HAUTE-ÉCOLE NAMUR-LIÈGE-LUXEMBOURG
Implantation IESN

IG/INTAR -UE157

OUTILS STATISTIQUES

Module 1 : Statistiques

Par les professeurs de mathématique de l'IESN

Année académique 2024-2025

PARTIE 1
STATISTIQUES DESCRIPTIVES À UNE DIMENSION

On peut revoir la différence entre les probabilités et les statistiques à travers l'histoire suivante.

Imaginez un naufragé sur une île perdue au milieu de nulle part. Appelons-le Robinson. Robinson a survécu de justesse à un naufrage et s'est réveillé sur une plage inconnue. En s'aventurant sur l'île, Robinson croise une peuplade d'indigènes qui parlent un langage qu'il ne reconnaît pas. Les indigènes se montrent plutôt accueillants à son égard : ils lui donnent à manger et une hutte où dormir.

Les premiers jours, Robinson écoute les indigènes parler. Il note dans son carnet les mots qu'ils prononcent quand ils s'adressent à lui. Petit à petit, Robinson se rend compte que certains mots reviennent plus souvent que d'autres. Ainsi, les indigènes disent quelque chose comme « Bouf » lorsqu'ils lui apportent à manger et ils prononcent « Pieute » quand vient l'heure de dormir.

Dans un premier temps, Robinson observe, écoute et classifie les mots. C'est la partie « **statistique descriptive** ».

Au bout d'un moment, Robinson émet quelques hypothèses quant à la signification de certains mots. Ainsi, « Bouf » signifierait-il nourriture ? Et « Pieute » voudrait-il dire dormir ? Peu à peu, Robinson construit ainsi un mini dictionnaire indigène-français.

Dans cette seconde phase, Robinson analyse ses observations et, sur base de celles-ci, il construit des hypothèses (celles qui semblent les plus « probables »). C'est la partie « **statistique inférentielle** ».

Finalement, un avion de passage aperçoit le gigantesque SOS que Robinson a dessiné sur la plage et envoie des secours. Au fil du temps passé chez les indigènes, Robinson a maîtrisé leur langage. Il le connaît tellement bien qu'il se met à l'enseigner. Il rédige un syllabus qui présente les règles du langage et, pendant les cours qu'il donne, il utilise ces règles pour construire de nouvelles phrases en indigène.

Dans cette partie de l'histoire, Robinson se base sur des règles pour produire des résultats (ici, construire de nouvelles phrases). Grâce aux règles qu'il a découvertes, il est capable de déduire les phrases que les indigènes vont utiliser pour parler de ceci ou de cela. C'est la partie « **probabilités** » de son histoire.

CHAPITRE 1 : DONNÉES STATISTIQUES

La statistique est la science de l'analyse des données. Cette définition générale peut s'appliquer dans de nombreux domaines : le terme « données » peut faire référence au sexe des nouveau-nés, aux choix politiques des Belges, au nombre d'accidents qui se produisent sur un tronçon routier, aux résultats scolaires des étudiants de la haute école, aux chiffres d'affaires des entreprises wallonnes, aux conséquences de l'absorption d'un nouveau médicament…

Plan du module

1 Population et échantillon

Bien souvent, quand on s'intéresse à une question (comme, par exemple « Parmi les nouveau-nés, y a-t-il autant de garçons que de filles ? »), on ne peut pas récolter toutes les données nécessaires pour y répondre. Pratiquement, il est impossible de comptabiliser tous les humains qui sont nés pour savoir s'il y a plus d'hommes ou de femmes.

Dans ce cas-là, on se contente de récolter une partie des données, partie qu'on espère aussi significative que possible. Dans le cadre de l'exemple, on pourrait se renseigner auprès de tous les hôpitaux belges pour connaître le sexe des nouveau-nés qui ont vu la vie au cours du mois de janvier 2024.

Ce procédé est appelé « **échantillonnage** ». Il consiste à se procurer seulement une partie des données (l'**échantillon**). Le but final reste d'obtenir des informations à propos du groupe plus large (ici, tous les humains). Ce groupe plus large est appelé la **population**.

Dans un premier temps, on analysera les données de l'échantillon : on les regroupera de manière lisible (tableaux récapitulatifs), on les présentera de manière claire (graphiques) et on en tirera quelques valeurs-clefs (appelés « paramètres »). Ces processus d'analyse des données font l'objet de la **statistique descriptive**.

On supposera ensuite que les résultats obtenus sur l'échantillon restent valables pour la population dans son ensemble. Ce procédé de généralisation, appelé l'**inférence**, est étudié par la **statistique inférentielle** (voir la 2e partie du cours).

2 Un peu de vocabulaire

La section précédente a déjà introduit quelques mots possédant une signification précise en statistique.

- **Population** : c'est l'ensemble des individus ou objets auxquels on s'intéresse (par exemple : les nouveau-nés).
- **Échantillon** : c'est l'ensemble des individus ou objets au sujet desquels on a obtenu des informations (par exemple : les bébés nés en Belgique en janvier 2024).

En voici quelques autres.

- **Individu** : c'est le nom générique donné aux objets qui constituent la population ; il peut s'agir de personnes, de choses, d'associations ou d'objets immatériels (dans l'exemple, il s'agit de nouveau-nés).
- **Effectif ou effectif total** : c'est le nombre d'individus dans l'échantillon (donc, le nombre de données récoltées). On le note souvent n.
- **Caractère, caractère étudié, variable ou variable statistique** : c'est l'aspect qu'on étudie (ici, le sexe des nouveau-nés).
- **Espace des observables** : c'est l'ensemble des réponses (résultats) possibles pour le caractère étudié (ici, l'ensemble $\Omega = \{\mathrm{M}, \mathrm{F}\}$)

Exercice

1) Dans chacune des situations suivantes, identifiez la population, l'échantillon, le caractère étudié, l'effectif et l'espace des observables.
   a) Une enquête a été réalisée dans toute la Wallonie pour déterminer le nombre d'armes à feu que chaque ménage possédait. L'équipe en charge de l'enquête a envoyé 5 000 questionnaires par courrier, sur lesquels 3 500 ont été complété et renvoyés.
   a) Pour connaître le temps que les directeurs de PME wallonnes consacrent chaque semaine aux tâches administratives, on en a choisi 200 au hasard et on les a interrogés à ce sujet.
   b) Une antenne de la Croix-Rouge placée à proximité de l'IESN désire savoir quelle quantité de sang de chaque groupe garder en stock. Elle organise une enquête auprès de 150 étudiants afin de connaître leur groupe sanguin.
   c) Les responsables d'un bureau de marketing font un sondage dans les rues de Namur afin de savoir ce que les Namurois pensent de la propreté de leur ville. Ils demandent aux 100 personnes interrogées de choisir l'un des qualificatifs suivants : très sale, sale, moyenne, propre, très propre.
   d) Les enseignants de l'IESN décident d'organiser un test sur ordinateur pour mieux cerner leurs futurs étudiants. Le test est constitué de 25 questions de culture générale de type vrai/faux valant chacune 1 point. Ils font appel à 50 volontaires pour effectuer ce test.
   e) Un enseignant étonné de voir ses étudiants s'endormir en cours décide d'en retenir 20 après la classe. Il leur demande à quelle heure ils ont été dormir la veille et note leurs réponses.

Types de données

La statistique peut traiter des données de divers types. On distingue tout d'abord les données **qualitatives** des données **quantitatives**. On parle de données quantitatives lorsqu'il s'agit de nombres qui mesurent une certaine grandeur (longueur, volume, cardinalité, durée…). Dans le cas contraire (lorsqu'il ne s'agit pas de valeurs numériques ou lorsque ces nombres ne sont pas des mesures), on parle de données qualitatives.

Exemples de données qualitatives : votre couleur préférée, le nom de votre acteur ou actrice préféré(e), le type de chaussures que vous portez, le sexe de votre meilleur(e) ami(e), votre chiffre préféré…

Parmi les données quantitatives, on distingue encore deux catégories : les données quantitatives **discrètes** et les données quantitatives **continues**. On parle de données discrètes quand on peut énumérer (citer) les valeurs possibles (typiquement, ces valeurs possibles seront des entiers). On parle de données continues quand les valeurs possibles ne sont pas énumérables (typiquement, quand il s'agit de réels).

Par exemple, le nombre d'années d'étude réussies est une donnée discrète, car la valeur ne peut être que 0, 1, 2, 3, 4, 5, … (les valeurs possibles sont énumérables – le terme exact est *dénombrables*). À l'inverse, la taille en centimètres des semelles de chaussures est une donnée continue car il pourrait s'agir de n'importe quel réel entre 0 cm et 40 cm.

Type de données		**Exemples**
Qualitatif		$\Omega = \{\text{oui, non}\}$ $\Omega = \{\text{CP, IG, TI, DR, MK}\}$
Quantitatif	**Discret**	$\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ $\Omega = \{0, 1, 2, 3, 4, 5, \dots\}$
	Continu	$\Omega = [0\text{ cm} \dots 200\text{ cm}]$ (taille) $\Omega = [-40° \dots 40°]$ (température)

Note. Dans la suite, on s'intéressera principalement aux caractères quantitatifs.

Exercice

2) Reprenez les situations de l'exercice précédent et indiquez à chaque fois le type du caractère étudié.

Les données brutes

Après la récolte des informations, le statisticien dispose de données brutes : une valeur pour chaque individu de l'échantillon. Cette longue liste de valeurs n'est pas forcément facile à utiliser telle quelle, mais on peut la transformer en tableaux plus aisés à lire (voir sections suivantes).

Exemple : résultats de 20 étudiants lors d'une interrogation cotée sur 10.

8	*5*	*2*	*9*	*10*	*6*	*7*	*5*	*4*	*2*
6	*9*	*8*	*4*	*1*	*4*	*5*	*8*	*5*	*9*

En mathématiques, lorsqu'on a affaire à plusieurs valeurs qui jouent le même rôle (comme les valeurs des données d'un échantillon), on utilise souvent une notation indicée (c'est-à-dire avec un indice). Dans la suite, on notera x_i les valeurs observées.

Cela signifie que la première valeur observée sera x_1 (qui vaut 8 dans l'exemple), que la seconde sera x_2 (qui vaut 5 dans l'exemple), et ainsi de suite jusqu'à la dernière valeur observée, x_{20} (qui vaut 9 dans l'exemple).

Les valeurs observées sont donc notées $x_1, x_2, \dots x_n$, où n est l'effectif total (c'est-à-dire le nombre d'individus dans l'échantillon mais également le nombre de valeurs récoltées).

Notation	Signification
x_i	le i^{e} résultat (i pouvant aller de 1 à n)
n	l'effectif total (= le nombre de valeurs dans l'échantillon)

Tableau des répétitions (de recensement)

Une première simplification consiste à repérer les valeurs qui se répètent. Plutôt que de les citer à de multiples reprises, il est plus clair d'indiquer le nombre de fois qu'elles apparaissent. Ce nombre est leur **répétition** ou **fréquence absolue**.

Dans un tableau des répétitions, on cite les différentes valeurs (on les appelle les **modalités**) et leurs répétitions. Pour plus de clarté, on peut même ajouter quelques valeurs qui sont possibles mais n'apparaissent pas dans l'échantillon (avec une répétition de 0).

Modalité X_i	1	2	3	4	5	6	7	8	9	10
Répétition r_i	1	2	0	3	4	2	1	3	3	1

Tout comme les données brutes, les modalités et les répétitions sont généralement représentées par une notation indicée. Pour les modalités, nous allons utiliser X_i (pour i allant de 1 jusqu'à un certain nombre p) et, pour les répétitions, r_i (pour i allant de 1 jusqu'à un certain nombre p). Dans le cas de l'exemple ci-dessus, p vaut 10.

Comme les modalités et les répétitions utilisent le même système d'indices, on peut affirmer que, pour toute valeur de l'indice i, le nombre r_i est la répétition qui correspond à la modalité X_i ; ceci est valable quelle que soit la valeur de i (entre 1 et p).

Notation	Signification
X_i	la i^{e} modalité (i pouvant aller de 1 jusqu'à p)
r_i	la répétition de la i^{e} modalité
p	le nombre de modalités

Note. Dans l'exemple, on a $X_i = i$ pour tout indice i mais ce n'est pas toujours le cas.

Exercices

3) Une étude concernant les femmes américaines d'origine mexicaine d'une ville donnée s'est intéressée au nombre d'enfants à qui elles avaient donné le jour. Ci-dessous se trouvent les informations données par celles qui ont accepté de répondre.

5 2 3 0 4 6 2 4 4 4 4

1	1	2	3	3	4	4	4	5	5	2
5	5	5	3	3	3	3	5	3		

a) Identifiez la population, l'échantillon et le caractère étudié.
b) Construisez le tableau des répétitions.
c) Que valent et que représentent X_1, X_4, r_1 et r_4 ?

4) Un biologiste s'intéresse à la longévité d'une race de papillons exotiques. Il a examiné un certain nombre d'entre eux et a noté combien d'heures ils ont vécu. Le tableau des répétitions est donné ci-dessous.

Durée de vie (heures)	27	28	29	30	31	32	33
Nombre de papillons	2	4	8	7	5	3	1

a) Quel est l'effectif total ?
b) Si on utilise les notations standards, que valent X_2, r_5 et X_4 ?
c) Que valent les expressions $X_3 + r_3$ et $X_1 r_4$?
d) Que valent les expressions $\sum X_i$, $\sum r_i^2$ et $\sum X_i r_i$?
e) De manière générale, comment peut-on calculer l'effectif total à partir des X_i et des r_i ?

5) On a recensé l'âge des étudiants d'une classe et on a obtenu les résultats suivants. Précisez l'espace des observables, construisez le tableau de recensement et calculez l'effectif total.

18	18	19	19	20	19	19	18	19	21
18	18	19	19	18	19	20	21	21	21
19	18	21	20	18	19	19	19	18	18

Groupement par classes

Le tableau des répétitions ne convient pas dans certains cas. Considérons l'exemple suivant, où on a interrogé plusieurs personnes et on leur a demandé le temps que leur prenait le trajet domicile-travail. Leurs réponses (en minutes) sont indiquées ci-dessous.

22.4	21.5	19.0	21.9	18.0
19.2	21.1	24.7	21.7	18.4
16.2	15.0	22.0	21.0	19.8
22.7	19.1	22.2	23.8	16.7

Ici, un tableau des répétitions n'aiderait en rien la lecture, car on se retrouverait avec 20 modalités différentes, chacune comportant 1 unique répétition : le tableau des répétitions serait aussi difficile à lire que la liste des données brutes. Dans ce genre de cas, on peut décider de perdre un peu de précision et de regrouper entre elles les valeurs qui sont proches les unes des autres.

Ainsi, comme les valeurs s'étendent de 15.0 (la plus petite) à 24.7 (la plus grande), on pourrait décider de les rassembler par tranche de 2 minutes : de 15 à 17 minutes, de 17 à 19 minutes, de 19 à 21 minutes, de 21 à 23 minutes et finalement de 23 à 25 minutes. Ces groupes de valeurs, appelés **classes**, regroupent ensemble toutes les données qui se trouvent entre deux valeurs limites.

On peut se demander où placer une valeur qui tomberait pile-poil entre deux classes. Par convention, on la place dans la seconde classe. Ainsi, dans l'exemple, la valeur 19.0 se placera dans la classe « de 19 à 21 minutes ». Cela revient à dire que la classe « de 19 à 21 minutes » reprend toutes les valeurs qui sont à la fois $\geq$ 19 minutes et $<$ 21 minutes.

Chacune des classes est délimitée par deux valeurs (qu'on notera a_i et b_i) et reprend toutes les données qui sont $\geq a_i$ et $< b_i$. Dans le tableau de recensement, on notera cette classe $[a_i, b_i[$ vu qu'il s'agit de l'intervalle des données qui sont rassemblées dans cette classe. Ainsi, la première classe sera $[a_1, b_1[$, la seconde $[a_2, b_2[$ et ainsi de suite.

Classe $[a_i, b_i[$ (durée en min)	**Centre c_i (min)**	**Répétition r_i**
[15,17[	16	3
[17,19[	18	2
[19,21[	20	4
[21,23[	22	9
[23,25[	24	2

Chaque classe $[a_i, b_i[$ possède une **répétition** r_i qui est tout simplement le nombre de données qu'elle rassemble.

On associe également à chaque classe son **centre** (qui est la valeur à mi-chemin entre a_i et b_i, c'est-à-dire $(a_i + b_i)/2$. Le centre de la classe, noté c_i, sert de valeur représentative pour toutes les données qui ont été rassemblées dans la classe : la valeur c_i « représente » (ou approxime) toutes les données de la classe. Rassembler les données en des classes implique une perte de précision : plutôt que de conserver les valeurs exactes, on fait comme si toutes les données de la classe $[a_i, b_i[$ valaient c_i.

Notation	**Signification**
a_i	la borne inférieure de la i^{e} classe
b_i	la borne supérieure de la i^{e} classe
c_i	le centre de la i^{e} classe (la valeur qui représente tous les résultats rassemblés dans cette classe) ; $c_i = (a_i + b_i)/2$
r_i	la répétition de la i^{e} classe (le nombre de données dans cette classe)
$b_i - a_i$	l'amplitude de la i^{e} classe, sa « largeur »

Note. On choisit généralement des classes d'amplitude identique, mais cela n'est pas obligatoire. Dans la suite, on supposera souvent que c'est bien le cas.

Note. Le choix des classes n'est pas toujours évident : il s'agit de réaliser un compris pour éviter d'une part des classes trop larges (perte de précision trop importante) et d'autre part des classes trop étroites (gain insuffisant en lisibilité). Les limites des classes sont souvent choisies en divisant l'étendue de l'échantillon (c'est-à-dire la différence entre la valeur la plus petite et la valeur la plus grande) par le nombre de classes désiré.

Exercices

6) Dans l'exemple présenté ci-dessus (trajets domicile-travail),
   a) Quel est l'effectif total ?
   b) Que valent a_2, b_5 et c_4 ?

c) Que signifie la valeur « 20 » inscrite en plein milieu du tableau ? Et le « 4 » à sa droite ?

7) On a mesuré la taille (en mètres) d'un échantillon de 25 personnes et on a obtenu les résultats suivants. Construisez un tableau de recensement en utilisant 4 classes de même amplitude recouvrant les tailles de 1,55 m à 1,95 m.

1,71	1,58	1,65	1,7	1,75	1,85	1,9	1,67	1,72	1,77	1,85
1,82	1,9	1,69	1,87	1,92	1,76	1,7	1,8	1,86	1,59	1,6
1,67	1,77	1,82								

8) Un artisan spécialisé dans le travail de l'argile a fabriqué 50 bâtons d'argile de 1 mètre de longueur, les a fait sécher puis les a mesurés à nouveau. Pour chaque bâton, il a noté le nombre de centimètres perdus. Les résultats sont donnés dans le tableau suivant. Répartir les données en classe d'amplitude de 1cm à partir de 13cm. Dresser le tableau de recensement.

18,2	21,2	23,1	18,5	15,6	20,8	19,4	15,4	21,2	13,4
16,4	18,7	18,2	19,6	14,3	16,6	24	17,6	17,8	20,2
17,4	23,6	17,5	20,3	16,6	19,3	18,5	19,3	21,2	13,9
20,5	19	17,6	22,3	18,4	21,2	20,4	21,4	20,3	20,1
19,6	20,6	14,8	19,7	20,5	18	20,8	15,8	23,1	17

9) Voici les notes sur 100 d'un groupe de 80 étudiants.

