,4	Laura	Centre	Salvatore	60,1	30,7	2
Friday	16-May-14	34,3	Laura	Centre	Salvatore	60,4	48,4	3
Friday	16-May-14	33,0	Laura	Centre	Bruno	60,4	43,1	4
Friday	16-May-14	31,8	Laura	West	Luigi	64,1	51,8	4
Saturday	17-May-14	36,1	Laura	East	Domenico	72,5	64,4	1
Saturday	17-May-14	37,6	Melissa	East	Domenico	68,9	30,8	1
Sunday	18-May-14	31,4	Laura	East	Salvatore	73,0	35,5	3
Monday	19-May-14	19,5	Laura	East	Bruno	78,8	20,9	2
Monday	19-May-14	34,4	Laura	Centre	Salvatore	64,6	56,8	6
Monday	19-May-14	35,9	Laura	West	Salvatore	62,6	36,4	2

Chapitre 1. Série statistique à une dimension (univariée)

A. Les représentations graphiques

1. Variables qualitatives

La description d'une population selon une variable qualitative est totalement résumée dans un tableau de pourcentages ou dans un **diagramme circulaire**, appelé aussi **diagramme en camembert**

Modalités	Modalité 1 (x_1)	Modalité 2 (x_2)	⋯	Modalité k (x_k)	⋯	Modalité j (x_j)	Total
Effectifs	$\boldsymbol{n_1}$	$\boldsymbol{n_2}$	⋯	$\boldsymbol{n_k}$	⋯	$\boldsymbol{n_j}$	$\boldsymbol{n}$

- n_k est le nombre de fois ou apparait la modalité k dans la série il est appelé **effectif** de la modalité k ;
- j est le nombre de modalités distinctes observées ;
- La série univariée est donc **résumée** par $\{(x_1, n_1); (x_2, n_2); \dots; (x_j, n_j)\}$;
- Les **fréquences** $f_k = \frac{n_k}{n}$ représentent la fréquence d'observations qui sont égales à x_k ;

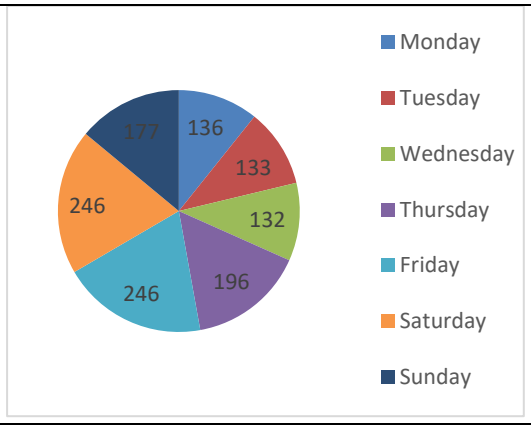
Modalités	Modalité 1 (x_1)	Modalité 2 (x_2)	⋯	Modalité k (x_k)	⋯	Modalité j (x_j)	Total
Effectifs	$\boldsymbol{n_1}$	$\boldsymbol{n_2}$	⋯	$\boldsymbol{n_k}$	⋯	$\boldsymbol{n_j}$	$\boldsymbol{n}$
Fréquence	$\boldsymbol{f_1}$	$\boldsymbol{f_2}$	⋯	$\boldsymbol{f_k}$	⋯	$\boldsymbol{f_j}$	$\mathbf{1}$

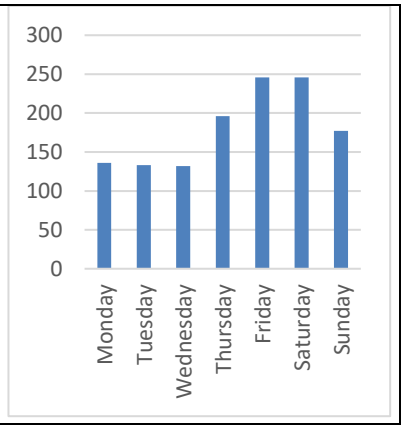
- Le **diagramme en secteurs** est une représentation graphique à l'aide d'un disque découpé en secteur. À chaque modalité observée correspond un secteur. Les angles α_k des secteurs sont proportionnels aux effectifs représentés. Ces angles sont donnés par $\alpha_k = 360 \times f_k$

- Le **diagramme en barres** est un graphique qui présente des variables catégorielles avec des barres rectangulaires avec des hauteurs ou des longueurs proportionnelles aux effectifs des modalités qu'elles représentent

