20 août 2013

Les statistiques

pour débutants

Stéphane PASQUET

MATHWEB.FR

Avant propos

Bonjour.

Cet ouvrage n'a nullement la prétention d'être le meilleur dans sa catégorie. Il est principalement destiné aux élèves et étudiants qui ont quelques difficultés à se familiariser avec les notions de statistiques vues en collège, lycée et quelques fois dans l'enseignement supérieur (selon les filières).

Il n'y a pas tout ce que l'on peut espérer trouver quand on parle de statistiques, mais si vous voyez des notions à ajouter, n'hésitez pas à me contacter pour me faire part de vos suggestions.

Bien entendu, je suis ouvert à toute critique, toute remarque constructive, donc n'hésitez pas à me contacter à l'adresse suivante si vous en estimez le besoin :

contact@mathweb.fr

En espérant que ce qui va suivre vous aidera,

Stéphane PASQUET

Sommaire

SOMMAIRE

Première partie

Statistiques à 1 variable

1 Vocabulaire

Nous allons voir ici le vocabulaire de base que l'on utilise lorsque l'on fait des statistiques. Mais avant tout, il faudrait se mettre d'accord sur ce que sont les statistiques.

Définition 1.1 (Statistique)

La **statistique** est un domaine des mathématiques qui a pour objectif d'observer, de collecter des informations et de les interpréter.

Par exemple, on pourra observer la répartition des élèves d'un collège suivant leur moyen de transport pour y venir. On pourra ensuite représenter les informations avec un graphique et éventuellement interpréter cette répartition et en tirer des conséquences sur certaines décisions à prendre.

Définition 1.2 (Série statistique)

Une **série statistique** est la donnée de toutes les informations observées.

Exemple 1.1

Observons les différents moyens de transports des élèves d'un collège pour s'y rendre. On peut représenter cette observation à l'aide d'un tableau :

Voiture	Bus	Métro	À pieds	Vélo
36	54	130	47	23

Ce tableau signifie que 36 élèves viennent au collège en voiture, 54 viennent en bus, etc.

Définition 1.3 (Caractère, variable ou modalités)

Le **caractère** (ou la **variable** ou la **modalité**) d'une série statistique est ce que l'on observe.

Exemple 1.2

Dans notre exemple, le caractère est le *moyen de transport* qu'utilisent les élèves pour venir au collège.

Il existe trois types de caractères :

- les caractères **qualitatifs**, c'est-à-dire les caractères qui sont représentés par des mots (comme dans l'exemple que nous avons pris).
- les caractères **quantitatifs discrets**, c'est-à-dire les caractères qui sont représentés par des nombres bien précis (par exemple, si l'on observe le nombre d'enfants par foyer, le caractère observé est quantitatif discret car il peut valoir « 0 », « 1 », « 2 », etc.) ; on parle alors de **modalités**.
- les caractères **quantitatifs continus**, c'est-à-dire les caractères qui sont représentés par des intervalles (par exemple, si l'on observe le salaire d'un ouvrier, il vaut mieux les regrouper. On peut donc voir ceux qui sont inférieurs à 800 €, ceux qui sont compris entre 800 € et 900 €, etc.) ; on parle alors de **classes**.

Définition 1.4 (Effectif)

Un **effectif** est un nombre quantifiant un caractère.

Exemple 1.3

Dans l'exemple que nous avons pris, « 36 » est l'effectif qui correspond au caractère « Voiture », « 54 » est celui qui correspond à « Bus », etc.

Définition 1.5 (Population)

La **population** d'une série statistique est l'ensemble dans lequel nous nous plaçons pour faire nos observations.

Exemple 1.4

Dans notre exemple, la population est le collège.

Définition 1.6 (Individu)

Un **individu** d'une série statistique est un élément de la population.

Exemple 1.5

Dans notre exemple, les individus sont les élèves.

Définition 1.7 (Classe)

Dans le cas d'une série statistique à caractère qualitatif continu, le caractère sera appelé une **classe**.

Exemple 1.6

Si l'on observe la répartition des salaires dans une entreprise, chaque intervalle est un mode. Par exemple, [800;900[est un mode, de même pour [900;1 000[, etc.

Définition 1.8 (Mode ou Classe modale)

Dans une série statistique, le **mode** est le caractère ayant l'effectif le plus grand. Dans le cas d'une série statistique à caractères qualitatifs continus, on parle de **classe modale**.

Exemple 1.7

Dans notre premier exemple, le mode est « Métro ».

Dans ce qui va suivre, nous allons nous pencher avant tout sur les séries statistiques à 1 variable, mais plus tard, nous verrons comment traiter les statistiques à 2 variables.

2 Représentations graphiques

Nous l'avons vu précédemment, un moyen assez naturel de représenter nos observations est d'utiliser un tableau. Il peut être *simple* :

Moyens de transport	**Voiture**	**Bus**	**Métro**	**À pieds**	**Vélo**
Effectifs	36	54	130	47	23

ou *à double entrée* :

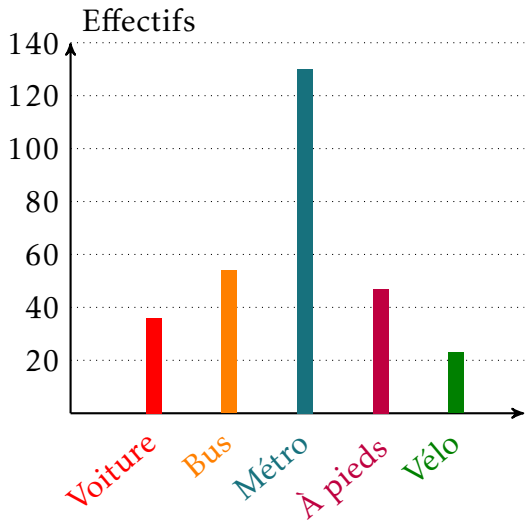
Langues / **Âges**	**Espagnol**	**Anglais**	**Allemand**
13 ans	1	7	3
14 ans	18	105	25
15 ans	1	8	2

Dans ce dernier tableau, nous observons la répartition des adolescents d'une colonie de vacances selon leur première langue vivante et leur âge.
« 105 » représente donc le nombre d'adolescents de 14 ans ayant l'anglais comme première langue vivante.

Mais afin de nous rendre compte visuellement des différentes répartitions, on peut utiliser différents diagrammes.

Le diagramme en tuyaux d'orgue

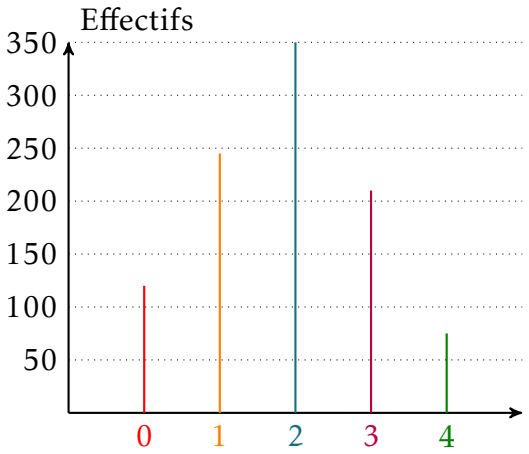
Il est utilisé pour représenter des caractères quantitatifs discrets, comme dans notre premier exemple. La hauteur des tuyaux est proportionnelle à l'effectif correspondant.