53	56	61	73	62	80	60	66	70	71	72
63	62	87	61	65	72	75	67	64	97	90
73	74	71	58	75	76	70	81	71	63	78
75	60	72	82	94	64	64	72	74	61	78
61	84	62	95	79	77	77	73	81	91	83
96	70	77	63	67	81	83	84	83	78	76
78	64	80	92	80	84	77	73	76	75	79
80	61	97								

a) Répartissez ces notes en classe d'amplitude 5, la première étant [50,55[puis dressez le tableau de recensement.
b) Faites de même en utilisant des classes d'amplitude 10 (la première commençant à 50).
c) Dans les deux cas, calculez le nombre d'étudiants ayant une note < 65, < 75 et >= 80.
d) Dans les deux cas, calculez la proportion d'étudiants ayant une cote < 87,5 et > 73.

CHAPITRE 2 : PRÉSENTATION DES DONNÉES

Les données brutes récoltées par une étude statistique sont bien souvent difficiles à interpréter directement. Les tableaux présentés dans le module précédent constituent un premier pas vers une lisibilité plus grande, mais on peut aller encore un peu plus loin.

Les méthodes pour « faire parler » les données se divisent principalement en deux catégories :

- le calcul de valeurs significatives qui décrivent les données (on peut penser à la moyenne par exemple) ; ce sujet sera traité dans les deux modules qui suivent ;
- la traduction des données en graphiques.

C'est aux représentations graphiques des données que ce module s'intéresse. Avant de pouvoir en parler cependant, il faut pousser un peu plus en avant l'analyse réalisée dans les tableaux de recensement.

Plan du module

1 Le tableau des répétitions complété

Les tableaux de recensement construits dans le module précédent se composent de deux colonnes : une colonne indiquant les modalités X_i ou les classes $[a_i, b_i[$ et une colonne indiquant les répétitions r_i.

On y ajoute souvent trois autres colonnes :

- la colonne des **répétitions cumulées** rc_i qui indique le total de toutes les répétitions jusque là ;
- la colonne des **fréquences relatives** f_i qui indique le pourcentage des résultats qui correspond à la modalité/classe en question ;
- la colonne des **fréquences relatives cumulées** g_i qui indique le total des fréquences relatives jusque là.

Les fréquences relatives permettent de « relativiser » les résultats de l'enquête par rapport à la taille de l'échantillon. Par exemple, l'affirmation « Dans ma classe, 10 étudiants ont réussi l'interrogation. »

possède une signification très différente selon que la classe en question comporte 10 ou 100 étudiants. Il serait plus porteur de sens de dire que « 10% des étudiants ont réussi l'interrogation » (ou « 100% des étudiants ont réussi l'interrogation »).

Si on ajoute ces trois colonnes au premier exemple du module précédent (côtes d'une interrogation), voici le tableau de recensement complété qu'on obtient.

Modalités x_i	**Répétitions** r_i	**Rép. cum.** rc_i	**Frqs. relatives (%)** f_i	**Frqs. cum. (%)** g_i
1	1	1	5	5
2	2	3	10	15
3	0	0	0	15
4	3	6	15	30
5	4	10	20	50
6	2	12	10	60
7	1	13	5	65
8	3	16	15	80
9	3	19	15	95
10	1	20	5	100

Il est important de savoir bien interpréter chacune des valeurs qui apparaissent dans le tableau. Voici quelques exemples d'interprétation pour les cases grisées du tableau ci-dessus.

- La répétition r_5 = 4 indique que l'échantillon comporte 4 individus qui correspondent à la modalité X_5 = 5, c'est-à-dire que 4 étudiants ont eu 5/10.
- La répétition cumulée rc_8 = 16 indique que l'échantillon comporte 16 individus qui ont eu une cote inférieure ou égale 8/10.
- La fréquence (relative) f_2 = 10 % indique que 10 % des individus de l'échantillon ont eu une cote de 2/10.
- La fréquence cumulée g_6 = 60 % indique que 60 % des individus de l'échantillon ont eu une cote inférieure ou égale à 6/10.

Notation	**Signification**	**Calcul**
rc_i	la répétition cumulée de la i^{e} modalité	$rc_i = \sum_{k=1}^{i} r_k$
f_i	la fréquence (relative) de la i^{e} modalité	$f_i = r_i/n$
g_i	la fréquence (relative) cumulée de la i^{e} modalité	$g_i = rc_i/n$

La formule donnée pour rc_i signifie que la répétition cumulée de la i^{e} modalité se calcule en effectuant la somme (Σ est le « S » majuscule en grec) des répétitions r_k pour k allant de 1 jusqu'à i. Il s'agit donc bien d'additionner toutes les répétitions jusqu'à la ligne concernée.

Remarque. D'un point de vue mathématique, les expressions « 0,3 », « 3/10 » et « 30% » sont entièrement identiques. Toutes peuvent convenir pour exprimer des fréquences relatives, même si on préfère généralement la forme de pourcentage.

Exercices (réflexion)

1) S'il y a p modalités dans le tableau, que vaut toujours la dernière répétition cumulée rc_p ? Pourquoi ?

2) S'il y a p modalités dans le tableau, que vaut toujours la dernière fréquence relative cumulée g_p ? Pourquoi ?

3) On pourrait calculer les fréquences cumulées en effectuant la somme des fréquences « jusque-là », c'est-à-dire en utilisant la formule

$$g_i = \sum_{k=1}^{i} f_k$$

ce qui devrait donner le même résultat que $g_i = rc_i/n$. Pourquoi, d'un point de vue pratique, est-il préférable d'utiliser la seconde formule cependant ?

4) Grâce au tableau de l'exemple, répondez aux questions suivantes (il peut y avoir plusieurs manières de trouver les réponses).
   a) Combien d'étudiants ont eu une note de 8/10 ?
   b) Quelle proportion représentent les étudiants qui ont eu une note de 9/10 ?
   c) Combien d'étudiants ont échoué (c'est-à-dire ont obtenu une note de 4/10 ou moins) ?
   d) Quelle proportion représentent les étudiants qui n'ont pas obtenu une dispense (c'est-à-dire qui ont obtenu une note de 5/10 ou moins) ?
   e) Combien d'étudiants ont eu une cote supérieure ou égale à 6/10 ?
   f) Quelle proportion représentent les étudiants qui ont réussi (c'est-à-dire ont obtenu une note de 5/10 ou plus) ?
   g) Combien d'étudiants ont eu une note située entre 3/10 et 7/10 (bornes y comprises) ?
   h) Quelle proportion représentent les étudiants qui ont eu une note située entre 4/10 et 8/10 (bornes non comprises) ?

5) Le service après-vente d'une entreprise fait établir des statistiques quant à la fréquence des appels reçus. À plusieurs reprises, des experts comptent le nombre d'appels reçus au cours d'une période de 1 minute. Voici les résultats qu'ils ont obtenus.

Nb d'appels au cours de la minute	0	1	2	3	4	5	6	7	8
Nb de relevés correspondant à cette valeur	93	261	416	393	308	174	93	42	20

   a) Complétez le tableau des répétitions
   b) Combien de relevés ont-ils effectués ?
   c) À quel point est-il fréquent que ce service après-vente ne reçoive aucun appel pendant 1 minute ?
   d) À quel point est-il fréquent que ce service après-vente reçoive plus de 4 appels par minute ?

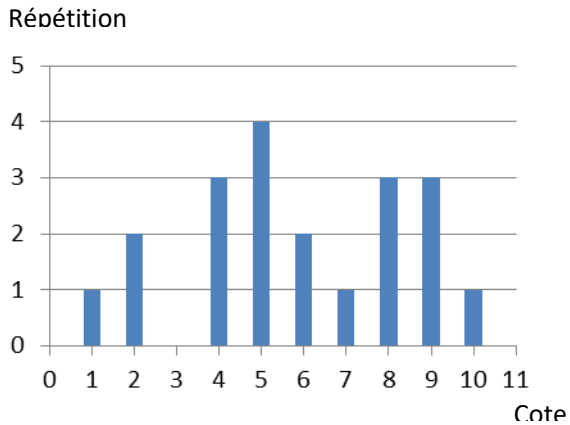
6) Complétez les tableaux de répétitions des exercices du module précédent.

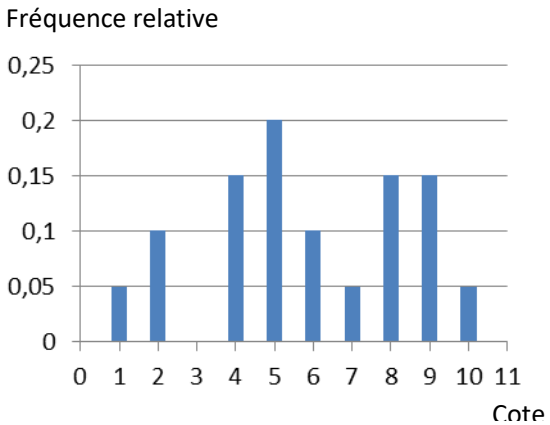
Les diagrammes en bâtons

Bien souvent, « une image vaut mieux qu'un long discours. » C'est pour cela qu'il existe toute une série de représentations graphiques standards pour les données statistiques. Dans le cadre de ce cours, nous nous contenterons d'en aborder quelques-unes.

Les **diagrammes en bâtons** permettent de représenter graphiquement les répétitions : à chaque modalité correspond un bâton dont la hauteur est proportionnelle à la répétition. Plus une modalité est représentée et plus son bâton sera long.

L'axe vertical du graphique peut être échelonné par les répétitions ou par les fréquences relatives. Dans les deux cas, le diagramme produit est le même, à l'échelle près.



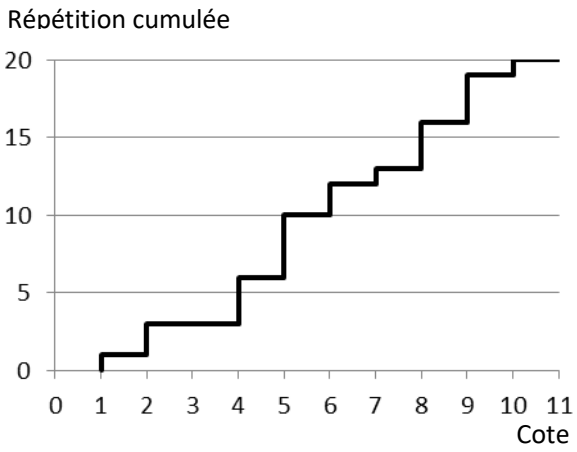


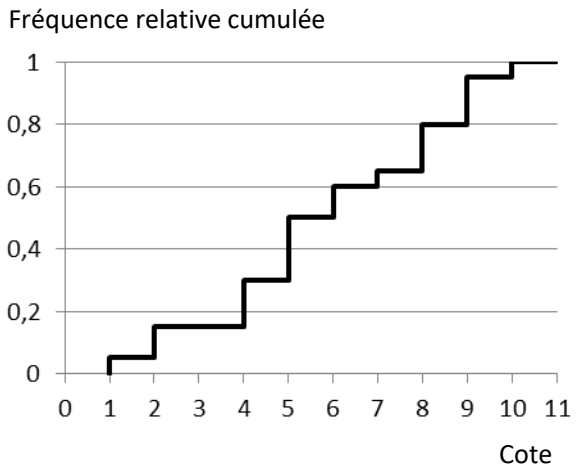
Les diagrammes en escaliers

Si les diagrammes en bâtons permettent de représenter les répétitions ou les fréquences relatives, les **diagrammes en escaliers**, eux, visent à présenter les répétitions et fréquences relatives cumulées. Ils se présentent sous la forme d'une ligne en escaliers dont chaque marche correspond à une modalité.

La ligne en escaliers part de 0 à gauche puis monte par paliers jusqu'à atteindre la valeur maximale (soit l'effectif total soit 100%). Chaque montée se situe au-dessus d'une modalité et a une longueur correspondant à la répétition ou à la fréquence relative de cette modalité.

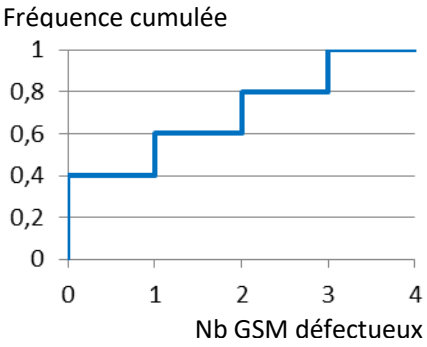
Ici encore, l'axe vertical peut être échelonné avec le nombre de répétitions cumulées ou avec les fréquences cumulées.





Exercice

7) Un ingénieur prélève un certain nombre de caisses à la sortie d'une usine et vérifie l'état des GSM qu'elles contiennent. Après avoir noté le nombre de GSM défectueux, il réalise le diagramme en escaliers suivants.



a) Sur les caisses examinées, quel pourcentage n'avait aucun GSM défectueux ?

b) Le patron estime qu'il est acceptable d'avoir au maximum 1 GSM défectueux par caisse. Quel pourcentage des caisses examinées satisfait aux exigences du patron ?

c) Sachant que l'ingénieur a contrôlé 40 caisses, combien parmi elles contenaient exactement 3 GSM défectueux ?

Le tableau de recensement complété (classes)

On peut également compléter les tableaux de recensement par classe. Les trois colonnes supplémentaires sont calculées et interprétées de manière similaire, comme le montre l'exemple suivant (durée du trajet domicile-travail).

Classe (durée en min)	**Centre c_i (min)**	**Répétition r_i**	**Rép. cum. rc_i**	**Fréquence f_i (%)**	**Frq. cum. g_i (%)**
[15,17[	16	3	3	15	15
[17,19[	18	2	5	10	25
[19,21[	20	4	9	20	45
[21,23[	22	9	18	45	90
[23,25[	24	2	20	10	100

Une fois encore, il est important de savoir interpréter chacun des nombres du tableau.

- Le centre c_2 = 18 indique que les résultats qui sont regroupés dans la classe [17,19[peuvent être représentés par la valeur 18 (minutes).

- La répétition r_5 = 2 indique que l'échantillon comporte 2 individus qui correspondent à la classe [23,25[, c'est-à-dire 2 personnes mettant entre 23 et 25 minutes pour effectuer le trajet domicile-travail.
- La répétition cumulée rc_3 = 9 indique que l'échantillon comporte 9 personnes mettant moins de 21 m pour effectuer le trajet domicile-travail.
- La fréquence (relative) f_1 = 15 % indique que 15 % des personnes de l'échantillon mettent entre 15 et 17 minutes pour effectuer le trajet domicile-travail.
- La fréquence cumulée g_4 = 90 % indique que 90 % des personnes de l'échantillon mettent moins de 23 minutes pour effectuer le trajet domicile-travail.

Remarque. Attention à l'interprétation correcte des répétitions et des fréquences cumulées : les valeurs indiquées correspondent aux individus de l'échantillon dont les résultats sont inférieurs à la borne supérieure de la classe, pas à son centre ! Ainsi, la répétition cumulée rc_5 = 20 signifie bien qu'il y a 20 individus qui mettent moins de 25 minutes pour effectuer le trajet (et pas moins de 24 minutes) !

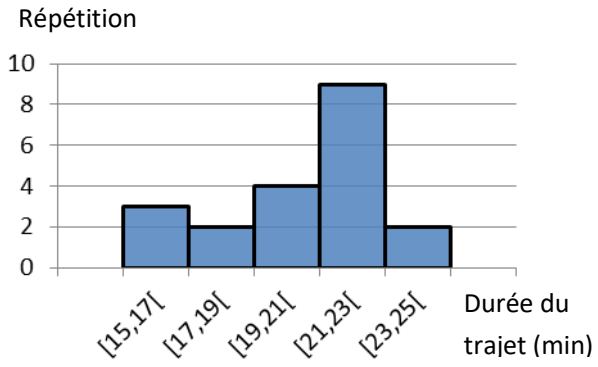
Exercices

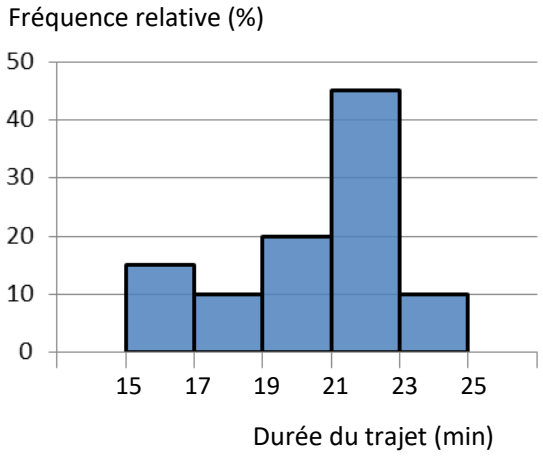
8) En vous aidant du tableau complété présenté ci-dessus, répondez aux questions suivantes.
   a) Parmi les personnes interrogées, combien mettent approximativement (à une minute près) 22 minutes pour se rendre de chez eux à leur lieu de travail ?
   b) Parmi les personnes interrogées, combien mettent moins de 23 minutes pour effectuer le trajet entre leur domicile et leur lieu de travail ?
   c) Quelle proportion des personnes interrogées mettent entre 21 et 23 minutes pour se rendre de chez eux à leur lieu de travail ?
   d) Quelle proportion des personnes interrogées mettent moins de 19 minutes pour aller de chez eux à leur lieu de travail ?
   e) Parmi les personnes interrogées, combien mettent plus de 19 minutes pour aller de chez elles à leur lieu de travail ?
   f) Quelle proportion des personnes interrogées mettent entre 17 et 23 minutes pour effectuer le trajet domicile-travail ?

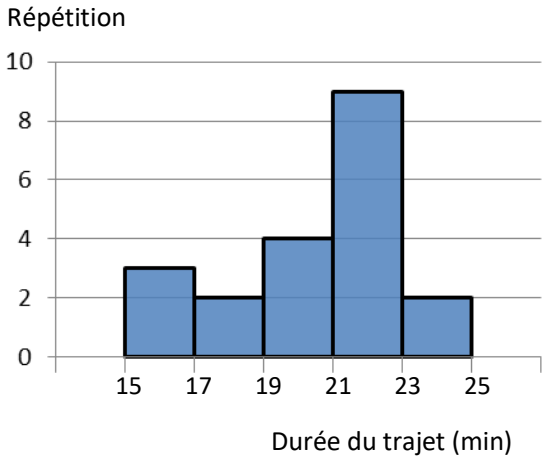
9) Complétez les tableaux de recensement par classes du module précédent.

Les histogrammes

Les **histogrammes** sont le pendant des diagrammes par bâtons dans le cas des classes. À chaque classe, on associe un rectangle dont la surface est proportionnelle à la répétition de la classe. Dans le cas où toutes les classes ont la même amplitude, cela revient à considérer que les hauteurs de ces rectangles doivent être proportionnelles aux répétitions.





