



Techniques de gestion - Licence 3

STATISTIQUES



SOMMAIRE

- ♦ INTRODUCTION
- ♦ COLLECTE DE DONNÉES
- ♦ STATISTIQUE DESCRIPTIVE À UNE DIMENSION
- ♦ STATISTIQUE DESCRIPTIVE À DEUX DIMENSIONS
- ♦ OUTILS INFORMATIQUE



INTRODUCTION

- ♦ OBJECTIFS DU COURS
- ♦ DÉFINITIONS
- ♦ HISTORIQUE
- ♦ CADRE GÉNÉRAL





— INTRODUCTION —



MOBILISER DES MOYENS ADAPTÉS À VOS ENJEUX

Investir le temps et/ou l'argent nécessaire à la réalisation d'un projet, connaître les retombées d'une action



CHERCHER ET RETENIR LES INFORMATIONS CRUCIALES

Réduire la quantité d'information pour ne garder que l'essentiel

PRENDRE LES BONNES DÉCISIONS AU BON MOMENT

Acheter ou vendre une action au moment opportun, signer un contrat plutôt qu'un autre, placer de l'argent et investir, faire la promotion d'un produit tendance, gérer le risque



DÉFINIR VOTRE STRATÉGIE

Visualiser les axes de travail cruciaux, définir une succession d'évènement et d'actions à prendre





CONNAÎTRE LE VOCABULAIRE DES STATISTIQUES

Pouvoir présenter ses idées et ses questions de manière claire et concise, être autonome dans l'apprentissage futur de nouvelles méthodes statistiques



S'ENTRAINER (BEAUCOUP)

« Practice makes perfect », s'entraîner à utiliser les outils pour développer des réflexes qui feront gagner bien du temps le moment venu, s'entraîner pour prendre de la hauteur sur vos questionnements

COMPRENDRE LES OUTILS STATISTIQUES UTILISÉS

Comprendre leur utilité, comprendre pourquoi et comment ils ont été développés, savoir les utiliser et interpréter leurs résultats



SE QUESTIONNER (SOUVENT)

Être capable d'avoir un sens critique sur ce que vous apprenez, garder le recul nécessaire sur vos conclusions





CONNAÎTRE ET SAVOIR MANIPULER LES TYPES DE VARIABLES

Pouvoir présenter ses idées et ses questions de manière claire et concise, être autonome dans l'apprentissage futur de nouvelles méthodes statistiques



STATISTIQUES DESCRIPTIVES

Maîtriser les statistiques univariées, comprendre les formalismes mathématiques sous-jacents

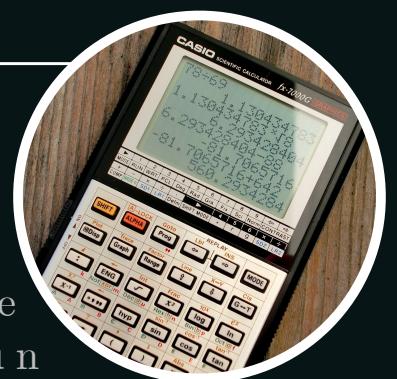
CONSTRUIRE DES TABLEAUX D'EFFECTIF

Comprendre leur utilité, comprendre pourquoi et comment ils ont été développés, savoir les utiliser et interpréter leurs résultats



L'OUTIL INFORMATIQUE

Savoir utiliser une calculatrice ou un ordinateur pour résoudre des problèmes de statistique.





— DÉFINITIONS —



STATISTIQUE n.f.

— du latin *STATUS* (état) —

- (1) Ensemble de données d'observation relatives à un groupe d'individus ou d'unités
 - ♦
- (2) Ensemble des méthodes qui ont pour objet la collecte, le traitement et l'interprétation de ces données
 - ♦
- (3) Variable aléatoire, fonction des observations, construite à partir d'un échantillon.

Source : *dictionary Larousse*



(1) Le plus souvent utilisé au pluriel, on parle par exemple des statistiques démographiques (natalité, mortalité, etc.), des statistiques du chômage, des statistique d'un joueur de football , etc.



(2) C'est à cette définition que nous nous intéressons dans ce cours. Nous suivrons pas à pas la démarche statistique, du recueil des données à leur analyse. L'interprétation fera l'objet d'un autre cours.



(3) Cette définition fera écho un peu plus loin dans ce cours. La moyenne d'un groupe de valeur rentre dans cette définition.



— HISTORIQUE —

1700

1800

1900

— STATISTIQUES ET PROBABILITÉS —

Discipline très récente, apparue au cours du 18^{ème} siècle et à la base appliquée à la connaissance d'un État (arithmétique politique). La théorie des probabilités se développe également en parallèle et stimule les plus grands esprits de l'époque.



PASCAL



LAPLACE



BERNOULLI



MOIVRE



GAUSS



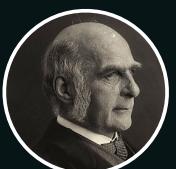
POISSON

— NAISSANCE DES STATISTIQUES MODERNES —

Née de la conjonction de l'arithmétique politique et des probabilités. On assiste à la création d'organismes nationaux et internationaux de statistiques. Les premiers recensement de population à grande échelle ont lieu.



QUETELET



GALTON



BABBAGE

1900

1950

2000

— L'ESSOR DE LA STATISTIQUE —

Développements de méthodes statistiques de plus en plus nombreux et ayant des secteurs d'application de plus en plus diversifiés. On pense notamment à la biologie, la psychologie, l'agronomie, l'économie, l'industrie, la gestion.



PEARSON



SPEARMAN



STUDENT



FISHER

— LES STATISTIQUES ET L'INFORMATIQUE —

Premiers ordinateurs commercialisés en 1950 et introduits dans les administrations et les universités. L'ordinateur facilite l'emploi des méthodes statistiques existantes et ouvre la voie à un nouveau domaine de recherche. L'analyse multidimensionnelle (dont le développement est bien antérieur à cette époque) est rendue aisée grâce aux capacités calculatrices de l'ordinateur. Le champ d'application s'étend également.

2000

— L'ÈRE DES *BIG DATA* —

À l'heure où la problématique n'est plus *COMMENT OBTENIR DES DONNÉES* ? mais bien *COMMENT TRAITER UN FLUX SI IMPORTANT DE DONNÉES* ? de nouveaux paradigmes sont mis en place. On peut citer des anglicismes désormais bien établis tel que *database*, *data center*, *data mining*, *data analyst*, *neural network*, *artificial intelligence*, etc. marquant bien le caractère mondial de cette science.



LE CUN



KOLLER



PATIL

2020



Les statistiques sont aujourd'hui utilisées dans à peu près tous les domaines imaginables :

-en **MÉDECINE**, comme cela a pu être démontré récemment avec les chiffres concernant l'efficacité d'un vaccin ,

-en **ÉCONOMIE**, pensez à la bourse ou aux compagnies d'assurance notamment,

- en **ENTREPRISE**, pour quantifier le bien-être des salariés, pour analyser les résultats financiers , pour contrôler la qualité des produits ,

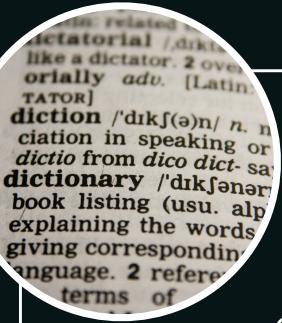
- au **QUOTIDIEN**, lorsque l'on vous donne des part des marché à la télé , la côte de popularité du Président de la République ,

- en **SCIENCES « DURES »**, pour déterminer l'effet d'une basket sur votre façon de courir, pour détecter des exoplanètes , pour mesurer la masse d'un proton ,

- en **SCIENCE SOCIALE**, pour évaluer le comportement de groupe de personnes, on pense facilement à **FACEBOOK**



— CADRE GÉNÉRAL —



ÉTUDE STATISTIQUE

Une étude statistique est composée de **DEUX PHASES** bien distinctes.

On procède d'abord au RASSEMBLEMENT ou à la **COLLECTE DES DONNÉES** qui servent à obtenir les informations sur lesquels on veut travailler.

S'en suit alors l'**ANALYSE STATISTIQUE**, qui peut elle même être séparée en deux étapes :

- la **STATISTIQUE DESCRIPTIVE** (c'est l'un des objets de ce cours)
- l'**INFÉRENCE STATISTIQUE**



COLLECTE DES DONNÉES

Peut être réalisé soit par simple **OBSERVATION** des phénomènes d'intérêt tels qu'ils se produisent naturellement, soit par l'**EXPÉRIMENTATION**, en provoquant volontairement l'apparition de phénomènes

Le recueil peut être **TOTAL**, on parle alors de **RECENSEMENT**, ou **PARTIEL**, on parle alors de **SONDAGE**.



ANALYSE STATISTIQUE

— STATISTIQUES DESCRIPTIVES —

Elles ont pour but de résumer et de présenter les données observées d'une manière intelligible et aisément compréhensible. Les tableaux et les graphiques sont les principales armes de cette méthode

— INFÉRENCES STATISTIQUES —

Elles ont pour but de généraliser les conclusions obtenues à l'aide des données mesurées sur une fraction des individus auxquels est porté intérêt.

— STATISTIQUE/STATISTIQUES —

Voir le chapitre [INTRODUCTION](#) que nous venons de passer.



— COLLECTE DE DONNÉES/OBSERVATION/EXPÉRIMENTATION —

Voir le chapitre suivant intitulé [COLLECTE DE DONNÉES](#) qui vous donnera les principales clés de ce qui constitue de manière générale la première étape de toute étude statistique.



— ANALYSE STATISTIQUE —



— STATISTIQUE DESCRIPTIVE —

Voir le chapitre [STATISTIQUE DESCRIPTIVE À UNE DIMENSION](#). Pour la STATISTIQUE DESCRIPTIVE À DEUX OU PLUSIEURS DIMENSIONS, vous devrez patienter un peu.

— INFÉRENCE STATISTIQUE —

Hors des limites de ce cours. Vous apprendrez les bases de l'inférence statistique plus tard, lorsque vous serez des champions de la statistique descriptive ! 🏆

statistiques
données
statistique
Collecte
observation
analyse
descriptive
expérimentation
inférence

COLLECTE DE DONNÉES

- ♦ [INTRODUCTION](#)
- ♦ [VOCABULAIRE DES STATISTIQUES](#)





— DÉFINITIONS —



— PRÉPARATION DE L'ÉTUDE STATISTIQUE —

— MAITRISER UN MAXIMUM D'ÉLÉMENTS —

À quelles questions doit répondre l'étude ? Que dois-je observer pour répondre à ces questions ? Comment m'y prends-je pour mesurer ces métriques ? Comment seront traiter les données ? Comment seront présentés les résultats ?



— ÉTUDE PAR ENQUÊTE —

Constitue un des deux moyen de prélever l'information. La PREMIÈRE partie de ce chapitre s'y intéresse.



— EXPÉRIMENTATION —

Constitue un des deux moyen de prélever l'information. La DEUXIÈME partie de ce chapitre s'y intéresse.



— ENREGISTREMENT ET TRAITEMENT —

À quels types de données aurais-je affaire ? Quelle quantité y en aura-t-il ? Les valeurs obtenues sont-elles plausibles à première vue ?

— VOCABULAIRE DES STATISTIQUES —





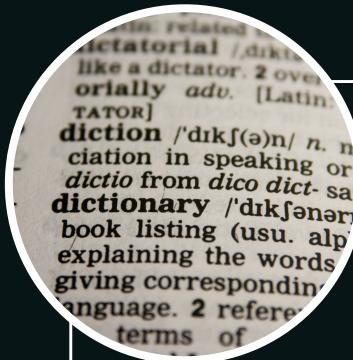
ENQUÊTE n.f.

— du latin INQUIRERE
(rechercher) —

Étude d'une question faite en réunissant des témoignages et des expériences

En statistiques, l'ENQUÊTE (ou INVENTAIRE) a pour but de COLLECTER DE FAÇON ORGANISÉE des INFORMATIONS relatives à un GROUPE D'INDIVIDUS ou d'éléments, observés dans leur MILIEU OU CADRE NATUREL.

Exemple : *mesure du diamètre des arbres d'une forêt.* 



EXPÉRIMENTATION

n.f. — du latin INQUIRERE
(rechercher) —

Méthode scientifique reposant sur l'expérience et l'observation contrôlée pour vérifier des hypothèses.

En statistiques, l'EXPÉRIMENTATION diffère de l'enquête par le fait que l'apparition des faits que l'on désire étudier est VOLONTAIREMENT PROVOQUÉE, dans des CONDITIONS QUE L'ON MAÎTRISE au moins partiellement.

Exemple : *mesure de la masse d'un électron.* 



INDIVIDU

Appelé aussi UNITÉ DE BASE ou UNITÉ STATISTIQUE, il représente l'élément indivisible d'une population statistique.

L'INDIVIDU peut être une personne, un groupe de personnes, un animal, un végétal ou un élément de toute nature (un ordinateur, une voiture, etc.)

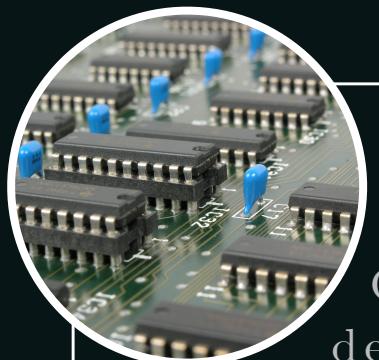


POPULATION

Appelée aussi UNIVERS ou ENSEMBLE STATISTIQUE, la POPULATION regroupe l'ensemble des individus auxquels on s'intéresse.

Quand on observe tous les INDIVIDUS d'une POPULATION, l'ENQUÊTE est dite COMPLÈTE ou EXHAUSTIVE. Elle s'appelle alors RECENSEMENT.

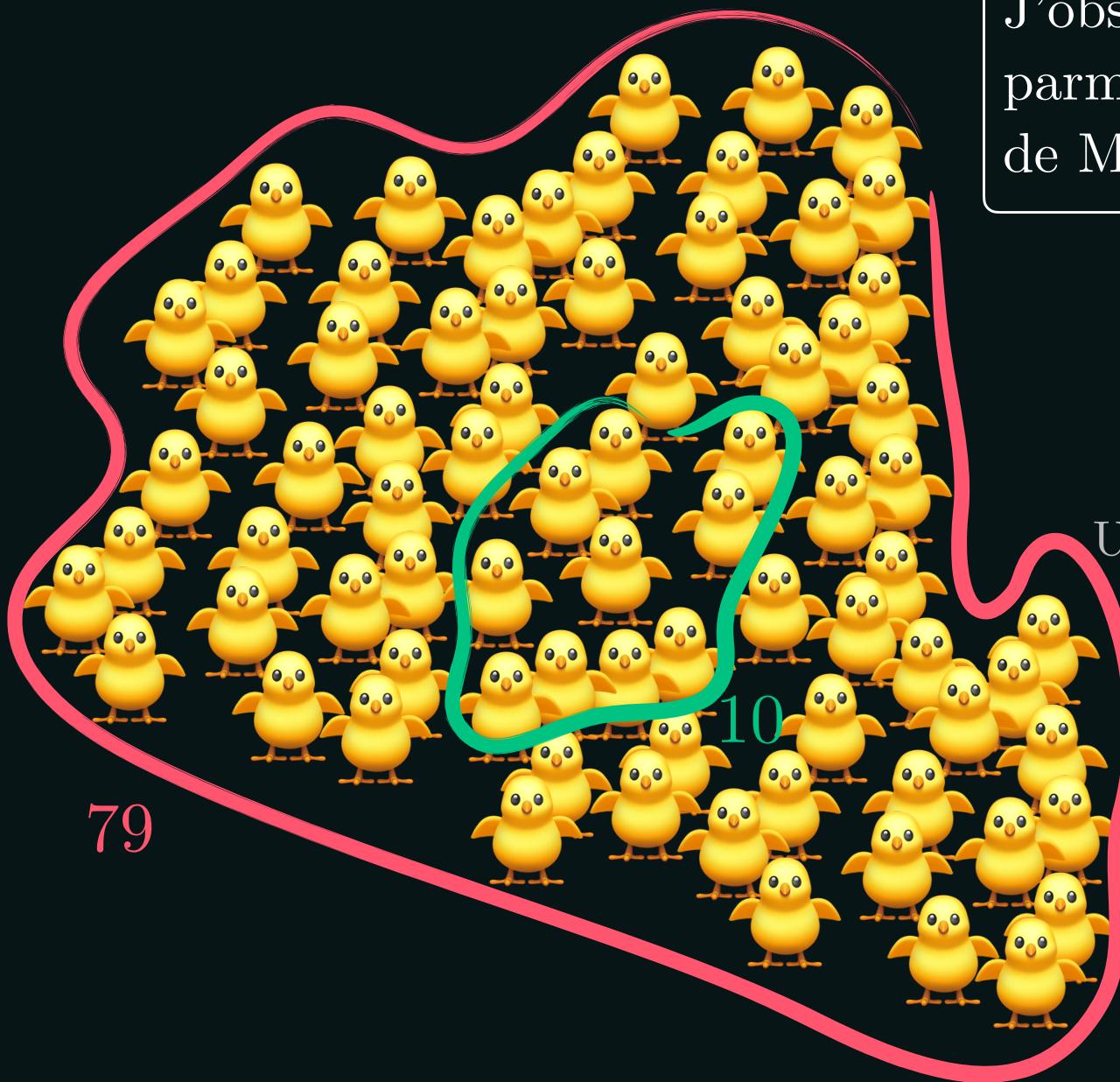
Quand on observe une partie de INDIVIDUS, l'ENQUÊTE est dite PARTIELLE ou À ÉCHANTILLONNAGE. On la nomme aussi SONDAGE.



VARIABLE

Grandeur susceptible de varier dans un ensemble donné. Également appelée CARACTÈRE, elle est l'essence même de la statistique.

Une VARIABLE peut se présenter sous plusieurs formes, que nous allons énumérer par la suite.