Le diagramme en bâtons

Il est utilisé pour représenter des caractères quantitatifs discrets.
Prenons l'exemple suivant qui représente le nombre d'enfants par foyer sur une population de 1 000 individus. La hauteur des bâtons est proportionnelle à l'effectif correspondant.

Nombres d'enfants par foyer	0	1	2	3	4
Effectifs	120	245	350	210	75



L'histogramme

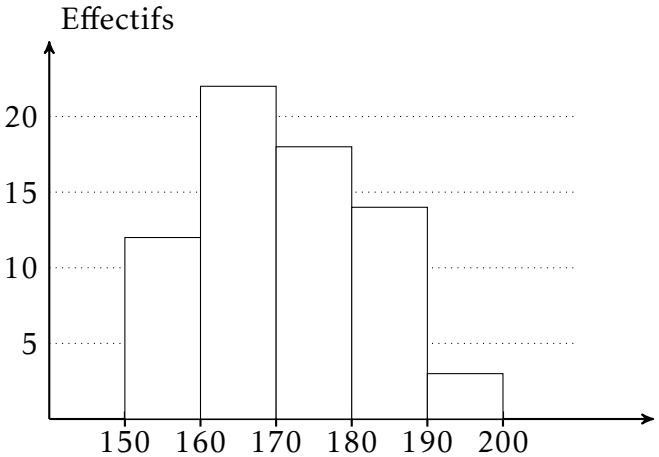
Il est utilisé pour des caractères quantitatifs continus.

Attention

Ce n'est pas la hauteur des rectangles qui est proportionnelle à l'effectif considéré, mais la surface du rectangle.

Prenons l'exemple suivant qui représente la répartition de certains individus d'une population suivant leur taille.

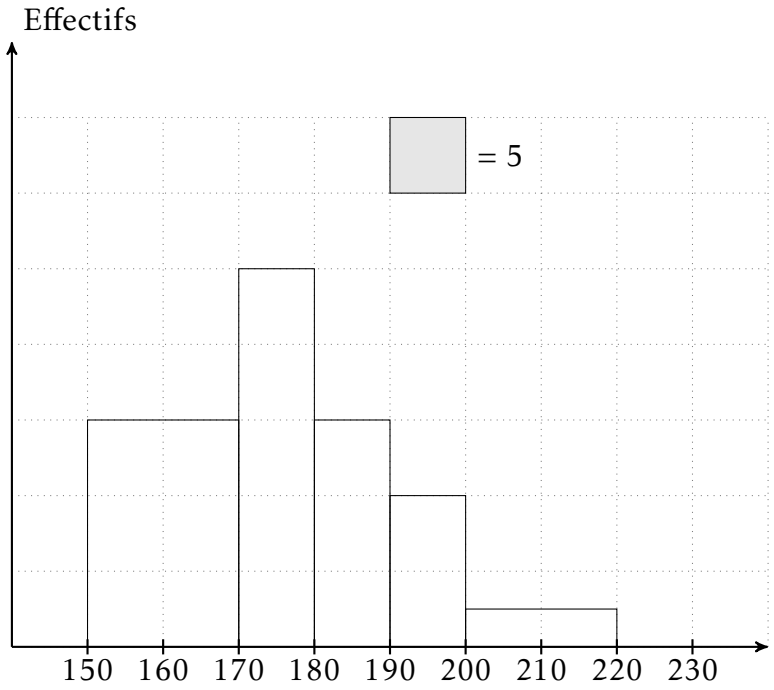
Tailles (en cm)	[150 ;160[	[160 ;170[	[170 ;180[	[180 ;190[	[190 ;200[
Effectifs	12	22	18	14	3



Ici, tous les intervalles ont la même *amplitude*, ce qui a pour conséquence que la hauteur de chaque rectangle est proportionnelle à chaque effectif.

Voyons un autre exemple où les classes n'ont pas la même amplitude.

Tailles (en cm)	**[150 ;170[**	**[170 ;180[**	**[180 ;190[**	**[190 ;200[**	**[200 ;230[**
Effectifs	30	25	15	10	5



Ici, j'ai décidé de représenter un effectif de 5 par un carré. Ainsi, les 30 personnes de la première classe sont représentés par 6 carrés et les 3 personnes de la dernière classe sont représentés par 1 rectangle dont la surface est égale à celle d'un carré.

Le diagramme circulaire

Il est en général adopté pour représenter des caractères qualitatifs. Reprenons donc notre exemple du début (les moyens de transports des collégiens). La population est représentée par un disque qui est partitionné en plusieurs secteurs représentant chacun des caractères. Ainsi, l'angle au centre de chaque secteur angulaire doit être proportionnel à l'effectif correspondant. Il faut donc ajouter une ligne au tableau :

Moyens de transport	**Voiture**	**Bus**	**Métro**	**À pieds**	**Vélo**	**Totaux**
Effectifs	36	54	130	47	23	290
Angles (en degrés)	45	67	161	58	29	360

Pour calculer l'angle correspondant à l'effectif « 36 », on trouve la quatrième proportionnelle du tableau suivant :

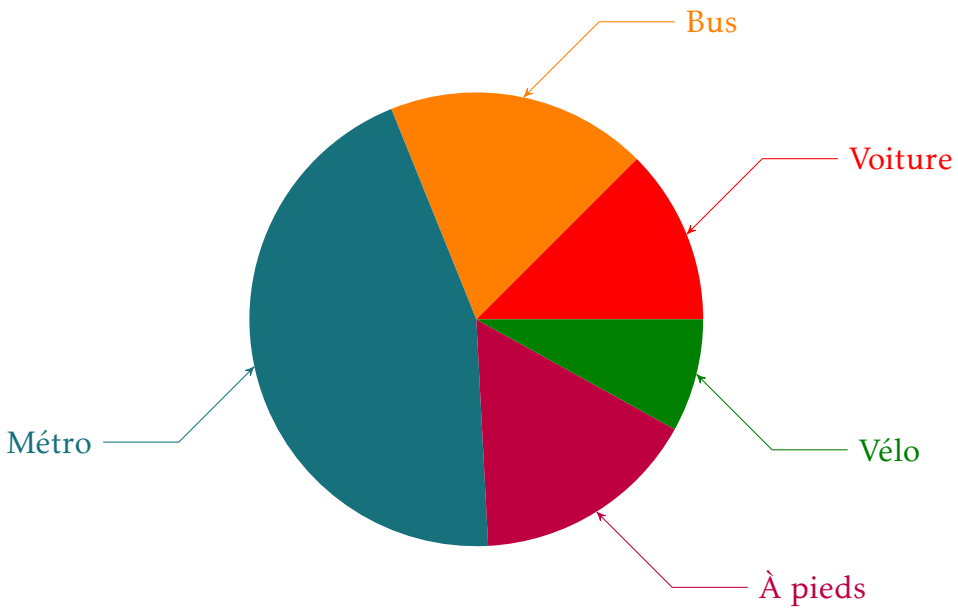
Effectifs	**290**	**36**
Angles (en degrés)	360	x

$$x = \frac{36 \times 360}{290}$$
$$x \approx 45$$

Attention

Il faudra toujours veiller à arrondir correctement de sorte que le total fasse bien 360°.