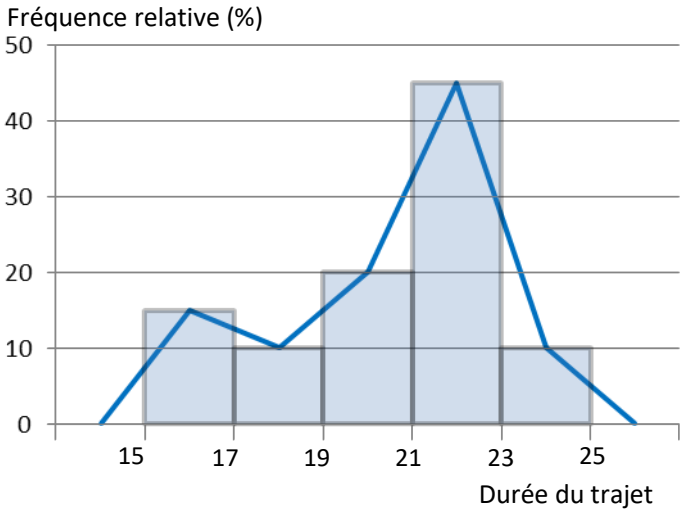


Les polygones des fréquences

On préfère parfois une version moins « carrée » des histogrammes appelée « **polygone des fréquences** ». Ce type de graphique s'obtient à partir de l'histogramme en réalisant les étapes suivantes :

(1) ajouter 2 classes vides, une à gauche de la première et une à droite de la dernière ;
(2) déterminer les points qui se trouvent au milieu des côtés supérieurs des rectangles ;
(3) relier ces points par des segments de lignes droites.

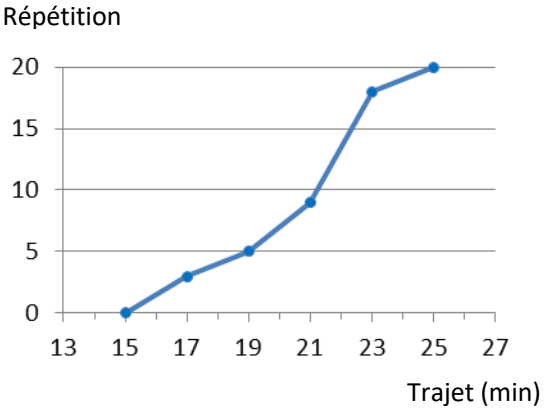
Voilà ce que cela donne sur l'exemple (l'histogramme est en grisé en arrière-fond).



Les diagrammes cumulatifs approchés

Finalement, les **diagrammes cumulatifs** (ou **diagramme des fréquences cumulées**) approchés visent à représenter sous forme graphique la manière dont des résultats statistiques regroupés par classes progressent, tout comme un diagramme en escaliers le fait pour des résultats discrets.

Dans le tableau recensé, on a associé à chaque classe $[a_i, b_i[$ une répétition cumulée rc_i qui indique le nombre d'individus de l'échantillon possédant un résultat inférieur à b_i. Pour construire le diagramme cumulatif approché, on représente tout d'abord tous les points de coordonnées (b_i, rc_i), puis on ajoute un point initial de coordonnées $(a_i, 0)$ et on relie tous ces points par des segments de droite.



Chaque segment de droite sur le diagramme cumulatif approché correspond à une classe $[a_i, b_i[$. Pour chaque limite de classe (sur l'axe horizontal), on trouve un point qui indique la répétition cumulée correspondante. (Alternativement, on peut également utiliser les fréquences relatives cumulées.)

Mais que se passe-t-il pour les autres valeurs, celles qui ne sont pas des limites de classes ? À ces autres valeurs-là correspond un point qui indique une approximation de la répétition (ou fréquence relative) cumulée. Il s'agit dans ce cas-ci d'une approximation linéaire : en traçant des segments de droite entre les points, on a supposé que les répétitions cumulées croissaient de manière constante / linéaire à l'intérieur de chacune des classes.

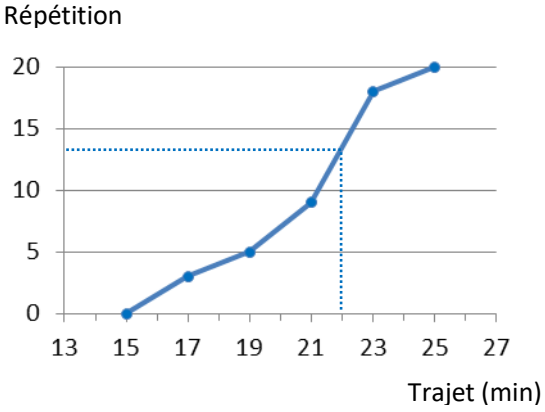
L'interpolation linéaire

L'observation mise en évidence à la fin de la section précédente indique qu'on suppose qu'au sein de chaque classe, les résultats sont distribués de manière homogène. Ainsi, au sein de la classe [21,23[, on suppose que les 9 résultats sont « bien éparpillés ». Concrètement, ce n'est quasiment jamais le cas, mais

cette perte de précision est un sacrifice inévitable lorsqu'on décide de passer des données brutes à des données rassemblées par classe.

On a déjà vu précédemment que les répétitions (et les fréquences relatives) cumulées permettaient de répondre à des questions telles que « Combien d'individus mettent moins de x minutes pour effectuer le trajet domicile-travail ? » lorsque la valeur x était une limite de classe. Et si x n'est pas une limite de classe ?

Dans ce cas-là, on peut utiliser l'interpolation linéaire, qui n'est rien d'autre que le pendant algébrique du fait de tracer des segments de droite dans le diagramme cumulatif approché. Dans l'exemple des trajets domicile-travail, considérons la question « Combien d'individus mettent moins de 22 minutes pour effectuer le trajet ? »



En utilisant le graphique, on peut trouver une réponse approximative. Mais on peut également obtenir une réponse plus précise (qui reste toutefois une estimation) en utilisant l'interpolation linéaire (qui ressemble fortement à une « règle de trois »).

Grâce au tableau, on sait que 9 personnes mettent moins de 21 minutes (pour faire le trajet domicile-travail) et que 18 personnes mettent moins de 23 minutes. Ainsi, si on passe de 21 minutes à 23 minutes (+2 minutes), on gagne 9 personnes.

Si on passe seulement de 21 minutes à 22 minutes, on ne fait que « la moitié du trajet » et donc, on ne gagne que 9/2 = 4,5 personnes. Ajoutées aux 9 personnes pour qui le trajet prend moins de 21 minutes, on trouve donc que 13,5 personnes ont un trajet de moins de 22 minutes.

Le fait que la réponse obtenue est « 13,5 personnes » montre bien qu'il s'agit là d'une estimation et non d'une réponse exacte. Pour trouver la réponse exacte, il faut retourner aux données brutes d'avant le regroupement en classes.

Exercices

10) Une étude portant sur un total de 1000 vaches laitières a rapporté que 400 d'entre elles produisaient 26 litres de lait par jour ou moins et que 600 d'entre elles produisaient 30 litres de lait par jour ou moins. Grâce à l'interpolation linéaire, estimez le nombre de vaches qui produisent 28 litres par jour ou moins.

11) On a réalisé une expérience visant à mesurer le temps de réaction de 500 conducteurs. Les résultats ont révélé que 200 d'entre eux ont mis moins de 8 dixièmes de seconde à réagir et que 350 d'entre eux ont mis moins de 13 dixièmes de secondes à réagir. Grâce à l'interpolation linéaire, estimez le nombre de conducteurs qui ont mis moins de 10 dixièmes de secondes à réagir.

Exercices (récapitulatifs)

12) Le service des loisirs de la Région Wallonne a réalisé récemment une enquête auprès de 500 ménages pour connaître le budget qu'ils allouaient chaque année aux vacances. Sur base de cette enquête, la répartition des ménages en fonction du montant consacré aux vacances est la suivante.

Montant alloué (× 100 €)	20	30	40	50	60	70
Nb de ménages	30	70	100	200	60	40

a) Indiquez le caractère étudié et dressez le tableau des répétitions complet.
b) Construisez le diagramme en bâtons et le diagramme en escaliers.
c) En utilisant les diagrammes et/ou le tableau de recensement, calculez le nombre de ménages dont le montant annuel alloué aux vacances est supérieur à 4 000 €.
d) En utilisant les diagrammes et/ou le tableau de recensement, calculez le nombre de ménages dont le montant annuel alloué aux vacances est au plus de 5 000 €.

13) Dans le cadre d'une enquête consacrée à la consommation, un enquêteur a demandé à plusieurs ménages d'indiquer le montant mensuel de leurs dépenses alimentaires. Il a reçu les réponses suivantes (en Euros).

540	642	392	428	472	372	450	620	542	464
482	572	398	532	626	472	566	604	508	320

a) Regroupez ces données en 5 classes et dressez le tableau des répétitions complet.
b) Construisez l'histogramme, le polygone des fréquences et le diagramme des fréquences cumulées.

14) Une usine produit des poutres de différentes longueurs. La production journalière est répartie comme suit.

Longueur (m)	[0, 2[	[2, 4[	[4, 6[	[6, 8[	[8, 10[
Nb de poutres	10	20	30	25	15

a) Dressez le tableau de recensement complet.
b) Représentez l'histogramme et le polygone des fréquences.
c) Chaque jour, combien l'usine produit-elle de poutres mesurant moins de 6 mètres ? Plus de 2 mètres ?
d) (Approximation linéaire) Combien l'usine produit-elle de poutres mesurant moins de 5 mètres ? Plus de 3 mètres ?

CHAPITRE 3 : PARAMÈTRES DE TENDANCE CENTRALE

Les sections précédentes traitaient de la récolte, du recensement et de la présentation des données statistiques. Jusqu'ici, il s'est agi avant tout de « mettre en forme » les résultats observés sur un échantillon. Dans cette section et la suivante, nous allons nous intéresser à l'analyse des données.

Cette analyse se fait sous la forme de valeurs significatives (appelées paramètres) calculées à partir des données. Chacune de ces valeurs donnera une indication sur la manière dont les données se comportent : les résultats sont-ils élevés ou faibles ? quel est le résultat qui revient le plus souvent ? les valeurs sont-elles fortement dispersées ou, au contraire, sont-elles très proches les unes des autres ?

On distingue deux grands types de paramètres : les paramètres de tendance centrale et les paramètres de dispersion. Les premiers, qui sont abordés dans cette section, visent à résumer l'ensemble des données à une seule valeur « représentative » (l'exemple le plus connu est sans doute celui de la moyenne). Les seconds, qui font l'objet de la section suivante, indiquent comment les résultats sont éparpillés autour de ces valeurs centrales.

Plan du module

1 Le mode

Le premier paramètre de tendance centrale, et sans doute le plus facile à définir, est le **mode**. Le mode est tout simplement la modalité (ou la classe) qui possède le plus grand nombre de répétitions ; c'est la modalité (ou la classe) qui est la plus représentée. Dans l'exemple des cotes d'interrogation (voir sections précédentes), le mode est la cote 5/10, avec une répétition de 4. Dans l'exemple des trajets domicile-travail, la classe modale est la classe [21,23[avec une répétition de 9.

Dans les cas d'ex-aequo, on peut se retrouver avec plusieurs modes (ou plusieurs classes modales). On parle alors de distribution et d'échantillon **multimodal(e)** (**bimodal(e)** s'il y en a deux, **trimodal(e)** s'il y en a trois…).

Pour que la notion de mode statistique soit véritablement significative, il faut que sa répétition soit nettement supérieure à celle des autres modalités.

Exercices (réflexion)

1) Quel(s) graphique(s) utiliser pour repérer le mode ou la classe modale ? Comment procéder ?

La moyenne

Si le mode est le paramètre central le plus facile à définir, la **moyenne** est, cependant, le plus connu. Son calcul dépend de la forme sous laquelle les données sont présentées.

Données brutes. À partir des données brutes x_1, x_2, …, x_n (dont certaines sont peut-être répétées), on calcule la moyenne grâce à la formule suivante.

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Données rassemblées par répétitions. Si on a dressé un tableau des modalités X_i et des répétitions r_i qui leur sont associées, la formule est légèrement différente. En effet, on doit tenir compte du nombre de fois que chacune des modalités apparaît.

$$\bar{x} = \frac{1}{n}\sum (X_i \times r_i) = \frac{X_1 r_1 + X_2 r_2 + \cdots + X_p r_p}{n}$$

Données rassemblées par classes. Finalement, si on a regroupé les résultats en des classes $[a_i, b_i[$ associées à des répétitions r_i, on calcule la moyenne en utilisant les centres c_i des classes. Il s'agit bien entendu d'une approximation car, au sein de chaque classe, les résultats ne sont pas forcément distribués « harmonieusement » autour du centre.

$$\bar{x} = \frac{1}{n}\sum (c_i \times r_i) = \frac{c_1 r_1 + c_2 r_2 + \cdots + c_p r_p}{n}$$

Contrairement au mode, la moyenne est difficile à déterminer précisément à partir des représentations graphiques (sauf dans des cas très particuliers).

Pratiquement, on peut calculer la moyenne dans les tableaux de recensement en ajoutant une colonne reprenant les produits $X_i r_i$ (pour chaque modalité) ou $c_i r_i$ (pour chaque classe), en calculant la somme des valeurs de cette colonne puis en divisant celle-ci par l'effectif n. Voici ce que ça donne en reprenant les exemples des sections précédentes (cotes d'interro et durée du trajet domicile-travail).

x_i	r_i	$x_i r_i$
1	1	1
2	2	4
3	0	0
4	3	12
5	4	20
6	2	12
7	1	7
8	3	24
9	3	27
10	1	10
		$\Sigma x_i r_i$ = 117
		$\bar{x}$ = 117/20 = 5,85

$[a_i, b_i[$	c_i	r_i	$c_i r_i$
[15,17[	16	3	48
[17,19[	18	2	36
[19,21[	20	4	80
[21,23[	22	9	198
[23,25[	24	2	48
			$\Sigma c_i r_i$ = 410
			$\bar{x}$ = 410/20 = 20,5

Exercices

2) Dans une entreprise, on constate que l'âge moyen des employées (femmes) est de 25 ans alors que l'âge moyen des employés (hommes) est de 27 ans. Peut-on en déduire que l'âge moyen des employés (tous sexes confondus) est de 26 ans ?

3) Calculez la moyenne de l'échantillon 2, 4, 4, 6, 7, 7, 8, 9, 9, 10.

4) Calculez la moyenne de l'échantillon décrit ci-dessous (résultats à une interrogation).

Note	5	6	7	8	9	10	11	12	13	14	15
Nb étudiants	1	3	9	4	3	1	4	1	2	1	3

5) Calculez la moyenne de l'échantillon présenté ci-dessous.

Classe	r_i
[1.5, 1.6[	3
[1.6, 1.7[	6
[1.7, 1.8[	12
[1.8, 1.9[	2
[1.9, 2[	1

Moyenne et transformation linéaire des données

Pour simplifier les calculs, on peut choisir de transformer les données avant de calculer la moyenne. Bien souvent, la transformation est d'un de ces deux types.

- On multiplie ou on divise toutes les données par une valeur constante : on remplace les données initiales x_i par les valeurs $a\ x_i$.
- On ajoute ou on soustrait une constante à toutes les données : on remplace les données initiales x_i par les valeurs $x_i + b$.

On peut aussi combiner ces transformations et remplacer les données x_i par les valeurs $a\ x_i + b$. On parle alors, de manière générale, d'une **transformation linéaire** des données.

Par exemple, une étude sur des salaires élevés pourrait produire des résultats tels que 100 000 €, 125 000 €, 115 000 €, 130 000 € et 145 000 €. Plutôt que de calculer la moyenne avec ces nombres à 6 chiffres, on pourrait choisir tout d'abord de leur soustraire 100 000 puis de diviser le reste par 5 000, c'est-à-dire de remplacer les x_i par

$$(x_i - 100\,000) / 5\,000 = (1/5)\ x_i - 20\,000.$$

Les valeurs deviennent alors 0, 5, 3, 6 et 9, des nombres beaucoup plus faciles à manipuler, dont la moyenne vaut 23/5.

Mais comment retrouver la « véritable » moyenne, celle des nombres de départ ? En observant que la moyenne a subi le même traitement que les données. Ainsi, si on note $\bar{x}$ la moyenne des salaires de départ, elle aussi a été remplacée par $(\bar{x} - 100\,000) / 5\,000 = 23/5$. En résolvant cette équation, on trouve $\bar{x}$ = 123 000 €.

De manière générale, si les données x_i ont été remplacées par $a\ x_i + b$, alors la moyenne $\bar{x}$ devient $a\ \bar{x} + b$.

Cette formule peut être utile non seulement pour simplifier les calculs mais également en cas de changement d'unités des données (pensez par exemple à la transformation d'un relevé de températures en degrés Fahrenheit en degrés Celsius).

Note. Ces transformations peuvent également s'opérer après avoir rassemblé les données par répétitions ou par classes : dans ce cas-là, c'est sur les modalités X_i ou sur les limites de classes a_i et b_i qu'on les effectue.

Exercices

6) En mesurant la quantité de lait produit quotidiennement par les vaches d'un troupeau, on a obtenu les résultats suivants : 24, 26, 26, 26, 28, 30, 30, 34. Calculez la production moyenne tout d'abord en utilisant les données telles quelles puis en opérant une transformation linéaire (ici, diviser par 2 puis soustraire 12).

7) Calculez la moyenne de l'échantillon décrit ci-dessous en utilisant une transformation linéaire.

X_i	r_i
18	6
19	10
20	4
21	2
22	2

8) En fin d'année, un professeur corrige un examen de Statistiques (côté sur 50 points) et se rend compte qu'une de ses classes obtient une moyenne de 22,5. Quelle moyenne aurait-il obtenu s'il avait d'abord pris la peine de ramener chacune des cotes sur 20 points ?

   Comme l'examen s'est déroulé dans de mauvaises conditions (bruits causés par des travaux de construction), le professeur décide d'ajouter 3 points à la cote (sur 50) de chacun des étudiants. Quelle moyenne obtiendra-t-il ?

9) Une étude révèle qu'en moyenne, les ménages d'une certaine ville belge remplissent 42 sacs poubelles par année. Si la ville en question décidait d'imposer une taxe annuelle de 0,5 € par sac augmentée d'une partie fixe de 3 € par ménage, combien celle-ci rapporterait-elle en moyenne et par ménage ?

10) On s'est intéressé aux sommes totales versées par 10 entreprises belges à leurs actionnaires sous la forme de dividendes au cours de l'année passée et on a obtenu les résultats suivants. On supposera que chacune de ces entreprises a calculé la part de ses profits à consacrer aux dividendes comme suit : une base forfaitaire de 1 000 € augmentée d'1/5 (donc 20%) des profits.

    Total des dividendes versés (× 1000 €) : 3, 2, 9, 2.5, 3.5, 8, 4.5, 7.5, 8, 5

    a) Calculez la somme moyenne consacrée aux dividendes par ces entreprises.

b) Calculez le profit moyen de ces entreprises.

c) Si on suppose que ces entreprises décident d'augmenter la base forfaitaire à 2 000 € et de diminuer la part des profits à 15%, déterminez la somme moyenne qu'elles consacreront aux dividendes cette année-ci, sachant que leurs profits auront augmenté de 5%.

La médiane

En géométrie des triangles, la médiane est une droite qui coupe un des côtés en deux parties égales. Dans le domaine des statistiques, cette idée se traduit comme suit : la **médiane** m est une valeur qui permet de diviser l'échantillon en deux parties égales.

Concrètement, ce n'est pas toujours possible de trouver une valeur qui coupe l'échantillon en deux parties parfaitement égales (c'est par exemple impossible si les données sont 2, 3, 3 et 4). On doit donc se ramener à une définition un peu plus souple.

Mathématiquement, la médiane est une valeur m telle que

- au moins 50% des observations sont ≤ m et
- au moins 50% des observations sont ≥ m.