J'observe le poids de **10 POUSSINS** parmi les **79 POUSSINS** de la ferme de M. Bertrand



Que nomme-t-on alors POPULATION ?

LES 79 POUSSINS de la ferme de M. Bertrand

Que nomme-t-on alors INDIVIDU ?

UN POUSSIN en particulier, celui là 🐥 par exemple

Quelle est la VARIABLE d'intérêt ?

Le POIDS (exprimé en g par exemple)

Que représente le groupe de 10 POUSSINS ?

Il constitue un ÉCHANTILLON de la population

À quel type d'ENQUÊTE avons-nous affaire ?

Il s'agit d'un SONDAGE

VARIABLES QUALITATIVES



Le résultat de l'observation d'une variable QUALITATIVE est un CARACTÈRE souvent représenté par un MOT. On distingue deux classes distinctes :

—VARIABLE ORDINALE—

Les CARACTÈRES présentent plusieurs niveaux différents pouvant être HIÉRARCHISÉS.

—VARIABLES NOMINALE—

Les CARACTÈRES présentent plusieurs niveaux différents ne pouvant PAS être ORDONNÉS DE MANIÈRE LOGIQUE.



VARIABLES QUANTITATIVES

Le résultat de l'observation d'une variable QUANTITATIVE est un NOMBRE. On distingue ici encore deux classes distinctes :

—VARIABLE DISCRÈTE—

Simple DÉNOMBREMENT ou COMPTAGE, son résultat R est un nombre entier non négatif

$$R \in \mathbb{N}^+$$

—VARIABLES CONTINUES—

MESURE ou MENSURATION, son résultat R est un nombre réel

$$R \in \mathbb{R}$$

Quel système d'exploitation utilise cet ordinateur ?

WINDOWS, LINUX, MAC OS

Est-il équipé d'un port FireWire ?

OUI / NON

Quel est son état général ?

NEUF, ÉTAT CORRECT, USÉ, CASSÉ



Donnez des exemples de variables qualitatives et quantitatives concernant cet ordinateur

Combien de port USB possède-t-il ?

0, 1, 2, ...

À quelle fréquence son processeur opère-t-il ?

Un nombre exprimé en GHz



SAISIE MANUSCRITE
ENREGISTREMENT

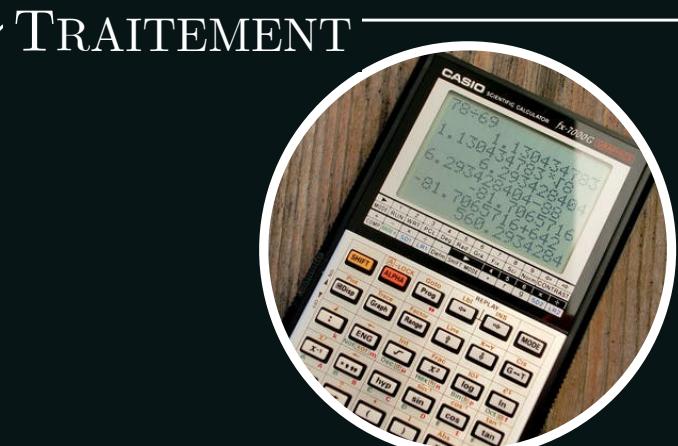
SAISIE MANUELLE
DANS UN TABLEUR

```
scrin(preg_replace('/\\\\\\/\r\n', '', $hex2rgb));
$captcha['config'] = serialize($captcha_config);
$captcha['code'] => $captcha_config['code'];
$image_src => $image_src;
```

```
if(function_exists('hex2rgb')) {
    $hex_str = preg_replace('/[a-f0-9a-f]/i', '', $hex_str); // Gets rid of separators
    $rgb_array = array();
    if(strlen($hex_str) == 6) {
        $color_val = hexdec($hex_str);
        $rgb_array[0] = $color_val & 0xFF;
        $rgb_array[1] = ($color_val >> 8) & 0xFF;
        $rgb_array[2] = ($color_val >> 16) & 0xFF;
    } else if(strlen($hex_str) == 3) {
        $rgb_array[0] = hexdec(str_repeat(substr($hex_str, 0, 1), 2));
        $rgb_array[1] = hexdec(str_repeat(substr($hex_str, 1, 1), 2));
        $rgb_array[2] = hexdec(str_repeat(substr($hex_str, 2, 1), 2));
    } else {
        return false;
    }
    return $return_string ? implode($separator, $rgb_array) : '';
}
```

ENREGISTREMENT
AUTOMATIQUE

—VÉRIFICATION RAPIDE DES DONNÉES—



CALCULATRICE

TABLEUR

```
scrin(preg_replace('/\\\\\\/\r\n', '', $hex2rgb));
$captcha['config'] = serialize($captcha_config);
$captcha['code'] => $captcha_config['code'];
$image_src => $image_src;
```

```
if(function_exists('hex2rgb')) {
    $hex_str = preg_replace('/[a-f0-9a-f]/i', '', $hex_str); // Gets rid of separators
    $rgb_array = array();
    if(strlen($hex_str) == 6) {
        $color_val = hexdec($hex_str);
        $rgb_array[0] = $color_val & 0xFF;
        $rgb_array[1] = ($color_val >> 8) & 0xFF;
        $rgb_array[2] = ($color_val >> 16) & 0xFF;
    } else if(strlen($hex_str) == 3) {
        $rgb_array[0] = hexdec(str_repeat(substr($hex_str, 0, 1), 2));
        $rgb_array[1] = hexdec(str_repeat(substr($hex_str, 1, 1), 2));
        $rgb_array[2] = hexdec(str_repeat(substr($hex_str, 2, 1), 2));
    } else {
        return false;
    }
    return $return_string ? implode($separator, $rgb_array) : '';
}
```

LANGAGE DE
PROGRAMMATION

recensement
échantillon
enquête
population
expérimentation

discrète
quantitative
continue
qualitative
nominale
ordinale

sondage

STATISTIQUE DESCRIPTIVE À UNE DIMENSION



- ♦ INTRODUCTION
- ♦ SÉRIE ET DISTRIBUTIONS
- ♦ REPRÉSENTATION GRAPHIQUES
- ♦ RÉDUCTION DE DONNÉES
- ♦ PARAMÈTRES DE POSITION
- ♦ PARAMÈTRES DE DISPERSION
- ♦ PARAMÈTRES DE CONCENTRATION
- ♦ PARAMÈTRES DE FORME



— INTRODUCTION —



— BUT DE LA STATISTIQUE DESCRIPTIVE —

Présenter les données sous une forme INTELLIGIBLE, FACILEMENT ABORDABLE. Elle peut s'intéresser à une VARIABLE UNIQUE, on parle alors de STATISTIQUE DESCRIPTIVE UNIVARIÉE ou à UNE DIMENSION, c'est l'objet de ce chapitre.



— TABLEAUX STATISTIQUES —

Présentent les données sous la forme numérique de distribution de fréquence.



— DIAGRAMMES —

Présentent les distribution de fréquence ou les données initiales sous forme GRAPHIQUE.



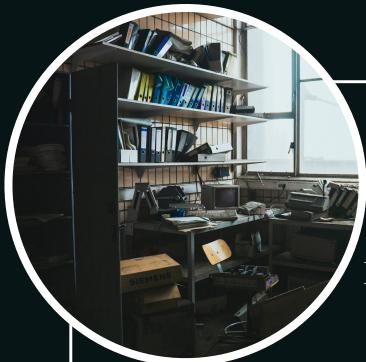
— PARAMÈTRES —

Valeurs typiques représentant la distribution statistique des données. Le calcul de ces paramètres constitue la RÉDUCTION DE DONNÉES.





SÉRIES ET DISTRIBUTIONS



— SÉRIE STATISTIQUE —

C'est la forme la plus ÉLÉMENTAIRE de présentation des données statistiques. Elle consiste en une simple ÉNUMÉRATION DES OBSERVATIONS :

$$S = \{x_1, x_2, \dots, x_n, \dots, x_N\}$$

avec $n \in [1; N]$ où $N = \text{Card}(S)$ désigne le NOMBRE TOTAL D'OBSERVATIONS, aussi appelé EFFECTIF.

x_n représente l'unique et $n^{\text{ième}}$ observation.

Les valeurs de S peuvent ne pas être uniques et peuvent vérifier l'égalité $x_i = x_j$.

Les observations peuvent également être rangées par ordre croissant tel que :

$$x'_1 \leq x'_2 \leq \dots \leq x'_n \leq \dots \leq x'_N$$

Que nomme-t-on
POPULATION ?

Les cyclistes du Tour

Que nomme-t-on
INDIVIDU ?
Un cycliste

Quelle est la
VARIABLE d'intérêt ?
La FTP

Quel est l'EFFECTIF
de l'étude ?
 $N = 25$

6.2	6.1	5.9	6.2	6.4
6.5	6.1	5.9	6.2	6.3
6.2	6	6.6	5.7	6.1
5.9	6.1	6.3	6.2	6.1
6.4	6.3	6.3	6.1	6.4

Seuils de puissance fonctionnelle (Fonctional Threshold Power — FTP) exprimé en W/kg de 25 cyclistes professionnels masculins participant au Tour de France.



— DISTRIBUTION GROUPÉE ET NON GROUPÉES —

Lorsque les observations sont NOMBREUSES, il est souvent utile de les CONDENSER sous la forme d'une DISTRIBUTION DE FRÉQUENCES. On distingue deux cas de DISTRIBUTIONS STATISTIQUES, les DISTRIBUTIONS NON GROUPÉES et les DISTRIBUTIONS GROUPÉES.

— DISTRIBUTIONS NON GROUPÉES —

On prend les éléments tels qu'ils sont, AUCUN REGROUPEMENT n'est effectué entre eux. Ce genre de distribution se retrouve généralement lorsque l'on a affaire à des VARIABLES DISCRÈTES.

'Autre'	'Kalenji'	'Asics'	'Adidas'	'Autre'
'Brooks'	'Autre'	'Adidas'	'Brooks'	'Adidas'
'Kalenji'	'Adidas'	'Nike'	'Kalenji'	'Brooks'
'Kalenji'	'Brooks'	'Kalenji'	'Nike'	'Autre'
'Autre'	'Kalenji'	'Autre'	'Brooks'	'Kalenji'
'Asics'	'Asics'	'Brooks'	'Adidas'	'Asics'

Marque de chaussure de 30 coureurs relevée au départ du marathon de Paris.

— DISTRIBUTION GROUPÉES —

On REGROUPE les éléments de la série en formant des CLASSES, des CATÉGORIES. Ce genre de distribution se retrouve généralement lorsque l'on a affaire à des VARIABLES CONTINUES.

4076.2	6767.2	5871.9	6793.5	6009.5	4198.2
3972.3	6592.7	3657.6	7373.3	6581.4	5575.4
5006.5	5731.2	12053.4	4775.1	1886.8	8830.2
6219.7	3704.2	10234.7	6101.9	5588.4	3444.3
2870.1	1463.2	4154.2	2893.8	6295.5	4364.7

Kilomètres annuels parcourus par 30 cadets du Vélo Club La Pomme Marseille.



— EFFECTIF —

L'EFFECTIF d'une MODALITÉ (valeur possible d'une variable statistique) est le nombre de fois n_i où la modalité i apparaît dans la SÉRIE STATISTIQUE. Dans l'exemple précédent, on a par exemple $n_{Asics} = 4$.

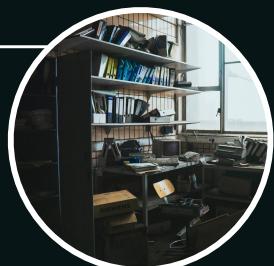
La relation $\sum_i n_i = N$ est également toujours vérifiée.

— FRÉQUENCE —

La FRÉQUENCE f_i de la modalité i est le rapport entre l'effectif de la modalité n_i et l'effectif total N tel que

$$f_i = \frac{n_i}{N}$$

La relation $\sum_i f_i = 1$ est également toujours vérifiée. Exprimé en pourcentage, cela devient $f_i(\%) = 100 \times \frac{n_i}{N}$



'Autre'	'Kalenji'	'Asics'	'Adidas'	'Autre'
'Brooks'	'Autre'	'Adidas'	'Brooks'	'Adidas'
'Kalenji'	'Adidas'	'Nike'	'Kalenji'	'Brooks'
'Kalenji'	'Brooks'	'Kalenji'	'Nike'	'Autre'
'Autre'	'Kalenji'	'Autre'	'Brooks'	'Kalenji'
'Asics'	'Asics'	'Brooks'	'Adidas'	'Asics'

	n_i	f_i
Nike	2	0,07
Adidas	5	0,17
Brooks	6	0,20
Asics	4	0,13
Kalenji	7	0,23
Autre	6	0,20



— ENONCÉ —

On effectue une enquête auprès d'une classe de L3 concernant leur genre cinématographique préféré en terme de séries. L'enquête prend cette forme simple :

Quel est votre genre cinématographique préféré en terme de séries ?

- comédie (🎭)
- fantastique (👽)
- action (🔫)
- thriller (😱)
- romantique (❤️)
- autre (💡)

Réponses brutes :

💡 🛸 🎭 ❤️ 💣 💡 ❤️ 😱 💣 🛸 🛸 🎭 😱 💡 💡 💡 💡 💡 💡 💡 } $N = 30$

x_i	n_i	f_i	$f_i(\%)$
🎭	4	4/30	100 x 4/30
🔫	4	4/30	100 x 4/30
❤️	3	3/30	100 x 3/30
👽	6	6/30	100 x 6/30
💡	5	5/30	100 x 5/30
😱	8	8/30	100 x 8/30

Tableau détaillé

x_i	n_i	f_i	$f_i(\%)$
🎭	4	0,13	13,33
🔫	4	0,13	13,33
❤️	3	0,10	10,00
👽	6	0,20	20,00
😱	5	0,17	16,67
💡	8	0,27	26,67

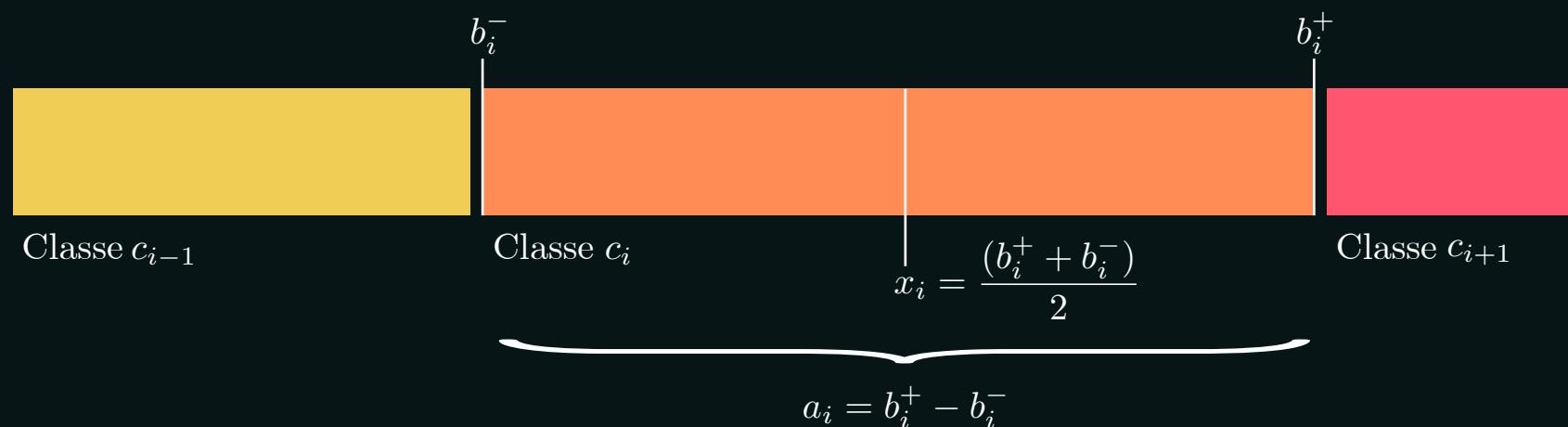


— DÉFINITION —

Comme nous l'avons vu dans l'exemple précédent, il est commode de REGROUER les données en CLASSES (ou modalités). Lorsque notre variable d'intérêt est QUANTITATIVE CONTINUE, les valeurs sont susceptibles de s'étaler sur l'ensemble des NOMBRE RÉELS \mathbb{R} .

Il est alors nécessaire de DÉFINIR nous même les classes, qui engloberont plusieurs observations en leur sein :

- le NOMBRE TOTAL de classe est noté I ,
- les BORNES INFÉRIEURE et SUPÉRIEURE de la classe i sont respectivement notées b_i^- et b_i^+ ,
- l'AMPLITUDE de la classe i est noté $a_i = b_i^+ - b_i^-$,
- le CENTRE de la classe se note $x_i = \frac{b_i^+ + b_i^-}{2}$.





— ENONCÉ —

Dans la même classe de L3, on demande aux élèves de CHRONOMÉTRER leur TEMPS DE VISIONNAGE pendant une semaine. On relève alors les résultats du chronométrage des 30 élèves exprimés EN MINUTES. Voici les résultats :

	464	259	194	380	450	301
	550	443	345	305	266	434
<u>Réponses brutes :</u>	329	98	297	476	213	280
	368	363	397	509	485	393
	443	253	295	262	144	431

On demande des CLASSES avec une AMPLITUDE de 100' ? Réalisez le tableau statistique (effectifs et fréquences).

CLASSES	x_i	n_i	f_i	$f_i(\%)$
[0;100[50	1	0,03	3,33
[100;200[150	2	0,07	6,67
[200-300[250	8	0,27	26,67
[300-400[350	9	0,30	30,00
[400-500[450	8	0,27	26,67
[500;600[550	2	0,07	6,67



— EFFECTIFS ET FRÉQUENCES CUMULÉS —

Calculée pour les variables qualitatives ordinaires ou les variables quantitatives, une grandeur (ici EFFECTIF ou FRÉQUENCE) cumulée d'une valeur observée x_i est la somme des grandeurs correspondant à cette valeur et à l'ensemble de valeurs inférieures. Nous notons N_i l'effectif cumulé et F_i la fréquence cumulée.