Le diagramme est donc le suivant :



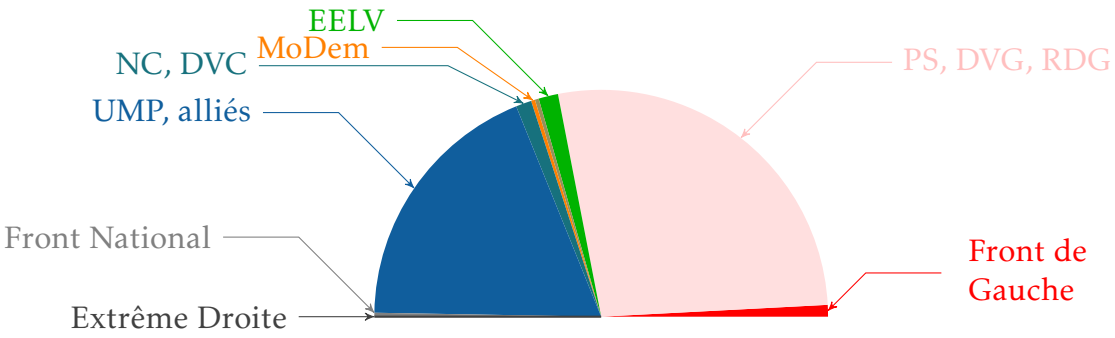
Remarque

On peut mettre les légendes comme sur l'exemple ci-dessus ou, si on a la place, directement dans les secteurs angulaires.

Le diagramme semi-circulaire

Le principe est le même que pour le diagramme circulaire sauf qu'ici, la population est représentée par un demi-disque. Ce diagramme est utilisé pour représenter, entre autre, la répartition des députés à l'Assemblée Nationale.

Partis politiques	Effectifs	Angles
Front de Gauche	10	3°
PS, Divers gauche, Radicaux de gauche	314	98°
EELV	17	5°
Indépendants martiniquais	2	1°
MoDem	2	1°
Nouveau centre et divers centre	14	4°
UMP et alliés	215	67°
Front National	2	1°
Extrême Droite	1	0°



Remarque

Ici, il a été difficile d'arrondir les angles car, comme vous pouvez le voir, on arrive à un total de 180° avant le dernier caractère. Mais celui-ci étant négligeable devant les autres, on peut le représenter par un simple trait (donc avec un angle de 0°).

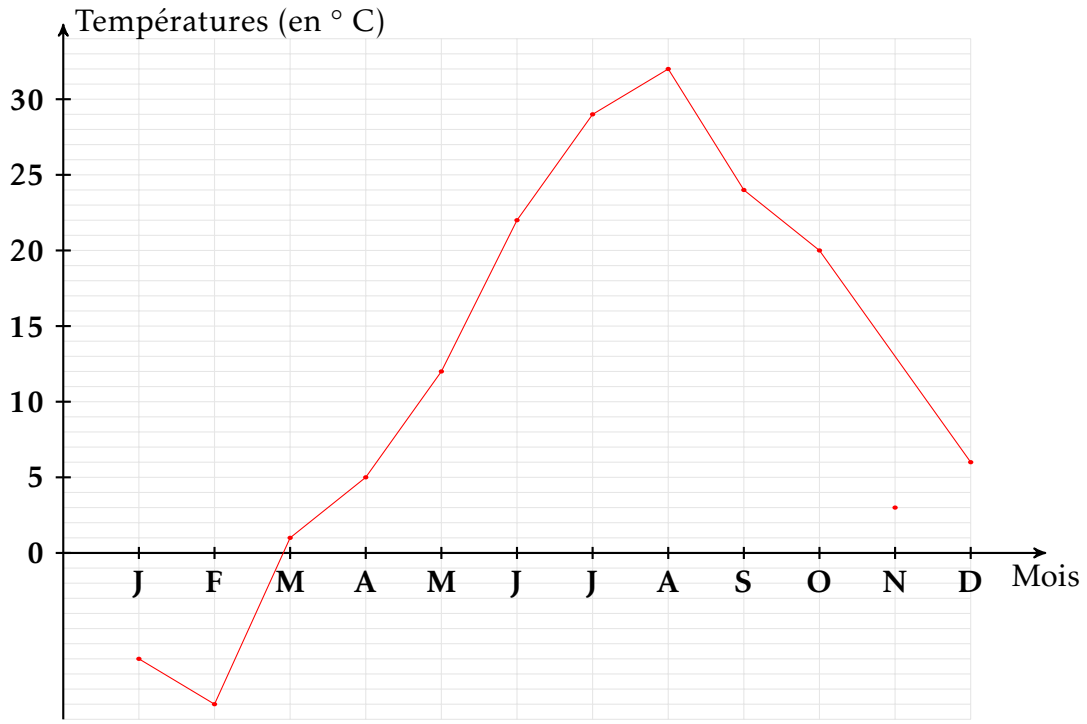
La courbe cartésienne

Dans certains cas (notamment quand il y a une notion de chronologie), il sera utile de représenter la série statistique par une courbe cartésienne. Voyons sur un exemple ce que cela donne.

Penchons-nous sur le tableau suivant qui donne les températures moyennes mensuelles (en ° C) relevées dans la ville imaginaire « Mathtown » au cours de l'année 2011 :

Mois	J	F	M	A	M	J	J	A	S	O	N	D
Temp. (° C)	−7	−10	1	5	12	22	29	32	24	20	13	6

Plaçons maintenant dans un repère cartésien le point d'ordonnée −7 en premier (qui correspond à la température moyenne relevée au cours du mois de janvier), puis −10 (qui correspond à celle du mois de février), etc.



3 Position et dispersion

Nous allons maintenant nous pencher sur un autre aspect de l'analyse d'une série statistique à une variable.

Fréquences

Définition 3.1

Soit $(x_i;n_i)_{1\leqslant i\leqslant N}$ une série statistique.
La **fréquence** du caractère x_i, notée f_i, est :

$$f_i = \frac{n_i}{n_1+n_2+n_3+\cdots+n_N}$$

Une fréquence peut être exprimée :

- sous forme de fraction
- en pourcentage
- sous forme de nombre décimal

Exemple 3.1

Reprenons le premier exemple du chapitre 1 :

Moyens de transport	Voiture	Bus	Métro	À pieds	Vélo	Total
Effectifs	36	54	130	47	23	290
Fréquences	$\frac{36}{290}=\frac{18}{145}$	$\frac{27}{145}$	$\frac{13}{29}$	$\frac{47}{290}$	$\frac{23}{290}$	1

Définition 3.2 (Cumulation)

Soit $(x_i;n_i)_{1\leqslant i\leqslant N}$ une série statistique.
On appelle **effectif cumulé croissant (e.c.c.)** du caractère x_i le nombre :

$$ecc_i = n_1+n_2+\cdots+n_i$$

On appelle **fréquence cumulée croissante (e.c.c.)** du caractère x_i le nombre :

$$fcc_i = f_1+f_2+\cdots+f_i$$

Exemple 3.2

Moyens de transport	Voiture	Bus	Métro	À pieds	Vélo
Effectifs	36	54	130	47	23
e.c.c.	36	$36+54=90$	$90+130=220$	$220+47=267$	$267+23=290$
Fréquences	$\frac{36}{290}=\frac{18}{145}$	$\frac{27}{145}$	$\frac{13}{29}$	$\frac{47}{290}$	$\frac{23}{290}$
f.c.c.	$\frac{36}{290}$	$\frac{90}{290}$	$\frac{220}{290}$	$\frac{267}{290}$	$\frac{290}{290}=1$

Définition 3.3 (Densité)

Soit une série statistique à caractère quantitatif continu.
La **densité d'effectif** d'une classe est le rapport de son effectif avec son amplitude.
La **densité de fréquence** d'une classe est le rapport de sa fréquence avec son amplitude.