La médiane dans le cas de données brutes

À partir des données brutes x_1, x_2, …, x_n (dont certaines sont peut-être répétées), on peut trouver la médiane m en accomplissant les étapes suivantes :

(1) ordonner les valeurs x_1, x_2, …, x_n de la plus petite à la plus grande ;

(2) si n est un nombre impair, la médiane est la valeur qui se trouve au milieu de la liste ;

(3) si n est un nombre pair, la médiane est la moyenne entre les deux valeurs centrales.

Considérons l'exemple suivant. On a mesuré la taille en cm des 10 enfants d'un groupe scolaire et on a obtenu les résultats suivants (déjà ordonnés par ordre croissant) :

108, 112, 120, 121, 122, 123, 123, 125, 126, 128.

Si on veut diviser le groupe en deux en mettant les plus petits d'un côté et les plus grands de l'autre, où fixer la limite ? Ici, on a affaire à un nombre pair de valeurs. La médiane sera donc la moyenne entre les deux valeurs au centre : 122 et 123. Elle vaudra donc 122,5 cm.

Si le groupe n'avait comporté que 9 enfants (oublions celui qui mesure 128 cm), la médiane aurait été la valeur centrale, à savoir 122 cm.

De manière générale, si les résultats x_1, x_2, …, x_n sont ordonnés par ordre croissant, la médiane m vaut

- $x_{(n+1)/2}$, le $\frac{n+1}{2}$-ème résultat si n est impair ; et
- $\frac{1}{2}(x_{n/2} + x_{n/2+1})$, la moyenne entre le $\frac{n}{2}$-ème résultat et le suivant si n est pair.

Exercice

11) Calculez la médiane des échantillons donnés ci-dessous.

a) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

b) 2, 3, 4, 5, 6, 7, 7, 8, 8, 8

c) 2, 4, 4, 6, 7, 7, 8, 9, 9, 10

La médiane dans le cas de données groupées par répétitions

Le principe de calcul est le même dans le cas de valeurs regroupées par répétitions. Il est alors souvent plus simple de se baser sur les répétitions cumulées.

X_i	r_i	rc_i
0	2	2
1	3	5
2	4	9
3	7	16
4	3	19

Dans l'exemple ci-contre, il y a un nombre impair ($n = 19$) de données. La médiane m est donc la valeur centrale, celle qui correspond à x_{10}.

La colonne des répétitions cumulées indique qu'il y a 9 valeurs inférieures ou égales à 2. La 10e valeur est donc un 3. On obtient m = 3.

Dans ce second exemple, il y a un nombre pair ($n = 20$) de données. La médiane m est donc la moyenne entre les deux valeurs centrales, qui correspondent aux indices 10 et 11.

La colonne des répétitions cumulées indique que x_{11} = x_{12} = 2. On obtient m = 2.

X_i	r_i	rc_i
0	2	2
1	5	7
2	6	13
3	4	17
4	3	20

X_i	r_i	rc_i
0	3	3
1	5	8
2	4	12
3	3	15
4	1	16

Dans ce troisième et dernier exemple, il y a également un nombre pair ($n = 16$) de données. La médiane m est donc la moyenne entre les deux valeurs centrales, qui correspondent aux indices 8 et 9.

La colonne des répétitions cumulées indique que x_8 = 1 et x_9 = 2. Ici, on obtient m = 1,5.

Exercices

12) Calculez la médiane des échantillons donnés ci-dessous (résultats à une interrogation).

Note	5	6	7	8	9	10	11	12	13	14	15
Nb étudiants	1	3	9	4	3	1	4	1	2	1	3

Note	5	6	7	8	9	10	11	12	13	14	15
Nb étudiants	1	3	8	4	3	1	4	1	3	6	3

13) Calculez la médiane des échantillons décrits ci-dessous.

X_i	r_i	rc_i	g_i
0	2	2	10%
1	4	6	30%
2	4	10	50%
3	7	17	95%
4	3	20	100%

X_i	r_i	rc_i	g_i
18	6	6	25%
19	10	16	67%
20	4	20	83%
21	2	22	92%
22	2	24	100%

La médiane dans le cas de données groupées par classes

Finalement, il nous reste à considérer le calcul de la médiane dans le cas de données regroupées par classes. Ici, le plus simple est sans doute d'examiner les fréquences relatives cumulées. Idéalement, la médiane m devrait correspondre à une fréquence relative cumulée de 50%.

Si la fréquence relative cumulée 50% apparaît dans le tableau, la médiane est tout simplement la valeur à laquelle elle est associée. Ainsi, par exemple, si une classe [10,14[a une fréquence relative cumulée g_i = 50%, cela signifie qu'il y a exactement 50% de l'échantillon en-dessous de 14. La moyenne est donc m = 14.

Cependant, la plupart du temps, le nombre 50% n'apparaît pas dans le tableau. Dans ce cas-là, il faut procéder par interpolation linéaire.

Classe (durée en min)	**Centre (min)**	**Frq. cum. (%)**
[15,17[	16	15
[17,19[	18	25
[19,21[	20	45
[21,23[	22	90
[23,25[	24	100

Si on reprend l'exemple des trajets domicile-travail (voir tableau ci-contre), on se rend compte que la médiane doit se trouver quelque part entre 21 et 23. En effet, la colonne des fréquences relatives cumulées indique qu'il y a 45% de l'échantillon en-dessous de 21 et 90% en-dessous de 23.

Si on place la limite à 21 minutes maximum, on se retrouve avec 45% de l'échantillon. Si on avance cette limite jusqu'à 23 minutes (+2 minutes), on passe à 90% de l'échantillon : on gagne alors 45% supplémentaires. Pour trouver la médiane, notre but est d'atteindre 50%, c'est-à-dire de ne gagner que 5% supplémentaires (soit 1/9). Au lieu d'avancer la limite de 2 minutes, on ne va donc l'avancer que de 2/9 minutes. La valeur ainsi obtenue, c'est-à-dire 21 minutes + 2/9 minutes = 21,22 min, est la médiane car 50% de l'échantillon se trouve avant elle.

Exercices

14) Calculez la médiane des échantillons présentés ci-dessous.

Classe	r_i	g_i **(%)**
[1.5, 1.6[	3	12.5
[1.6, 1.7[	6	37.5
[1.7, 1.8[	12	87.5
[1.8, 1.9[	2	95.83
[1.9, 2[	1	1

Classe	r_i
[10, 20[	5
[20, 30[	10
[30, 40[	15
[40, 50[	20
[50, 60[	15
[60, 70[	5

Classe	r_i
[10, 20[	70
[20, 30[	30
[30, 50[	180
[50, 90[	240
[90, 100[	40

15) À partir de quel(s) graphique(s) peut-on retrouver la médiane ? Comment procéder ?

16) Une étude s'est intéressée aux dépenses annuelles en emplois de personnel à domicile pour un certain nombre de ménages (tableau ci-dessous).
Déterminez la médiane (a) sur un graphique, (b) par calcul.

Dépenses (€)	[300, 400[	[400, 500[	[500, 600[	[600, 700[	[700, 800[	[800, 1000[

Nb ménages	5	60	15	95	30	5

Exercices (paramètres de tendance centrale)

17) Déterminez le mode, la moyenne et la médiane pour les échantillons suivants.

2 4 5 5 6 6 6 6 7
8 8 8 9 9 10 10 11

x_i	r_i
2	4
3	5
4	8
5	0
6	4
7	3

$[a_i, b_i[$	r_i
[10,20[	5
[20,30[	10
[30,40[	15
[40,50[	20
[50,60[	15
[60,70[	5

$[a_i, b_i[$	r_i
[10,20[	70
[20,30[	30
[30,50[	180
[50,90[	240
[90,100[	40

18) Un magasin de chaussures étudie les ventes effectuées au cours d'une semaine, et plus particulièrement les pointures de chaque paire vendue. Voici les résultats de cette observation. Calculez le mode, la moyenne et la médiane. Lequel des trois paramètres est le plus significatif pour le responsable du magasin qui doit choisir quelle pointure commander pour renouveler son stock ?

Pointure	37	38	39	40	41	42	43	44	45
Nb ventes	6	5	6	5	8	5	4	2	2

19) Un physicien réalise une expérience pratique sur la dilatation des métaux en mesurant (aussi précisément que possible) un bâtonnet de fer chauffé à 60°. Pour éviter les erreurs de lecture et les variations dues aux conditions de l'expérience, il répète celle-ci 6 fois. Voici les mesures qu'il obtient : 10.321 cm, 10.324 cm, 10.323 cm, 10.324 cm, 10.325 cm, 10.327 cm. Calculez le mode, la moyenne et la médiane. Lequel des trois paramètres est le plus significatif dans ce cas-ci ?

20) Le tableau suivant reprend le taux d'intérêts (en %) des dépôts d'épargne proposés par 20 banques. Construisez le tableau de recensement puis trouvez le mode, la moyenne et la médiane. Interprétez ensuite (par une courte phrase) chacun de ces résultats.

4 6 7 4 10 8 7 5 7 5
5 9 6 5 4 7 5 8 5 6

21) Le tableau suivant reprend la répartition des ouvriers d'une entreprise en fonction de leur salaire net.

Salaire horaire net (€)	[16,17[	[17,18[	[18,19[	[19,20[	[20,21[	[21,22[	[22,23[
Nb d'employés	4	14	18	28	20	12	4

a) Dressez le tableau de recensement complet.
b) Calculez le mode, la moyenne et la médiane puis interprétez chacun de ces paramètres.

c) Dans une entreprise concurrente, chaque employé reçoit un salaire horaire 20% plus élevé et perçoit, en plus, un bonus fixe de 2€ par heure. Quel est le salaire horaire moyen d'un employé dans cette entreprise concurrente ?

22) On a observé les taux de chômage (en pourcentage de la population active) dans 24 secteurs d'activité. Voici les résultats obtenus. Dressez le tableau de recensement, puis calculez les paramètres de tendance centrale et interprétez-les.

8	8	9	8	9	8	9	8	14	8	8	9
6	11	6	9	9	8	8	9	14	9	11	10

23) En étudiant le Q.I. de 480 enfants de 5 ans, on a obtenu les résultats suivants. Dressez un tableau de recensement complet, calculez les paramètres de tendance centrale puis interprétez-les.

QI	Nombre	QI	Nombre	QI	Nombre
70	4	90	80	110	27
74	9	94	66	114	18
78	16	98	77	118	11
82	28	102	54	122	5
86	45	106	38	126	2

24) On a demandé aux employés d'une petite entreprise d'indiquer leur salaire mensuel brut. Trois personnes ont répondu 2 000 €, cinq personnes ont répondu 2 500 € et deux personnes, 3 000 €. Que valent le mode, la moyenne et la médiane ?

Si on ajoute le salaire du patron de l'entreprise (10 000 €), comment évoluent les paramètres de tendance centrale ? Quels sont les paramètres qui sont très sensibles aux valeurs extrêmes (outliers) et ceux qui le sont moins ?

25) Un individu passe son temps en lançant les 4 pièces de monnaie qu'il a dans sa poche et en notant sur un bout de papier combien d'entre elles retombent sur « pile ». Il effectue ainsi 16 lancers et obtient les résultats suivants : 1 fois « 0 pile », 4 fois « 1 pile », 6 fois « 2 pile », 4 fois « 3 pile » et 1 fois « 4 pile ». Dans le cas de cette distribution parfaitement symétrique, que valent le mode, la moyenne et la médiane ?

26) Un professeur termine la correction d'un examen coté sur 100. La moyenne de la classe est de 65 points.
   a) Comme les questions étaient particulièrement difficiles, il décide d'ajouter 5 points à tout le monde. Que devient la moyenne ?
   b) Comme le cours vaut 2 crédits, il doit encoder les points sur un maximum de 20. Que devient la moyenne ?

27) Le tableau suivant représente la distribution du nombre de kilomètres parcourus par les camions d'une compagnie de transport pendant l'année passée.

Kilométrage	Nb de camions
[10000, 14000[	5
[14000, 18000[	10

[18000, 22000[	12
[22000, 26000[	24
[26000, 30000[	26
[30000, 34000[	36
[34000, 38000[	39
[38000, 42000[	48

a) Dressez le tableau de recensement complet.
b) Quel est le nombre de camions qui ont parcouru moins de 25 000 km ?
c) Calculez et interprétez la médiane de cet échantillon puis confirmez le résultat en traçant le graphique adéquat.
d) En effectuant un changement de variable approprié, calculez la moyenne arithmétique de cet échantillon.

CHAPITRE 4 : PARAMÈTRES DE DISPERSION

Les paramètres de tendance centrale présentés dans le module précédent permettent de résumer un échantillon en une valeur : celle qui revient le plus souvent (le mode), celle qui permet de séparer les résultats en deux groupes de même taille (la médiane) ou celle qui approxime au mieux l'ensemble des résultats (la moyenne).

La moyenne, aussi significative soit-elle, ne permet pas de savoir comment les données sont distribuées autour d'elles : sont-elles toutes agglutinées tout près de la valeur centrale ou, au contraire, sont-elles éparpillées à bonne distance de celle-ci ?

Considérons par exemple les deux situations suivantes (les résultats d'une interrogation sur 20 points), qui sont bien différentes mais qui, pourtant, possèdent la même moyenne.

Interro 1	9	9	9	9	10	10	10	11	11	11	11
Interro 2	0	2	4	6	8	10	12	14	16	18	20

C'est pour déceler ce genre de différence qu'on introduit les paramètres de dispersion.

Plan du module

1. L'étendue
2. Les quantiles (données groupées par classes)
3. La boîte à moustaches
4. La variance et l'écart-type
5. Le coefficient de variation

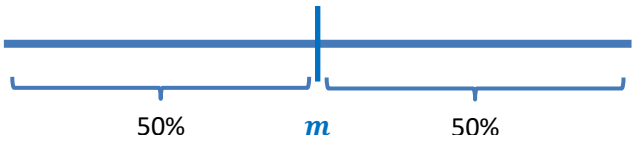
1 L'étendue

La mesure la plus intuitive de la dispersion est sans doute l'**étendue** de l'échantillon. Il s'agit tout simplement de la « longueur » sur laquelle les valeurs de l'échantillon s'étendent, à savoir la différence entre la valeur maximale et la valeur minimale. Ainsi, dans l'exemple donné ci-dessus, le premier échantillon (interro 1) a une étendue de 2 alors que le second (interro 2) a une étendue de 20.

L'étendue permet de savoir à quel point les données s'éloignent de la moyenne et de connaître l'intervalle de valeurs qu'elles recouvrent, mais ce paramètre n'apporte aucune information sur la manière dont les résultats se répartissent au sein de cet intervalle : les données pourraient être éparpillées de manière régulière ou toutes concentrées en un même endroit.

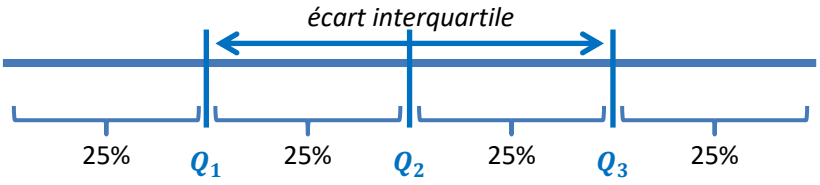
2 Les quantiles

Les **quantiles** (aussi appelés **fractiles**) sont une généralisation de la médiane. Pour rappel, la médiane m d'un échantillon est une valeur censée permettre de diviser l'échantillon en deux parties, avec (au moins) 50% des résultats avant la médiane et (au moins) 50% des résultats après.



Mais rien ne nous oblige à nous arrêter à une découpe de l'échantillon en deux parties !

Quartiles et écart interquartile. On pourrait choisir de le découper en 4 parties contenant chacune (grosso modo) 25% de l'échantillon. Les valeurs qui permettent cette découpe sont les **quartiles** Q_1, Q_2 et Q_3, qui correspondent respectivement aux fréquences relatives cumulées 25%, 50% et 75%. Le second quartile, Q_2, n'est rien d'autre que la médiane m.



On définit également l'écart interquartile comme la distance entre le premier et le troisième quartile, c'est-à-dire comme $Q_3 - Q_1$. C'est la longueur de l'intervalle dans lequel on retrouve 50% des résultats, les résultats les plus « centraux ».

Déciles. Les **déciles** D_1, D_2, …, D_9 sont 9 valeurs qui permettent de diviser l'échantillon en 10 parties correspondant chacune à (plus ou moins) 10% des résultats. Ils correspondent donc aux fréquences relatives cumulées 10%, 20%, …, 90%. Le cinquième décile, D_5, n'est rien d'autre que la médiane m.

Centiles. On pousse parfois la découpe encore plus loin en définissant les **centiles** C_1, C_2, …, C_{99}, des valeurs qui découpent l'échantillon en portions de 1%. Le 50e centile, C_{50}, n'est rien d'autre que la médiane m.

Quantile d'ordre α. Les quartiles, déciles et centiles présentés ci-dessus sont des cas particuliers de la notion de quantile. De manière générale, on parle de **quantile d'ordre α** pour désigner la valeur Qu_α telle que

- il y a ait (au moins) une proportion α de l'échantillon située avant Qu_α et
- il y ait (au moins) une proportion $1-\alpha$ de l'échantillon située après Qu_α.

Par exemple, la valeur $Qu_{60\%}$ sera située après 60% de l'échantillon (il s'agira en fait de D_6) alors que la valeur $Qu_{95\%}$ sera telle que 95% de l'échantillon lui sera inférieur ou égal.

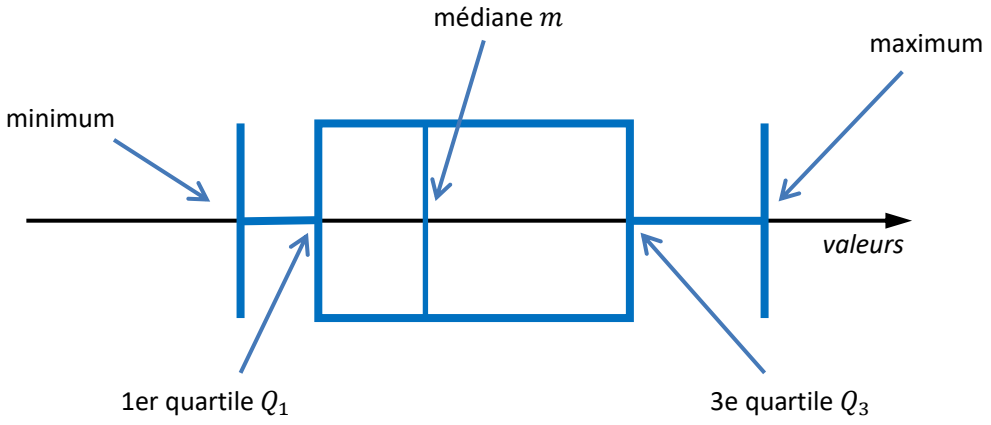
Notation	Signification
Q_1, Q_2, Q_3	les quartiles ($Q_2 = m$) situés à 25%, 50% et 75%
$D_1, \dots, D_9$	les déciles ($D_5 = m$) situés à 10%, 20%, …, 90%
$C_1, \dots, C_{100}$	les centiles ($C_{50} = m$) situés à 1%, 2%, …, 99%
Qu_α	le quantile d'ordre α situé à $(\alpha \times 100)$ %

Calcul des quantiles. Dans tous les cas, les quantiles se calculent exactement comme la moyenne. Ils sont surtout utilisés avec des données regroupées par classes, où le calcul s'effectue par interpolation linéaire. On peut également approximer la valeur des quantiles en utilisant un diagramme cumulatif.

3 La boîte à moustaches

La **boîte à moustaches** est un résumé graphique reprenant les informations suivantes : le minimum et le maximum de l'échantillon, sa médiane, ses quartiles et son écart interquartile. Toutes ces valeurs sont repérées sur une droite représentant les valeurs possibles.