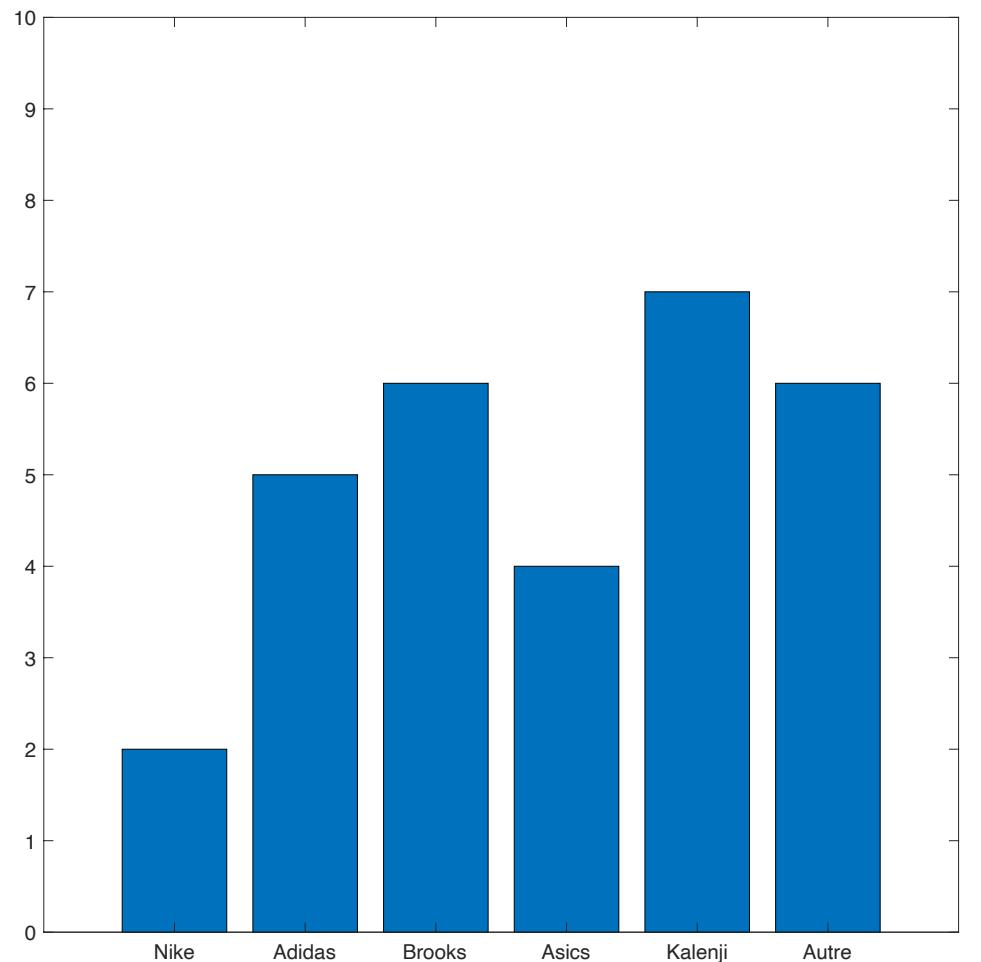
Mathématiquement, on a $N_i = \sum_{j=1}^i n_j$ et $F_i = \sum_{j=1}^i f_j$

CLASSES	x_i	n_i	N_i	f_i	F_i	$f_i(\%)$	$F_i(\%)$
[0;100[50	1	1	0,03	0,03	3,33	3,33
[100;200[150	2	3	0,07	0,10	6,67	10,00
[200-300[250	8	11	0,27	0,37	26,67	36,67
[300-400[350	9	20	0,30	0,67	30,00	66,67
[400-500[450	8	28	0,27	0,93	26,67	93,33
[500;600[550	2	30	0,07	1,00	6,67	100,00

Application à l'énoncé précédent

— REPRÉSENTATIONS GRAPHIQUES —



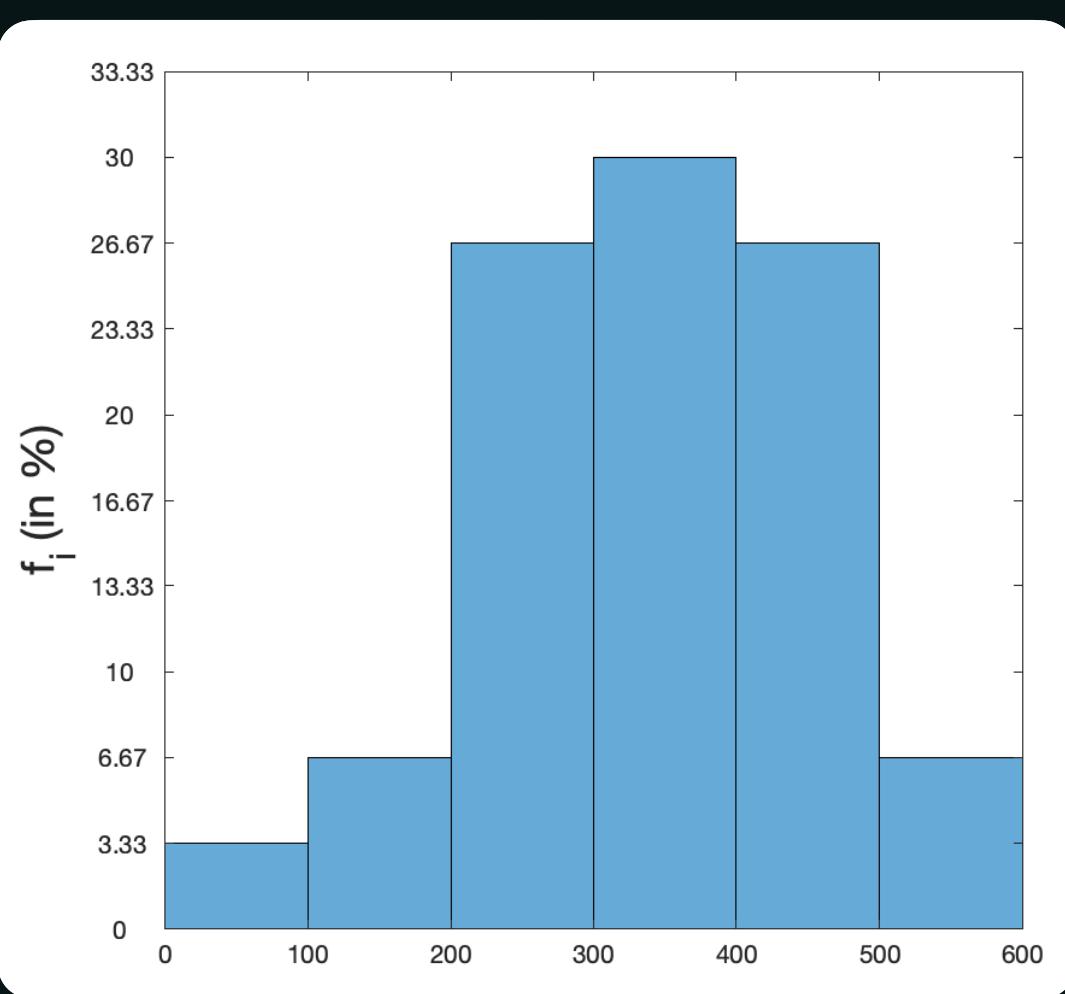


— DIAGRAMMES EN BÂTONS —

Les DIAGRAMMES EN BÂTONS sont particulièrement bien adaptés à la représentation graphique de DISTRIBUTIONS NON GROUPÉES.

Ils sont construits en traçant en regard de chaque valeur observée x_i une barre ayant comme longueur la fréquence f_i de cette valeur.

	n_i	f_i
Nike	2	0,07
Adidas	5	0,17
Brooks	6	0,20
Asics	4	0,13
Kalenji	7	0,23
Autre	6	0,20

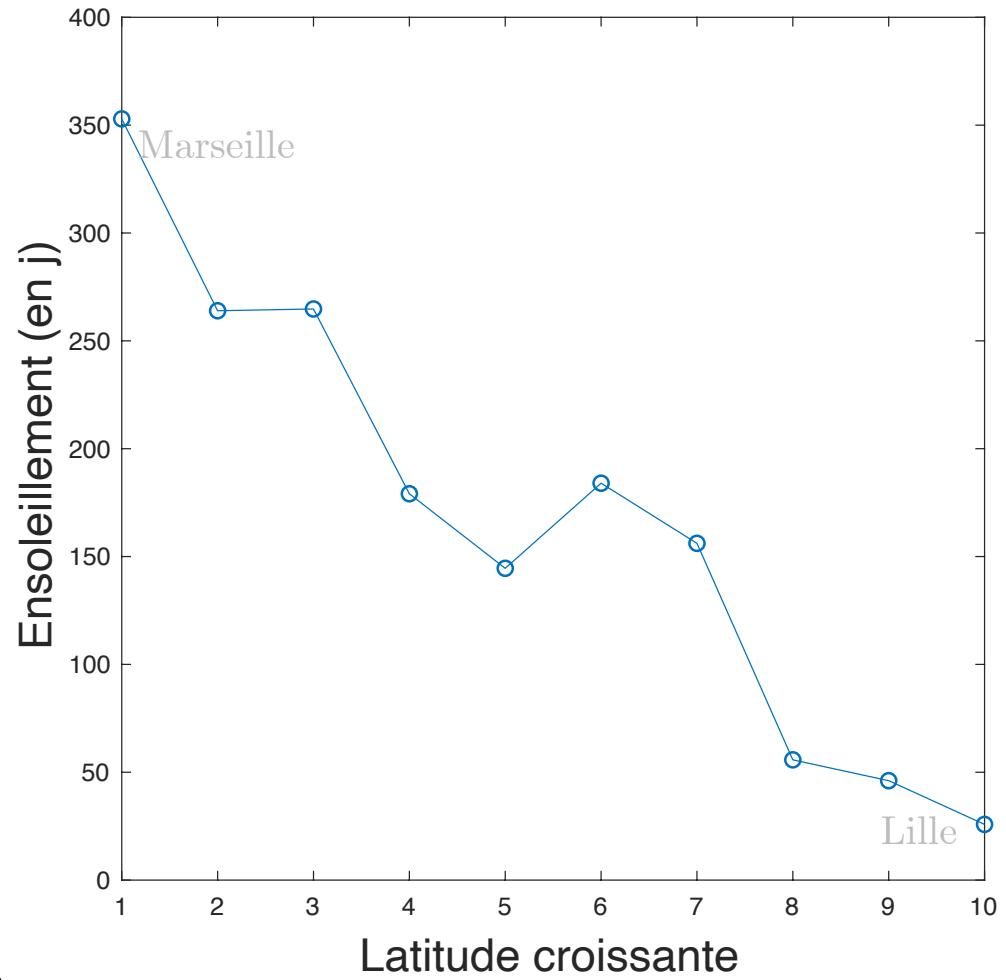


— HISTOGRAMMES —

Les HISTOGRAMMES sont particulièrement bien adaptés à la représentation graphique de DISTRIBUTIONS GROUPÉES.

Ils sont construits en traçant des rectangles contigus ayant comme limites les bornes de chacune des classes et dont la hauteur correspond aux modalités ou fréquences respectives.

CLASSES	x_i	n_i	f_i	$f_i(%)$
[0;100[50	1	0,03	3,33
[100;200[150	2	0,07	6,67
[200-300[250	8	0,27	26,67
[300-400[350	9	0,30	30,00
[400-500[450	8	0,27	26,67
[500;600[550	2	0,07	6,67

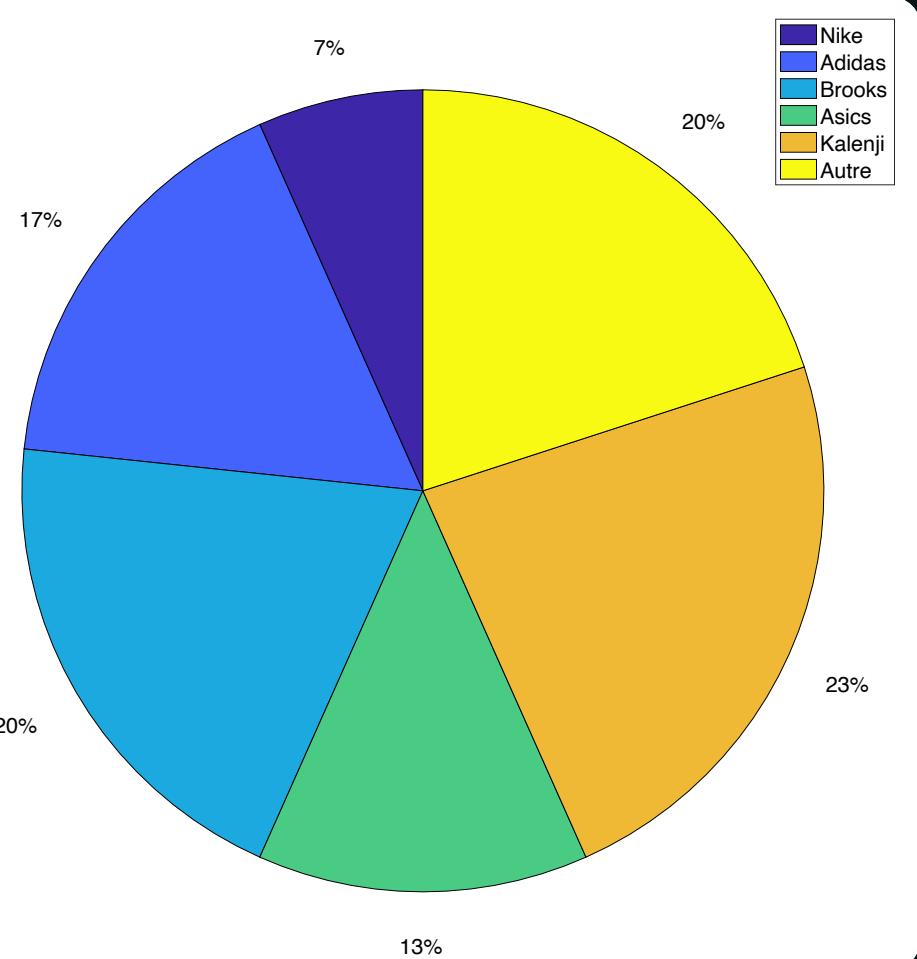


— POLYGONES DE FRÉQUENCES —

Les POLYGONES DE FRÉQUENCES sont une manière de représenter complémentaire aux DIAGRAMMES EN BÂTONS.

Ils sont également appréciés pour des représentations temporelles.

	n_i	$n_i/365$
Marseille	353	0,97
Avignon	264	0,72
Valence	265	0,73
Lyon	179	0,49
Macôn	145	0,40
Bourges	184	0,50
Orléan	156	0,43
Paris	56	0,15
Reims	46	0,13
Lille	26	0,07

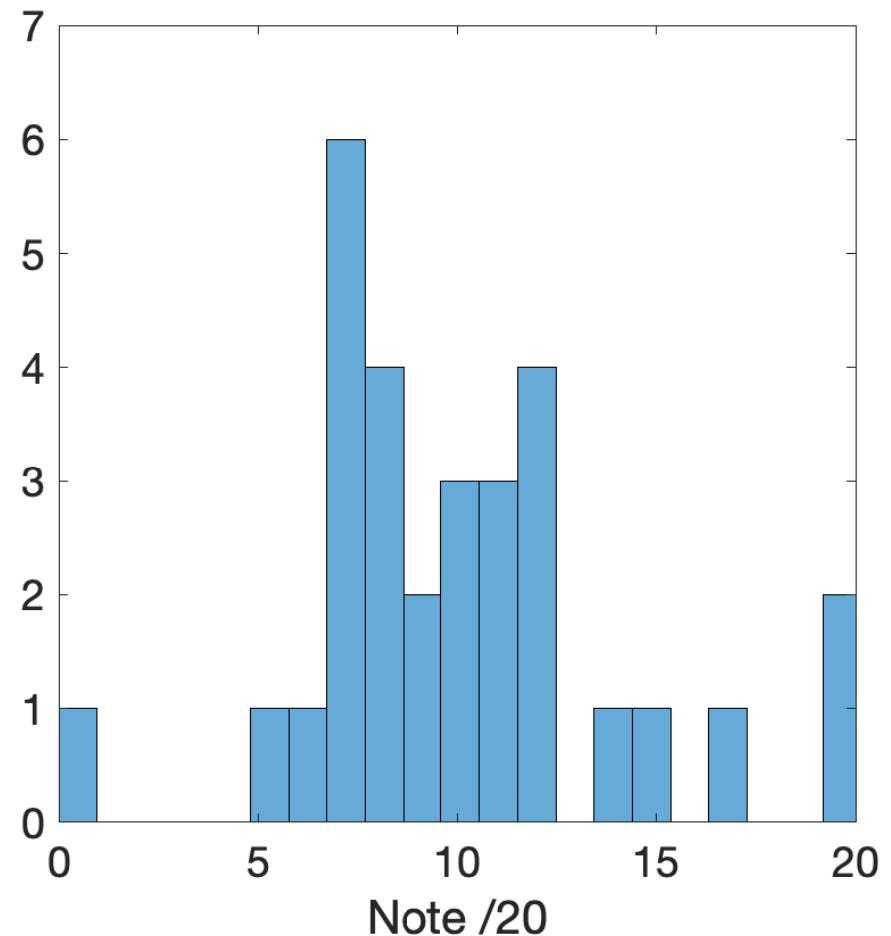


— DIAGRAMMES CIRCULAIRES —

Les DIAGRAMMES CIRCULAIRES sont particulièrement bien adaptés à la représentation graphique de DONNÉES QUALITATIVES .