Exemple 3.3

Le tableau suivant donne les différentes tailles observées au sein d'une population.

Tailles (en cm)	[140 ;170[	[170 ;180[	[180 ;190[	[190 ;210[
Effectifs	13	25	21	30
Densité d'effectif	$\frac{13}{30}\approx 0,43$	$\frac{25}{10}=2,5$	$\frac{21}{10}=2,1$	$\frac{30}{20}=1,5$

Paramètres de position

Définition 3.4

Les paramètres de position d'une série statistique à une variable sont sa **moyenne**, son **mode** et ses **quantiles** (médiane, quartiles, déciles, centiles).

Remarque

Ces paramètres sont des valeurs numériques qui « résument » la série statistique en donnant un ordre de grandeur des observations.
Ils s'expriment toujours en la même unité que les observations.

3.2.1 La moyenne

Définition 3.5

Soit la série statistique $(x_i;n_i)_{1\leqslant i\leqslant N}$. Alors, sa **moyenne** est le nombre défini par :

$$\overline{x}=\frac{1}{N}\sum_{i=1}^{N}x_i=\frac{x_1+x_2+x_3+\cdots+x_N}{N}$$

Exemple 3.4

1 Reprenons l'exemple des températures relevées à « Mathtown » en 2011

Mois	J	F	M	A	M	J	J	A	S	O	N	D
Temp. (° C)	−7	−10	1	5	12	22	29	32	24	20	13	6

Alors, la température moyenne sur toute l'année est :

$$\overline{x} = \frac{(-7)+(-10)+1+5+12+\cdots+24+20+13+6}{12} = 12,25$$

2 Considérons maintenant le tableau suivant :

Tailles (en cm)	[160 ;170[	[170 ;180[	[180 ;190[	[190 ;200[
Effectifs	13	25	21	9

Pour calculer la taille moyenne de la population, on considère le milieu de chaque intervalle :

$$\begin{aligned}\overline{x} &= \frac{165\times 13+175\times 25+185\times 21+195\times 9}{13+25+21+9}\\ &\approx 179\end{aligned}$$

Définition 3.6 (Moyenne pondérée)

Soit $(x_i;n_i)_{1\leqslant i\leqslant N}$ une série statistique, où les x_i sont des caractères quantitatifs. On appelle **moyenne pondérée** de la série statistique le nombre défini par :

$$\overline{x} = \frac{n_1x_1+n_2x_2+n_3x_3+\cdots+n_Nx_N}{n_1+n_2+n_3+\cdots+n_N}$$

Exemple 3.5

On a relevé dans le tableau suivant les notes d'une classe lors d'un devoir sur table :

Notes (x_i)	3	4	6	7	9	10	11	12	14	15	18	19
Effectifs (n_i)	1	1	2	5	4	3	4	5	3	2	1	1

La moyenne de la classe à ce contrôle est donc :

$$\begin{aligned}\overline{x} &= \frac{3\times 1+4\times 1+6\times 2+7\times 5+9\times 4+\cdots+15\times 1+18\times 1+19\times 1}{1+1+2+5+4+3+4+5+3+2+1+1}\\ \overline{x} &= \frac{333}{32}\\ \overline{x} &\approx 10,4\end{aligned}$$

Proposition 3.1 (Moyenne de moyennes)

Soient $(x_i;n_i)_{1\leqslant i\leqslant N}$ et $(y_i;m_i)_{1\leqslant i\leqslant M}$ deux séries statistiques de moyennes respectives $\overline{x}$ et $\overline{y}$.
Alors, la moyenne de la série regroupant ces deux séries est :

$$\overline{m} = \frac{\overline{x}\sum_{i=1}^{N} n_i + \overline{y}\sum_{i=1}^{M} m_i}{\sum_{i=1}^{N} n_i + \sum_{i=1}^{M} m_i}$$

Démonstration

Par hypothèse, on a :

$$\overline{x} = \frac{x_1 n_1 + \cdots + x_N n_N}{n_1 + \cdots + n_N} \quad ; \quad \overline{y} = \frac{y_1 m_1 + \cdots + y_M m_M}{m_1 + \cdots + m_M}$$

On peut alors écrire :

$$\overline{x}\sum_{i=1}^{N} n_i = \sum_{i=1}^{N} x_i n_i \quad \text{et} \quad \overline{y}\sum_{i=1}^{M} m_i = \sum_{i=1}^{M} y_i m_i$$

Ainsi :

$$\overline{x}\sum_{i=1}^{N} n_i + \overline{y}\sum_{i=1}^{M} m_i = \sum_{i=1}^{N} x_i n_i + \sum_{i=1}^{M} y_i m_i$$

Et donc, en divisant les deux membres de cette égalité par $\sum_{i=1}^{N} n_i + \sum_{i=1}^{M} m_i$:

$$\frac{\overline{x}\sum_{i=1}^{N} n_i + \overline{y}\sum_{i=1}^{M} m_i}{\sum_{i=1}^{N} n_i + \sum_{i=1}^{M} m_i} = \underbrace{\frac{\sum_{i=1}^{N} x_i n_i + \sum_{i=1}^{M} y_i m_i}{\sum_{i=1}^{N} n_i + \sum_{i=1}^{M} m_i}}_{=\overline{m}}$$

On obtient alors la formule de la propriété. ■

Exemple 3.6

Dans la même classe que précédemment, les notes lors du contrôle suivant on été :

Notes (x_i)	2	3	5	7	9	10	11	12	14	15	18	20
Effectifs (n_i)	2	1	2	3	5	2	8	2	3	2	1	1

La moyenne de cette classe à ce contrôle est donc :

$$\overline{y} = \frac{4+3+10+21+45+20+88+24+42+30+18+20}{32}$$
$$\overline{y} \approx 9,84$$

Ainsi, la moyenne de cette classe sur les deux contrôles est :

$$\overline{m} = \frac{10,4 \times 32 + 9,84 \times 32}{32+32}$$
$$\overline{m} \approx 10,12$$

Proposition 3.2 (Linéarité de la moyenne)

Soit $(x_i;n_i)_{1\leqslant i\leqslant N}$ une série statistique de moyenne $\overline{x}$. Alors, la moyenne de la série $(ax_i+b;n_i)_{1\leqslant i\leqslant N}$, où a et b sont deux réels non nuls, est :

$$\overline{y} = a\overline{x} + b$$

Démonstration

$$\begin{aligned}\frac{\sum_{i=1}^{N} n_i(ax_i+b)}{\sum_{i=1}^{N} n_i} &= \frac{\sum_{i=1}^{N} an_ix_i + \sum_{i=1}^{N} bn_i}{\sum_{i=1}^{N} n_i} \\ &= \frac{a\sum_{i=1}^{N} n_ix_i + b\sum_{i=1}^{N} n_i}{\sum_{i=1}^{N} n_i} \\ &= a\frac{\sum_{i=1}^{N} n_ix_i}{\sum_{i=1}^{N} n_i} + b \\ &= a\overline{x} + b\end{aligned}$$

■

Exemple 3.7

Les notes du second contrôle de la classe précédemment observée étant décevante, le professeur décide de multiplier toutes les notes par 1,1 et d'ajouter 0,5.
Ainsi, la série statistique obtenue est $(1,1x_i + 0,5)_{i\leqslant i\leqslant 32}$. La nouvelle moyenne est donc :

$$\overline{y} \approx 1,1 \times 9,84 + 0,5 \qquad \text{soit} \qquad \overline{y} \approx 11,32$$

3.2.2 Le mode

Définition 3.7

Le **mode** d'une série est le caractère pour lequel l'effectif est le plus grand.
Si le caractère est quantitatif continu, on parle de **classe modale**.