Elle prend la forme d'une boîte rectangulaire s'étendant du premier quartile au troisième quartile et divisée en deux par un repère situé au niveau de la médiane, le tout accompagné d'un « T » à gauche indiquant le minimum et d'un « T » à droite indiquant le maximum.



Exercices

1) On a examiné le montant de l'épargne de 100 ménages lors du dernier mois de janvier (c'est-à-dire la différence entre leurs revenus et leurs dépenses de consommation). Le tableau suivant présente les données recueillies.

Épargne (€)	Nb de ménages
[-1000, -500[	5
[-500, 0[	20
[0, 500[	20
[500, 1000[	40
[1000, 1500[	15

a) Complétez le tableau de recensement.
b) Calculez la médiane, les quartiles et l'écart interquartile.
c) Dessinez la boîte à moustache.
d) Calculez le second décile. Que représente ce nombre ?
e) Les autorités souhaitent envoyer un dépliant contenant divers conseils sur l'épargne à certains des ménages concernés par l'étude de l'exercice précédent. Suite à une erreur d'impression, seuls 40 exemplaires du dépliant ont été imprimés. Si on souhaite cibler les ménages ayant la moins bonne épargne, à quelle valeur doit-on fixer la limite (on enverra un dépliant à tous les ménages dont l'épargne se situe sous cette limite) ?
f) Une banque souhaite envoyer des propositions de placement aux 25 ménages ayant réalisé la meilleure épargne au cours du dernier mois de janvier, c'est-à-dire à tous les ménages dont l'épargne est supérieure ou égale à un certain montant-limite. Calculez ce montant-limite.

2) On a mesuré la taille des étudiants de deux classes de 24 étudiants. Calculez les quartiles dans chaque cas et dessinez les boîtes à moustaches sur un même axe.

Taille (m)	Nb étudiants (classe 1)	Nb étudiants (classe 2)
[1.5, 1.6[	3	10
[1.6, 1.7[	6	1
[1.7, 1.8[	12	0
[1.8, 1.9[	2	0
[1.9, 2[	1	13

3) Une étude portant sur 100 entreprises belges a révélé que, parmi celles-ci, 20 payaient un impôt des sociétés de moins de 15 000 € ; 10 payaient un impôt situé entre 15 000 € et 30 000 € ; 40 payaient un impôt situé entre 30 000 € et 45 000 € et que le 30 autres payaient un impôt entre 45 000 € et 60 000 €.
a) Dressez un tableau de recensement complet.
b) Calculez l'impôt moyen pour ces 100 entreprises.
c) Déterminez le troisième quartile. Qu'est-ce que ce nombre représente ?
d) Complétez la phrase suivante au plus juste : « Sur les 100 entreprises examinées, 50 payaient un impôt des sociétés inférieur ou égal à …€ »

4 La variance et l'écart-type

Pour savoir à quel point les données x_i de l'échantillon s'éloignent de la moyenne $\bar{x}$, on aurait pu s'intéresser aux écarts $|x_i - \bar{x}|$. On aurait pu alors calculer la moyenne de tous ces écarts afin de savoir, « en moyenne », à quel point les résultats s'éloignent de $\bar{x}$.

Malheureusement, d'un point de vue mathématique, les valeurs absolues ne permettent pas les développements nécessaires aux parties plus avancées des statistiques. Alors, plutôt que de considérer la moyenne des écarts entre les données et $\bar{x}$, on s'intéresse plutôt à la moyenne des *carrés* de ces écarts.

Définition de la variance et de l'écart-type

L'écart entre une donnée x_i et la moyenne $\bar{x}$ vaut $|x_i - \bar{x}|$. Si on calcule le carré de cet écart, on peut oublier la valeur absolue : cela donne $(x_i - \bar{x})^2$. La moyenne des carrés des écarts, qu'on appelle la **variance**, vaut donc

$$\frac{1}{n}\sum (x_i - \bar{x})^2$$

Pour revenir à une valeur plus proche de la « moyenne des écarts », on calcule souvent la racine carrée de la variance, qu'on appelle l'**écart-type**.

L'écart-type n'est donc pas exactement la moyenne des écarts entre les données x_i et la moyenne $\bar{x}$ mais c'est un bon indicateur de cette valeur. Plus l'écart-type est grand et plus les valeurs sont éloignées autour de la moyenne. Plus l'écart-type est petit et plus les valeurs sont concentrées près de la moyenne.

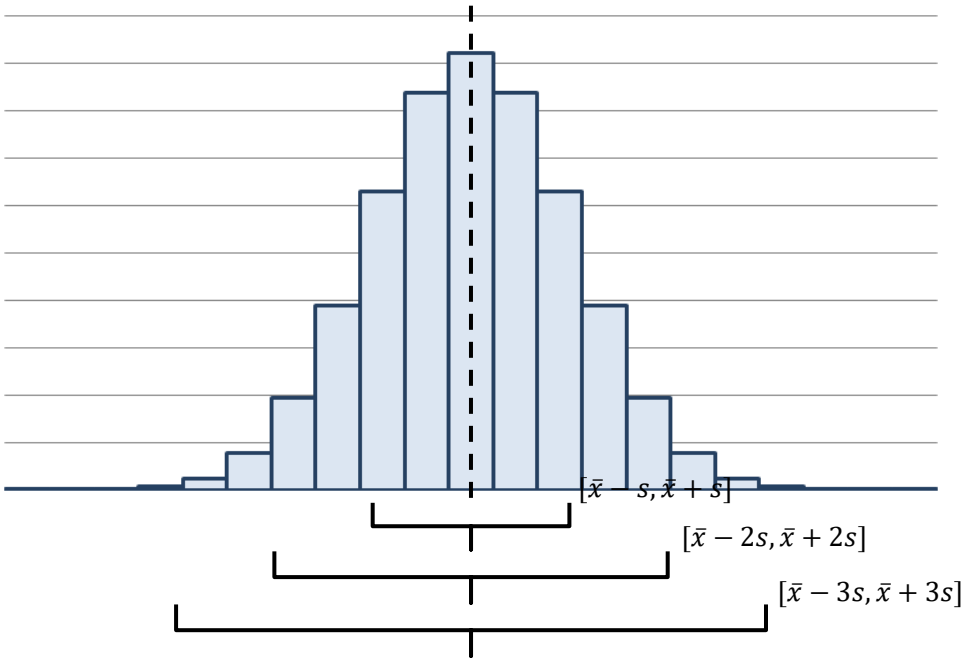
On note généralement s (ou s_x) l'écart-type et donc, tout naturellement, s^2 (ou s_x^2) la variance. En tant que moyenne des écarts au carré, la variance est toujours un nombre positif : si vous trouvez une variance négative, c'est qu'il y a une erreur dans les calculs !

Notation	Signification	Calcul
s_x^2 ou s^2	la variance	$s_x^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$
s_x ou s	l'écart-type	$s_x = \sqrt{s_x^2}$

Si les données sont exprimées dans une unité, l'écart-type (qui représente des différences entre des données) utilise lui aussi cette unité. La variance, quant à elle, utilise le carré de cette unité. Ainsi, si les données sont formulées en mètres, la variance s'écrira en mètres-carrés et l'écart-type, en mètres.

On prend souvent comme référence une distribution dite « normale » ou « de Gauss » (voir seconde partie du cours de Statistiques). Il s'agit d'une distribution symétrique et parfaitement équilibrée qui se développe autour d'une valeur modale centrale qui est également sa moyenne (voir l'histogramme suivant). De nombreux échantillons statistiques correspondent plus ou moins à une distribution de ce genre.

$\bar{x}$



Dans la distribution normale, on constate que près de 70% des données se trouvent à moins d'un écart-type de distance de la moyenne, c'est-à-dire dans l'intervalle $[\bar{x} - s, \bar{x} + s]$. Si on agrandit cet intervalle pour accepter les valeurs qui se trouvent à 2 écarts-types de la moyenne (à gauche ou à droite), on englobe alors à peu près 95% des données ! Finalement, si on prend toutes les valeurs situées à au plus 3 écarts-types de la moyenne, on recouvre presque 99,8% des données, la quasi-totalité de l'échantillon.

Dans une distribution normale, on peut donc interpréter l'écart-type comme l'écart (par rapport à la moyenne) que ne dépassent pas 70% des données. Bien évidemment, ces valeurs numériques ne sont valables que pour les distributions normales parfaites, mais elles peuvent tout de même donner une idée de l'importance de l'écart-type pour les distributions proches de cette courbe parfaite.

Calcul de la variance

Si on reprend les deux exemples donnés au début de cette section, on peut calculer leur variance pas à pas comme suit.

I1	x_i	9	9	9	9	10	10	10	11	11	11	11	
	$x_i - \bar{x}$	-1	-1	-1	-1	0	0	0	1	1	1	1	Somme
	$(x_i - \bar{x})^2$	1	1	1	1	0	0	0	1	1	1	1	8
I2	x_i	0	2	4	6	8	10	12	14	16	18	20	
	$x_i - \bar{x}$	-10	-8	-6	-4	-2	0	2	4	6	8	10	Somme
	$(x_i - \bar{x})^2$	100	64	36	16	4	0	4	16	36	64	100	440

Pour la première interrogation, on trouve donc une variance de 8/11 = 0,73 et un écart-type de 0,85. Pour la seconde interrogation, on obtient une variance de 440/11 = 40 et un écart-type de 6,32. L'écart-type, beaucoup plus grand dans le second cas, montre bien que les cotes de la seconde interrogation sont plus dispersées que celles de la première.

Une formule alternative permet de faciliter le calcul de la variance :

$$s^2 = \overline{x^2} - \bar{x}^2$$

Ou, en français, par « La variance vaut la moyenne des carrés des données *moins* le carré de la moyenne. » Pour l'utiliser, il faut

- calculer le carré de chacune des données,
- puis la moyenne de ces carrés
- et finalement soustraire à cette valeur le carré de la moyenne.

Dans le cas des deux interrogations, on trouve :

I1	x_i	9	9	9	9	10	10	10	11	11	11	11	Somme
	x_i^2	81	81	81	81	100	100	100	121	121	121	121	1108
I2	x_i	0	2	4	6	8	10	12	14	16	18	20	Somme
	x_i^2	0	4	16	36	64	100	144	196	256	324	400	1540

Pour la première interrogation, on trouve $\overline{x^2}$ = 1108/11 = 100,73 et, pour la variance, 100,73 – 10^2 = 0,73. Pour la seconde interrogation, on a $\overline{x^2}$ = 1510/11 = 140 et, pour la variance, 140 – 10^2 = 40. On retombe donc bien sur les mêmes résultats que ci-dessus.

La formule indiquée ci-dessus est valable dans tous les cas, qu'on utilise des données brutes, des données rassemblées par répétitions ou des données regroupées par classe. Dans tous les cas, il suffit de calculer la moyenne des valeurs (des modalités, des centres de classe) au carré puis de soustraire à ce résultat le carré de la moyenne.

Concrètement, cela peut se faire de manière similaire au calcul de la moyenne :

- on ajoute une colonne où on calcule les produits $x_i^2 r_i$ (pour les modalités) ou $c_i^2 r_i$ (pour les classes) ;
- on calcule la somme de ces valeurs ;
- on divise cette somme par l'effectif n pour obtenir la moyenne des carrés $\overline{x^2}$;
- on calcule la variance en utilisant la formule $s^2 = \overline{x^2} - \bar{x}^2$.

Voici ce que cela donne dans le cas des deux exemples des sections précédentes (côtes d'interro et durée du trajet domicile-travail).

x_i	r_i	$x_i r_i$	$x_i^2 r_i$
1	1	1	1
2	2	4	8
3	0	0	0
4	3	12	48
5	4	20	100
6	2	12	72
7	1	7	49
8	3	24	192
9	3	27	243
10	1	10	100
$\Sigma x_i r_i$ = 117 donc $\bar{x}$ = 117/20 = 5,85			
$\Sigma x_i^2 r_i$ = 813 ; $\overline{x^2}$ = 813/20 = 40,65			
Donc s^2 = 40,65 – $5{,}85^2$ = 6,43			

$[a_i, b_i[$	c_i	r_i	$c_i r_i$	$c_i^2 r_i$
[15,17[	16	3	48	768
[17,19[	18	2	36	648
[19,21[	20	4	80	1600
[21,23[	22	9	198	4356
[23,25[	24	2	48	1152
$\Sigma c_i r_i$ = 410 donc $\bar{x}$ = 410/20 = 20,5				
$\Sigma c_i^2 r_i$ = 8524 ; $\overline{x^2}$ = 8524/20 = 426,2				
Donc s^2 = 426,2 – $20{,}5^2$ = 5,95				

Notez que, si on a déjà ajouté une colonne $x_i r_i$ ou $c_i r_i$ pour le calcul de la moyenne, il suffit de multiplier les valeurs de cette colonne par x_i pour par c_i pour obtenir $x_i^2 r_i$ ou $c_i^2 r_i$: inutile de reprendre les calculs à partir de zéro !

Variance, écart-type et transformation linéaire

Le module précédent a introduit la notion de transformation linéaire des données, où on remplaçait les valeurs x_i par $a\ x_i + b$ (soit pour faciliter les calculs soit pour effectuer un changement d'unité ou autre). On peut prouver mathématiquement que, si on effectue une telle transformation, la variance s^2 devient $a^2\ s^2$ et l'écart-type, s, devient $a\ s$.

L'écart-type est donc simplement multiplié par la constante a. Le déplacement « $+b$ » ne modifie pas l'écart-type (ni la variance). C'est assez proche de l'intuition car ce déplacement revient à accroître ou à décroître toutes les données de la même valeur, ce qui ne modifie en rien les écarts entre les données.

5 Le coefficient de variation

Si on dispose de deux échantillons, on peut les comparer en utilisant différents critères. Prenons l'exemple de deux produits distincts dont on relève le prix dans divers magasins.

Si on compare les moyennes, on tente de situer les valeurs d'un des échantillons par rapport à celles de l'autres. Cela permet de voir si l'un des produits est significativement plus cher que l'autre ou si leurs prix sont « en moyenne » similaires.

Si on compare les variances (ou les écarts-types), on tente de voir à quel distance les résultats s'éloignent de la moyenne. Si les variances sont quasiment identiques, cela signifie que les fourchettes de prix pour les deux produits sont de même taille (par exemple, que le prix varie dans un intervalle de 20€ autour de la moyenne).

Dans certains cas, ces comparaisons ne sont pas suffisamment pertinentes. Imaginons que le premier article étudié soit un jouet dont le prix moyen est de 10,00 € et que le second soit un composant d'ordinateur dont le prix moyen est de 150,00 €.

Article 1 (€)	10,00	9,00	12,00	8,50	10,50
Article 2 (€)	150,00	149,00	152,00	148,50	150,50

Si les résultats obtenus sont ceux du tableau ci-dessus, on remarque que les deux échantillons ont exactement la même variance (1,5 €^{2}) et le même écart-type, à savoir 1,22 €. La « fourchette des prix » a la même taille dans les deux cas : de -1,50 € à +2 € autour de la moyenne. Cependant, comme l'article 2 coûte beaucoup plus cher que l'article 1, cette fourchette indique en fait une variabilité beaucoup plus grande pour l'article 1 (de -15% à +20%) que pour l'article 2 (de -1% à +1,3%).

Dans ce genre de situation, il est préférable de considérer non pas les écart-types mais plutôt les rapports entre les écarts-types et les moyennes. Cette valeur est appelée le **coefficient de variation** et s'exprime généralement sous la forme d'un pourcentage. Pour l'article 1, on obtient un coefficient de variation valant 12,25% alors que pour l'article 2, on a un coefficient de variation de 0,8%.

Notation	Signification	Calcul
CV	le coefficient de variation	$CV = \frac{s_x}{\bar{x}}$

Exercices

4) Calculez la variance, l'écart-type et le coefficient de variation des échantillons suivants.

X_i	r_i
0	2
1	4
2	4
3	7
4	3

X_i	r_i
18	6
19	10
20	4
21	2
22	2

Épargne (€)	Nb de ménages
[-1000, -500[	5
[-500, 0[	20
[0, 500[	20
[500, 1000[	40
[1000, 1500[	15

5) On a mesuré la taille des étudiants de deux classes. Calculez les coefficients de variation et.

Taille (m)	Nb étudiants (classe 1)	Nb étudiants (classe 2)
[1.5, 1.6[	3	10
[1.6, 1.7[	6	1
[1.7, 1.8[	12	0
[1.8, 1.9[	2	0
[1.9, 2[	1	13

6) Calculez la moyenne, la variance, l'écart-type et le coefficient de variation des échantillons suivants (utilisez des transformations linéaires adéquates).

Prix de vente	Nb de ventes
25000	13
50000	41
75000	58
100000	22
125000	12
150000	4

Salaire journalier (€)	Nb d'employés
[75, 87.5[	6
[87.5, 100[	10
[100, 112.5[	16
[112.5, 125[	12
[125, 137.5[	4
[137.5, 150[	2

7) Voici les résultats d'une étude portant sur le montant alloué par divers ménages aux activités extrascolaires au cours de l'année passée. Calculez la moyenne, la variance et l'écart-type.

Montant (€)	Nb de ménages
[100, 170[	6
[170, 240[	13
[240, 310[	21
[310, 380[	46
[380, 450[	49
[450, 520[	42
[520, 590[	14
[590, 660[	6
[660, 730[	3

8) On a relevé les températures suivantes ces 10 derniers jours : 12°, 25°, 14°, 13°, 24°, 23°, 28°, 16°, 17°, 20°. Calculez la moyenne, la variance, l'écart-type et le coefficient de variation.

   Ces températures sont exprimées en degrés Celsius. Sachant que, pour les convertir en degrés Fahrenheit, il faut utiliser la formule $t_F = (9/5)\ t\_C\ + 32$, quelle moyenne, quelle variance,

quel écart-type et quel coefficient de variation aurait-on obtenu si on était parti de données exprimées en degrés Fahrenheit ?

9) Une étude portant sur la longueur (en centimètres) des poutres fabriquées dans une usine a indiqué une moyenne de 150 et une variance de 36.
   a) Précisez les unités de ces deux valeurs.
   b) Calculez également l'écart-type et le coefficient de variation (en précisant leurs unités).
   c) Si les mesures avaient été effectuées en mètres, quelle moyenne, quelle variance, quel écart-type et quel coefficient de variation aurait-on obtenu ?
   d) Si, avant la mesure, on avait raboté les extrémités de chacune des poutres de 2 cm (pour une perte totale de 4 cm), quelle moyenne, quelle variance, quel écart-type et quel coefficient de variation aurait-on obtenu ?

10) Une étude portant sur les chiffres d'affaires en Belgique et aux USA de diverses entreprises implantées dans les deux pays a révélé les résultats suivants. Y a-t-il plus de variabilité en Belgique ou aux États-Unis ?