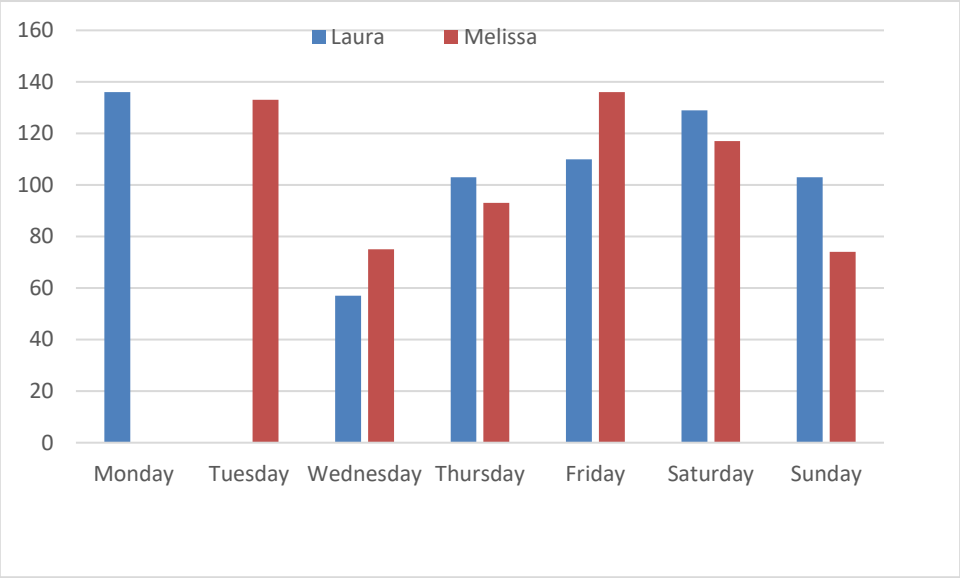
Exemple : le tableau suivant donne le nombre d'admis au baccalauréat en Mauritanie en 2020 par série

Jour	Nombre de commandes
Monday	136
Tuesday	133
Wednesday	132
Thursday	196
Friday	246
Saturday	246
Sunday	177
Total général	**1266**





On peut utiliser le diagramme en barres pour comparer la distribution d'une variable sur deux populations. Par exemple comparer la distribution de la variable série du bac sur les populations des admis et admises au bac.



2. Variables discrètes

Considérons une variable observée sur une population de *n* individus. Si la variable *X* prend *k* valeurs ou ensembles de valeurs (appelés dans ce qui suit, modalités), le premier traitement des données brutes consiste à compter le nombre n_i (appelé **effectif**) d'individus qui présentent la *i -ème* modalité
Soit le **tableau récapitulatif** de la variable

Modalités	x_1	x_2	⋯	x_k	⋯	x_j	Total
Effectifs	$\boldsymbol{n_1}$	$\boldsymbol{n_2}$	⋯	$\boldsymbol{n_k}$	⋯	$\boldsymbol{n_j}$	$\boldsymbol{n}$
Fréquence	$\boldsymbol{f_1}$	$\boldsymbol{f_2}$	⋯	$\boldsymbol{f_k}$	⋯	$\boldsymbol{f_j}$	$\mathbf{1}$
Effectif cumulé	$\boldsymbol{N_1}$	$\boldsymbol{N_2}$	⋯	$\boldsymbol{N_k}$	⋯	$\boldsymbol{n}$	
Fréquence cumulée	$\boldsymbol{F_1}$	$\boldsymbol{F_2}$	⋯	$\boldsymbol{F_k}$	⋯	$\mathbf{1}$	
Effectif cumulé à droite	$\boldsymbol{n}$	$\boldsymbol{N_2^*}$	⋯	$\boldsymbol{N_k^*}$	⋯	$\boldsymbol{N_j^*}$	
Fréquence cumulée à droite	$\mathbf{1}$	$\boldsymbol{F_2^*}$	⋯	$\boldsymbol{F_k^*}$	⋯	$\boldsymbol{F_j^*}$	

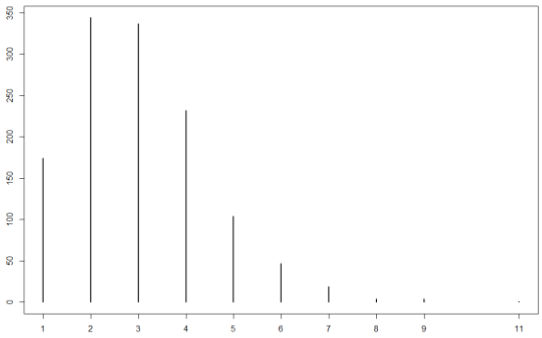
- $f_i = \frac{n_j}{n}$: fréquence
- $N_i = \sum_{j=1}^{i} n_j$: l'effectif cumulé
- $F_i = \frac{N_i}{n}$: fréquence cumulée
- $N_i^* = \sum_{j=i}^{k} n_j$: l'effectif cumulé à droite

- **Diagramme en bâtons :**

  Il est construit dans un système d'axes rectangulaires ; les valeurs de la variable statistique *X* sont portées en abscisse à partir de chaque modalité on trace un segment de droite vertical et dont la hauteur est proportionnelle à l'effectif correspondant. On peut retenir indifféremment une échelle qui explicite les effectifs ou une échelle qui explicite les fréquences.