- Si les classes sont de même amplitude, la classe modale sera la classe où l'effectif est le plus élevé ;
- Si les classes ne sont pas de même amplitude, la classe modale sera la classe où la densité d'effectif (ou la densité de fréquence) est la plus élevée.

S'il existe deux caractères pour lesquels l'effectif est maximal,

- si les caractères se suivent, on parle d'**intervalle modal** ;
- si les caractères ne se suivent pas, on dit que la série est **plurimodale**.

Exemple 3.8

1 Dans l'exemple suivant :

Notes (x_i)	2	3	5	7	9	10	11	12	14	15	18	20
Effectifs (n_i)	2	1	2	3	5	2	8	2	3	2	1	1

Le mode est « 11 ».

2 Dans l'exemple suivant :

Notes (x_i)	2	3	5	7	9	10	11	12	14	15	18	20
Effectifs (n_i)	2	1	2	3	5	2	8	8	3	2	1	1

L'intervalle modal est « [11;12] ».

3 Dans l'exemple suivant :

Tailles (en cm)	[160 ;170[	[170 ;180[	[180 ;190[	[190 ;200[
Effectifs	13	25	21	9

La classe modale est « [170 ;180[».

Exemple 3.8 (suite)

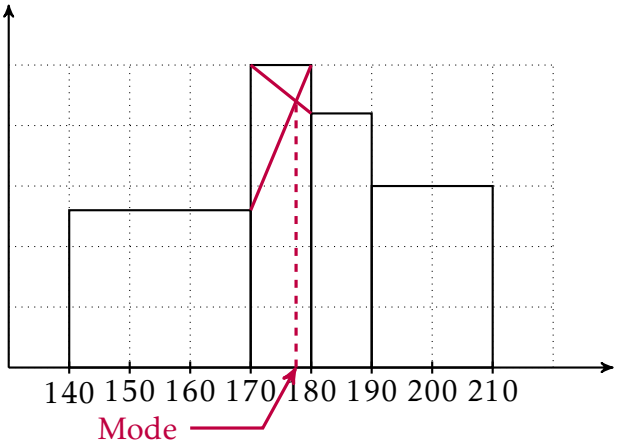
4 Dans l'exemple suivant :

Tailles (en cm)	[140 ;170[	[170 ;180[	[180 ;190[	[190 ;210[
Effectifs	13	25	21	30
Densité d'effectif	$\frac{13}{30} \approx 0,43$	$\frac{25}{10} = 2,5$	$\frac{21}{10} = 2,1$	$\frac{30}{20} = 1,5$

La classe modale est « [170;180[».

Méthode

Dans le cas des caractères quantitatifs continus, on peut déterminer avec précision le mode en joignant les bornes inférieures et supérieures des classes qui précèdent et succèdent la classe modale.



3.2.3 Les quantiles

La médiane

Définition 3.8

La **médiane** d'une série statistiques est le caractère qui la partage en deux séries de même effectif. On la note M_e.

Méthode

Il est souvent très pratique d'avoir les effectifs (ou fréquences) cumulé(e)s croissant(e)s pour déterminer la médiane.

Exemple 3.9

Notes (x_i)	2	3	5	7	9	10	11	12	14	15	18	20
Effectifs (n_i)	2	1	2	3	5	2	8	8	3	2	1	1
E.c.c.	2	3	5	8	13	15	23	31	34	36	37	38

La moitié de « 38 » est « 19 », nombre atteint (sur la dernière ligne) pour le caractère « 11 » ; ainsi, $M_e = 11$.

Remarque

Lorsque l'effectif total de la série est pair et que la moitié de cet effectif total est exactement atteint pour un caractère, la médiane est égale à la moyenne de ce caractère et du suivant.

Exemple 3.10

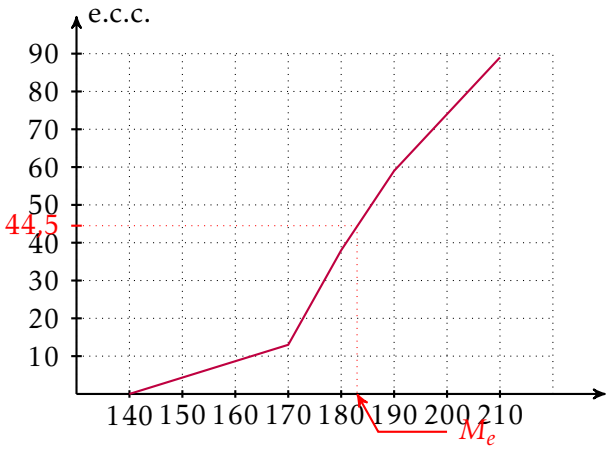
Notes (x_i)	2	3	5	7	9	10	11	12	14	15	18	20
Effectifs (n_i)	2	1	2	3	5	2	4	12	3	2	1	1
E.c.c.	2	3	5	8	13	15	19	31	34	36	37	38

La moitié de « 38 » est « 19 », nombre atteint exactement pour le caractère « 11 » ; ainsi, $M_e = 11,5$ (moyenne de 11 et 12).

Méthode

Quand on a une série statistique à caractères quantitatifs continus, on détermine la médiane à l'aide du **polygone des e.c.c. (ou f.c.c.)**.

Tailles (en cm)	[140 ;170[	[170 ;180[	[180 ;190[	[190 ;210[
Effectifs	13	25	21	30
E.c.c.	13	38	59	89



Remarque

On pourrait calculer avec précision la médiane en trouvant l'équation du segment sur [180 ;190[.
En effet, ce segment passe par les points de coordonnées (180;38) et (190;59). Donc son coefficient directeur est :

$$m = \frac{59-38}{190-180} = \frac{21}{10}$$

L'équation est donc de la forme :

$$y = \frac{21}{10}x + p$$

Donc :

$$p = 59 - \frac{21}{10} \times 190 = 59 - 21 \times 19 = -340$$

Donc, l'équation du segment sur [180 ;190[est :

$$y = 2,1x - 340$$

Ainsi, nous savons que :

$$44,5 = 2,1M_e - 340$$

Donc :

$$M_e = \frac{10}{21}(44,5 + 340) \approx 183$$

La médiane est donc à peu près égale à 183 cm.

Les quartiles

Définition 3.9

Le **premier quartile** d'une série statistique (noté Q_1) est le caractère qui partage la série en deux séries d'effectifs 25% et 75%.
Le **troisième quartile** d'une série statistique (noté Q_3) est le caractère qui partage la série en deux séries d'effectifs 75% et 25%.

Exemple 3.11

Notes (x_i)	2	3	5	7	9	10	11	12	14	15	18	20
Effectifs (n_i)	2	1	2	3	5	2	4	12	3	2	1	1
E.c.c.	2	3	5	8	13	15	19	31	34	36	37	38

25% de 38 est égal à 9,5, atteint pour « 9 » ; donc $Q_1 = 9$.
75% de 38 est égal à 28,5, atteint pour « 12 » ; donc $Q_3 = 12$.