	Belgique	**États-Unis**
CA moyen	1 566 569,98 €	$ 1 212 046,41
Écart-type	475 832,76 €	$ 368 149,14

Exercices récapitulatifs

11) Une enquête auprès d'un certain nombre de chômeurs leur demandait d'indiquer approximativement depuis combien de jours ils étaient au chômage. Quatre réponses étaient proposées : 120 (choisie par 10 chômeurs), 150 (choisie par 40 chômeurs), 180 (choisie par 20 chômeurs) et 240 (choisie par 5 chômeurs).
   a) Dressez le tableau de recensement complet.
   b) Représentez le diagramme en bâtons et le diagramme en escaliers.
   c) Calculez l'écart-type de cet échantillon.
   d) Déterminez son coefficient de variation.
   e) Chacun de ces chômeurs a reçu une allocation de 10€ par jour. Si on étudie le total des allocations perçues par chacun de ces chômeurs, quel écart-type obtiendra-t-on ? Et quel coefficient de variation ?
   f) On suppose maintenant que chacun des chômeurs a dû payer une taxe forfaitaire de 80€ sur le total des allocations perçues. Si on étudie la somme nette perçue par les chômeurs (en décomptant donc cette taxe), quel écart-type obtiendra-t-on ? Et quel coefficient de variation ?

12) On a réalisé une enquête auprès d'un groupe d'étudiants en leur demandant combien de films ils étaient allés voir au cinéma pendant le mois écoulé. Voici les résultats obtenus.

Nb de films	0	1	2	3	4
Nb d'étudiants	5	9	6	4	1

   a) Dressez le tableau des répétitions complet.

b) Représentez le diagramme en bâtons et le diagramme en escaliers.
c) En utilisant les diagrammes, repérez le mode et la médiane.
d) Calculez la moyenne, la variance, l'écart-type et le coefficient de variation.

13) Une usine belge produit des poutres de différentes longueurs. La production journalière est répartie comme suit.

Longueur (m)	[0, 2[	[2, 4[	[4, 6[	[6, 8[	[8, 10[
Nb de poutres	10	20	30	25	15

a) Calculez et interprétez les paramètres suivants : moyenne, mode, 8^{e} décile, écart-type, coefficient de variation.
b) Si on avait mesuré les poutres en centimètres (au lieu de mètres), que seraient devenues les valeurs calculées au point précédent ?
c) Quelle est la proportion de poutres produites dont la longueur ne s'écarte pas plus d'un écart-type de la moyenne ? De deux écarts-types ?
d) Comparez la variabilité entre cette usine belge et une usine anglaise pour laquelle on a obtenu une moyenne de 6,5 yards et un écart-type de 1,9 yards.

14) Une compagnie de taxis vous soumet la distribution statistique suivante concernant les distances hebdomadaires parcourues par des taxis : 5 taxis ont parcouru entre 300 et 400 km ; 6 entre 600 et 700 km ; 5 entre 700 et 900 km ; 2 entre 900 et 1000 km.

a) Identifiez la variable étudiée et son type.
b) Calculez et interprétez le mode, la médiane et la moyenne.
c) Calculez l'écart-type et le coefficient de variation.

15) Une enquête s'est intéressée aux salaires mensuels des employés dans deux entreprises A et B.

Salaires (× 1000 €)	**Nb employés (entreprise A)**	**Nb employés (entreprise B)**
[1, 1.5[	35	15
[1.5, 2[	125	83
[2, 2.5[	173	251
[2.5, 3[	109	79
[3, 3.5[	48	28

a) Dressez un tableau de recensement complet pour chacune des deux entreprises.
b) Construisez, dans un même graphique, les histogrammes correspondant aux deux entreprises.
c) Construisez, dans un second graphique, les polygones des fréquences des deux entreprises.
d) Construisez, dans un troisième graphique, les diagrammes cumulatifs pour les deux entreprises.
e) Pour chaque entreprise, déterminez la classe modale. Quel(s) graphique(s) utiliser pour la repérer et comment procéder ?
f) Sur base des graphiques, approximez la médiane pour chacune des entreprises. Quel graphique utilisez-vous et comment procédez-vous ?

g) Calculez précisément la médiane pour chacune des entreprises, en utilisant le tableau de recensement.
h) Calculez la moyenne, la variance, l'écart-type et le coefficient de variation pour chacune des entreprises. Interprétez comparativement chacune de ces valeurs.
i) Déterminez l'étendue et l'écart interquartile pour chacune des entreprises.
j) Représentez, sur un même axe, la boîte à moustaches de chacune des entreprises.
k) Pour chaque entreprise, déterminez la proportion d'employés dont le salaire se situe à moins d'un écart-type de la moyenne (c'est-à-dire dans l'intervalle $[\bar{x} - s, \bar{x} + s]$).
l) Pour chaque entreprise, déterminez la proportion d'employés dont le salaire se situe dans l'intervalle $[\bar{x} - 3s, \bar{x} + 3s]$.

PARTIE 2
STATISTIQUES DESCRIPTIVES À DEUX DIMENSIONS

Dans la partie précédente, nous nous sommes contentés d'étudier un caractère statistique à la fois. Les traitements statistiques abordés ont permis de répondre à des questions portant sur les valeurs moyennes, les valeurs les plus souvent rencontrées (mode), les valeurs qui permettent de diviser la population en deux (médiane) ou en plusieurs parties (quartiles) et sur la manière dont les valeurs sont réparties autour de la moyenne (variance et écart-type).

Or, la statistique dispose également d'outils permettant d'examiner deux caractères à la fois et, surtout, de déceler les liens qui pourraient les unir. Dans cette partie, nous nous intéressons à des questions telles que : y a-t-il un lien entre le nombre d'années d'étude et le salaire ? comment peut-on quantifier le lien entre la quantité d'engrais déposé sur un champ et la production ? à partir d'une étude de marché, comment modéliser le lien entre le prix d'un article et le volume de ventes qu'on peut espérer ? à partir de cette même étude de marché, comment estimer le volume de ventes qu'on peut espérer en fonction du prix choisi ?

Comme dans l'étude de la statistique à une dimension, nous nous penchons tout d'abord sur la manière de présenter des données liant deux caractères, tant sous la forme de tableau (tableaux de contingence) que de graphiques (nuages de points).

Nous abordons ensuite le calcul des paramètres permettant d'analyser les caractères de manière séparée (distributions marginales) et de manière combinée (covariance et coefficient de corrélation linéaire), en insistant tout particulièrement sur l'interprétation de ces résultats.

Puis, finalement, nous examinons la méthode des moindres carrés pour élaborer un modèle linéaire à partir des données, de manière à pouvoir réaliser des interpolations ou extrapolations lorsque le modèle semble suffisamment pertinent.

CHAPITRE 1 : TABLEAUX DE CONTINGENCE

Jusqu'ici, nous ne nous sommes intéressés qu'à une question statistique (un seul caractère, une seule variable) à la fois. En statistiques dites « à deux dimensions », on considère deux questions en même temps et on récolte donc deux données pour chaque individu de l'échantillon. Cela permet d'étudier non seulement chacun des caractères examinés mais aussi de pencher sur les liens qui pourraient les unir.

Plan du module

1 Données brutes et nuage de points

Ce module et le suivant vont s'appuyer sur les deux exemples présentés ci-dessous.

Dans le premier exemple, on considère 30 employés d'une société, à qui on a demandé leur âge et leur salaire mensuel (en Euros). Les résultats obtenus sont présentés dans le tableau ci-dessous.

Âge	20	21	25	24	30	30	27	33	33	35
Salaire	1200	1700	1400	2000	1400	2600	2000	1700	2200	1400

Âge	35	35	37	37	38	43	42	45	45	47
Salaire	2500	3100	2000	2600	2000	2000	3000	3600	3500	3300

Âge	47	50	50	49	52	53	55	55	57	59
Salaire	3000	2600	3000	3500	3600	3000	3500	3400	3100	3600

Le second exemple concerne une étude entreprise par une compagnie d'assurances qui a examiné l'âge et le nombre d'accidents rapportés au cours de l'année précédente pour un certain nombre de ses clients. Comme l'étude porte sur 4 000 conducteurs, seule une partie des données sont présentées dans le tableau ci-dessous.

Numéro	729	1321	2501	2714	2690	2999	3115	3472	3849
Âge	58	48	52	30	22	61	38	53	51
Nb accidents	3	1	0	1	3	0	0	0	2

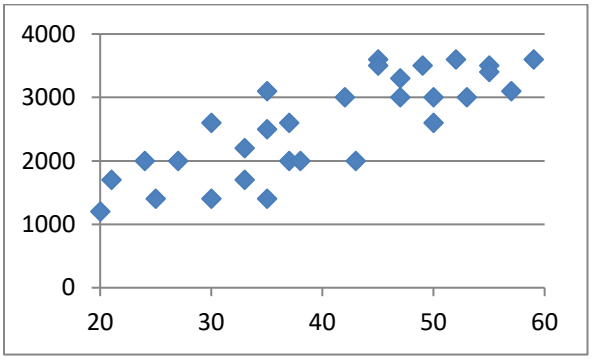
Lors d'une étude statistique à deux dimensions, pour chacun des individus de l'échantillon (il y en a n au total, n étant l'effectif total), on dispose de deux valeurs. Dans le cadre de la statistique à une dimension, on notait les valeurs $x_1, x_2, \ldots, x_n$. Ici, pour chaque individu i, on disposera de deux valeurs : x_i et y_i.

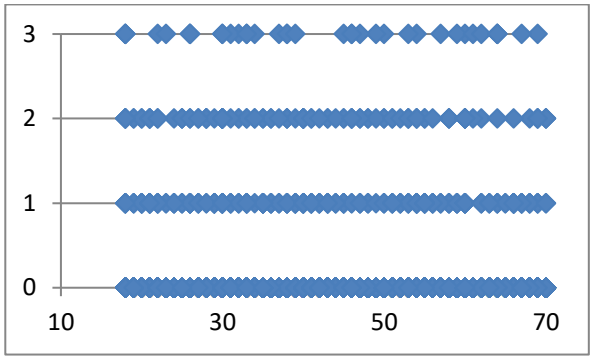
Il n'y a pas de règle précise pour choisir qui va être appelé « x » et qui va être appelé « y » mais, en règle générale, quand on suppose qu'un des caractères est plutôt un état et l'autre une conséquence de celui-ci, on utilise x pour l'état et y pour la conséquence supposée. Ainsi, dans le cadre des exemples, on notera x_i l'âge du i^{e} salarié et y_i son salaire mensuel (pour le premier cas) et x_i l'âge du i^{e} conducteur et y_i le nombre d'accidents qu'il (ou elle) a rapportés au cours de l'année précédente.

Notation	**Signification**
x_i	la première valeur associée à l'individu i (i allant de 1 à n)
y_i	la seconde valeur associée à l'individu i (i allant de 1 à n)

Comme, pour chacun des individus, on dispose de deux valeurs, on peut représenter chacun d'eux par un point de coordonnées (x_i, y_i). Le graphique ainsi obtenu est appelé un **nuage de points**. Les nuages de points permettent d'avoir une idée générale dont les valeurs sont distribuées.

Voici les nuages de points pour les deux exemples présentés ci-dessus.





Dans le cas du premier exemple, on devine une certaine forme qui pourrait laisser penser que, en règle générale, le salaire est de plus en plus élevé avec l'âge (on verra plus loin divers outils pour vérifier ce genre d'hypothèses de manière plus formelle). Le second graphique, par contre, est relativement peu utile : vu le nombre de points représentés, on peine à voir autre chose que quatre « lignes horizontales » de points.

Modalités/classes en statistique à 2 dimensions

Comme en statistique à une dimension, les tableaux présentant les données brutes sont généralement difficiles à lire : quand on dispose d'un très grand nombre de données, il est extrêmement difficile d'avoir une vue d'ensemble sur celles-ci.

Une première étape pour améliorer la présentation des données consiste à examiner, pour chacun des caractères étudiés, quelles sont les modalités observées et s'il semble nécessaire de regrouper les valeurs par classe. Le procédé est identique à ce qui a été réalisé en statistique à une dimension.

Dans le premier exemple (âge et salaire), on observe que les âges vont de 20 à 59 ans et que les salaires varient entre 1200 et 3600. Il semble donc assez naturel de former des classes. On a choisi de réaliser 4 classes de 10 ans d'amplitude pour les âges (de 20 à 30, de 30 à 40, de 40 à 50 et de 50 à 60) et 3 classes de 1000 € d'amplitude pour les salaires (de 1000 à 2000, de 2000 à 3000 et de 3000 à 4000).

On introduit les notations suivantes : on note $X_1, X_2, \dots, X_p$ les modalités pour la variable x et $Y_1, Y_2, \dots, Y_q$ celles pour la variable y. Si on a effectué un regroupement en classes pour l'une des variables (ou les deux), ces notations désignent alors les centres des classes.

Ainsi, dans l'exemple des âges et des salaires, on aura

- pour l'âge : $p = 4$, $X_1 = 25$, $X_2 = 35$, $X_3 = 45$ et $X_4 = 55$;
- pour les salaires : $q = 3$, $Y_1 = 1500$, $Y_2 = 2500$ et $Y_3 = 3500$.

Dans le second exemple (âge et nombre d'accidents), on remarque que les âges s'étendent entre 18 (le minimum pour un conducteur) et 70 alors que le nombre d'accidents rapportés varie entre 0 et 3. Pour les accidents, on peut donc se contenter de garder les données telles quelles : les modalités observées sont 0, 1, 2 et 3. Pour l'âge, par contre, on décide de rassembler les données par classe. On choisit de faire 5 classes comme suite : [18,30[, [30,40[, [40,50[, [50,60[et [60,70[. Notez que, pour éviter de faire une classe [10,20[qui n'aurait repris que les 18-20 ans, on a ajouté ces conducteurs à la classe des 20-30 ans.

Dans l'exemple des conducteurs, on aura

- pour l'âge : $p = 5$, $X_1 = 24$, $X_2 = 35$, $X_3 = 45$, $X_4 = 55$ et $X_5 = 65$;
- pour le nombre d'accidents rapportés : $q = 4$, $Y_1 = 0$, $Y_2 = 1$, $Y_3 = 2$ et $Y_4 = 3$.

Notation	Signification
X_i	la i^{e} modalité ou le centre de la i^{e} classe correspondant à la variable x (pour i allant de 1 à p)
p	le nombre de modalités/classes pour la variable x
Y_j	la j^{e} modalité ou le centre de la j^{e} classe correspondant à la variable y (pour j allant de 1 à q)
q	le nombre de modalités/classes pour la variable y

Tableau de contingence

Une fois qu'on a décidé des modalités et/ou des classes, on peut construire un **tableau de contingence**. Il s'agit d'un tableau à double entrée où

- chaque ligne correspond à une modalité/classe pour x ;

- chaque colonne correspond à une modalité/classe pour y ;
- et chaque case indique le nombre d'individus qui correspondent à une modalité/classe pour x et à une modalité/classe pour y.

Voici ce qu'on obtient pour le premier exemple.

Âge \ Salaire	**[1000, 2000[**	**[2000, 3000[**	**[3000, 4000[**
[20, 30[	3	2	0
[30, 40[	3	6	1
[40, 50[	0	1	6
[50, 60[	0	1	7

On peut remarquer que chacune des lignes correspond bien à une classe de la variable « x », à savoir de l'âge alors que chacune des colonnes correspond à une classe de la variable « y », le salaire. Dans les cases centrales, on retrouve les répétitions pour chaque cas.

Par exemple, le nombre « 1 » indiqué à l'intersection de la ligne [40,50[et la colonne [2000,3000[signifie que 1 seul des salariés sondés a un âge entre 40 et 50 ans et un salaire entre 2000 et 3000 €. C'est donc la répétition qui correspond à la modalité $X_3 = 45$ et $Y_2 = 2500$. C'est pour cela qu'on note ce nombre $r_{3,2}$ ou, plus simplement, r_{32}.

Notation	**Signification**
$r_{i,j}$ ou r_{ij}	la répétition qui correspond à la i^{e} modalité/classe de x et à la j^{e} modalité/classe de y (i allant de 1 à p et j allant de 1 à q)

De manière théorique, un tableau de contingence devrait donc ressembler à ceci :

x \ y	$\boldsymbol{Y_1}$	$\boldsymbol{Y_2}$	…	$\boldsymbol{Y_q}$
$\boldsymbol{X_1}$	r_{11}	r_{12}	…	r_{1q}
$\boldsymbol{X_2}$	r_{21}	r_{22}	…	r_{2q}
⋮	⋮	⋮	⋱	⋮
$\boldsymbol{X_p}$	r_{p1}	r_{p2}	…	r_{pq}

En réalisant le même exercice pour le second exemple, voici le tableau qu'on obtient.

Âge \ Nb acc	**0**	**1**	**2**	**3**
[18, 30[	718	65	31	9
[30, 40[	821	50	24	10
[40, 50[	752	41	22	6
[50, 60[	699	46	16	5
[60, 70[	623	40	15	7

Comme dans le cas des tableaux de recensement, il est très important de savoir exprimer précisément la signification de chacune des valeurs reprises dans un tableau de contingence.

Distributions marginales

On complète généralement les tableaux de contingence en ajoutant une colonne et une ligne reprenant les totaux pour chacune des modalités/classes des variables.

Voici ce que cela donne dans le cas du premier exemple (âge et salaire).

Salaire	**[1000, 2000[**	**[2000, 3000[**	**[3000, 4000[**

Âge				
[20, 30[	3	2	0	5
[30, 40[	3	6	1	10
[40, 50[	0	1	6	7
[50, 60[	0	1	7	8
	6	10	14	**30**

Le nombre 5 écrit en haut de la dernière colonne a été calculé en additionnant les répétitions r_{11}, r_{12} et r_{13}, c'est-à-dire toutes les répétitions qui concernent la classe d'âge [20,30[. Il s'agit donc du nombre de salariés interrogés ayant entre 20 et 30 ans.

On peut réaliser une interprétation similaire pour les autres nombres de cette colonne. Ces valeurs correspondent donc à la distribution de la variable âge (x) si on décide de ne plus tenir compte des salaires. C'est ce qu'on appelle la **distribution marginale** de x.

De la même manière, le nombre 6 écrit au début de la dernière ligne correspond à la somme des répétitions r_{11}, r_{21}, r_{31} et r_{41}, c'est-à-dire de toutes les répétitions concernant la classe de salaires [1000, 2000[. C'est donc bien le nombre de salariés interrogés touchant entre 1000 et 2000 Euros par mois. Les valeurs de la dernière ligne (6, 10 et 14) correspondent donc à la distribution marginale des salaires.

Chacun des nombres des distributions marginales a été calculé en effectuant la somme des répétitions d'une même ligne ou d'une même colonne. Au niveau des notations, cela revient à calculer une somme telle que $r_{i1} + r_{i2} + \cdots + r_{iq}$ (en suivant la i^{e} ligne) ou telle que $r_{1j} + r_{2j} + \cdots + r_{pj}$ (en suivant la j^{e} colonne). On a généralement l'habitude de noter ces valeurs r_{i*} et r_{*j}, l'astérisque représentant les indices sur lesquels on a calculé la somme.

De la même manière, quand on fait la somme de tous les r_{ij}, le résultat obtenu est parfois noté r_{**} pour indiquer qu'on effectue la somme sur les deux indices. Ce n'est rien d'autre que n, l'effectif total. C'est le nombre indiqué en gras tout en bas à droite dans le tableau ci-dessus.