Exemple. Le tableau suivant synthétise les informations liées à la variable nombre de pizzas par commande de e l'exemple introductif

x_k	n_k	f_k	N_k	F_k	N_k^*	F_k^*
1	174	0,137	174	0,137	1266	1,000
2	344	0,272	518	0,409	1092	0,863
3	337	0,266	855	0,675	748	0,591
4	232	0,183	1087	0,859	411	0,325
5	104	0,082	1191	0,941	179	0,141
6	47	0,037	1238	0,978	75	0,059
7	19	0,015	1257	0,993	28	0,022



8	4	0,003	1261	0,996	9	0,007
9	4	0,003	1265	0,999	5	0,004
11	1	0,001	1266	1,000	1	0,001

3. Variables continues

Le domaine de variation d'une *variable statistique continue X* est partagé en *k parties*. L'intervalle $[l_k^-; l_k^+[$ fermé à gauche, ouvert à droite, est appelé *k*-ème **classe** (*k* = 1,2*, …,j*) ; son amplitude est égale à : $a_k = l_k^+ - l_k^-$.

Il arrive que l'**amplitude** des classes extrêmes soit indéterminée : la première classe étant définie par « *moins de…* », et la dernière par « *plus de…* ». Le choix des *extrémités* des classes se fait à partir des données brutes.

Le nombre de classe doit être modéré entre 4 et 10.

Une classe est un intervalle $[l_k^-; l_k^+[$; son centre ou milieu est $c_k = \frac{l_k^- + l_k^+}{2}$. Le nombre d'observations dans la classe k est noté n_k et est appelé effectif de la classe. On résume les données dans le **tableau récapitulatif** de la variable

Classe	$[l_1^-; l_1^+[$	$[l_2^-; l_2^+[$	⋯	$[l_k^-; l_k^+[$	⋯	$[l_j^-; l_j^+[$	Total
Centre	c_1	c_2	⋯	c_k	⋯	c_j	
Effectifs	$\boldsymbol{n_1}$	$\boldsymbol{n_2}$	⋯	$\boldsymbol{n_k}$	⋯	$\boldsymbol{n_j}$	**N**
Fréquence	$\boldsymbol{f_1}$	$\boldsymbol{f_2}$	⋯	$\boldsymbol{f_k}$	⋯	$\boldsymbol{f_j}$	**1**
Effectif cumulé	$\boldsymbol{N_1}$	$\boldsymbol{N_2}$	⋯	$\boldsymbol{N_k}$	⋯	$\boldsymbol{n}$	
Fréquence cumulée	$\boldsymbol{F_1}$	$\boldsymbol{F_2}$	⋯	$\boldsymbol{F_k}$	⋯	**1**	
Effectif cumulé à droite	$\boldsymbol{n}$	$\boldsymbol{N_2^*}$	⋯	$\boldsymbol{N_k^*}$	⋯	$\boldsymbol{N_j^*}$	
Fréquence cumulée à droite	**1**	$\boldsymbol{F_2^*}$	⋯	$\boldsymbol{F_k^*}$	⋯	$\boldsymbol{F_j^*}$	

- **Histogramme**. À la *k-ème* classe, correspond un rectangle dont la base est l'intervalle $[l_k^-; l_k^+[$ et dont la surface est proportionnelle à la fréquence f_k(ou à l'effectif n_k).

Si les classes ont toutes la même amplitude, les hauteurs des rectangles sont proportionnelles aux fréquences. Dans le cas où les classes sont d'amplitudes inégales, la hauteur du rectangle correspondant à la *k-ème* classe d'amplitude a_k sera $h_k = \frac{f_k}{a_k}$. La surface du rectangle représentant la *k-ème* classe sera ainsi égale à f_k.

- **Polygone** relie le centre de l'histogramme par des segments
- **Courbe cumulative** est une fonction croissante telle que :

- $F(x) = 0$ pour tout $x \leq \min x_i$
- $F(x) = 1$ pour tout $x \geq \max x_i$
- Si $x \in [l_k^-; l_k^+[$,

$$F(x) = \frac{F_k}{a_k}(x - l_k^-) + \frac{F_{k-1}}{a_k}(l_k^+ - x) = F_{k-1} + \frac{f_k}{a_k}(x - l_k^-)$$