Les déciles

Définition 3.10

Le **premier décile** d'une série statistique (noté d_1) est le caractère qui partage la série en deux séries d'effectifs 10% et 90%.
Le **neuvième décile** d'une série statistique (noté d_9) est le caractère qui partage la série en deux séries d'effectifs 90% et 10%.

Exemple 3.12

Notes (x_i)	2	3	5	7	9	10	11	12	14	15	18	20
Effectifs (n_i)	2	1	2	3	5	2	4	12	3	2	1	1
E.c.c.	2	3	5	8	13	15	19	31	34	36	37	38

10% de 38 est égal à 3,8, atteint pour « 5 » ; donc $d_1 = 5$.
90% de 38 est égal à 34,2, atteint pour « 14 » ; donc $d_9 = 14$.

Remarque

On pourrait définir n'importe quel quantile d'ordre α % comme le premier caractère tel que α % des valeurs soient inférieures à ce caractère.

3.2.4 Diagramme en boîte

Étendue

Définition 3.11

L'**étendue** d'une série statistique à caractères quantitatifs est la différence entre la plus grande et la plus petite des valeurs des caractères.

$$e = x_{\max} - x_{\min}$$

Exemple 3.13

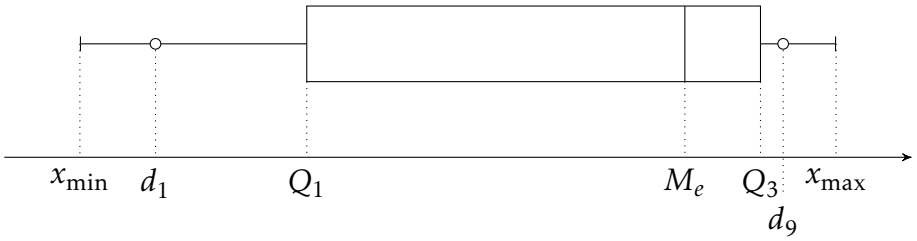
Dans l'exemple précédent (notes), la plus petite valeur est « 2 » et la plus grande, « 20 ». Donc :

$$e = 20 - 2 = 18$$

La boîte à moustaches

La représentation graphique d'une série statistique à une variable est appelée **diagramme en boîte** ou **boîte à moustaches**.

Elle à l'allure suivante :



Remarque

En mettant plusieurs boîtes à moustaches les unes en-dessous des autre, on peut comparer visuellement plusieurs séries statistiques.

Les paramètres de dispersion

3.3.1 Variance et écart-type

Définition 3.12

Soit $(x_i;n_i)_{1\leqslant i\leqslant N}$ une série statistique.
La **variance** de cette série est le nombre :

$$V = \frac{1}{n_1+n_2+\cdots+n_N}\sum_{i=1}^{N} n_i\,(x_i-\overline{x})^2$$

où $\overline{x}$ représente la moyenne de la série.
L'**écart-type** de la série statistique est le nombre :

$$\sigma = \sqrt{V}$$

Exemple 3.14

Notes (x_i)	2	3	5	7	9	10	11	12	14	15	18	20
Effectifs (n_i)	2	1	2	3	5	2	8	2	3	2	1	1

On avait calculé précédemment : $\overline{x} \approx 9,84$. Donc :

$$V = \frac{1}{38}[2(2-9,84)^2+1(3-9,84)^2+2(5-9,84)^2+\cdots+1(20-9,84)^2]$$
$$V = \frac{539,4192}{38}$$
$$V \approx 14,2$$

D'où :

$$\sigma \approx \sqrt{14,2} \approx 3,77$$

Théorème 3.1 de KOENIG (OU KÖNIG)

Soit $(x_i;n_i)_{1\leqslant i\leqslant N}$ une série statistique.
Alors,

$$V=\frac{\sum_{i=1}^{N}n_ix_i^2}{\sum_{i=1}^{N}n_i}-(\overline{x})^2$$

Démonstration

Par définition, on a :

$$V=\frac{\sum_{i=1}^{N}n_i\left(x_i-\overline{x}\right)^2}{\sum_{i=1}^{N}n_i}$$

$$=\frac{\sum_{i=1}^{N}n_i\left(x_i^2-2x_i\overline{x}+\overline{x}^2\right)}{\sum_{i=1}^{N}n_i}$$

$$V=\frac{\sum_{i=1}^{N}n_ix_i^2}{\sum_{i=1}^{N}n_i}-2\overline{x}\frac{\sum_{i=1}^{N}n_ix_i}{\sum_{i=1}^{N}n_i}+\overline{x}^2\frac{\sum_{i=1}^{N}n_i}{\sum_{i=1}^{N}n_i}$$

$$V=\frac{\sum_{i=1}^{N}n_ix_i^2}{\sum_{i=1}^{N}n_i}-2\overline{x}\times\overline{x}+\overline{x}^2$$

$$V=\frac{\sum_{i=1}^{N}n_ix_i^2}{\sum_{i=1}^{N}n_i}-\overline{x}^2$$

■

Exemple 3.15

Notes (x_i)	2	3	5	7	9	10	11	12	14	15	18	20
Effectifs (n_i)	2	1	2	3	5	2	8	2	3	2	1	1

D'après le théorème de Koenig, on a :

$$V = \frac{2\times 2^2 + 1\times 3^2 + 2\times 5^2 + \cdots + 1\times 20^2}{38} - (9,84)^2$$
$$V = \frac{8+9+50+147+\cdots+400}{38} - 96,8256$$
$$V = \frac{3\,837}{38} - 96,8256$$
$$V \approx 4,15$$

Remarque

Nous n'obtenons pas ici exactement la même variance que précédemment car trop arrondir implique qu'au final, la valeur approchée obtenue est peu précise.

3.3.2 Les intervalles interquantiles

Définition 3.13

Les **intervalles interquantiles** d'ordre α % contiennent α % des valeurs observées centrales.

- L'intervalle interquartile est l'intervalle $[Q_1; Q_3]$ (ou $[Q_{25}; Q_{75}]$) ; il contient 50 % des valeurs observées centrales ;
- L'intervalle interdécile est l'intervalle $[d_1; d_9]$; il contient 80 % des valeurs observées centrales ;
- L'intervalle intercentile est l'intervalle $[c_1; c_{99}]$; il contient 98 % des valeurs observées centrales.

4 Échantillonnage

Nous allons ici aborder un chapitre délicat.
Pour l'introduire, nous allons nous baser sur un exemple unique :

« On lance un dé équilibré à 6 faces 25 fois de suite. »

Définition 4.1

Un **échantillonnage** de taille n est constitué des résultats d'une expérience répétée de façon indépendante n fois.

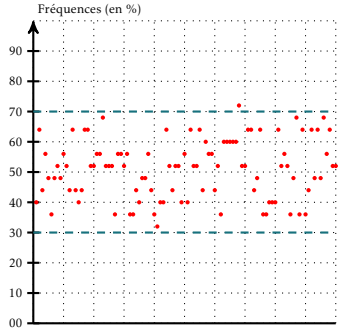
Dans notre exemple, $n = 25$. L'expérience répétée est le lancer du dé.
On peut relever les fréquences d'apparition des différentes faces du dé : on obtiendra alors ce que l'on appelle la **distribution des fréquences**.

Maintenant, si on considère 100 échantillons de même taille, il est fort à parier que les fréquences vont varier d'un échantillon à l'autre.