Notation	**Signification**
r_{i*}	la somme $r_{i1} + r_{i2} + \cdots + r_{iq}$ en suivant la i^{e} ligne ; c'est la répétition qui correspond à la modalité/classe X_i
r_{*j}	la somme $r_{1j} + r_{2j} + \cdots + r_{pj}$ en suivant la j^{e} colonne ; c'est la répétition qui correspond à la modalité/classe Y_j
r_{**}	la somme de toutes les répétitions r_{ij}, à savoir n, l'effectif total

Voici une version complétée du tableau de contingence théorique faisant le point sur toutes les notations introduites jusqu'ici.

y / x	$\boldsymbol{Y_1}$	$\boldsymbol{Y_2}$	…	$\boldsymbol{Y_q}$	
$\boldsymbol{X_1}$	r_{11}	r_{12}	…	r_{1q}	r_{1*}
$\boldsymbol{X_2}$	r_{21}	r_{22}	…	r_{2q}	r_{2*}
⋮	⋮	⋮	⋱	⋮	⋮
$\boldsymbol{X_p}$	r_{p1}	r_{p2}	…	r_{pq}	r_{p*}
	r_{*1}	r_{*2}	…	r_{*q}	$\boldsymbol{r_{**}}$

Finalement, voici le tableau qu'on obtient en ajoutant les distributions marginales dans le cadre de l'exemple 2 sur les conducteurs.

Âge \ Nb acc	**0**	**1**	**2**	**3**	
[18, 30[	718	65	31	9	823
[30, 40[	821	50	24	10	905
[40, 50[	752	41	22	6	821
[50, 60[	699	46	16	5	766
[60, 70[	623	40	15	7	685
	3613	242	108	37	**4000**

Notez que les distributions marginales permettent de calculer tous les paramètres de tendance centrale et de dispersion propres à chacune des variables, entre autres les moyennes, variances et écarts-types. Voici les notations qu'on utilise généralement pour ces valeurs.

Notation	**Signification**
$\bar{x}$	la moyenne de la distribution marginale de x
$\bar{y}$	la moyenne de la distribution marginale de y
s_x^2	la variance de la distribution marginale de x
s_y^2	la variance de la distribution marginale de y
s_x	l'écart-type de la distribution marginale de x
s_y	l'écart-type de la distribution marginale de y

Toutes ces valeurs seront fort nécessaires lors du calcul de paramètres permettant d'étudier le rapport entre les deux variables.

Distributions conditionnelles

En effectuant les totaux en suivant les lignes ou les colonnes, on retombe sur les distributions propres à un seul des deux caractères étudiés. À partir de ces distributions, on peut calculer divers paramètres tels que la moyenne, la variance ou l'écart-type.

On peut également effectuer des calculs similaires en partant des autres lignes (ou colonnes) du tableau de contingence. Dans ce cas-là, on parle de **distribution conditionnelle**. L'adjectif « conditionnel » est utilisé ici dans un sens très proche de celui rencontré dans le chapitre Probabilités conditionnelles. En effet, cela revient à se poser des questions du type « Que vaut la variable y quand je me restreins à une des modalités/classes de x ? »

Reprenons par exemple le cas des âges et des salaires et concentrons-nous sur la ligne correspondant à la classe d'âge [30,40[. Si on occulte les autres lignes, on se concentre uniquement sur les 10 salariés qui ont entre 30 ans et 40 ans et on peut étudier leurs salaires.

Âge \ Salaire	**[1000, 2000[**	**[2000, 3000[**	**[3000, 4000[**
[30, 40[	3	6	1

Par exemple, en calculant la moyenne (3 × 1500 + 6 × 2500 + 1 × 3500) /10 = 2300, on trouve le salaire moyen *pour un salarié âgé entre 30 et 40 ans*. Il ne s'agit pas du salaire moyen de tous les employés mais seulement de ceux qui vérifient la condition « avoir entre 30 et 40 ans ».

On peut effectuer des calculs similaires en considérant une ligne et, dans ce cas-là, étudier l'âge des salariés qui vérifient la condition « avoir un salaire mensuel entre a et b Euros. »

Exercices

1) Sur base du tableau de contingence de l'exemple « âge des conducteurs et nombre d'accidents rapportés », répondez aux questions suivantes.

Âge \ Nb acc	**0**	**1**	**2**	**3**	
[18, 30[	718	65	31	9	823
[30, 40[	821	50	24	10	905
[40, 50[	752	41	22	6	821
[50, 60[	699	46	16	5	766
[60, 70[	623	40	15	7	685
	3613	242	108	37	**4000**

   a) Parmi les conducteurs sondés, combien n'ont rapporté aucun accident ?
   b) Parmi les conducteurs sondés, combien sont au moins quarantenaires ?
   c) En moyenne, combien d'accidents ont rapportés les conducteurs trentenaires (30-40 ans) ?
   d) En moyenne, combien d'accidents les conducteurs sondés ont-ils rapportés ?
   e) Quelle est la moyenne d'âge des conducteurs sondés ?

2) À l'approche d'une interrogation, un professeur de statistiques a réparti ses étudiants en quatre groupes. Les étudiants du premier groupe ont pu étudier la matière pendant 4 heures ; ceux du second groupe, pendant 15 heures ; ceux des deux derniers groupes, respectivement pendant 20 et 30 heures. Après l'interrogation, il a examiné les cotes obtenues par chacun des étudiants et a rassemblé les données dans le tableau suivant.

Nb heures \ Cote	**[4, 6[**	**[6, 10[**	**[10, 14[**	**[14, 20[**
4 heures	33	2	1	0
15 heures	0	18	11	2
20 heures	3	6	24	1
30 heures	1	1	22	35

   a) Que représente la valeur « 18 » inscrite dans le tableau ?
   b) Combien d'étudiants se trouvent dans cette classe ?
   c) Combien d'étudiants ont pu étudier exactement 20 heures ? 15 heures ou plus ?
   d) Combien d'étudiants ont réussi (cote de 10 ou plus) l'interrogation ?
   e) Quelle est la cote moyenne des étudiants qui ont étudié 20 heures ? 4 heures ?
   f) Quelle est la cote moyenne de la classe ?
   g) En moyenne, combien d'heures les étudiants qui ont réussi ont-ils passées à étudier ?

3) Voici les résultats d'une enquête concernant le temps passé sur Internet chaque semaine (en heures) et les dépenses en ligne mensuelles (en Euros).

Nb heures \ Dépenses	**[0, 100[**	**[100, 200[**	**[200, 300[**	**[300, 400[**
[0, 6[	8	8	0	2
[6, 12[	10	0	9	4
[12, 18[	8	9	11	8
[18, 24[	10	2	10	12
[24, 30[	3	3	0	3

a) Combien de personnes ont été interrogées ?
b) Que vaut r_{3*} ? Que représente ce nombre ?
c) Combien de temps les personnes interrogées passent-elles chaque semaine sur Internet (en moyenne) ? Combien dépensent-elles sur Internet (en moyenne) ?
d) Calculez l'écart-type du temps passé chaque semaine et des dépenses mensuelles.

Conseils de mise en page pour faciliter les calculs

Que vous utilisiez Excel ou que vous ayez à réaliser les calculs à la main, il existe une mise en page standard pour faciliter les calculs des distributions marginales et des paramètres (moyennes, variances et écart-types) de ces distributions.

Celle-ci consiste à ajouter, pour le caractère x, des colonnes à droite du tableau de contingence et, pour le caractère y, des colonnes en dessous du tableau. Ces colonnes permettront d'effectuer les calculs suivants :

- les répétitions pour chacune des modalités (distribution marginale), donc les r_{i*} et r_{*j} ;
- les produits entre les répétitions et les valeurs (modalités ou centres), à savoir les $X_i r_{i*}$ et les $Y_j r_{*j}$;
- les produits entre les répétitions et le carré des valeurs (modalités ou centres), à savoir les $X_i^2 r_{i*}$ et les $Y_j^2 r_{*j}$ (qui peuvent s'obtenir en multipliant les valeurs de la 2e ligne/colonne par les modalités/centres de classe).

En effectuant la somme de ces colonnes/lignes, on obtient (respectivement)

- pour les r_{i*} et les r_{*j} : l'effectif total n ;
- pour les $X_i r_{i*}$ et les $Y_j r_{*j}$: des nombres qu'il suffit de diviser par n pour obtenir les moyennes $\bar{x}$ et $\bar{y}$;
- les produits entre les répétitions et le carré des valeurs (modalités ou centres), à savoir les $X_i^2 r_{i*}$ et les $Y_j^2 r_{*j}$: des nombres qu'il suffit de diviser par n pour obtenir les moyennes $\overline{x^2}$ et $\overline{y^2}$ (qui sont utilisées dans le calcul des variances et écarts-types).

Le diagramme suivant résume toutes ces indications. Les flèches signifient que la valeur de la case devant la flèche se calcule en réalisant la somme des cases situées derrière elle.

x \ y	$\boldsymbol{Y_1}$	$\boldsymbol{Y_2}$	…	$\boldsymbol{Y_q}$				
$\boldsymbol{X_1}$	r_{11}	r_{12}	…	r_{1q}	⇨	r_{1*}	$X_1 r_{1*}$	$X_1^2 r_{1*}$
$\boldsymbol{X_2}$	r_{21}	r_{22}	…	r_{2q}	⇨	r_{2*}	$X_2 r_{2*}$	$X_2^2 r_{2*}$

⋮	⋮	⋮	⋱	⋮	⇨	⋮	⋮	⋮
$\boldsymbol{X_p}$	r_{p1}	r_{p2}	…	r_{pq}	⇨	r_{p*}	$X_p r_{p*}$	$X_p^2 r_{p*}$
	⇩	⇩		⇩		⇩	⇩	⇩
	r_{*1}	r_{*2}	…	r_{*q}	⇨	$\boldsymbol{r_{**} = n}$	$\boldsymbol{M_x}$	$\boldsymbol{V_x}$
	$Y_1 r_{*1}$	$Y_2 r_{*2}$	…	$Y_q r_{*q}$	⇨	$\boldsymbol{M_y}$		
	$Y_1^2 r_{*1}$	$Y_2^2 r_{*2}$	…	$Y_q^2 r_{*q}$	⇨	$\boldsymbol{V_y}$		

Les valeurs obtenues en fin de lignes/colonnes peuvent être utilisées pour calculer

- les moyennes : $\bar{x} = M_x/n$ et $\bar{y} = M_y/n$;
- les moyennes des carrés : $\overline{x^2} = V_x/n$ et $\overline{y^2} = V_y/n$;
- les variances : $s_x^2 = \overline{x^2} - \bar{x}^2$ et $s_y^2 = \overline{y^2} - \bar{y}^2$;
- les écarts-types : $s_x = \sqrt{s_x^2}$ et $s_y = \sqrt{s_y^2}$.

Voici ce que cela donne dans le cadre de l'étude relative aux salaires et aux âges.

Âge \ Salaire	**[1000, 2000[**	**[2000, 3000[**	**[3000, 4000[**			
[20, 30[	3	2	0	5	125	3125
[30, 40[	3	6	1	10	350	12250
[40, 50[	0	1	6	7	315	14175
[50, 60[	0	1	7	8	440	24200
	6	10	14	**30**	**1230**	**53750**
	9000	25000	49000	**83000**		
	13500000	62500000	171500000	**247500000**		

Pour l'âge : $\bar{x} = 1230/30 = 41$, $\overline{x^2} = 53750/30 = 1791{,}67$, $s_x^2 = 110{,}67$, $s_x = 10{,}52$

Pour les salaires : $\bar{y} = 83000/30 = 2766{,}67$, $\overline{y^2} = 247500000/30 = 8250000$, $s_x^2 = 595555{,}56$, $s_x = 771{,}72$

CHAPITRE 2 : CORRÉLATION ET RÉGRESSION LINÉAIRE

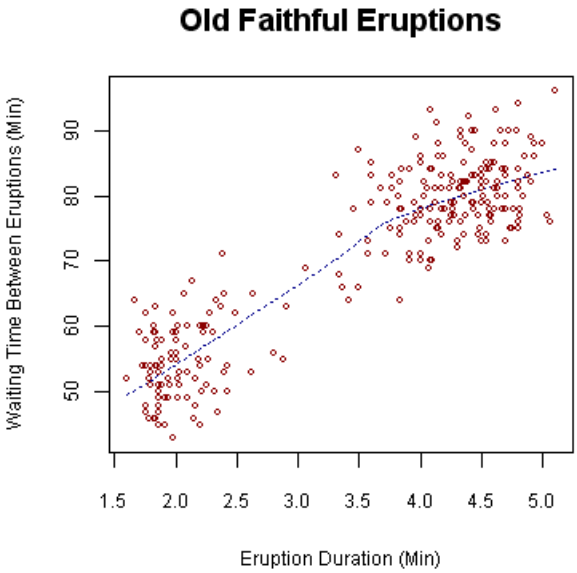
Ce dernier module aborde divers outils mathématiques permettant de jauger de la solidité du lien entre les deux caractéristiques étudiées. Dans le cadre des exemples présentés plus haut, ces outils devraient nous aider à répondre à des questions telles que : les salariés plus âgés touchent-ils des salaires plus élevés ? l'âge du conducteur a-t-il une influence sur le nombre d'accidents rapportés à l'assurance ?

Plan du module

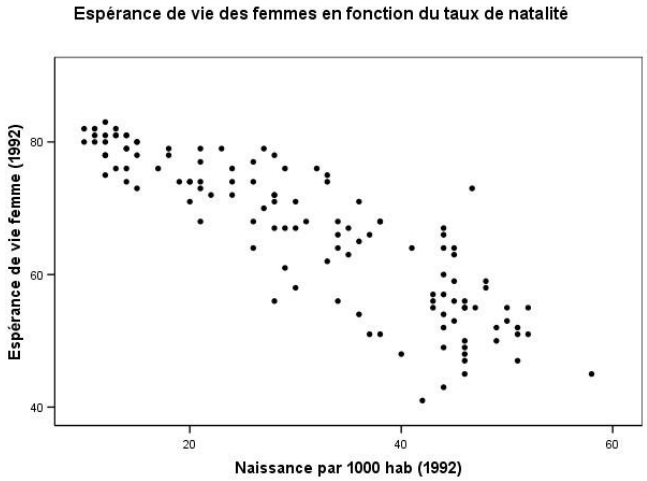
1 Sens de variation

Quand on étudie deux caractères en parallèle, on peut se demander si leurs valeurs évoluent dans le même sens, dans des sens opposés, ou de manière indépendante. Autrement dit, si l'un des caractères augmente, le second augmente-t-il également ? Ou diminue-t-il ?

Pour répondre à cette question, on peut se baser sur les nuages de points. Le graphique ci-contre (source : Wikipedia) a été réalisé lors de l'étude des éruptions d'un geyser américain (le « Old Faithful »). Pour chaque éruption, on a examiné deux caractères : x, la durée de l'éruption et y, le temps de calme qui a suivi l'éruption (le délai avant l'éruption suivante).

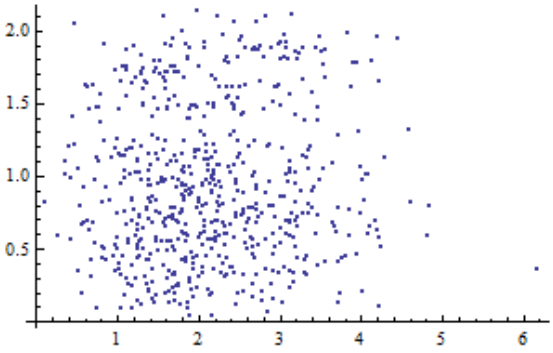


Le graphique tend à indiquer que plus une éruption a duré longtemps, plus la période de calme qui a suivi a été longue. Autrement dit, plus x est grand, plus y l'est également : les deux caractères varient dans le même sens.



Le second graphique, à gauche, représente l'espérance de vie des femmes (en année) en fonction du taux de natalité (nombre de naissances par 1000 habitants) dans divers pays. Ici, plus x est grand (plus le taux de natalité est grand) et plus y est petit (plus l'espérance de vie des femmes est limitée).

Les deux caractères varient dans des sens différents.



Finalement, dans le troisième nuage de points présenté, il semble qu'il n'y ait aucun lien évident entre les caractères x et y étudiés.

Covariance

Formellement, pour étudier le sens de la variation de deux caractères x et y, on se réfère aux moyennes $\bar{x}$ et $\bar{y}$ et on observe comment les différentes données recueillies (à savoir les x_i et les y_i) s'y rapportent. Plus précisément, pour chaque individu i, on examine le signe des expressions $x_i - \bar{x}$ et $y_i - \bar{y}$.

Si $x_i - \bar{x}$ est positif (c'est-à-dire si $x_i > \bar{x}$), cela signifie que l'individu a une valeur en x plus grande que la moyenne. Dans ce cas-là, si $y_i - \bar{y}$ est également positif (c'est-à-dire si $y_i > \bar{y}$), l'individu a également une valeur en y plus grande que la moyenne : pour cet individu, les données indiquent que x et y varient dans le même sens. En revanche, si $y_i - \bar{y}$ est négatif (c'est-à-dire si $y_i < \bar{y}$), les données indiquent que x et y varient dans des sens différents. On pourrait mener une réflexion similaire dans le cas où $x_i - \bar{x}$ est négatif.

On peut résumer la situation ainsi :

- Si $x_i - \bar{x}$ et $y_i - \bar{y}$ ont le même signe (tous les deux positifs ou tous les deux négatifs), les caractères x et y semblent évoluer dans le même sens.
- Si $x_i - \bar{x}$ et $y_i - \bar{y}$ ont des même signes différents (un positif l'autre négatif), les caractères x et y semblent évoluer dans des sens différents.

On pourrait encore condenser cette observation en étudiant le signe du produit

$$(x_i - \bar{x}) \times (y_i - \bar{y}).$$

De manière générale, si ce produit est positif pour un grand nombre d'individus, les caractères x et y évoluent dans le même sens. Et si ce produit est négatif, les caractères évoluent dans des directions différentes.

Comme il faut considérer tous les individus de l'échantillon, on s'intéressera à la moyenne de ces produits, c'est-à-dire à la **covariance**

$$\mathrm{Cov}_{xy} = \frac{1}{n}\sum (x_i - \bar{x}) \times (y_i - \bar{y})$$

Pour résumer les observations réalisées ci-dessus, on peut retenir que :

- Une covariance positive $\mathrm{Cov} > 0$ indique que les caractères x et y évoluent dans le même sens ;
- Une covariance négative $\mathrm{Cov} < 0$ indique que les caractères x et y évoluent dans des sens contraires ;
- Une covariance nulle/proche de 0 ne permet pas d'affirmer quoi que ce soit au sujet des sens d'évolution de x et de y.

Si on a regroupé les données par modalités/classes, la formule devient naturellement

$$\mathrm{Cov}_{xy} = \frac{1}{n}\sum r_{ij} \times (X_i - \bar{x}) \times (Y_i - \bar{y})$$

Calcul de la covariance

Pour calculer la covariance, on n'utilise que très rarement la formule de sa définition (celle donnée plus haut). Tout comme pour la variance, on utilise plutôt une formule de calcul généralement plus simple à mettre en place :

$$\mathrm{Cov}_{xy} = \overline{xy} - \bar{x} \times \bar{y}$$

En français : la covariance entre les caractères x et y est égale à la moyenne des produits *moins* le produit des moyennes. Notez que cette formule est très similaire à celle utilisée pour la variance ($s^2 = \overline{x^2} - \bar{x}^2$), à ceci près qu'on étudie deux caractères plutôt qu'un seul.

Le terme $\bar{x} \times \bar{y}$ est simplement le produit des moyennes : on peut le calculer à partir des distributions marginales. L'autre terme, $\overline{xy}$, la moyenne des produits, peut se calculer en observant, pour chacune des cases du tableau de contingence, (1) ce que le produit vaut et (2) le nombre de répétitions correspondant à la case.