Ils sont construits en séparant un disque en secteurs dont la surface est proportionnelle à la modalité de chaque classe.

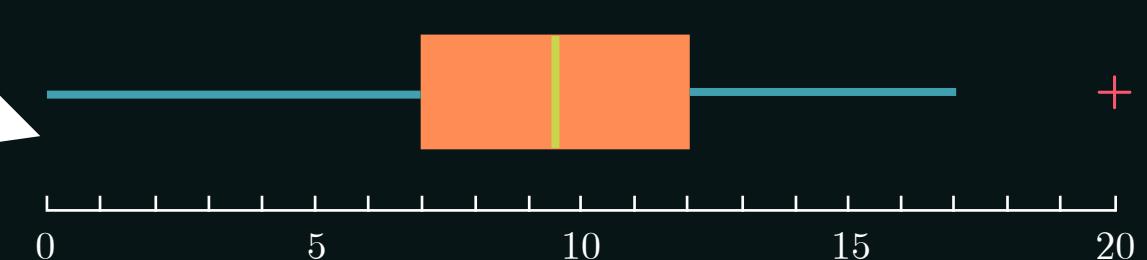
	n_i	f_i
Nike	2	0,07
Adidas	5	0,17
Brooks	6	0,20
Asics	4	0,13
Kalenji	7	0,23
Autre	6	0,20

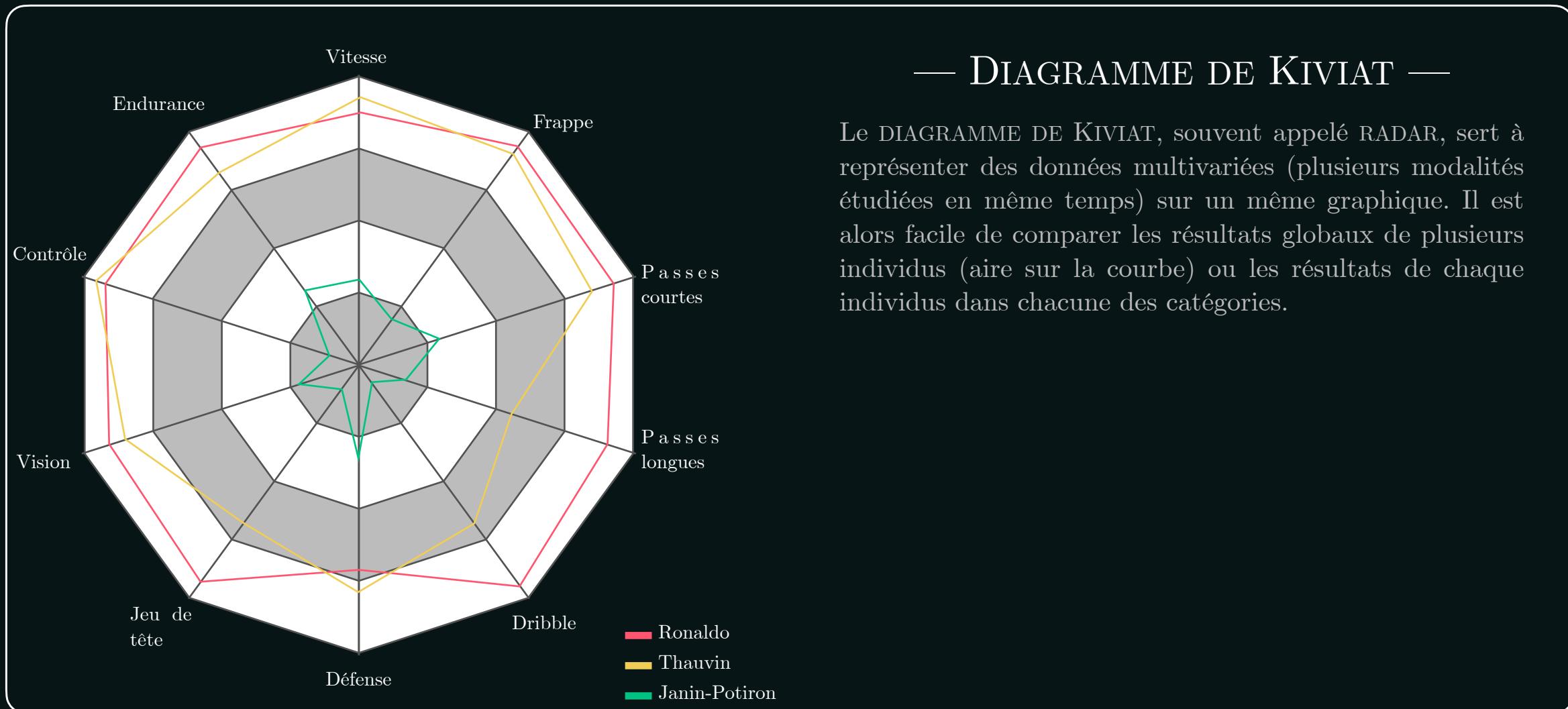


— BOÎTES À MOUSTACHES —

Les BOÎTES À MOUSTACHES permettent de condenser l'information contenues dans des données. Il s'agit de réduction de donnée, et cela est le propos des prochaines planches.

Une boîte à moustache est composée d'un RECTANGLE CENTRAL borné par le PREMIER ET TROISIÈMES QUARTILE. Une BARRE VERTICALE est ajoutée aux coordonnées de la MÉDIANE. Les MOUSTACHES bornent les PREMIER ET NEUVIÈME DÉCILES. Les données HORS DE CES LIMITES sont représentées par des CROIX.





— RÉDUCTION DE DONNÉES —





— RÉDUCTION DE DONNÉES —

Le but de la réduction de données est de CONDENSER L'INFORMATION contenue dans une série statistique afin que celle-ci soit plus INTELLIGIBLE ET AISÉMENT MANIPULABLE par le·a statisticien·ne. On distingue TROIS TYPES DE PARAMÈTRES STATISTIQUES couramment utilisés.

— PARAMÈTRES DE POSITION —

Également appelés VALEURS CENTRALES, ils servent à caractériser l'ordre de grandeur des observations. On compte parmi ces paramètres la MOYENNE — qu'elle soit ARITHMÉTIQUE, GÉOMÉTRIQUE ou HARMONIQUE —, la MÉDIANE, la MÉDIALE ou encore le MODE.

— PARAMÈTRES DE DISPERSION —

Ils permettent de chiffrer la variabilité des observations autour d'un paramètre de position. On compte parmi ces paramètres de dispersion la VARIANCE, l'ÉCART-TYPE, le COEFFICIENT DE VARIATION, l'ÉCART MOYEN ABSOLU, l'ÉCART MÉDIAN, etc.

— PARAMÈTRES DE FORME —

Comme leur nom l'indique, ils permettent de caractériser la forme d'une distribution. On compte parmi ces paramètres les COEFFICIENTS D'ASYMÉTRIE, de KURTOSIS, de FISHER, de PEARSON, etc.

— PARAMÈTRES DE POSITION —



— MOYENNE ARITHMÉTIQUE —

La MOYENNE ARITHMÉTIQUE est la moyenne que « tout le monde » connaît. Généralement notée \bar{x} elle est le résultat de la somme des valeurs observées $x_1, \dots, x_n, \dots, x_N$ divisée par le nombre d'observations N tel que

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n$$

Dans le cas des DISTRIBUTIONS DE FRÉQUENCE, on doit évidemment compter toutes les occurrences d'une valeur x_i . La formule précédente devient alors

$$\bar{x} = \frac{1}{N} \sum_{i=1}^I n_i x_i$$

où I est le nombre de modalité total et i l'indice parcourant toutes les modalités.

— MOYENNE ARITHMÉTIQUE - EXEMPLES —

On mesure en centimètre la distance par rapport au centre d'une cible de 30 fléchettes lancées à l'aveugle. On cherche alors à calculer la distance moyenne \bar{d} .

$$\mathcal{D} = \{-13.35, -2.04, -3.57, -12.34, 20.04, 16.20, -6.06, 14.34, -36.54, 39.41, -14.60, 13.62, -8.32, 34.24, -6.99, N = 30 \\ 4.22, -21.46, -36.74, 5.21, 17.55, 19.56, 22.52, -26.25, 24.65, -24.16, 45.62, -18.48, 19.03, -18.81, 8.83\}$$

$$\bar{d} = \frac{1}{30} \sum_{n=1}^{30} d_n \\ = (1/30)(-13.35 - 2.04 - 3.57 - 12.34 + 20.04 + 16.20 - 6.06 + 14.34 - 36.54 + 39.41 - 14.60 + 13.62 - 8.32 + 34.24 - 6.99 + \\ 4.22 - 21.46 - 36.74 + 5.21 + 17.55 + 19.56 + 22.52 - 26.25 + 24.65 - 24.16 + 45.62 - 18.48 + 19.03 - 18.81 + 8.83) \\ \simeq 1.84 \text{ cm}$$

p_i	n_i	f_i
5,9	3	0,12
6	2	0,08
6,1	3	0,12
6,2	4	0,16
6,3	8	0,32
6,4	2	0,08
6,5	3	0,12

À partir du tableau de fréquence des FTP exprimé en W/kg de 25 cyclistes professionnels masculins participant au Tour de France, calculez la FTP moyenne.

$$\bar{P} = \frac{1}{25} \sum_{i=1}^7 n_i p_i$$

$$\bar{P} = (1/25) \times (3 \times 5,9 + 2 \times 6,0 + \dots + 3 \times 6,5) \simeq 6,2 \text{ W/kg}$$

— MOYENNE GÉOMÉTRIQUE —

La MOYENNE GÉOMÉTRIQUE est utilisée pour les situations dans lesquelles on rencontre des valeurs qui sont destinées à être MULTIPLIÉES entre elles. On pense par exemple au TAUX D'INTÉRÊT d'un investissement financier ou encore au TAUX DE CROISSANCE d'une population.

La moyenne géométrique de N observations s'exprime comme la RACINE N^{IEME} DU PRODUIT DES OBSERVATIONS, soit

$$\bar{x}_g = \left(\prod_{n=1}^N x_n \right)^{1/N} = \sqrt[N]{x_1 x_2 \dots x_N}$$

Dans le cas des DISTRIBUTIONS DE FRÉQUENCE, on doit évidemment compter toutes les occurrences d'une valeur x_i . La formule précédente devient alors

$$\bar{x}_g = \left(\prod_{i=1}^I x_i^{n_i} \right)^{1/N}$$

— MOYENNE GÉOMÉTRIQUE - EXEMPLES —

Je place 1000€ sur un livret bancaire qui rapporte selon les conditions suivantes : 3% les 3 premières années, 4% de les 4 et 5ième années, 5% des années 6 à 8, 2% des années 9 à 12 et finalement 0.1% des années 13 à 20, date à laquelle le compte doit être clôturé. Je souhaite calculer le taux moyen de ce placement ainsi que la somme finale que j'aurai au bout des 20 ans.

y_i	t_i	n_i
1-3	3 %	3
4-5	4 %	2
6-8	5 %	3
9-12	2 %	4
13-20	0.1 %	8

- (1) Construire un tableau contenant les données du problème

- (2) Exprimer la somme d'argent moula($n + 1$) présente sur le compte à l'année $n + 1$ en fonction du taux $t(n)$ exprimé en pourcentage et de la somme moula(n) sur le compte à l'année n .

$$\text{moula}(n + 1) = \text{moula}(n) \times \left(1 + \frac{t(n)}{100}\right)$$

- (3) Calculer le taux moyen sur vingt ans

$$\bar{t}_g = (1,03^3 + 1,04^2 + 1,05^3 + 1,02^4 + 1,001^8)^{1/20} = 1,0202 \text{ soit } 2,02 \%$$

$$\bar{t} = (1,03 \times 3 + 1,04 \times 2 + 1,05 \times 3 + 1,02 \times 4 + 1,001 \times 8) = 1,0282 \text{ soit } 2,82 \%. \text{ On voit clairement que } \bar{t} \neq \bar{t}_g.$$

- (4) Calculer la somme présente sur le compte après les vingt années de placement.

$$\text{moula}(20) = 1000 \times (1,0202)^{20} \simeq 1492,9 \text{ €}$$

— MOYENNE HARMONIQUE —

La MOYENNE HARMONIQUE est généralement utilisée pour des calculs de moyenne mettant en jeu de RAPPORTS DE VALEURS. On pense notamment au calcul d'une vitesse moyenne sur un parcours donné ou d'un Price-Earning Ratio (PER — rapport entre le prix d'un titre boursier et son bénéfice) moyen sur un ensemble de titres boursiers.

La MOYENNE HARMONIQUE de N variables s'exprime comme l'inverse de la moyenne arithmétique des inverses :

$$\bar{x}_h = \left(\frac{\sum_{n=1}^N 1/x_n}{N} \right)^{-1} = \frac{N}{1/x_1 + 1/x_2 + \dots + 1/x_N}$$

Dans le cas des DISTRIBUTIONS DE FRÉQUENCE, on doit évidemment compter toutes les occurrences d'une valeur x_i . La formule précédente devient alors

$$\bar{x}_h = \left(\frac{\sum_{i=1}^I n_i/x_n}{N} \right)^{-1}$$

— MOYENNE HARMONIQUE - EXEMPLES —

Vous effectuez un trajet de 1 km en faisant l'aller à pied, à la vitesse de 3 km/h puis le retour en vélo à la vitesse de 30 km/h.

(1) Calculer la moyenne arithmétique de la vitesse.

$$\bar{v} = (30 + 3)/2 = 16,5 \text{ km/h}$$

(2) Calculer la moyenne harmonique

$$\bar{v}_h = 2/(1/30 + 1/3) \simeq 5,45 \text{ km/h}$$

(3) Assurez-vous de votre réponse en exprimant et calculant la vitesse moyenne harmonique d'une autre manière

$$\bar{v}_h = \frac{\sum_i d_i}{\sum_i t_i} = \frac{\sum_i d_i}{\sum_i d_i/v_i} = \frac{2 \times 1}{1/3 + 1/30}$$

— MOYENNE HARMONIQUE - EXEMPLES —

Lors du Tour de France 2020, nous mesurons le temps de Julian Alaphilippe sur des 5 dernières étapes et reportons cela dans un tableau contenant également les distances respectives des étapes.

- (1) Calculer la vitesse moyenne de Julian sur chacune des cinq dernières étapes

Étape i	d _i	t _i	v _i
17	170	5h 7min 13s	33,20
18	175	5h 19min 5s	32,91
19	166,5	3h 44min 11s	44,56
20	36,2	1h 2min 57s	34,50
21	122	2h 55min 2s	41,82

$$v_i \text{ (en km/h)} = \frac{d_i \text{ (en km)}}{t_i \text{ (en h)}} \text{ et } t_i \text{ (en h)} = h_i + m_i/60 + s_i/3600$$

- (2) Calculer la vitesse moyenne sur ces 5 dernières étapes en utilisant la moyenne harmonique pondérée

$$\bar{v}_h = \frac{5}{\frac{170}{33,20} + \frac{175}{32,91} + \frac{166,5}{44,56} + \frac{36,2}{34,50} + \frac{122}{41,82}} = 36,92$$

- (3) Montrer numériquement que la moyenne harmonique est différente de la moyenne arithmétique

$$\bar{v} = \frac{33,20 + 32,91 + 44,56 + 34,50 + 41,82}{5} = 37,40 \neq 36,92$$

- (4) Vérifier le résultat de la question (2) en utilisant les distances et temps totaux

Étape i	d _i	t _i	v _i
17	170	5h 7min 13s	33,20
...
21	122	2h 55min 2s	41,82
Total	669,7	18:8:28	36,92



— MÉDIANE —

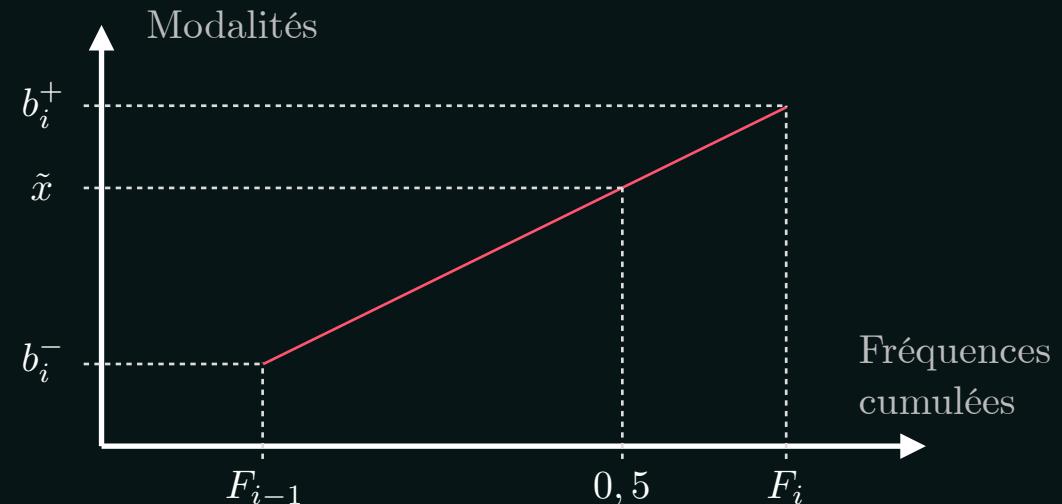
La MÉDIANE, notée \tilde{x} est le paramètre de position tel que la moitié des observations lui sont inférieures ou égales et l'autre moitié supérieures ou égales.