Définition 4.2

On appelle **fluctuation d'échantillonnage** la variation de la distribution des fréquences d'un échantillonnage à un autre.

On lance 25 fois fois un dé à 6 faces et on note maintenant la fréquence d'apparition des faces portant un nombre premier (2, 3 ou 5).
On répète 100 fois cette expérience et on représente les résultats par le graphique ci-contre :



On peut remarquer que les fréquences se situent aux alentours de 50 % (probabilité d'obtenir 2, 3 ou 5). On s'intéresse alors à un intervalle qui contiendrait 95 % des fréquences.
Sur le graphique, on peut considérer l'intervalle $[30;70]$ et on pourrait le considérer comme étant l'intervalle $[50-20;50+20]$.

Proposition 4.1

Considérons une population dans laquelle nous savons que la probabilité d'apparition d'un caractère A est égale à p, où $0,2 \leqslant p \leqslant 0,8$.
Considérons les échantillons de taille $n \geqslant 25$.
Dans 95 % des échantillons de taille n, la fréquence d'apparition du caractère A dans l'échantillon se trouve dans l'intervalle :

$$f \in \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right].$$

Ce dernier intervalle est appelé **un intervalle de fluctuation au seuil de** 95 %.

Remarque

Cela signifie qu'il est rare (5 %) de trouver un échantillon dans lequel la fréquence d'apparition du caractère A n'est pas dans l'intervalle.
On pourra alors considérer qu'un tel échantillon n'est pas « conforme à nos attentes », avec un risque de se tromper de 5 %.

Comment utiliser ce résultat ? Voyons un exemple classique :

Exemple 4.1

Sachant que dans la population française il y a autant d'hommes que de femmes (à peu près), on voudrait savoir si deux entreprises respectent le principe de parité dans leur recrutement.
L'une a recruté 42 femmes sur 100 employés, l'autre 1 175 femmes sur 2 500 employés.
À première vue, il semble que la seconde entreprise, qui a recruté 47 % de femmes, soit plus respectueuse du principe de parité que la première, qui n'a recruté que 42 % de femmes.
Est-ce exact ?

On va supposer ici :

- que chaque entreprise, au moment du recrutement de chaque employé, a le choix entre un grand nombre de candidats (ce qui n'est pas toujours le cas) ;
- que si l'on respecte le principe de parité, le recrutement est parfaitement aléatoire quant au genre du futur employé (masculin ou féminin).

On peut alors supposer que chaque entreprise est un échantillon de la population totale des candidats au recrutement.

- Pour la première entreprise.
  L'échantillon est de taille 100. La propriété précédente nous permet de dire que 95 % des échantillons de taille 100 ont une proportion de femmes comprise dans l'intervalle :
  $$\left[0,5 - \frac{1}{\sqrt{100}}; 0,5 + \frac{1}{\sqrt{100}}\right] = [0,4; 0,6].$$

Exemple 4.1 (suite)

Avec une proportion de femmes de 0,42, cette entreprise peut donc être considérée comme un échantillon « conforme à nos attentes » d'un recrutement aléatoire quant au genre des employés.

- Pour la seconde entreprise.
  L'échantillon est de taille 2 500. La propriété précédente nous permet de dire que 95 % des échantillons de taille 2 500 ont une proportion de femmes comprise dans l'intervalle :

$$\left[0,5-\frac{1}{\sqrt{2500}};0,5+\frac{1}{\sqrt{2500}}\right]=[0,48;0,52].$$

  Avec une proportion de femmes de 0,47, cette entreprise ne semble pas être un échantillon « conforme à nos attentes » d'un recrutement aléatoire quant au genre des employés.

C'est donc pour la seconde entreprise qu'il faut rejeter l'idée qu'elle ait respectée le principe de parité, avec un risque de se tromper de 5 %.

Proposition 4.2

Considérons maintenant une population dont la proportion p d'un caractère A est inconnue.
Considérons alors un échantillon de taille $n \geqslant 25$ de cette population, et appelons f la fréquence du caractère A dans cet échantillon. Alors,

$$p \in \left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right].$$

Ce dernier intervalle est appelé **un intervalle de confiance** avec un **niveau de confiance** de 95 %, souvent appelé **fourchette de sondage au seuil de** 95 %.

Démonstration

La démonstration de cette propriété s'appuie sur la propriété précédente.
En effet, pour 95 % des échantillons de taille n,

$$p-\frac{1}{\sqrt{n}} \leqslant f \leqslant p+\frac{1}{\sqrt{n}}$$

Or,

$$p-\frac{1}{\sqrt{n}} \leqslant f \Leftrightarrow p \leqslant f+\frac{1}{\sqrt{n}}$$

et

$$f \leqslant p+\frac{1}{\sqrt{n}} \Leftrightarrow p \geqslant f-\frac{1}{\sqrt{n}}.$$

D'où le résultat. ■

Remarque

Cela signifie qu'il est rare (5 %) de trouver un échantillon dans lequel la fréquence du caractère A est telle que p n'est pas dans la fourchette de sondage. On pourra donc considérer que p est dans la fourchette de sondage, avec un risque de se tromper de 5 %.

Exemple 4.2

Un sondage auprès de 1 000 personnes donne, pour le second tour d'une élection présidentielle, 51 % d'intentions de vote pour un candidat.
L'intervalle de confiance au niveau de confiance de 95 % est :

$$\left[0,51-\frac{1}{\sqrt{1\,000}};0,51+\frac{1}{\sqrt{1\,000}}\right]=[0,479;0,541]$$

Donc, on peut estimer que ce candidat aura entre 47,9 % et 54,1 % des votes ; il n'est donc pas certain de gagner cette élection.

Méthode

Pour vérifier si une hypothèse de proportion p (appelée **hypothèse nulle** et souvent notée H_0) est vraie,on prélève un échantillon aléatoire de taille n dans la population sur lequel on observe une fréquence :

- si $f \in \left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$, alors on accepte l'hypothèse nulle ;
- sinon, on rejette l'hypothèse nulle.

Exemple 4.3

Un fournisseur de pièces électronique affirme que 98 % de ces pièces ne comporte aucun défaut.
Pour vérifier cela, on prélève un échantillon de 5 000 pièces. On trouve, dans cet échantillon, 187 pièces défectueuses. Sous l'hypothèse $p = 0,98$ (proportion théorique de pièces non défectueuses), l'intervalle de fluctuation au seuil de 95 % est :

$$\left[0,98-\frac{1}{\sqrt{5\,000}};0,98+\frac{1}{\sqrt{5\,000}}\right]\approx[0,9659;0,9941]$$

La fréquence observée de pièces non défectueuses est :

$$f=\frac{5\,000-187}{5\,000}\approx 0,9626\notin[0,9659;0,9941]$$

On peut donc rejeter l'hypothèse : ce fournisseur ment sur la proportion de pièces non défectueuses (avec un risque de se tromper de 5 %).

Définition 4.3

Une **épreuve de Bernoulli** (du nom du mathématicien suisse Daniel Bernoulli) est une expérience à seulement deux issues possibles, que l'on appelle « succès » et « échec ».
Un **échantillon de Bernoulli de taille n** est la répétition de n épreuves de Bernoulli.

Exemple 4.4

L'expérience :

> « On lance un dé équilibré à 6 faces et on regarde quand on obtient un nombre premier »

est une épreuve de Bernoulli, où le succès est : « obtenir un nombre premier » et où l'échec est son contraire : « obtenir un nombre qui n'est pas premier ».
Considérer 25 fois cette expérience constitue un échantillon de Bernoulli de taille 25.