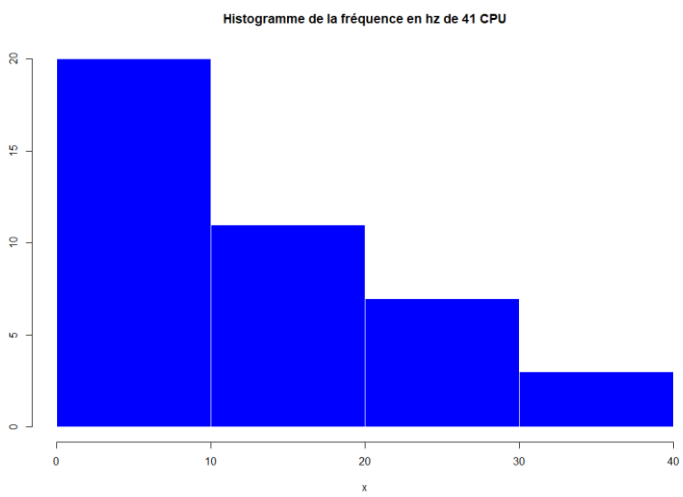
- **Courbe cumulative** à droite

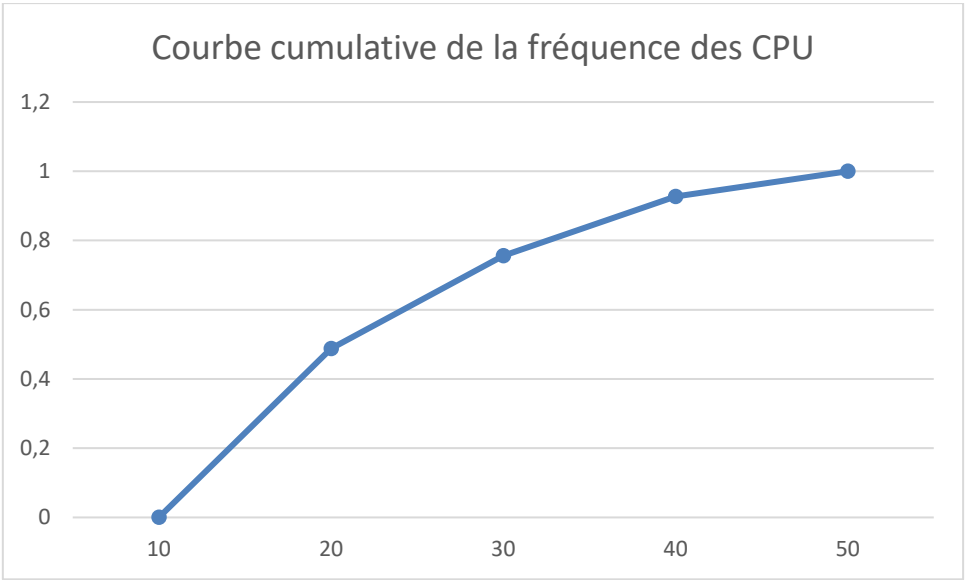
$$x \in [l_k^-; l_k^+[; \qquad F^*(x) = F_k^* + \frac{f_k}{h_k}(x - l_k^-)$$

$F(x)$= proportion des observations inférieurs ou égal à x

Exemple. Les fréquences en hz de 41 CPU sont donnée dans la liste suivante : {3 ; 6 ; 7 ; 7 ; 7 ; 9 ; 3 ; 3 ; 3 ; 40 ; 20 ; 1 ; 10 ; 3 ; 6 ; 7 ; 7 ; 18 ; 5 ; 1 ; 5 ; 40 ; 1 ; 20 ; 20 ; 9 ; 11 ; 17 ; 17 ; 21 ; 26 ; 9 ; 18 ; 18 ; 18 ; 18 ; 26 ; 17 ; 20 ; 33 ; 10}. En les regroupant en 4 classes d'amplitudes égales on obtient

$[l_k^-; l_k^+[$	c_k	n_k	f_k	N_k	F_k	N_k^*	F_k^*
[0-10[	5	20	0,49	20	0,49	41	1,00
[10-20[	15	11	0,27	31	0,76	21	0,51
[20-30[	25	7	0,17	38	0,93	10	0,24
[30-40]	35	3	0,07	41	1,00	3	0,07





B. Représentations numériques

Le tableau de distribution d'une variable statistique présente l'information recueillie sur cette variable. Une représentation graphique en fournit un portrait pour appréhender plus facilement la globalité de l'information. On peut désirer aller plus loin en cherchant à caractériser la représentation visuelle par des éléments synthétiques sur :

- la valeur de la variable située au « centre » de la distribution : la *tendance centrale* et, plus généralement, un *indicateur de position* non nécessairement centrale, liée à un rang donné ;
- la variation des valeurs : la *dispersion* ;
- la *forme* de la distribution ;

- les aspects particuliers : valeurs *extrêmes*, *groupes* de valeurs…

Ces indicateurs étant exprimés dans les unités de la variable étudiée. Pour comparer plusieurs distributions entre elles il est intéressant de calculer des *caractéristiques de dispersion relative.*

1. Tendance centrale (position)

- **Les moyennes**

a. Moyenne arithmétique $\bar{x}$

- Se calcule pour les variables quantitatives comme suit :

Les données brutes	**Les variables discrète**	**Les données regroupées en classe**
$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $= \frac{1}{n}(x_1 + \cdots + x_n)$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{j} n_i x_i$ $= \frac{1}{n}(n_1 x_1 + \cdots + n_j x_j)$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{j} n_i c_i$ $= \frac{1}{n}(n_1 c_1 + \cdots + n_j c_j)$

- Sensible face aux points aberrants
- La somme des valeurs observées est égale à $n\bar{x}$
- Si $Y = aX + b$ alors $\bar{y} = a\bar{x} + b$
- Soit p séries de tailles $n_1, \ldots, n_p$ et de moyennes respectives $\bar{x}_1, \ldots, \bar{x}_p$ alors la moyenne de la série obtenue en regroupant toute les séries est donnée par

$$\bar{x} = \frac{n_1\bar{x}_1 + \cdots + n_p\bar{x}_p}{n_1 + \cdots + n_p}$$

- La série centrée est obtenu en retranchant de chaque valeur la moyenne, c-à-d, $\{x_i - \bar{x}\}$. La moyenne de la série centrée vaut toujours 0

b. Moyenne géométrique est utilisé uniquement pour les variables quantitatives positives