Pour les SÉRIES STATISTIQUES et LES DISTRIBUTIONS NON-GROUPÉES, la médiane possède deux expressions distinctes selon que le nombre d'observations est PAIR ou IMPAIR :

$$\tilde{x} = \frac{x_{N/2} + x_{N/2+1}}{2} \text{ si } N \text{ est pair}$$

$$\tilde{x} = x_{(N+1)/2} \text{ si } N \text{ est impair}$$

Pour les DISTRIBUTIONS GROUPÉES, la CLASSE MÉDIANE est CELLE QUI CONTIENT LA MÉDIANE. Cette CLASSE MÉDIANE est celle pour laquelle la FRÉQUENCE CUMULÉE DÉPASSE 50%. Pour obtenir la valeur médiane, on suppose généralement que les individus sont répartis uniformément et la MÉDIANE est calculée en faisant une INTERPOLATION LINÉAIRE ENTRE LES LIMITES DE LA CLASSE MÉDIANE.



— INTERPOLATION LINÉAIRE ENTRE LES LIMITES DE LA CLASSE MÉDIANE —

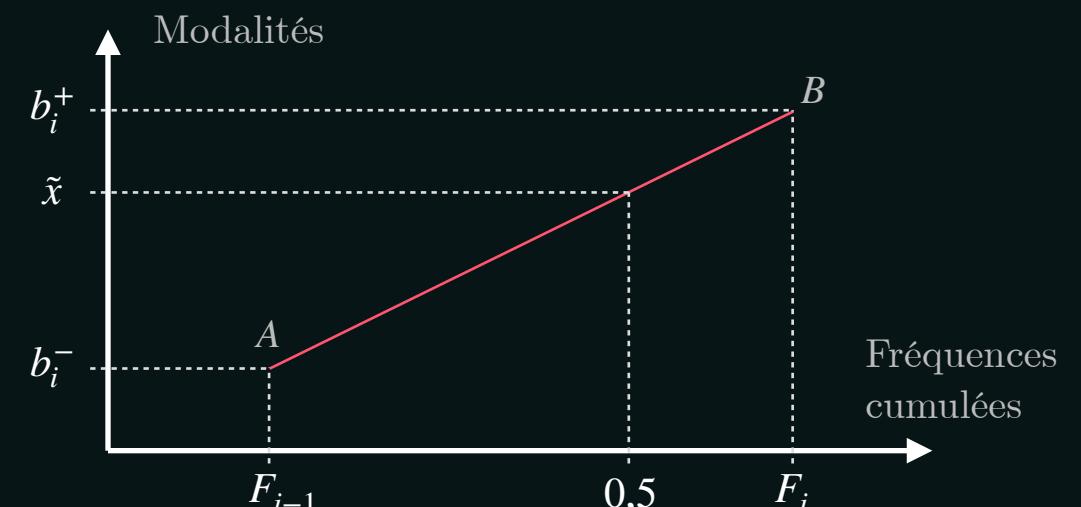
L'interpolation linéaire entre les limites de la classe moyenne ne signifie rien de plus que de trouver l'équation d'une droite de la forme $y = \alpha x + \beta$. Elle s'effectue de la manière suivante :

- (1) On détermine le coefficient directeur α de la droite joignant les bornes de la classe (voir figure, points A et B) dans le plan où l'abscisse représente les fréquences cumulées et l'ordonnée les valeurs des bornes.

$$\alpha = \frac{y_B - y_A}{x_B - x_A} = \frac{b_i^+ - b_i^-}{F_i - F_{i-1}} = \frac{a_i}{F_i - F_{i-1}}$$

- (2) On détermine l'ordonnée du point dont l'abscisse vaut 0,5 (la médiane) :

$$\tilde{x} = b_i^- + \alpha(0,5 - F_{i-1}) = b_i^- + \frac{a_i(0,5 - F_{i-1})}{F_i - F_{i-1}}$$



— MÉDIANE - EXEMPLES —

Après l'examen final de statistique des L3-MS, le professeur s'intéresse aux résultats obtenus par groupe de TD-1 constitué de 30 élèves. Il souhaite calculer la note médiane, aidez-le, il ne sait pas comment faire ...

$$\eta_1 = \{5; 10; 7; 12; 12; 7; 7; 12; 11; 10; 8; 8; 9; 11; 7; 8; 8; 11; 6; 10; 0; 7; 12; 9; 7; 20; 14; 17; 20; 15\}$$

(1) Ordonnez les notes par ordre croissant

$$\eta_1 = \{0; 5; 6; 7; 7; 7; 7; 7; 8; 8; 8; 9; 9; 10; 10; 10; 11; 11; 11; 12; 12; 12; 14; 15; 17; 20; 20\}$$

(2) Calculez la médiane $\tilde{\eta}_1$ du groupe de TD

N est pair ($N = 30$). Les 15 premières valeurs vont de 0 à 9 et les 15 dernières de 10 à 20. La médiane vaut donc
 $\tilde{\eta}_1 = (9 + 10)/2 = 9,5$

Le professeur a oublié un élève lors de son précédent relevé ... Il s'agit de (à compléter selon votre choix)
qui a obtenu une note s'élevant à 3 ... belle perf' !

(1) Insérer la note de dans la série η_2

$$\eta_2 = \{0; 3; 5; 6; 7; 7; 7; 7; 8; 8; 8; 9; 9; 10; 10; 10; 11; 11; 11; 12; 12; 12; 14; 15; 17; 20; 20\}$$

(2) Calculez la nouvelle médiane du groupe de TD

N est maintenant impair ($N = 31$). La médiane est donc égale à la valeur dont l'indice est $(N+1)/2$ soit 16 dans notre cas. On a donc $\tilde{\eta}_2 = 9$

(3) Comparer les valeurs de $\bar{\eta}_1, \bar{\eta}_2, \tilde{\eta}_1, \tilde{\eta}_2$

$$\bar{\eta}_1 = 10 / \tilde{\eta}_1 = 9,5 / \bar{\eta}_2 = 9,7742 / \tilde{\eta}_2 = 9$$

— MÉDIANE - EXEMPLES —

On reprend les mêmes données que lors de l'exercice précédent mais en créant cette fois-ci les classes suivantes : [0,7[, [7,10[, [10,13[, [13,18[, [18,20].

- (1) Construisez le tableau statistique incluant les classes x_i , les amplitudes de classe a_i , les effectifs n_i et fréquences f_i ainsi que les effectifs cumulés N_i et fréquences cumulées F_i .

Note n_i	a_i	n_i	N_i	f_i	F_i
[0,7[7	4	4	0,13	0,13
[7,10[3	12	16	0,39	0,52
[10,13[3	10	26	0,32	0,84
[13,18[5	3	29	0,10	0,94
[18,20]	3	2	31	0,06	1,00

- (2) Indiquez la classe médiane

La classe médiane est la classe pour laquelle on atteint $F_i \geq 50\%$. Dans ce cas, $\tilde{\eta} = [7,10[$.

- (3) Indiquez pour la classe médiane les paramètres suivant : bornes inférieure et supérieure b_i^- et b_i^+ ainsi que les fréquences cumulées F_{i-1} et F_i

$$b_i^- = 7 / b_i^+ = 10 / F_{i-1} = 0,13 / F_i = 0,52$$

- (4) En vous reportant à la [DÉFINITION DE LA MÉDIANE POUR UNE DISTRIBUTION GROUPÉE](#), calculez la médiane de cet exemple

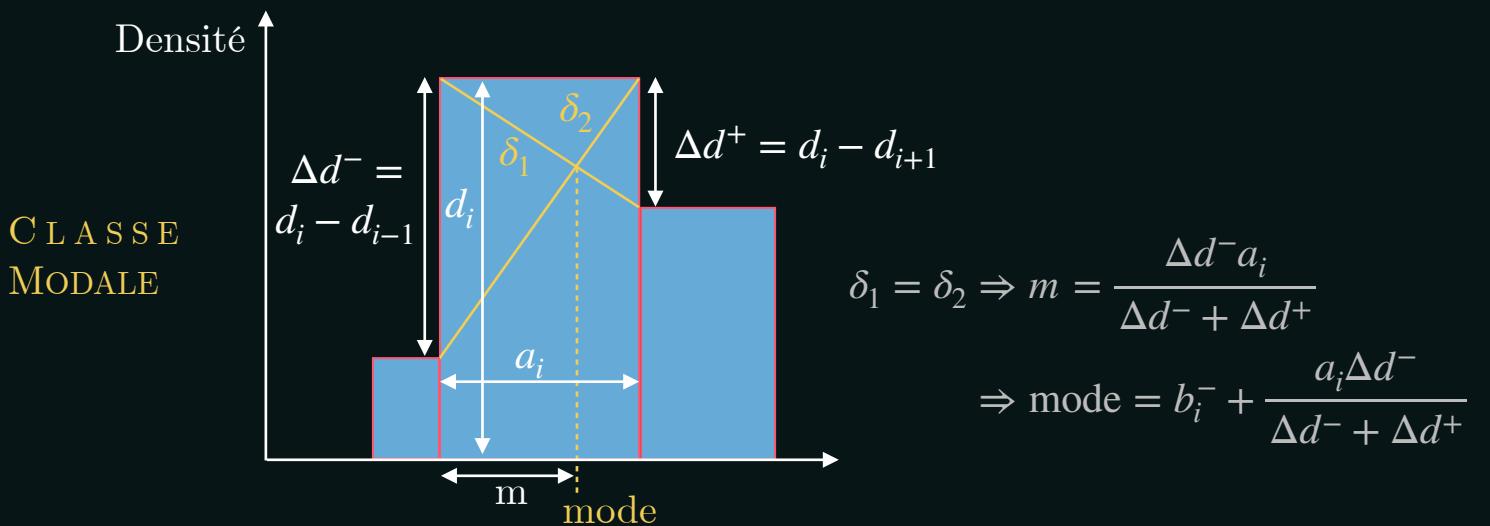
$$\tilde{\eta} = b_i^- + (0,5 - F_{i-1}) \left(\frac{a_i}{F_i - F_{i-1}} \right) = 7 + (0,5 - 0,13)(3/(0,52 - 0,13)) = 9,8462$$

— MODE —

Le MODE d'une distribution non-groupée est la MODALITÉ DE FRÉQUENCE MAXIMALE. Les distributions possédant un seul maximum sont appelées UNIMODALES et opposées aux distributions PLURIMODALES (bimodales, trimodales, etc.)

Pour les distributions groupées, on peut calculer la CLASSE MODALE qui est la CLASSE CONTENANT LE PLUS D'INDIVIDUS. Lorsque les classes ne sont PAS DE MÊMES AMPLITUDES, on se base sur la DENSITÉ DE CLASSE $d_i = f_i/a_i$ pour trouver la classe modale. On peut également calculer le MODE qui est l'INTERSECTION DES DROITES δ_1 ET δ_2 sur le schéma ci-dessous.

Note n_i	a_i	n_i	f_i	d_i
[0,6[6	4	0,13	0,02
[6,10[4	12	0,39	0,10
[10,13[3	10	0,32	0,11
[13,18[5	3	0,10	0,02
[18,20]	3	2	0,06	0,02

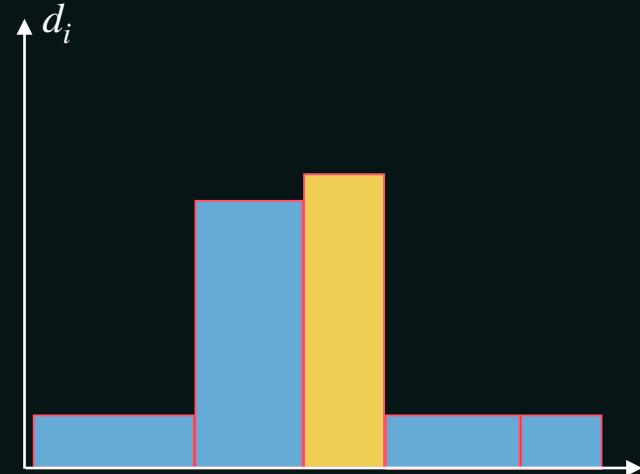
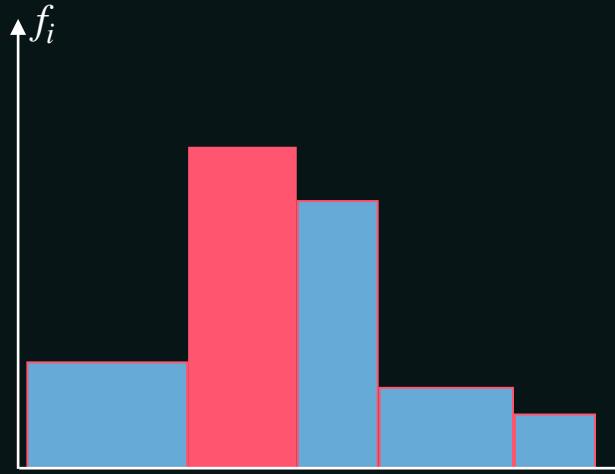


Dans cet exemple (modifié par rapport à l'exemple précédent), la classe modale est [10,13[puisque sa densité est la plus élevée. Le mode vaut 10,3.

— MODE —

Lorsque les classes n'ont pas la même amplitude, on calcule le mode en se basant sur les densités $d_i = f_i/a_i$. On peut alors se trouver dans des cas pour lesquels $f_i > f_j$ mais $d_i < d_j$ si $a_i > a_j$ comme le montre le tableau ci-dessous accompagné des histogrammes correspondant.

Note n_i	a_i	n_i	f_i	d_i
[0,6[6	4	0,13	0,02
[6,10[4	12	0,39	0,10
[10,13[3	10	0,32	0,11
[13,18[5	3	0,10	0,02
[18,20]	3	2	0,06	0,02



— PARAMÈTRES DE DISPERSION —



— EXEMPLE —

On souhaite comparer les résultats des deux groupes de TD de L3-MS dans l'UE Statistiques. On relève les notes des 15 élèves constituant chacun des groupes et on obtient alors :

Groupe 1	2	4	4	5	5	6	7	10	12	14	15	15	17	17	17
-----------------	---	---	---	---	---	---	---	----	----	----	----	----	----	----	----

Groupe 2	8	9	9	9	10	10	10	10	10	10	10	11	12	12
-----------------	---	---	---	---	----	----	----	----	----	----	----	----	----	----

- (1) Calculez la moyenne de chacun des deux groupes

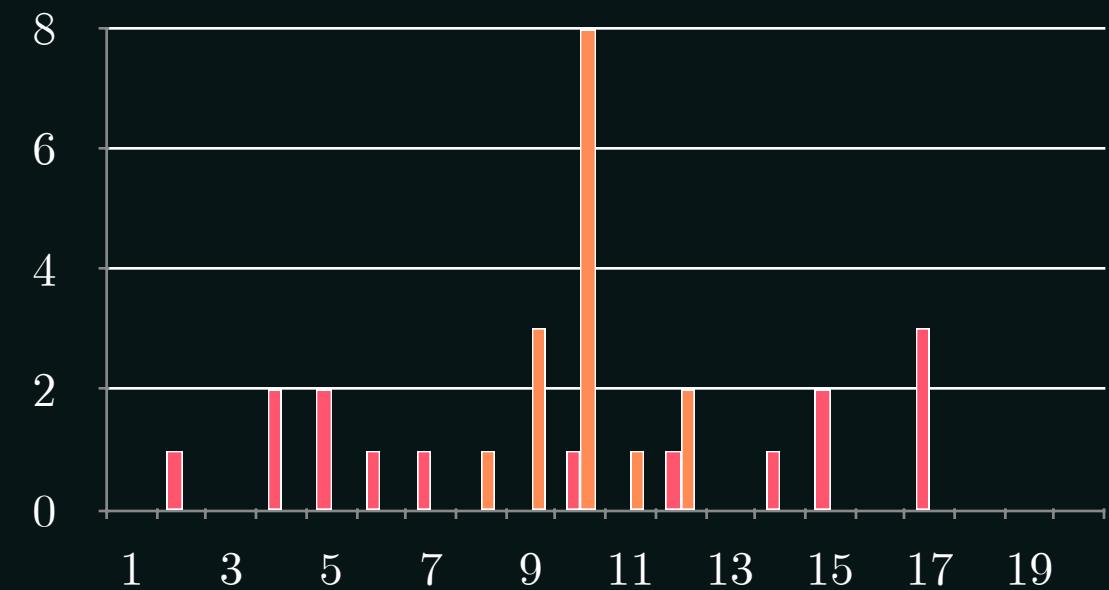
Les deux groupes ont la même moyenne, à savoir 10.

- (2) Calculez la médiane de chacun des deux groupes

Les deux groupes ont la même médiane, à savoir 10, la huitième [$N = 15$ étant impair, $\tilde{x} = (N + 1)/2$] note de la série.

- (3) Diriez-vous que les niveaux de ces deux groupes sont équivalents ? Pour répondre, tracez les deux séries sous forme de diagramme en bâtons.

Les niveaux sont clairement non équivalents. Les notes du groupe 1 sont très étalées alors que celles du groupe 2 sont très resserrées autour de 10.





Comme nous l'avons vu dans l'exemple précédent, les PARAMÈTRE DE POSITION ne suffisent pas à eux seul à analyser entièrement une série statistique. En effet, ils ne donnent pas d'information sur la répartition des valeurs dans la série, ce qui est dans la plupart des cas une information très importante à connaître.

C'est pourquoi nous introduisons les PARAMÈTRES DE DISPERSIONS afin d'aller plus loin dans notre analyse. Ils nous indiquent la manière dont sont dispersées nos séries statistiques, leur propension à s'étendre en quelque sorte.



— L'AMPLITUDE —

L'AMPLITUDE ou ÉTENDUE d'une série statistique S , souvent désignée par le symbole w , est l'écart entre les valeurs extrêmes de cette séries, soit

$$w = \max(S) - \min(S).$$

Si la série $S = \{x_1, \dots, x_N\}$ est rangée par ordre croissant, on a également

$$w = x_N - x_1$$

Groupe 1	2	4	4	5	5	6	7	10	12	14	15	15	17	17	17
Groupe 2	8	9	9	9	10	10	10	10	10	10	10	10	11	12	12

En vous reprenant l'exemple du début du chapitre, calculez l'amplitude des notes des deux groupes de TD.