Âge \ Salaire	**[1000, 2000[**	**[2000, 3000[**	**[3000, 4000[**	
[20, 30[	3	2	0	5
[30, 40[	3	6	1	10
[40, 50[	0	1	6	7
[50, 60[	0	1	7	8
	6	10	14	**30**

Dans le cas de l'exemple âge/salaire introduit dans le module précédent, on peut calculer la moyenne du produit xy en observant qu'il y a 3 cas où le produit vaut 25 × 1500 = 37500, 2 cas où le produit vaut 25 × 2500 = 62500, et ainsi de suite. Pour obtenir la moyenne, il faudra effectuer la somme de tous ces produits répétitions × valeur de x × valeur de y (c'est-à-dire $r_{ij}X_iY_j$) puis diviser le résultat par l'effectif total.

Pratiquement, cela peut se faire en construisant un nouveau tableau reprenant les produits $r_{ij}X_iY_j$ pour chacune des cases puis en additionnant les résultats et en divisant le tout par n :

3 × 25 × 1500 = 112500	2 × 25 × 2500 = 125000	0
3 × 35 × 1500 = 157500	2 × 35 × 2500 = 525000	1 × 35 × 3500 = 122500
0	1 × 45 × 2500 = 112500	6 × 45 × 3500 = 945000
0	1 × 55 × 2500 = 137500	7 × 55 × 3500 = 1347500

Dans le cas de l'exemple, le total vaut 3585000 et donc, $\overline{xy} = 119500$.

Pour obtenir la covariance, il ne reste plus qu'à calculer les moyennes $\bar{x} = 41$ ans et $\bar{y} = 2766{,}67$ € pour obtenir

$$\mathrm{Cov} = \overline{xy} - \bar{x} \times \bar{y} = 119500 - 41 \times 2766{,}67 = 6066{,}67$$

Coefficient de corrélation linéaire

On définit le **coefficient de corrélation linéaire** entre les caractères x et y par la formule suivante : la covariance des deux caractères divisés par leurs écarts-types.

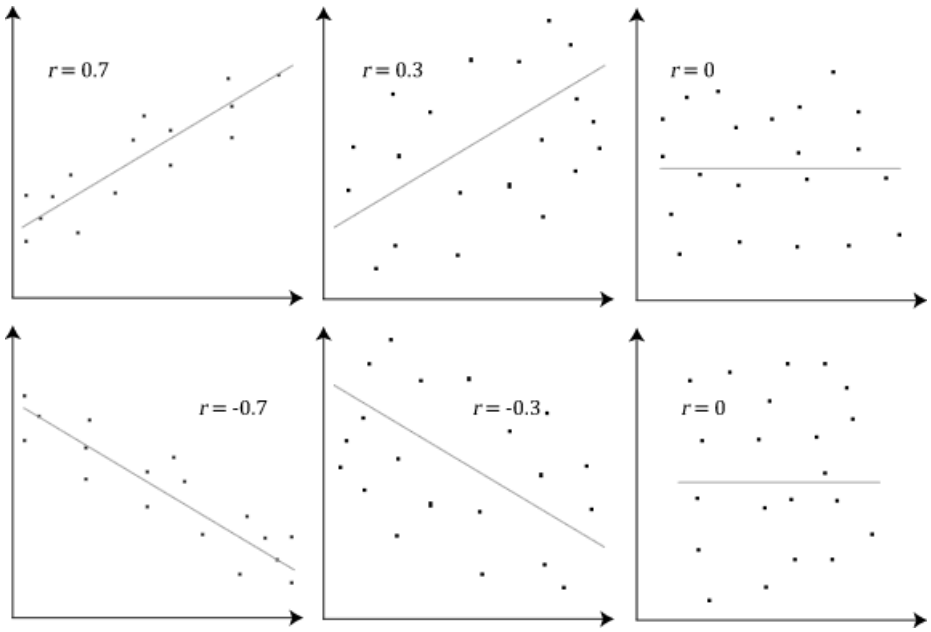
$$r_{xy} = \frac{\mathrm{Cov}_{xy}}{s_x s_y}$$

Comme les écarts-types sont des valeurs positives, le coefficient de corrélation r_{xy} a le même signe que la covariance. On peut donc immédiatement en déduire que :

- $r_{xy} > 0$ signifie que les caractères x et y varient dans le même sens ;
- $r_{xy} < 0$ signifie que les caractères x et y varient dans des sens opposés.

Mais le coefficient de corrélation linéaire permet également d'en apprendre plus au sujet du lien éventuel entre les caractères x et y, et plus particulièrement de voir si ce lien est de nature linéaire. On parle de lien linéaire si le graphique qui le représente a la forme d'une droite. Mathématiquement, cela revient à dire que l'équation qui lie les valeurs de y et de x s'écrit $y = a\,x + b$ pour certains réels a et b.

Voici quelques nuages de points et les valeurs des coefficients de corrélation linéaire correspondants (source : statistics.laerd.com).



On peut prouver que la valeur du coefficient de corrélation linéaire est toujours comprise entre -1 et 1. Sur les nuages de points ci-dessus, on peut observer que, plus r_{xy} est proche de 1 ou de -1, plus les points se condensent en une droite. À l'inverse, plus r_{xy} est proche de 0 et plus les points sont « dispersés ».

De manière générale, on peut distinguer 5 situations possibles (dont les limites sont plus ou moins floues) :

- Si $-1 \leq r_{xy} < -0.7$, les caractères y et x ont un lien (très) proche du linéaire et varient en sens opposés.
- Si $-0.7 \leq r_{xy} < -0.3$, les caractères y et x varient en sens opposés mais leur lien n'est pas clairement linéaire.
- Si $-0.3 \leq r_{xy} \leq 0.3$, on ne peut pas dire grand-chose.
- Si $0.3 < r_{xy} \leq 0.7$, les caractères y et x varient dans le même sens mais leur lien n'est pas clairement linéaire.
- Si $0.7 < r_{xy} \leq 1$, les caractères y et x varient dans le même sens et possèdent un lien (très) proche du linéaire.

Les nuages de points suivants (source : Wikipedia) illustrent graphiquement l'interprétation qu'on peut réaliser des diverses valeurs que r_{xy} peut prendre (entre -1 et 1).

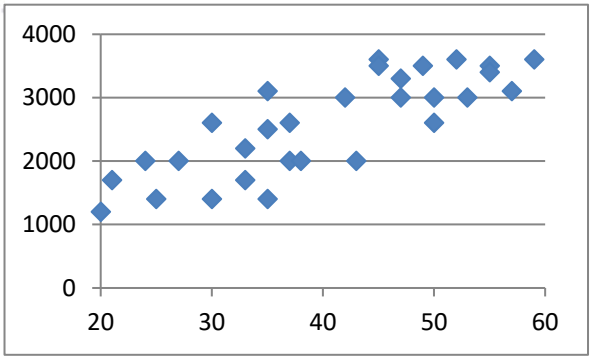
Dans le cas de l'exemple des âges et des salaires, on obtient

$$r_{xy} = \frac{\mathrm{Cov}_{xy}}{s_x s_y} = \frac{6066{,}67}{10{,}52 \times 771{,}72} = 0{,}75,$$

Ce qui semble indiquer une relation linéaire où l'âge et les salaires varient dans le même sens. Cela semble cohérent avec le nuage de points obtenu (qui est repris ci-contre).

Il faut toutefois rester prudent dans l'interprétation, comme le soulignent les deux remarques suivantes.

Tout d'abord, le coefficient de corrélation linéaire n'a pas de lien direct avec la pente de la droite liant x et y : il indique seulement si les divers points (x_i, y_i) du nuage de points sont « bien alignés », mais pas vraiment dans quelle direction. Le seul indicateur est le signe de r_{xy}, une valeur proche de 1 indiquant une pente positive (droite qui « monte ») et une valeur proche de -1 indiquant une pente négative (droite qui « descend »). Mais, dans tous les cas, si le nuage de points forme une droite quasi-parfaite, le coefficient de corrélation sera proche de -1 ou de 1, quelle que soit la pente de la droite.



Finalement, il faut rester prudent dans le cas où le coefficient de corrélation linéaire est proche de 0. Cela ne signifie pas qu'il n'y a pas de lien entre x et y mais seulement qu'il n'y a pas de *lien linéaire* apparent. Les liens autres que linéaires ne sont pas repérés par le coefficient de corrélation linéaire, comme le montrent les exemples suivants, où $r_{xy} = 0$.

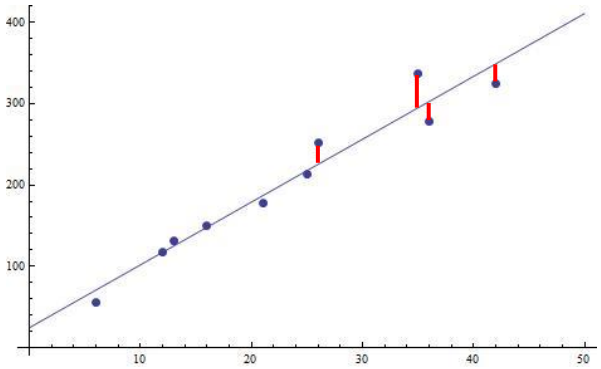
Régression linéaire

Si le coefficient de corrélation linéaire est proche de 1 en valeur absolue (c'est-à-dire s'il est proche de -1 ou de +1), on peut supposer qu'il existe un lien linéaire entre les caractères x et y. Mathématiquement, cela signifie qu'il est possible de décrire ce lien sous la forme d'une équation de droite $y = a\,x + b$.

Comme il est très rare que les points du nuage de points forment une droite parfaite (c'est-à-dire qu'ils soient parfaitement alignés), le problème revient à trouver la droite qui « colle au mieux » à ces points. La

notion de « coller au mieux » est vague et pourrait être définie de plusieurs manières. La plupart du temps, on utilise la définition dite « des **moindres carrés** ».

Si on dispose d'une droite censée approximer un nuage de points, on peut mesurer sa précision (sa pertinence) en examinant les écarts entre les points et la droite. Plus précisément, pour chacun des points (x_i, y_i), on va observer la différence entre y_i, la véritable valeur, et l'approximation que la droite donne pour x_i. Graphiquement, cela revient à mesure, pour chaque point, la longueur des segments présentés en rouge sur la représentation ci-contre.



Dans la méthode des moindres carrés, on va jauger de la pertinence d'une droite donnée en calculant la somme des carrés de tous ces écarts. Le but du jeu est évidemment de trouver une droite pour laquelle cette somme est la plus petite possible : plus la somme est petite, plus les écarts sont petits et mieux la droite « colle » aux points. Cette définition tire son nom du fait qu'on va tenter de minimiser, d'amoindrir la somme des carrés des écarts.

L'étude des moindres carrés aboutit à la formule suivante, donnant l'équation de la « **droite des moindres carrés** », aussi appelé **droite de régression linéaire**.

$$y - \bar{y} = \alpha\,(x - \bar{x}) \quad \text{où} \quad \alpha = \mathrm{Cov}_{xy}/s_x^2$$

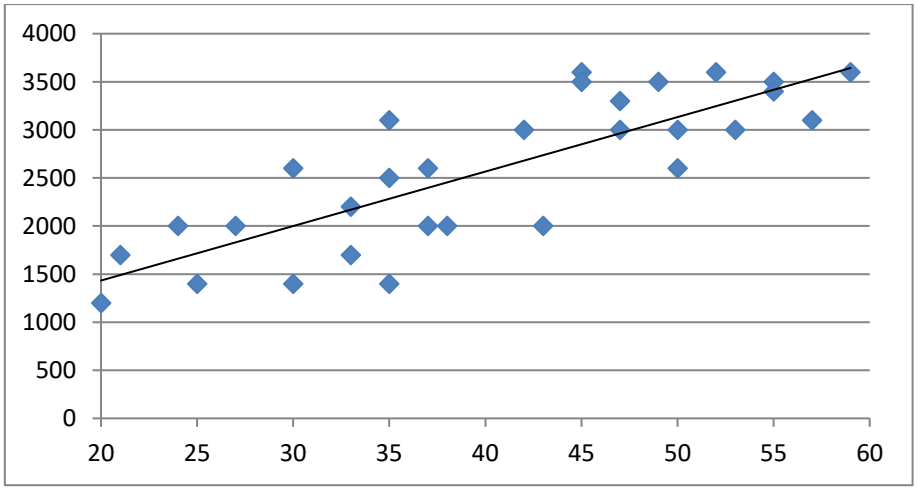
Si on applique cette formule au cas des âges et des salaires, on obtient l'équation

$$y - 2766{,}67 = \frac{6066{,}67}{110{,}66}\,(x - 41),$$

c'est-à-dire après transformation et réécriture sous forme conventionnelle,

$$y = 54{,}82x + 519{,}08$$

Si on rapporte cette droite sur le nuage de points, on obtient le graphique présenté ci-dessous.



Le coefficient de corrélation linéaire (0,75 dans le cas de l'exemple) indique à quel point le modèle donné par la droite décrit précisément les données. Plus le coefficient de corrélation linéaire r_{xy} est proche de 1

en valeur absolue, plus les données sont alignées et donc plus la régression linéaire décrit précisément la situation. À l'inverse, si le coefficient de corrélation linéaire est proche de 0, l'équation de la droite de régression linéaire n'aura que peu de valeur.

L'équation de la droite de régression permet également de réaliser des interpolations ou des extrapolations, c'est-à-dire d'estimer la valeur de y pour des valeurs de x qui n'ont pas été observées.

Ainsi, dans le cas des âges/salaires, aucune des personnes interrogées n'avait exactement 40 ans mais, grâce à l'équation de la droite de régression linéaire, on peut estimer le salaire d'un individu de 40 ans. Pour cela, il suffit de remplacer x par 40 dans l'équation, ce qui donne $y = 54{,}82 \times 40 + 519{,}08 = 2711{,}85$ €.

Exercices sur des petits échantillons (aucun groupement nécessaire)

1) Pendant dix années, un marchand de parapluies a noté le nombre de jours de pluie pendant l'été ainsi que le nombre de parapluies vendus pendant cette même période. Il a rassemblé les résultats obtenus dans le tableau suivant.

Année	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034
Jours de pluie	42	26	16	12	25	35	6	36	21	13
Ventes	325	252	150	118	213	337	55	278	178	132

   a) Calculez le nombre de jours de pluie moyen sur ces 10 étés.
   b) Calculez le nombre de parapluies vendus en moyenne au cours de ces 10 étés.
   c) Calculez les variances et écarts-types de ces distributions.
   d) Calculez la covariance et le coefficient de corrélation et interprétez-les.
   e) Déterminez l'équation de la droite de régression linéaire.
   f) Estimez le nombre de ventes qui seront réalisées au cours d'un été comportant exactement 30 jours de pluie. Est-ce une bonne approximation ?

2) Afin de comparer l'efficacité de deux aliments pour bovins appelés X et Y, on a nourri 8 vaches avec du X pendant un mois puis avec du Y pendant 1 mois et on a mesuré leur production quotidienne moyenne de lait (en kg). Voici les résultats obtenus.

Vache	1	2	3	4	5	6	7	8
Avec X	27,6	23,4	25,2	28,2	28,8	25,8	27,0	27,0
Avec Y	28,8	25,6	26,4	28	31,2	27,2	28,8	28

   a) Calculez les moyennes, variances et écarts-types des deux caractères.
   b) Calculez la covariance et le coefficient de corrélation.
   c) Les valeurs calculées au point b permettent-elles de conclure que le produit Y est plus efficace que le produit X ?
   d) Calculez l'équation de la droite de régression.
   e) D'après cette équation, combien pourrait produire une vache actuellement nourrie avec du produit X et donnant 30 kg de lait par jour si on utilisait plutôt du produit Y ?
   f) L'équation permet-elle de conclure que le produit Y est plus efficace que la produit X ?

3) Une étude sur la quantité de déchets ménagers produits par les habitants d'une commune wallonne a donné les résultats suivants, exprimés en fonction du nombre de membres composant les ménages.

Nb membres	1	2	3	4	5	6	7
Déchets (kg/an)	123	152	178	209	255	300	349

   a) Écrivez l'équation de la droite de régression linéaire exprimant la quantité de déchets produits par année en fonction de la composition du ménage.
   b) En vous basant sur ce modèle, estimez la quantité de déchets produit par un ménage de 9 personnes.
   c) Le modèle linéaire est-il pertinent ? Justifiez votre réponse.
   d) Confirmez vos réponses précédentes en représentant sur un même graphique le nuage de points et la droite de régression linéaire.

4) En analysant la comptabilité nationale d'un petit pays européen, on observe les dépenses de consommation et le revenu national à prix courants exprimés en milliards d'Euros. L'étude porte sur 10 années.

Revenu national	88	92	96	103	111	119	126	129	132	134
Consommation	64	66	70	72	78	81	85	86	88	90

   a) Calculez l'équation de la droite de régression (et représentez-la).
   b) Estimez la consommation pour la 11^{e} année sachant que le revenu national sera de 140 milliards d'Euros.
   c) Est-ce une bonne approximation ?

5) Une entreprise informatique a relevé chez dix de ses clients le nombre de pannes réseau (sur un mois) et le nombre d'appareils installés. Voici les résultats de leur étude.

Nb pannes	2	4	6	7	10	11	14	15	20	21
Nb d'appareils	24	31	40	39	55	53	61	65	78	64

   a) Donnez un modèle linéaire estimant le nombre de pannes en fonction du nombre d'appareils.
   b) Donnez une approximation du nombre de pannes auxquels il faut s'attendre pour une installation de 100 appareils.
   c) Cette approximation est-elle judicieuse ? Justifiez.

exercices avec groupement par modalités/classes

6) Reprenez l'exemple du module précédent étudiant les âges de certains conducteurs et le nombre d'accidents rapportés à l'assurance. Complétez le tableau de contingence puis calculez la covariance et le coefficient de corrélation et interprétez ces résultats. Cela a-t-il un sens de calculer l'équation de la droite de régression dans ce cas-ci ?

7) Reprenez les exercices 2 et 3 du module précédent. Calculez la covariance et le coefficient de corrélation ainsi que l'équation de la droite de régression.

8) Lors d'un stage en agence immobilière, on vous demande d'étudier l'éventuelle corrélation existant entre le nombre de rendez-vous que les agents immobiliers ont pris (en centaines) et le nombre de maisons individuelles qu'ils ont vendues. Les résultats sont résumés dans le tableau suivant.

Nb rdv \ Nb ventes	**1**	**2**	**3**	**4**
200	4	3	2	0
300	2	3	4	1
600	0	4	5	3
700	0	5	7	7

a) Calculez et interprétez les moyennes des deux caractères.
b) Calculez et interprétez la covariance.
c) Y a-t-il une relation linéaire entre les deux caractères étudiés ?

9) On a fait passer un test de connaissances générales (coté sur 15 points) à un certain nombre d'étudiants de première année en baccalauréat. Les résultats sont repris dans le tableau suivant.

Résultat \ Âge	**18**	**19**	**20**
[5, 7[	3	12	11
[7, 9[	18	21	18
[9, 11[	44	43	27
[11, 13[	30	29	18
[13, 15]	7	17	8

a) Combien d'étudiants a-t-on interrogé ?
b) Que représente le nombre 29 indiqué dans le tableau ?
c) Dans un premier temps, on ne considère que l'âge des étudiants. Quel est le mode ? Que vaut sa fréquence relative ? Indiquez ce que représentent ces valeurs en termes simples.
d) Calculez l'âge moyen des étudiants interrogés.
e) Calculez la covariance.
f) Justifiez (en vous appuyant sur des résultats statistiques) pourquoi il n'y a pas lieu de penser qu'il existe une relation linéaire entre l'âge et les résultats à ce test.