Les données brutes	**Les variables discrètes**	**Les données regroupées en classe**
$g = \sqrt[n]{\prod_{i=1}^{n} x_i}$	$g = \sqrt[n]{\prod_{i=1}^{j} {x_i}^{n_i}}$	$g = \sqrt[n]{\prod_{i=1}^{j} {c_i}^{n_i}}$

c. Moyenne harmonique

Les données brutes	**Les variables discrètes**	**Les données regroupées en classe**

$h = \dfrac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$	$h = \dfrac{n}{\sum_{i=1}^{j} \frac{n_i}{x_i}}$	$h = \dfrac{n}{\sum_{i=1}^{j} \frac{n_i}{c_i}}$

Remarque : On a toujours $h \leq g \leq \bar{x}$

- **Mode.** Pour obtenir une mesure de la tendance centrale non influencée par les valeurs extrêmes de la distribution, on peut prendre la valeur – ou la classe de valeurs – du caractère pour laquelle le diagramme en bâtons – respectivement l'histogramme – présente son *maximum* : c'est le *mode* – respectivement l'*intervalle modal* – de la distribution ; dans le cas où le diagramme en bâtons ou l'histogramme – présente aussi un maximum local, il y a deux modes –respectivement deux classes modales.

La présence de plusieurs modes peut être due au fait que les données sont issues de populations différentes

- **Médiane $x_{0.5}$ et quantiles**

De façon générale le quantile d'ordre $p \in]0,1[$ est la valeur x_p telle que $F(x_p) = p$

a. Si la série est numérique non groupée alors la médiane est l'observation au centre de la série ordonnée (ou la moyenne des deux observations au centre si n est paire)

b. Si la série est groupé il faut d'abord déterminer la classe médiane qui est la classe $[l_k^-; l_k^+[$ pour laquelle $F_{k-1} \leq 0.5$ et $F_k > 0.5$. Ensuite la médiane est donnée par

$$x_{0.5} = l_k^- + h_k \frac{\frac{n}{2} - N_{k-1}}{n_k} = l_k^- + h_k \frac{0{,}5 - F_{k-1}}{f_k}$$

c. Pas influencée par des valeurs extrêmes

d. Pour des distributions dissymétriques, la médiane offre une meilleure représentation que la moyenne

e. La médiane représente la valeur au centre : la moitié est supérieur à cette valeur et la moitié lui est inférieur

- **Quantiles x_p**

- Soit $p \in]0,1[$ le quantile d'ordre p est la valeur x_p pour laquelle $N(x_p) \geq p$ et $N^*(x_p) \geq n(1-p)$

- **Quartiles** : médiane $x_{0.5}$ quantile d'ordre 0.5 ; $x_{0.25}$ **premier quartile** ; $x_{0.75}$ **troisième quartile.** Ils sont notés $\boldsymbol{Q_1}$ **et** $\boldsymbol{Q_3}$

- **Déciles** notés $D_1, \dots, D_{10}$ sont les quantiles pour $p = 0.10, 0.20, \dots, 0.90$

- **Boite à moustache**

On trace un rectangle de largeur fixée à priori et de longueur $EIQ = (Q_3 - Q_1)$, et on y situe la médiane par un segment positionné à la valeur Q_2, par rapport à Q_3 et Q_1 ; on a alors la *boîte* ;
On détermine x_h la plus grande valeur inférieure à $Q3 + 1{,}5\,EIQ$ et x_b la plus petite valeur inférieure à $Q_1 - 1{,}5\,EIQ$;
On trace deux lignes allant des milieux des largeurs du rectangle aux valeurs x_b et x_h
S'il y a des valeurs qui sont inférieures à x_b ou supérieures x_h elles sont représentées par des points et sont considérées comme valeurs aberrantes ou extrêmes
Exemple. Retour à l'exemple précédent
On remarque $F_1 = 0{,}49 < 0{,}5$ et que $F_2 = 0{,}76 \geq 0{,}5$ donc

$$Q_2 = l_2^- + h_2 \frac{0{,}5 - F_1}{f_2} = 10 + 10 \times \frac{0{,}5 - 0{,}49}{0{,}27} = 10{,}37$$

De même on $F_1 = 0{,}49 \geq 0{,}25$ donc

$$Q_1 = l_1^- + h_1 \frac{0{,}25 - F_0}{f_1} = 0 + 10 \times \frac{0{,}25 - 0}{0{,}49} = 5{,}10$$