On a $w_1 = 17 - 2 = 15$ et $w_2 = 12 - 8 = 4$, ce qui confirme notre conclusion précédente.

— LES QUANTILES —

Les QUANTILES sont des paramètres qui divisent la série $S = \{x_1, \dots, x_N\}$ en portions égales. Ils permettent de mesurer la répartition des valeurs dans la série.

Les QUANTILES sont calculés sur le même principe que la MÉDIANE. On cherche à séparer la série en Q morceaux d'effectifs égaux, on parle alors de quantile d'ordre Q (pour la médiane, $Q = 2$) que l'on note x_p où p est généralement une fraction (pour la médiane, $p = 1/2$). On note qu'il existe un nombre $Q - 1$ de quantile d'ordre Q .

Les quantiles d'ordres Q vérifient alors la règle suivante

$x_p = x_{([Np])}$ si Np N'EST PAS UN NOMBRE ENTIER, et où $[y]$ est l'arrondi à la valeur supérieure,

$x_p = x_{Np}$ ou $x_p = \frac{x_{Np} + x_{Np+1}}{2}$ si Np EST UN NOMBRE ENTIER, et ce selon la convention adoptée. Nous choisissons la deuxième solution dans ce cours.

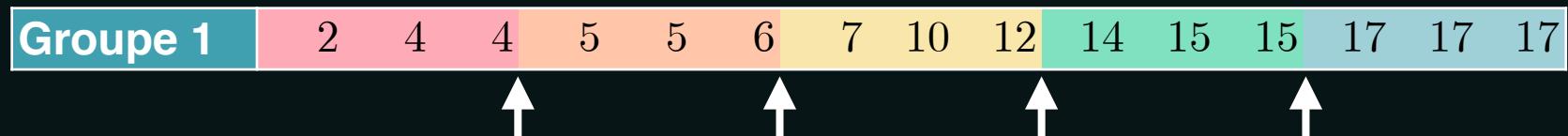
Les quantiles les plus souvent utilisés sont la médiane ($Q = 2$), les quartiles ($Q = 4$), les déciles ($Q = 10$) et le centiles ($Q = 100$)

— EXEMPLE —

Groupe 1	2	4	4	5	5	6	7	10	12	14	15	15	17	17	17
-----------------	---	---	---	---	---	---	---	----	----	----	----	----	----	----	----

(1) En reprenant l'exemple du début du chapitre, calculez les quatre quintiles du groupe 1 ($Q = 5$)

On a $N = 15$ et $Q = 5$. Donc p est de la forme $p = kN/Q = 3k$ et est un entier. Les quatre quintiles sont alors $x_{1/5} = \frac{x_3 + x_4}{2} = 4.5$, $x_{2/5} = \frac{x_6 + x_7}{2} = 6.5$, $x_{3/5} = \frac{x_9 + x_{10}}{2} = 13$, $x_{4/5} = \frac{x_{12} + x_{13}}{2} = 16$



(2) Calculez les trois quartiles ($Q = 4$)

On a $N = 15$ et $Q = 4$. Donc p est de la forme $p = kN/Q = 15k/4$ qui n'est pas entier pour $k \in [1,3]$. Les quatre quintiles sont alors $x_{1/4} = x_{[15/4]} = x_{[3.75]} = x_4 = 5$, $x_{1/2} = x_{[15/2]} = x_{[7.5]} = x_8 = 10$, $x_{3/4} = x_{[45/4]} = x_{[11.25]} = x_{12} = 15$.



— LES QUANTILES POUR LES DISTRIBUTIONS CONTINUES —

Reprenez l'[EXEMPLE](#) du temps de visionnage en minute d'une classe de 30 élèves et calculez les quartiles de cette distribution continue.

CLASSES	x_i	n_i	N_i	f_i	F_i	$f_i(\%)$	$F_i(\%)$
[0;100[50	1	1	0,03	0,03	3,33	3,33
[100;200[150	2	3	0,07	0,10	6,67	10,00
[200-300[250	8	11	0,27	0,37	26,67	36,67
[300-400[350	9	20	0,30	0,67	30,00	66,67
[400-500[450	8	28	0,27	0,93	26,67	93,33
[500;600[550	2	30	0,07	1,00	6,67	100,00



Les trois quartiles $x_{1/4}$, $x_{1/2}$ et $x_{3/4}$ séparent les données aux points où les fréquences cumulées valent respectivement 25%, 50% et 75%. Comme pour le calcul de la médiane dans ces conditions (voir [INTERPOLATION LINÉAIRE ENTRE LES LIMITES DE LA CLASSE MÉDIANE](#)), on utilise l'interpolation linéaire pour calculer la place des quartiles. On a alors

$$x_{1/4} = b_i^- + \frac{a_i(0,25 - F_{i-1})}{F_i - F_{i-1}} = 200 + 100 \times \frac{0,25 - 0,10}{0,37 - 0,10} \simeq 255,6$$

$$x_{1/2} = 300 + 100 \times \frac{0,5 - 0,37}{0,67 - 0,37} \simeq 343,3$$

$$x_{3/4} = 400 + 100 \times \frac{0,75 - 0,67}{0,93 - 0,67} \simeq 430,8$$

— ÉCART INTERQUARTILE —

L'écart interquartile est la distance entre le premier et le troisième quartile. On le note soit sous la forme d'un intervalle

$$[x_{1/4}; x_{3/4}]$$

soit sous la forme d'un scalaire (un nombre) défini comme la différence entre le premier et le troisième quartile

$$\text{IQR} = x_{3/4} - x_{1/4} = Q_3 - Q_1$$

Dans la même optique, on parle d'écart interdécile lorsque l'on calcule la distance entre le premier et le neuvième décile

$$\text{IDR} = x_{9/10} - x_{1/10} = D_9 - D_1$$

Groupe 1	2	4	4	5	5	6	7	10	12	14	15	15	17	17	17
Groupe 2	8	9	9	9	10	10	10	10	10	10	10	10	11	12	12

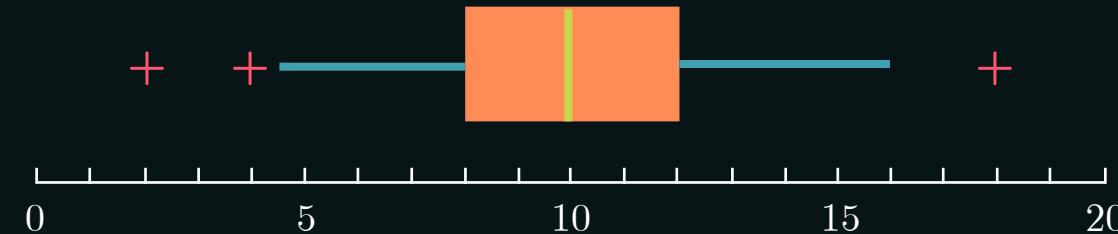
En reprenant l'exemple du début du chapitre, calculez l'écart interquartile pour les deux groupes de TD.

Pour le groupe 1, on a $Q_1^{G1} = 5$ et $Q_3^{G1} = 15$, ce qui implique $\text{IQR}^{G1} = Q_3^{G1} - Q_1^{G1} = 15 - 5 = 10$. Pour le groupe 2, on a $Q_1^{G2} = 9$ et $Q_3^{G2} = 10$, ce qui implique $\text{IQR}^{G2} = Q_3^{G2} - Q_1^{G2} = 10 - 9 = 1$. Cela confirme les conclusions précédentes.

— BOÎTES À MOUSTACHE —

Les BOÎTES À MOUSTACHES permettent de condenser l'information contenues dans des données. Il s'agit de réduction de donnée, et cela est le propos des prochaines planches.

Une boîte à moustache est composée d'un RECTANGLE CENTRAL borné par le PREMIER ET TROISIÈMES QUARTILE. Une BARRE VERTICALE est ajoutée aux coordonnées de la MÉDIANE. Les données HORS DE CES LIMITES sont représentées par des CROIX. Il existe plusieurs convention pour les MOUSTACHES. Elle peuvent être les bornes des (1) PREMIER ET NEUVIÈME DÉCILES, des (2) PREMIER ET 99ÈME CENTILES, ou encore être limitées à (3) UNE FOIS ET DEMI L'ÉCART INTERQUARTILE — $1,5 \times IQR$.



— EXEMPLE —

G1+G2

1 4 4 5 5 6 7 8 9 9 10 10 10 10 10 10 10 11 12 12 12 14 15 15 17 17 17

- 1) On prend en compte les deux groupes de TD dans l'exemple du début du chapitre pour ne former qu'une seule série. Un élève a vu sa note baisser pour rendre l'exercice plus amusant. Calculez alors la médiane

$$\text{On a } N = 30, \text{ donc } x_{1/2} = \frac{x_{15} + x_{16}}{2} = 10$$

- 2) Calculez les premier et troisième quartiles ainsi que l'écart interquartile

$$x_{1/4} = 8 \text{ et } x_{3/4} = 12, \text{ ce qui donne IQR} = 4$$

- 3) Calculez les premier et neuvième déciles

$$x_{1/10} = 4,5 \text{ et } x_{9/10} = 16$$

- 4) Quelles sont les valeurs hors des limites des premier et neuvième déciles ? Quelle sont les valeurs hors de l'intervalle $[Q_1 - 1,5 \times \text{IQR}; Q_3 + 1,5 \times \text{IQR}]$?

Valeurs hors de $[D_1; D_9]$: 1,4,4,17,17,17. Valeurs hors de $[Q_1 - 1,5 \times \text{IQR}; Q_3 + 1,5 \times \text{IQR}] = [2; 18]$: 1

- 5) Tracer la boîte à moustache de cette distribution avec des moustaches aux premier et neuvième déciles (1) puis changez les moustaches avec la convention $1,5 \times \text{IQR}$ (2)



— ÉCART ABSOLU MOYEN —

L'ÉCART ABSOLU MOYEN, noté généralement e_m , quantifie la fluctuation de la série statistique par rapport à la moyenne et est égal à la moyenne arithmétique des valeurs absolues des écarts par rapport à la moyenne soit

$$e_m = \frac{1}{N} \sum_{n=1}^N |x_n - \bar{x}|$$

Dans le cas des distributions de fréquence, la formule devient naturellement

$$e_m = \frac{1}{N} \sum_{i=1}^I [n_i |x_i - \bar{x}|]$$

On peut éventuellement remplacer la moyenne \bar{x} par la médiane \tilde{x} . Cela est précisé lorsque c'est le cas.

CLASSES	a_i	x_i	n_i	f_i	$x_i f_i$	$n_i x_i - \bar{x} $	$f_i x_i - \bar{x} $
[0;100[100	50	1	0,03	1,67	290,00	9,67
[100;200[100	150	2	0,07	10,00	380,00	12,67
[200-300[100	250	8	0,27	66,67	720,00	24,00
[300-400[100	350	9	0,30	105,00	90,00	3,00
[400-500[100	450	8	0,27	120,00	880,00	29,33
[500;600[100	550	2	0,07	36,67	420,00	14,00
			30	1,00	340,00	92,67	92,67

La moyenne de la série :
 $\sum x_i f_i$

L'écart absolu moyen de la
 série : $\frac{1}{30} \sum n_i |x_i - \bar{x}|$ ou
 $\sum f_i |x_i - \bar{x}|$

— ÉCART MÉDIAN —

Sur le même principe que l'écart absolu moyen, l'ÉCART MÉDIAN, que nous noterons e_{med} , est égal à la médiane des valeurs absolues des écarts par rapport à la moyenne soit

$$e_{med} = \text{Me}(|x_n - \bar{x}|) \text{ ou Me est la médiane}$$

On peut éventuellement remplacer la moyenne \bar{x} par la médiane \tilde{x} . Cela est précisé lorsque c'est le cas.

CLASSES	a_i	x_i	n_i	f_i	$x_i f_i$	$ x_i - \bar{x} $
[0;100[100	50	1	0,03	1,67	290,00
[100;200[100	150	2	0,07	10,00	190,00
[200-300[100	250	8	0,27	66,67	90,00
[300-400[100	350	9	0,30	105,00	10,00
[400-500[100	450	8	0,27	120,00	110,00
[500;600[100	550	2	0,07	36,67	210,00
			30	1,00	340,00	

En prenant l'exemple ci-dessus, calculez l'écart médian de la série statistique regroupant les heures de visionnage Netflix de 30 élèves.

La série des écarts à la moyenne est [10 10 10 10 10 10 10 10 10 10 90 90 90 90 90 90 90 90 110 110 110 110 110 110 110 110 190 190 210 210 290]. Sa médiane est la moyenne entre le 15ème et le 16ème terme, soit 90.

— VARIANCE, ÉCART-TYPE ET COEFFICIENT DE VARIATION —

La VARIANCE d'une série statistique ou d'une distribution de fréquence, notée généralement v ou s^2 , est la moyenne arithmétique des carrés des écarts par rapport à la moyenne. Elle s'exprime respectivement

$$s^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \text{ pour les séries statistiques et } s^2 = \frac{1}{N} \sum_{i=1}^I [n_i(x_i - \bar{x})^2] \text{ pour les distributions de fréquence}$$

L'ÉCART-TYPE, également appelé écart quadratique moyen ou déviation standard (*standard deviation* en anglais) et la racine carrée de la variance. Elle s'exprime comme

$$s = \sqrt{s^2} = \sqrt{v}$$

Le COEFFICIENT DE VARIATION, notée CV, est le rapport entre l'écart-type et la moyenne, soit

$$CV = \frac{s}{\bar{x}}$$

— EXEMPLE —

En reprenant le même exemple que précédemment, calculez l'écart-type et le coefficient de variation.

CLASSES	a_i	x_i	n_i	f_i	$x_i f_i$	$f_i(x_i - \bar{x})^2$
[0;100[100	50	1	0,03	1,67	2803,33
[100;200[100	150	2	0,07	10,00	2406,67
[200-300[100	250	8	0,27	66,67	2160,00
[300-400[100	350	9	0,30	105,00	30,00
[400-500[100	450	8	0,27	120,00	3226,67
[500;600[100	550	2	0,07	36,67	2940,00
			30	1,00	340,00	13566,67

On calcule $f_i(x_i - \bar{x})^2$ et on somme sur les 6 classes. On obtient alors la variance $v \simeq 13566,7$, ce qui donne un écart-type $s = 116,5$.

Le coefficient de variation vaut alors $CV = s/\bar{x} \simeq 116,5/340 \simeq 0,34$.

ATTENTION : un logiciel de calcul donnerait une déviation standard environ égale à 118,5. Cela est dû à la méthode de calcul employée qui remplace l'effectif N par la valeur $N - 1$ afin de produire un estimateur non-biaisé. Cette considération est hors des limites de ce cours cependant.

— PARAMÈTRES DE CONCENTRATION —



⚠ La mesure de la concentration concerne les caractères statistiques quantitatifs représentant une valeur positive cumulable.
Il s'agit d'étudier la densité des données autour de la valeur centrale.

— MASSES —

Étant donné une série statistique comportant N observations ordonnées dans un tableau statistique (x_i, n_i) , représentant I modalités, on appelle MASSE associée à la modalité x_i d'effectif n_i , la quantité définie par

$$x_i n_i,$$

et MASSE RELATIVE associée à la modalité x_i notée q_i , la quantité définie par

$$q_i = \frac{x_i n_i}{\sum_{k=1}^I n_k x_k}.$$

Il est également utile de définir la MASSE RELATIVE CUMULÉE comme

$$q_i^{cc} = \sum_{k=1}^i q_k.$$

— MÉDIALE —

La MÉDIALE, noté Ml , est la valeur tel que la somme des observations qui lui sont inférieures et la somme des observations qui lui sont supérieures sont égales. Elle sépare en deux parties égales la masse totale de la variable. Comme pour le calcul de la médiane, la médiale est la valeur de la modalité pour laquelle la MASSE RELATIVE CUMULÉE vaut 50%.

Note n_i	x_i	n_i	f_i	$F_i (%)$	$x_i n_i$	q_i	$q_i^{cc} (%)$
[0,4[2,00	1,00	0,03	0,03	2,00	0,01	0,01
[4,8[6,00	6,00	0,20	0,23	36,00	0,11	0,12
[8,12[10,00	14,00	0,47	0,70	140,00	0,44	0,56
[12,16[14,00	6,00	0,20	0,90	84,00	0,27	0,83
[16,20]	18,00	3,00	0,10	1,00	54,00	0,17	1,00
	30,00	1,00			316,00	1,00	

$$\tilde{x} = 8 + (0,5 - 0,23)(4/(0,70 - 0,23)) \simeq 10,30$$

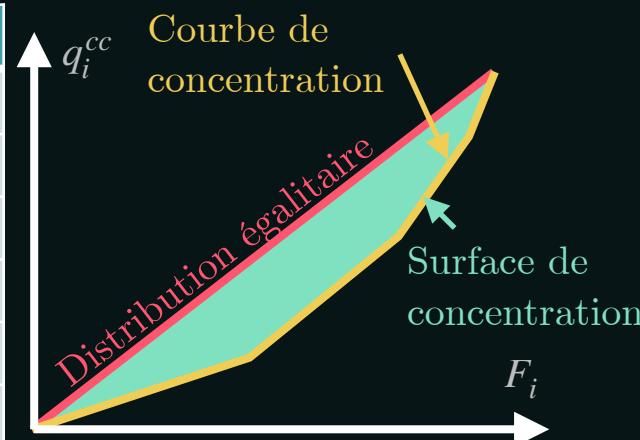
$$Ml = 8 + (0,5 - 0,12)(4/(0,56 - 0,12)) \simeq 11,45$$

$$Ml \neq \tilde{x}$$

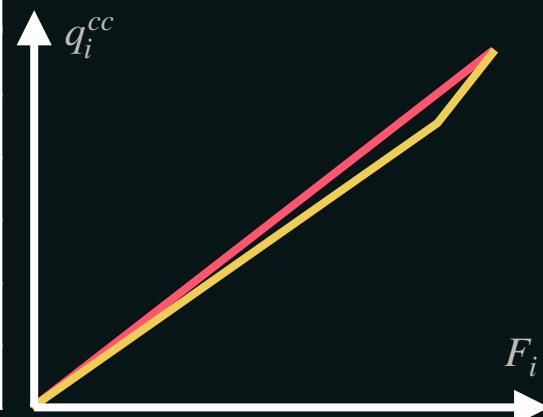
— COURBE DE CONCENTRATION —

La COURBE DE CONCENTRATION se trace sur un graphique à deux dimension en reportant les valeurs des masses relatives cumulées q_i^{cc} en fonction des valeurs des fréquences cumulées F_i , soit la courbe $q_i^{cc} = f(F_i)$.