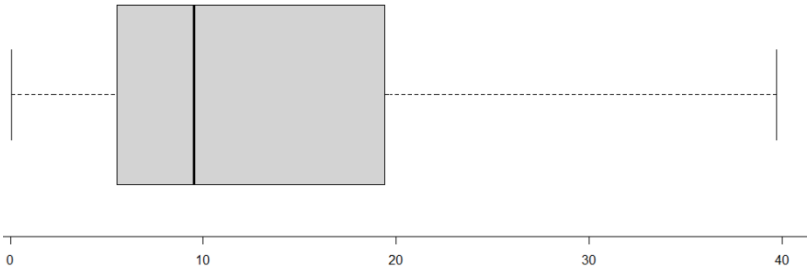
$F_1 = 0{,}49 < 0{,}75$ et que $F_2 = 0{,}76 \geq 0{,}25$ donc

$$Q_3 = l_2^- + h_2 \frac{0{,}75 - F_1}{f_2} = 10 + 10 \times \frac{0{,}75 - 0{,}49}{0{,}27} = 19{,}63$$

On a $Q3 + 1{,}5\,EIQ = 19{,}63 + 1{,}5\,(19{,}63 - 5{,}10)$ =41,42

et $Q1 - 1{,}5\,EIQ = 19{,}63 - 1{,}5\,(19{,}63 - 5{,}10)$ =-2,16

donc $x_h = 40$ et x_b =0



2. Dispersion

- **Etendue** $x_{max} - x_{min}$
- **Ecart interquartile** $EIQ = Q_3 - Q_1$; écart-interdécile $D_9 - D_1$
- **Variance** ;

Les données brutes	Les variables discrètes	Les données regroupées en classe
$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ $= \frac{1}{n}\sum_{i=1}^{n} {x_i}^2 - \bar{x}^2$	$\sigma^2 = \frac{1}{n}\sum_{i=1}^{j} n_j(x_i - \bar{x})^2$ $= \frac{1}{n}\sum_{i=1}^{j} n_j {x_i}^2 - \bar{x}^2$	$\sigma^2 = \frac{1}{n}\sum_{i=1}^{j} n_j(c_i - \bar{x})^2$ $= \frac{1}{n}\sum_{i=1}^{j} n_j {c_i}^2 - \bar{x}^2$

Soit une population de taille *n* composée de deux sous-populations de taille n_1 et de taille n_2. Soit *X*, une variable statistique observée sur la population, alors

$$\sigma^2 = \frac{1}{n}(n_1{\sigma_1}^2 + n_2{\sigma_2}^2 + n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2)$$

- **L'écart-type** σ est la racine carrée de la variance. Si $Y = aX + b$, alors $\sigma_Y = |a|\sigma_X$

- **L'écart absolu moyen** $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$. On note que l'écart absolue moyen est toujours inférieur à l'écart-type. Il est moins utilisé car plus difficile à calculer.

- **Les caractéristiques de dispersion relative**
  Ces caractéristiques permettent de *comparer* les distributions statistiques de plusieurs sous-ensembles d'une même population, ou de faire des comparaisons dans le temps ou dans l'espace.
  Le ***coefficient de variation*** $\frac{\sigma}{\bar{x}}$ est défini pour des variables *positives*
  L'***interquartile relatif*** $\frac{Q_3 - Q_1}{Q_2}$
  L'***interdécile relatif*** $\frac{D_9 - D_1}{D_5}$

Liaisons entre 2 variables

Liaison entre deux variables qualitatives

La répartition des n observations, ou distribution conjointe, suivant les modalités de X et Y se présente sous forme d'un tableau à double entrée, appelé tableau de contingence

X \ Y	y_1	$\cdots$	y_j	$\cdots$	y_c
x_1	n_{11}		n_{1j}		n_{1c}
$\vdots$					$\vdots$
x_i	n_{i1}		n_{ij}		n_{ic}
$\vdots$					$\vdots$
x_l	n_{l1}	$\cdots$	n_{lj}	$\cdots$	n_{lc}

Ici n_{ij} est le nombre d'unité statistique possédant simultanément la modalité x_i de la variable X et la modalité y_j de la variable Y, il est appelé effectif conjoint

Exemple :

On reprend les données des livraisons de pizzas et on croise les variables : succursale et operateur

Succursale\operateur	Laura	Melissa	Total général
Centre	213	208	421
East	209	201	410
West	216	219	435
Total général	**638**	**628**	**1266**

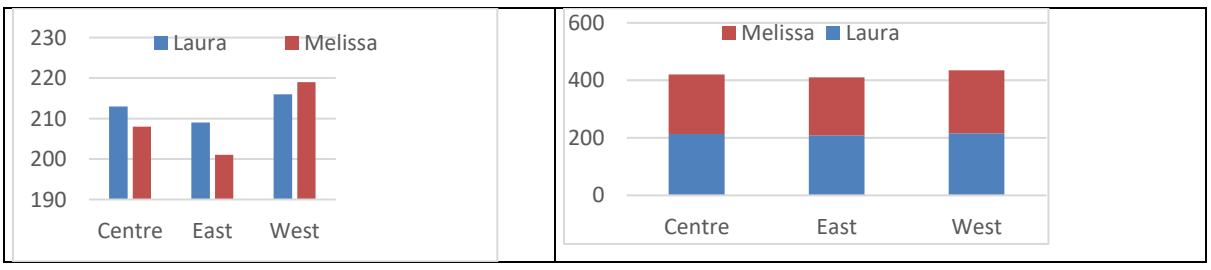
L'effectif 213 correspond au nombre de livraison opérées par Laura à partir de la succursale centre.

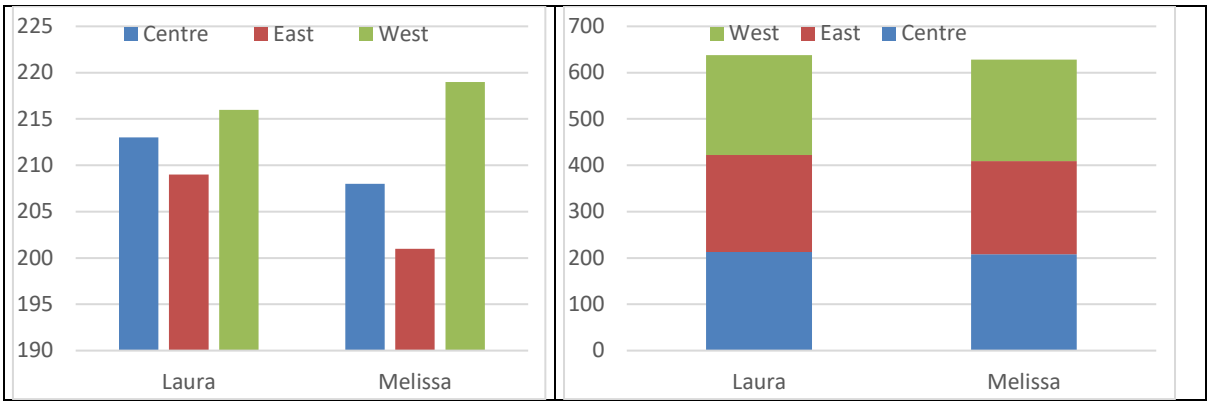
La somme des effectifs conjoints de chaque ligne ou de chaque colonne est appelé effectif marginal :

$$n_{i.} = \sum_j n_{ij} \quad et\ n_{.j} = \sum_i n_{ij}$$

Exemple : 213+208=421 correspond au nombre de livraison effectuées à partir du centre. 208+201+219=628 correspond au nombre de livraisons opérées par Melissa

Plusieurs représentations graphiques peuvent se faire





En cas d'indépendance entre les variables X et Y, on aurait pour tout i et j $f_{ij} = f_{i.} \times f_{.j}$. Ceci impliquerait que les effectifs conjoints seraient égaux $n_{i.} \times n_{.j}/n$. On appelle ces valeurs les effectifs théoriques. L'écart entre ses effectifs ceux observés, n_{ij}, mesure l'écart à la situation d'indépendance. Une mesure de l'écart à l'indépendance que l'on appelle coefficient du χ^2 est donnée par

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - t_{ij})^2}{t_{ij}}$$

où $t_{ij} = n_{i.} \times n_{.j}/n$.

Plus χ^2 est grand plus la liaison entre les variables est forte. Cependant χ^2 dépend de nombre d'observations et des nombres de modalités des 2 variables. On peut démontrer que :$\chi^2 \leq n(\min(l, c) - 1)$. C'est pour cela que l'on a défini le V de Cramèr

$$V = \sqrt{\frac{\chi^2}{n(\min(l, c) - 1)}}$$

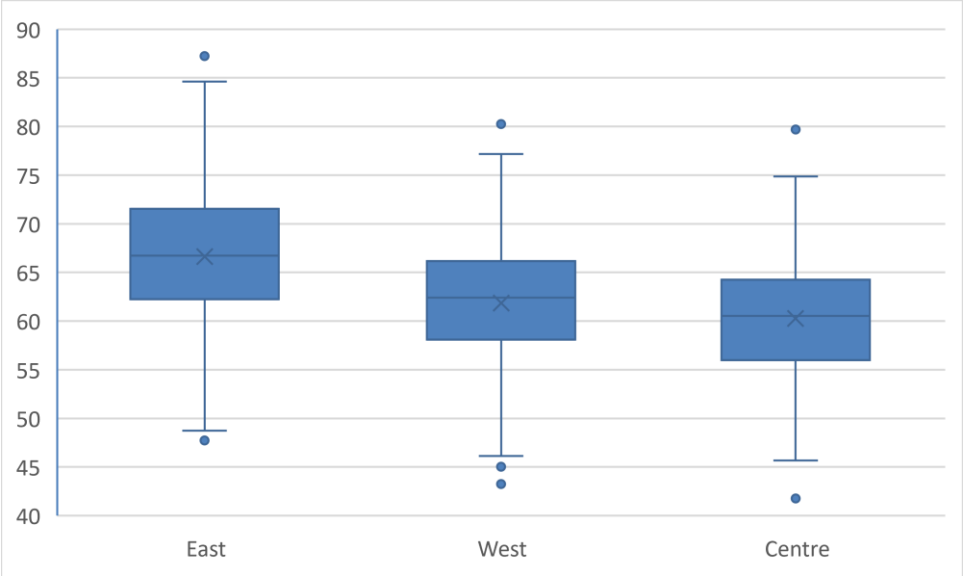
qui est toujours compris entre 0 et 1.