Note n_i	x_i	n_i	$f_i (\%)$	$F_i (\%)$	$x_i n_i$	$q_i (\%)$	$q_i^{cc} (\%)$
[0,4[2,00	1,00	6,67	6,67	2,00	1,27	1,27
[4,8[6,00	6,00	40,00	46,67	36,00	22,78	24,05
[8,12[10,00	1,00	6,67	53,33	10,00	6,33	30,38
[12,16[14,00	4,00	26,67	80,00	56,00	35,44	65,82
[16,20]	18,00	3,00	20,00	100,00	54,00	34,18	100,00
	15,00	100,00			158,00	100,00	



Note n_i	x_i	n_i	$f_i (\%)$	$F_i (\%)$	$x_i n_i$	$q_i (\%)$	$q_i^{cc} (\%)$
[0,4[2,00	0,00	0,00	0,00	0,00	0,00	0,00
[4,8[6,00	0,00	0,00	0,00	0,00	0,00	0,00
[8,12[10,00	13,00	86,67	86,67	130,00	82,28	82,28
[12,16[14,00	2,00	13,33	100,00	28,00	17,72	100,00
[16,20]	18,00	0,00	0,00	100,00	0,00	0,00	100,00
	15,00	100,00			158,00	100,00	



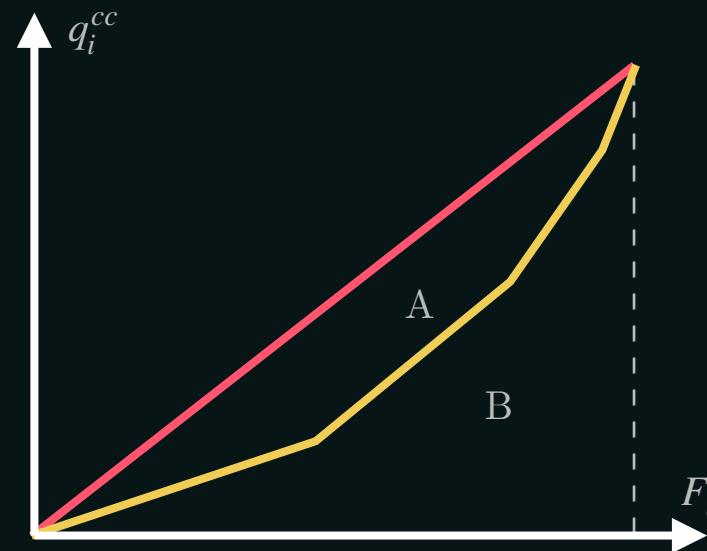
La distribution égalitaire signifie dans notre cas : tout le monde a la même note.

— INDICE DE GINI —

L'indice de Gini, noté I_G est le rapport entre l'aire de la surface de concentration et l'aire de la surface du triangle rectangle délimité par la distribution théorique et l'axe des abscisses. De manière équivalente, l'indice de Gini vaut également

$$I_G = 2A = 1 - 2B,$$

où A est la surface de concentration et B est l'aire sous la courbe de concentration.



L'indice de Gini vaut 0 pour une distribution totalement égalitaire (tout le monde reçoit la même note) et 1 pour une distribution totalement inégalitaire (tout le monde reçoit 0 sauf une personne). Cet exemple se comprend mieux lorsque l'on partage une ressource finie comme par exemple une somme d'argent.

— PARAMÈTRES DE FORME —



— EXEMPLE —

On souhaite comparer les résultats d'une année à l'autre entre deux promos de L3-MS dans l'UE Statistiques. On relève les notes des 90 élèves constituant chaque promo et on obtient alors :

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Promo 1	0	0	0	0	0	2	4	6	11	16	15	13	10	7	3	2	1	0	0	0	0
Promo 2	0	0	0	0	1	2	3	7	10	13	15	16	11	6	4	2	0	0	0	0	0

- (1) Calculez la moyenne de chacune des deux promos

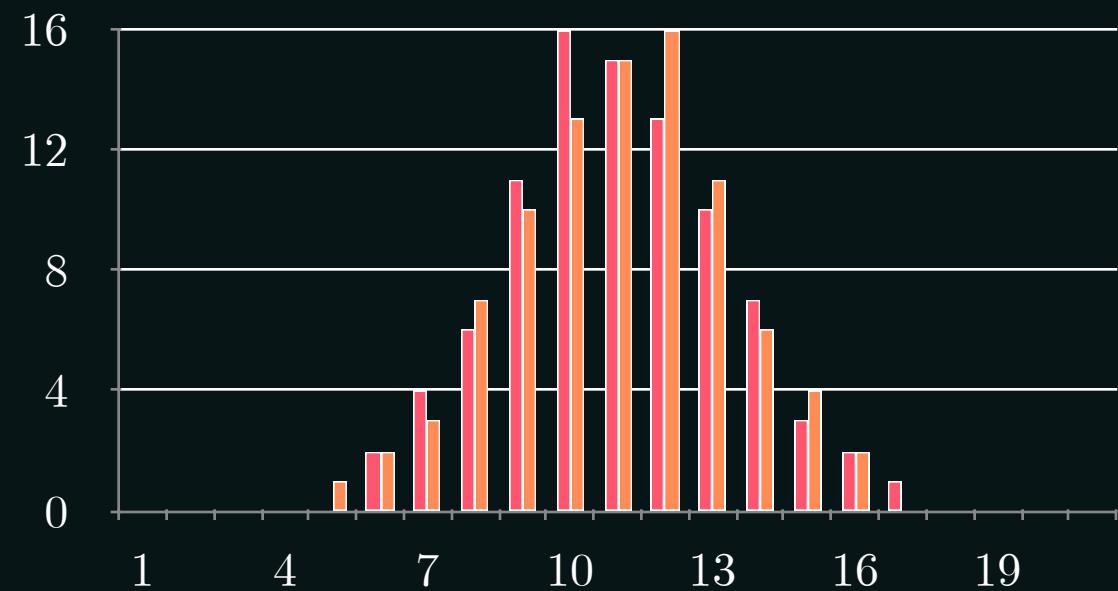
Les deux groupes ont la même moyenne, à savoir 10.

- (2) Calculez l'écart-type de chacune des deux promos

Les deux groupes ont le même écart-type qui vaut $\sigma \simeq 2,30$.

- (3) Diriez-vous que les niveaux de ces deux promos sont équivalents ? Pour répondre, tracez les deux séries sous forme de diagramme en bâtons.

Les niveaux sont clairement non équivalents. Les notes du groupe 1 tirent plus vers les notes faibles alors que celles du groupe 2 tirent plus vers les notes hautes.



— MOMENTS —

Les MOMENTS d'ordre k par rapport au point c sont une généralisation mathématiques de concept que nous avons vus précédemment dans ce cours : la moyenne (moment d'ordre $k = 1$ au point $c = 0$) et la variance (moment d'ordre $k = 2$ au point $c = \bar{x}$). Le moment d'ordre k par rapport au point c s'exprime comme :

$$\frac{1}{N} \sum_{n=1}^N (x_n - c)^k \text{ pour les séries statistiques et } \frac{1}{N} \sum_{i=1}^I n_i(x_i - c)^k \text{ pour les distributions de fréquence.}$$

En pratique, on utilise généralement $c = 0$, que l'on appelle les moments non centrés :

$$a_k = \frac{1}{N} \sum_{n=1}^N x_n^k \text{ ou } a_k = \frac{1}{N} \sum_{i=1}^I n_i x_i^k ,$$

ou $c = \bar{x}$, que l'on appelle les moments centrés :

$$m_k = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^k \text{ ou } m_k = \frac{1}{N} \sum_{i=1}^I n_i (x_i - \bar{x})^k.$$

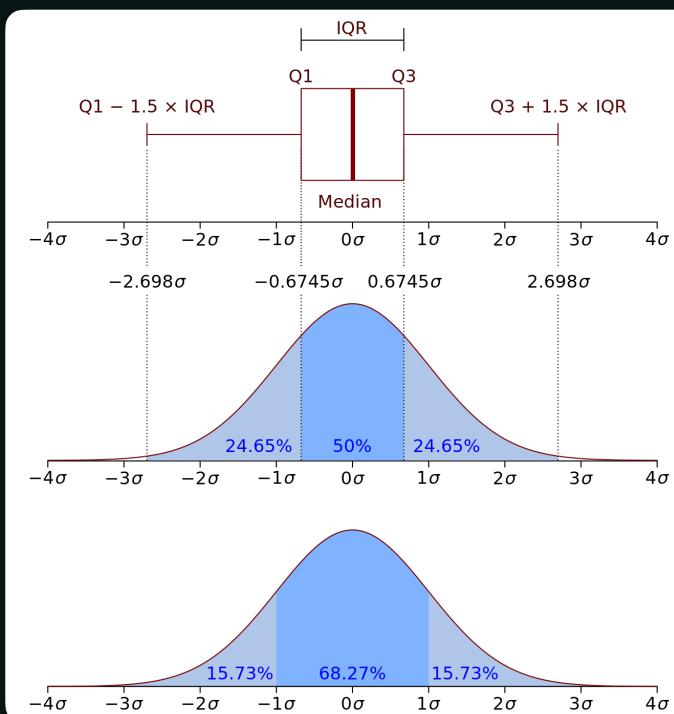
On retrouve alors que le moment non centré d'ordre 1 équivaut à la moyenne ($a_1 = \bar{x}$) et que le moment centré d'ordre 2 équivaut à la variance ($m_2 = s^2$).



— DISTRIBUTION OU LOI NORMALE —



La LOI NORMALE, appelée également LOI GAUSSIENNE, est un modèle de distribution continue. Elle se rencontre dans de nombreux cas en statistique et fait partie des LOIS DE PROBABILITÉ LES PLUS UTILISÉES pour modéliser des phénomènes naturels issus de plusieurs événements aléatoires. De nombreux caractères suivent une loi normale, parmi lesquels : la taille des êtres humains, leur poids, leur pressions artérielles, leur taux de cholestérol, etc.

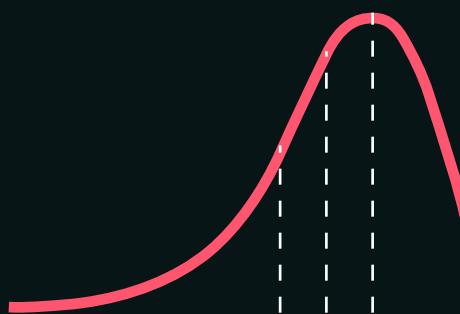


La REPRÉSENTATION GRAPHIQUE d'une loi normale est une COURBE EN CLOCHE, appelée plus communément GAUSSIENNE. La loi normale étant omniprésente en statistique, il est habituel de lui comparer les données qui nous intéressent. Nous aborderons dans ce cours deux paramètres :

- celui d'ASYMÉTRIE — la loi normale est symétrique par rapport à sa moyenne, on la compare ainsi à nos données,
- celui d'APLATISSEMENT — la loi normale possède un aplatissement pris comme référence, on compare l'aplatissement de nos données par rapport à celle-ci.

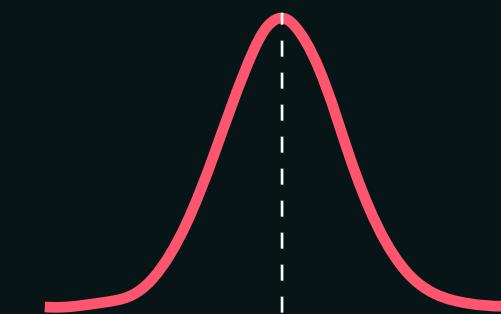
Fun fact : on retrouve par exemple la distribution normale lorsque l'on regarde la proportion de face obtenu à pile ou face sur un grand nombre de lancer et un grand nombre d'essai. On la retrouve aussi lorsque l'on additionne les valeurs obtenues lors d'un grand nombre de lancers de dés. Voir également la [PLANCHE DE GALTON](#).

— COMPARAISON À LA SYMÉTRIE DE LA LOI NORMALE —



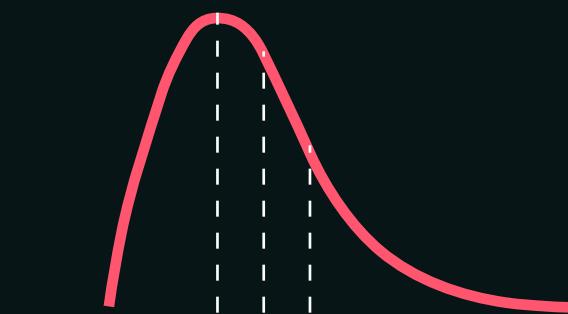
Moyenne < Médiane < Mode

ASYMÉTRIE NÉGATIVE —
DISTRIBUTION OBLIQUE À
DROITE —



Moyenne = Médiane = Mode

COURBE NORMALE



Mode < Médiane < Moyenne

ASYMÉTRIE POSITIVE —
DISTRIBUTION OBLIQUE À
GAUCHE —

— EXEMPLE —

On observe encore les notes obtenues par trois groupes d'élèves composés chacun de 37 individus. Voici les tableaux statistiques associés. Leurs distributions sont-elles symétriques \Leftrightarrow comment sont ordonnés leurs modes, moyennes et médianes ?

x_i	n_i	f_i	F_i	$x_i f_i$
7	3	0,08	0,08	0,57
8	5	0,14	0,22	1,08
9	6	0,16	0,38	1,46
10	9	0,24	0,62	2,43
11	6	0,16	0,78	1,78
12	5	0,14	0,92	1,62
13	3	0,08	1,00	1,05
37	1,00			10,00

La moyenne vaut $\bar{x} = 10$, la médiane vaut $\tilde{x} = 10$ et le mode vaut $x_m = 10$.

La distribution est bien symétrique car la moyenne, la médiane et le mode sont égaux.

x_i	n_i	f_i	F_i	$x_i f_i$
7	1	0,03	0,03	0,19
8	1	0,03	0,05	0,22
9	3	0,08	0,14	0,73
10	4	0,11	0,24	1,08
11	7	0,19	0,43	2,08
12	10	0,27	0,70	3,24
13	11	0,30	1,00	3,86
37	1,00			11,41

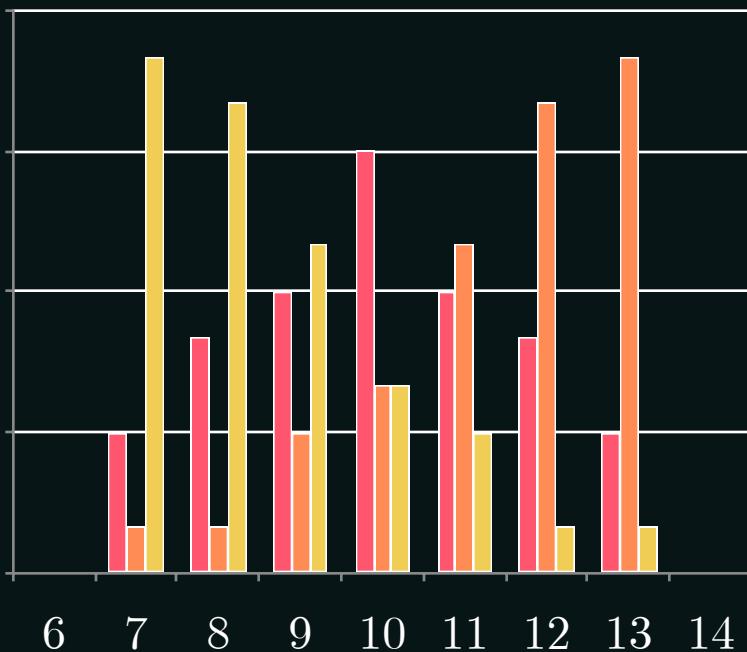
La moyenne vaut $\bar{x} = 11,41$, la médiane vaut $\tilde{x} = 12$ et le mode vaut $x_m = 13$.

La distribution n'est pas symétrique. Il s'agit d'une distribution à asymétrie négative.

x_i	n_i	f_i	F_i	$x_i f_i$
7	11	0,30	0,30	2,08
8	10	0,27	0,57	2,16
9	7	0,19	0,76	1,70
10	4	0,11	0,86	1,08
11	3	0,08	0,95	0,89
12	1	0,03	0,97	0,32
13	1	0,03	1,00	0,35
37	1,00			8,59

La moyenne vaut $\bar{x} = 8,59$, la médiane vaut $\tilde{x} = 8$ et le mode vaut $x_m = 7$.

La distribution n'est pas symétrique. Il s'agit d'une distribution à asymétrie positive.



— COEFFICIENTS D'ASYMÉTRIE DE PEARSON ET DE FISHER —

Les coefficients d'ordre 1 de Pearson et de Fisher permettent de caractériser le degré de symétrie d'une distribution. Ils s'expriment respectivement comme :

$$\beta_1 = \frac{m_3^2}{m_2^3} = \frac{m_3^2}{s^6} \text{ pour le coefficient de Fisher, où } m_2 \text{ et } m_3 \text{ sont les moments d'ordre 2 et 3 respectivement,}$$

$$\gamma_1 = \frac{m_3}{m_2^{3/2}} = \frac{m_3}{s^3} \text{ pour le coefficient de Pearson.}$$

$\beta_1 = 0$ pour une distribution symétrique, $\beta_1 > 0$ et $m_3 > 0$ pour une distribution étalée à droite et $\beta_1 > 0$ et $m_3 < 0$ pour une distribution étalée à gauche.