Liaison entre une variable qualitative et une variable quantitative

Soient n observations portant simultanément sur une variable qualitative X à k modalités $x_1, \cdots x_i, \cdots, xk$ et sur une variable quantitative Y. L'objectif ici est déterminer s'il y a un lien entre les variables X et Y. La variable X partitionne la population en k groupes : le groupe d'individus possédant la modalité x_1, le groupe d'individus possédant la modalité x_2, etc... Lorsqu'il y un lien entre les variables, la mesure de la variable quantitative devrait être différente d'un groupe à l'autre. L'objectif est donc de déterminer une façon de mesurer cette différence.

Exemple.

On trace les boites à moustache de la température des pizzas (Données livraisons pizzas) en fonction de la succursale



On observe que les livraisons effectuées à partir de la succursale Est sont ceux qui arrivent les plus chaudes.

Pour mesurer le degré de liaisons on remarque que la variance de la variable quantitative peut se décomposer ainsi :

$$\sigma^2 = \frac{1}{n}\sum_i n_i\sigma_i^2 + \frac{1}{n}\sum_i n_i(\bar{y}_i - \bar{y})^2$$

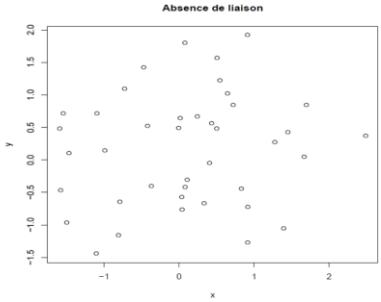
Ce qui se traduit par : la variance totale est égale à la variance inter groupes + variance entre groupe. Plus la variance entre les groupes est importante plus la liaison des variables est forte. Une mesure de la force de liaison entre les variables est mesurée par le rapport de corrélation donnée par

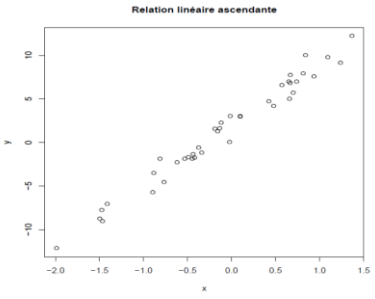
$$\rho = \frac{Variance\ inter}{Varince\ totale}$$

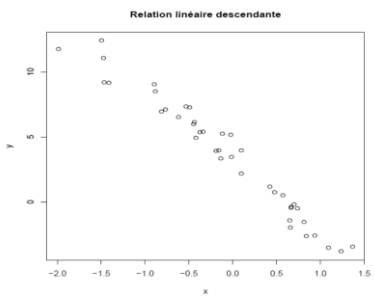
Ce rapport est compris entre 0 et 1 : 0 signifie Independence et 1 signifie liaison fonctionnelle parfaite

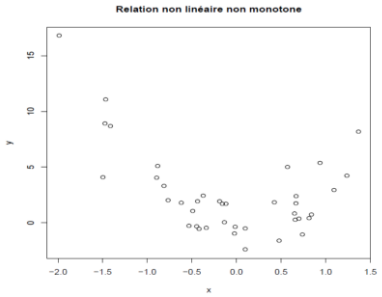
Liaison entre des variables quantitatives

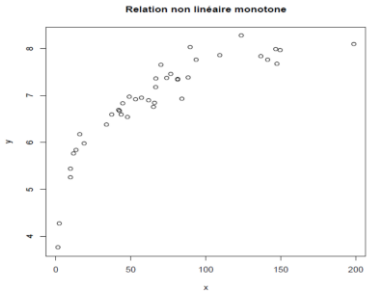
On considère le cas deux variables quantitatives. L'objet ici est de présenter des mesures statistiques rendant compte du sens et de la force de la liaison mathématique qui peut exister entre deux variables quantitatives X et Y. Graphiquement on peut visualiser cette liaison à l'aide d'un diagramme de dispersion ou graphique nuage de points. Un coup d'œil sur ce graphique permet de saisir la nature du lien éventuel. On peut distinguer plusieurs cas











On définit la covariance entre les variables X et Y par

$$\sigma_{xy} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y}$$

Si les variables sont indépendantes alors la covariance est nulle. Cependant on peut avoir une covariance nulle entre deux variables liées mais cette dépendance ne serait pas une liaison linéaire.
La covariance est donc une mesure de la liaison linéaire qui peut exister entre les deux variables. Plus cette covariance est grande en valeurs absolue, plus la liaison est forte.
On définit le coefficient de corrélation linéaire par

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Ce nombre est toujours compris entre -1 et 1. Lorsqu'il est proche de 1 cela signifie une liaison linéaire ascendante, proche de -1 : liaison linéaire descendante, et proche de 0 signifie absence de liaison linéaire.