$\gamma_1 = 0$ pour une distribution symétrique, $\gamma_1 > 0$ pour une distribution étalée à droite et $\gamma_1 < 0$ pour une distribution étalée à gauche.

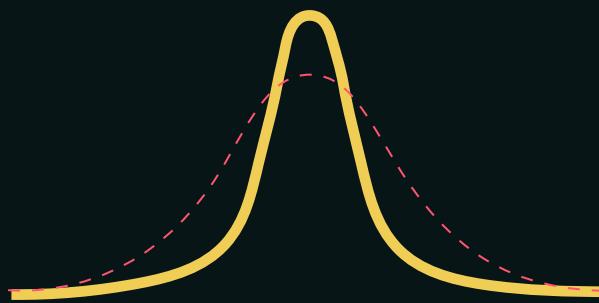
— COEFFICIENTS D'ASYMÉTRIE DE YULE —

Le coefficient de Yule, noté C_y , mesure l'asymétrie d'une distribution en tenant compte de la répartition des quartiles par rapport à la médiane.

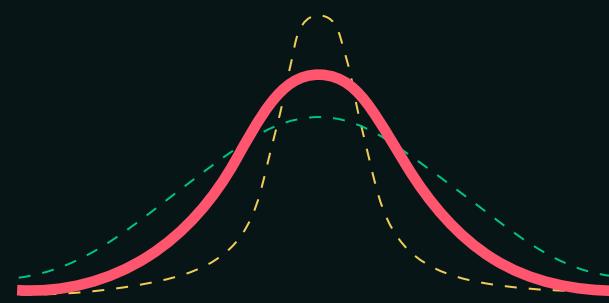
$$C_y = \frac{Q_1 + Q_3 - 2\tilde{x}}{Q_3 - Q_1}.$$

$C_y = 0$ pour une distribution symétrique, $C_y < 0$ pour une distribution étalée à droite et $C_y > 0$ pour une distribution étalée à gauche.

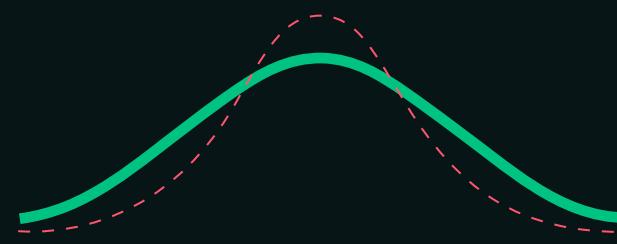
— COMPARAISON DE L'APLATISSEMENT À LA LOI NORMALE —



DISTRIBUTION
LEPTOCURTIQUE



DISTRIBUTION NORMALE
(MÉSOKURTIQUE)



DISTRIBUTION
PLATICURTIQUE

— COEFFICIENTS D'APLATISSEMENT DE PEARSON ET DE FISHER —

Les coefficients d'ordre 2 de Pearson et de Fisher permettent de caractériser le degré d'aplatissement d'une distribution. Ils s'expriment respectivement comme :

$$\beta_2 = \frac{m_4}{m_2^2} = \frac{m_4}{s^4} \text{ pour le coefficient de Fisher, où } m_2 \text{ et } m_4 \text{ sont les moments d'ordre 2 et 4 respectivement,}$$

$$\gamma_2 = \frac{m_4}{m_2^2} - 3 = \frac{m_4}{s^4} - 3 \text{ pour le coefficient de Pearson.}$$

$\beta_2 = 3$ pour une distribution normale, $\beta_2 > 3$ pour une distribution leptocurtique et $\beta_2 < 3$ pour une distribution platicurtique.

$\gamma_2 = 0$ pour une distribution normale, $\gamma_2 > 0$ pour une distribution leptocurtique et $\gamma_2 < 0$ pour une distribution platicurtique gauche.

On remarque aisément que $\beta_2 = \gamma_2 + 3$.

STATISTIQUE DESCRIPTIVE À DEUX DIMENSIONS

- ♦ ORGANISATION DES DONNÉES
- ♦ DISTRIBUTIONS MARGINALES
- ♦ DISTRIBUTIONS CONDITIONNELLES
- ♦ FRÉQUENCES
- ♦ PARAMÈTRES STATISTIQUES MARGINAUX
- ♦ PARAMÈTRES STATISTIQUES CONDITIONNELS
- ♦ COVARIANCE, CORRÉLATION



— EXEMPLE —

On souhaite comparer les résultats obtenus en français et en mathématique par une classe de seconde. On vous donne ci-dessous les couples de notes obtenus par les 30 élèves composant la classe sous la forme (français, maths).

$S=\{(18,2),(11,4),(13,5),(12,7),(15,7),(8,11),(9,11),(11,11),(10,9),(11,8),(11,10),(10,9),(13,11),(15,10),(14,8),(0,15),(5,12),(4,15),(7,15),(6,13),(11,15),(8,12),(9,13),(8,13),(13,14),(14,15),(3,16),(2,20),(5,19),(7,18)\}.$

F \ M	[0,4[[4,8[[8,12[[12,16[[16,20]
[0,4[0	0	0	1	2
[4,8[0	0	0	4	2
[8,12[0	1	7	4	0
[12,16[0	3	3	2	0
[16,20]	1	0	0	0	0

Français



- (1) Peut-on trouver une relation entre les notes obtenues en français et celles obtenues en mathématique ?

Pour cela, nous allons calculer le coefficient de corrélation. Ou bien, plus simplement, représenter graphiquement les notes obtenues en français en fonction de celles obtenues en maths. On voit que les élèves forts en maths ont tendance à avoir des mauvaises notes en français, et inversement. On appelle cela une corrélation négative.

— NOTATIONS ET TABLEAU DE CONTINGENCE —

Nous incluons à présent dans nos analyses deux variables distinctes. Elles seront généralement notées X et Y . On note les MODALITÉS DE CES DEUX VARIABLES

x_i avec $i \in [1, I]$ pour la variable X , où I est le nombre de modalités de X

y_j avec $j \in [1, J]$ pour la variable Y , où J est le nombre de modalités de Y .

Les variables X et Y sont mesurées simultanément sur les N individus formant la population. On définit alors l'EFFECTIF CORRESPONDANT AU COUPLE (x_i, y_j) comme

$$n_{i,j}$$

On appelle DISTRIBUTION JOINTE DES EFFECTIFS DE X ET DE Y l'ensemble des informations $(x_i, y_j, n_{i,j})$.

Le TABLEAU DE CONTINGENCE est un tableau à double entrée qui regroupe les informations de la distribution jointe des effectifs de X et de Y . Il se présente sous la forme ci-contre.

On a toujours la relation :

$$N = \sum_{i=1}^I \sum_{j=1}^J n_{i,j}$$

X\Y	y₁	y₂	...	y_J
x₁	n _{1,1}	n _{1,2}	...	n _{1,J}
...
x_I	n _{I,1}	n _{I,2}	...	n _{I,J}

— DISTRIBUTIONS MARGINALES —

On ajoute au tableau de contingence les données correspondant au totaux en ligne $n_{i,:}$ et en colonne $n_{:,j}$.

X\Y	y₁	...	y_j	...	y_J	Totaux
x₁	n _{1,1}	...	n _{1,j}	...	n _{1,J}	n _{1,:}
...
x_i	n _{i,1}	...	n _{i,j}	...	n _{i,J}	n _{i,:}
...
x_I	n _{I,1}	...	n _{I,j}	...	n _{I,J}	n _{I,:}
Totaux	n _{:,1}	...	n _{:,2}	...	n _{:,J}	n _{:, :}

Mathématiquement, on peut écrire

$$n_{:,j} = \sum_{i=1}^I n_{i,j} \text{ pour la colonne } j, \quad n_{i,:} = \sum_{j=1}^J n_{i,j} \text{ pour la ligne } i,$$

Les I couples $(x_i, n_{i,:})$ définissent la DISTRIBUTION MARGINALE de la variable X et les J couples $(y_j, n_{:,j})$ définissent la DISTRIBUTION MARGINALE de la variable Y . Elles correspondent aux distributions que nous obtiendrions si l'on observait seulement le caractère X (respectivement Y).

— DISTRIBUTIONS CONDITIONNELLES —

Les distributions conditionnelles sont obtenues en fixant la valeur de l'une des deux variables. On parle alors de la valeur de la variable X sachant que $Y = y_j$ (respectivement la valeur de la variable Y sachant que $X = x_i$), que l'on note

$$X|_{Y=y_j} \text{ (respectivement } Y|_{X=x_i}).$$

F \ M	[0,4[[4,8[[8,12[[12,16[[16,20]
[0,4[0	0	0	1	2
[4,8[0	0	0	4	2
[8,12[0	1	7	4	0
[12,16[0	3	3	2	0
[16,20[1	0	0	0	0

Dans l'exemple du début de chapitre, déterminez la distribution conditionnelle de la note de français sachant que l'élève a obtenu une note de maths entre 8 et 12.

On fixe $y = y_3$, ce qui donne $X|_{Y=y_3} = \{11, 11, 11, 9, 8, 10, 9, 11, 10, 8\}$. On pourrait également présenter le résultats sous forme de la colonnes $[8,12[$ du tableau de contingence.

— FRÉQUENCES —

Tout comme nous l'avons fait pour la statistique univariée, il est possible de calculer pour les statistiques bivariées trois types de fréquences. On distingue donc

- $f_{i,j} = \frac{n_{i,j}}{N}$ appelée FRÉQUENCE DU COUPLE (x_i, y_j) ,
- $f_{i,:} = \frac{n_{i,:}}{N}$ appelée FRÉQUENCE MARGINALE de x_i , respectivement $f_{:j} = \frac{n_{:,j}}{N}$ appelée FRÉQUENCE MARGINALE de y_j ,
- $f_i|_{Y=y_j} = \frac{n_{i,j}}{n_{:,j}}$ appelée FRÉQUENCE CONDITIONNELLE de x_i sachant que $Y = y_j$, respectivement $f_j|_{X=x_i} = \frac{n_{i,j}}{n_{i,:}}$ appelée FRÉQUENCE CONDITIONNELLE de x_i sachant que $Y = y_j$

— PARAMÈTRES STATISTIQUES MARGINAUX —

Tout comme nous l'avons fait pour la statistique univariée, il est possible de calculer pour les statistiques bivariées les paramètres statistiques marginaux. On présente ici les formules pour la moyenne et l'écart-type.

- $\bar{x} = \frac{1}{N} \sum_{i=1}^I n_{i,:} x_i$ pour la moyenne marginale de X et $\bar{y} = \frac{1}{N} \sum_{j=1}^J n_{:,j} y_j$ pour la moyenne marginale de Y
- $V(X) = \frac{1}{N} \sum_{i=1}^I n_{i,:} (x_i - \bar{x})^2$ pour la variance marginale de X , $V(Y) = \frac{1}{N} \sum_{j=1}^J n_{:,j} (y_j - \bar{y})^2$ pour la variance marginale de Y

— PARAMÈTRES STATISTIQUES CONDITIONNELS —

Tout comme nous l'avons fait pour la statistique univariée, il est possible de calculer pour les statistiques bivariées les paramètres statistiques conditionnels. On présente ici les formules pour la moyenne et l'écart-type.

- $\bar{x}|_{Y=y_j} = \frac{1}{n_{:,j}} \sum_{i=1}^I n_{i,j} x_i$ pour la moyenne conditionnelle de X et $\bar{y}|_{X=x_i} = \frac{1}{n_{i,:}} \sum_{j=1}^J n_{i,j} y_j$ pour la moyenne conditionnelle de Y
- $V(X|_{Y=y_j}) = \frac{1}{n_{:,j}} \sum_{i=1}^I n_{i,j} (x_i - \bar{x}|_{Y=y_j})^2$ pour la variance conditionnelle de X et
 $V(Y|_{X=x_i}) = \frac{1}{n_{i,:}} \sum_{j=1}^J n_{i,j} (y_j - \bar{y}|_{X=x_i})^2$ pour la variance conditionnelle de Y

— COVARIANCE, CORRÉLATION —

La COVARIANCE et la CORRÉLATION sont des outils pour quantifier la dépendance linéaire entre deux caractères quantitatifs X et Y . On définit la COVARIANCE des variables X et Y comme

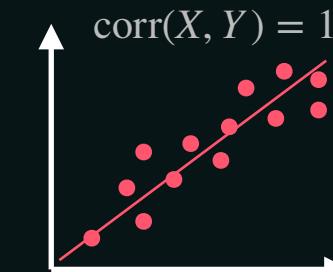
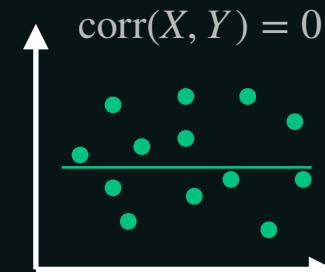
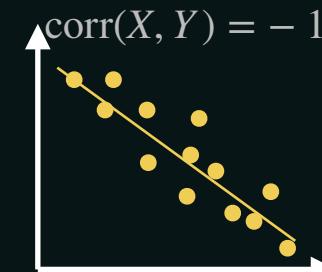
$$\text{cov}(X, Y) = \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^J n_{i,j} (x_i - \bar{x})(y_j - \bar{y}).$$

On peut noter que $\text{cov}(X, Y) = \text{cov}(Y, X)$ et que $\text{cov}(X, X) = V(X)$.

On définit également le COEFFICIENT DE CORRÉLATION LINÉAIRE qui comme étant la version normée de la covariance soit

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma(X)\sigma(Y)} \text{ où } \sigma \text{ représente l'écart-type.}$$

On note que $\text{corr}(X, Y) \in [-1, 1]$. Un coefficient de corrélation égal à ± 1 indique une relation linéaire entre X et Y . Un coefficient de corrélation égal à 0 indique une non corrélation entre X et Y , l'indépendance de X et Y est possible mais pas certaine. Voir ce site pour [UN MAX DE CORRÉLATIONS](#).



OUTILS INFORMATIQUES

♦ FONCTION DE BASES DANS EXCEL





— FONCTION DE BASE EXCEL —

— FORMULES DE BASE —

	A	B	C	D	E	F	G	H	I	J
1	10	2	3	4	5	4	3	2	1	1
2	15									
3	20									
4	25									

=SOMME(A1:A4) effectue la somme des cellules A1 à A4 (en colonnes).

=SOMME(A1:J1) effectue la somme des cellules A1 à J1 (en ligne).

SOMME(A1:A4)

70

SOMME(A1:J1)

35

=MOYENNE(A1:A4) effectue la moyenne des cellules A1 à A4 (en colonnes).

=MOYENNE(A1:J1) effectue la moyenne des cellules A1 à J1 (en ligne).

MOYENNE(A1:A4)

17,50

MOYENNE(A1:J1)

3,50

=MIN(A1:A4) donne la valeur minimale des cellules A1 à A4 (en colonnes).

=MIN(A1:J1) donne la valeur minimale des cellules A1 à J1 (en ligne).

MIN(A1:A4)

10,00

MIN(A1:J1)

1,00

=MAX(A1:A4) donne la valeur maximale des cellules A1 à A4 (en colonnes).

=MAX(A1:J1) donne la valeur maximale des cellules A1 à J1 (en ligne).

MAX(A1:A4)

25,00

MAX(A1:J1)

10,00

— FORMULES DE BASE —

Fréquence : =FREQUENCE(*série*)

Médiane : =MEDIANE(*série*)

Mode : =MODE(*série*)

Écart moyen : =ECART.MOYEN(*série*)

Écart-type : =STD(*série*) ou =ECARTYPE(*série*)

Variance : =VAR (*série*)

Quartile : =QUARTILE(*série* ; *Q*). Par exemple =QUARTILE(A1:J10, 3) pour avoir le troisième quartile.

Centile : =CENTILE(*série* ; *C*). Par exemple =DECILE(A1:J10, 99) pour avoir le quatre-vingt-dix-neuvième décile.

Coefficient de corrélation : =COEFFICIENT.CORRELATION(*série*)

Covariance : =COVARIANCE(*série*)

Voir [CE SITE](#) pour la liste exhaustive des fonctions Excel et [CE SITE](#) pour leurs équivalents sur Numbers.

— COMPTER DES OCCURRENCES —

	A	B	C	D	E	F	G	H	I	J
1	10	2	3	4	5	4	3	2	1	1
2	15	1	2	3	4	5	8	9	0	7
3	20	2	7	8	6	4	9	3	7	0
4	25	3	1	1	2	7	4	9	3	0

=NB(A1:A4) donne le nombre d'éléments des cellules A1 à A4 (en colonnes).

=NB(A1:J1) donne le nombre d'éléments des cellules A1 à J1 (en ligne).

NB(A1:A4)

4

NB(A1:J1)

10

=NB.SI(B1:B4;3) donne le nombre d'éléments égaux à 3 dans les cellules B1 à B4 (en colonnes).

=NB.SI(A1:J1;3) donne le nombre d'éléments égaux à 3 dans les cellules A1 à J1 (en ligne).

NB.SI(A1:A4)

1,00

NB.SI(A1:J1)

2,00

=NB.SI.ENS(A1:A4;"<=20";A1:A10;">11") donne le nombre d'éléments compris entre 11 et 20 dans les cellules A1 à A4 (en colonnes).

=NB.SI.ENS(A1:J1,"<=4";A1:A10;">=2") donne le nombre d'éléments compris entre 2 et 4 dans les cellules A1 à J1 (en ligne).

=NB.SI.ENS(A1:J4,"<=4";A1:A10;">=2") donne le nombre d'éléments compris entre 2 et 4 dans tout le tableau (lignes et colonnes).

NB.SI.ENS(A1:A4)

2,00

MIN(A1:J1)

6,00

MIN(A1:J1)

16,00