Définition 4.4 (Intervalle de fluctuation dans le cas d'une loi binomiale)

Soit X la variable aléatoire représentant le nombre de succès ; on suppose que X suit la loi binomiale $\mathscr{B}(n;p)$ (dans un échantillon de Bernoulli de taille n). On considère alors la variable aléatoire $\dfrac{X}{n}$ représentant la fréquence du succès. L'intervalle de fluctuation au seuil de 95 % de $\dfrac{X}{n}$ est $\left[\dfrac{a}{n};\dfrac{b}{n}\right]$, où :

- a est le plus petit entier tel que $P(X \leqslant a) > 0,025$
- b le plus petit entier tel que $P(X \leqslant b) \geqslant 0,975$.

Remarque

Si on voulait un intervalle de fluctuation au seuil de 90 %, on considèrerait les nombres a', plus petit entier tel que $P(X \leqslant a') > 5\%$ et b', plus petit entier naturel tel que $P(X \leqslant b') \geqslant 95\%$.

Exemple 4.5

On lance un dé à 6 face et on s'intéresse à l'événement :

> « Obtenir le 2 ou le 5 »

La probabilité de cet événement est égale à $p = \frac{1}{3}$.

On observe, sur un échantillon de taille 50, la fréquence de cet événement et on trouve $f = 0,4$.
On se pose la question de savoir si cette fréquence est « compatible » avec l'hypothèse émise (à savoir que $p = \frac{1}{3}$).

Exemple 4.5 (suite)

Le tableau ci-dessous correspond à la loi binomiale $\mathscr{B}\left(50; \frac{1}{3}\right)$.

k	$P(X \leqslant k)$
9	0.012709
10	0.028443
11	0.057051
12	0.103537
⋮	⋮
21	0.924433
22	0.957616
23	0.977814
24	0.989175

On voit ici que $a = 10$ et $b = 23$. L'intervalle de fluctuation au seuil de 95 % est donc $\left[\frac{10}{50}; \frac{23}{50}\right]$, soit $[0,2; 0,46]$.

Ainsi, 0,4 appartient à l'intervalle de fluctuation au seuil de 95 %, donc l'hypothèse selon laquelle $p = \frac{1}{3}$ est acceptée.

Règle de décision

Si la fréquence observée appartient à l'intervalle $\left[\frac{a}{n}; \frac{b}{n}\right]$, on considère que l'hypothèse selon laquelle la proportion est p dans la population n'est pas remise en question. Sinon, on rejette l'hypothèse (au risque de se tromper de 5 %).

Remarque

Si $n \geqslant 25$ et $0,2 \leqslant p \leqslant 0,8$, on pourra observer que $\left[\frac{a}{n}; \frac{b}{n}\right]$ est sensiblement le même que $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$. Mais dans un cas général, il n'y a pas de restriction sur n et p. dans le cas d'une loi binomiale.

Afin de déterminer les nombres a et b, on se sert souvent d'un logiciel, d'un tableur ou d'une calculatrice.

Exemple 4.6

Pour faire un contrôle de qualité dans une usine automobile, on étudie les défauts de peinture « grains ponctuels sur le capot » (quasiment invisibles pour le client). Lorsque le processus est sous contrôle, on a 10 % de ce type de défaut, c'est-à-dire que la proportion de capots présentant ce défaut de peinture dans la production totale est $p = 0,10$.
On choisit un échantillon de 200 capots et on observe 26 capots défectueux.
Faut-il s'en inquiéter ?
(Un constructeur d'automobiles français a effectivement pratiqué ce type de contrôle de qualité).

Exemple 4.6 (suite)

On a représenté ci-dessous les valeurs pour la loi binomiale $\mathscr{B}(200;0,1)$:

k	$P(X \leqslant k)$
10	0.008071
11	0.016790
12	0.032047
13	0.056562
14	0.092946
⋮	⋮
28	0.972908
29	0.983673
30	0.990492
31	0.994646

On a alors $a = 12$ et $b = 29$.

On en déduit que l'intervalle de fluctuation est $[0,06;0,145]$.

26 capots défectueux sur 200 correspondent à une fréquence de 0,13, et cette valeur appartient à l'intervalle de fluctuation.

Il n'y a donc pas raison de s'inquiéter.

Voici un algorithme qui permet de trouver les valeurs de a et b :

Algorithme

```
> Entrées
    Variable n (entière)
    Variable p (entière)
    Variable S (réelle)
    Variable k (entière)
    Variables a et b (entières)
> Traitement
    S←0
    k←0
    Tant que S⩽0,025
        S←S+(k n) p^k (1−p)^(n−k)
        k←k+1
    Fin du Tant que a←k-1
    Tant que S<0,975
        S←S+(k n) p^k (1−p)^(n−k)
        k←k+1
    Fin du Tant que b←k-1
> Sortie
    Afficher a et b
```

Voici maintenant deux programmes de calculatrice, l'un pour TI, l'autre pour CASIO :

TI	CASIO
PROGRAM:INTFLUC	INTFLUCT
:Prompt N	?→N ↵
:Prompt P	?→P ↵
:0→I	0→I ↵
:binompdf(N,P,0)→C	BinomialPD(0,N,P)→C ↵
:While C≤0.025	While C≤0.025 ↵
:I+1→I	I+1→I ↵
:C+binompdf(N,P,I)→C	C+BinomialPD(I,N,P)→C ↵
:End	WhileEnd ↵
:I→A	I→A ↵
:While c<0.975	While C<0.975 ↵
:I+1→I	I+1→I ↵
:C+binompdf(N,P,I)→C	C+BinomialPD(I,N,P)→C ↵
:End	WhileEnd ↵
:I→B	I→B ↵
:Disp A	A ◢
:Disp B	B ◢

Nous allons maintenant parler de la loi normale, et nous verrons qu'elle a son utilité dans ce chapitre.

Définition 4.5 (Variable aléatoire centrée réduite)

On dit d'une variable aléatoire qu'elle est une **variable centrée réduite** quand son espérance mathématique est nulle et son écart-type égal à 1.

Proposition 4.3 (Changement de variable aléatoire)

Soit X une variable aléatoire d'espérance m et d'écart-type $\sigma \neq 0$.
Alors, la variable aléatoire Z définie par :

$$Z = \frac{X - m}{\sigma}$$

est une variable aléatoire centrée réduite.

Démonstration

Rappelons les formules :

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b \quad ; \quad \sigma(aX + b) = |a|\sigma(X).$$

Alors,

$$\begin{aligned}\mathbb{E}(Z) &= \mathbb{E}\left(\frac{1}{\sigma}X - \frac{m}{\sigma}\right)\\ &= \frac{1}{\sigma}\mathbb{E}(X) - \frac{m}{\sigma}\\ &= \frac{1}{\sigma} \times m - \frac{m}{\sigma}\\ &= 0\end{aligned}$$

Démonstration (suite)

et

$$\begin{aligned}\sigma(Z) &= \frac{1}{\sigma}\sigma \\ &= 1.\end{aligned}$$

Z est donc bien une variable aléatoire centrée réduite. ■

Théorème 4.1 (Théorème de Moivre-Laplace)

Soit $p \in]0;1[$ et soit, pour $n \in \mathbb{N}$, X_n une variable aléatoire suivant la loi binomiale $\mathscr{B}(n;p)$. On pose alors :

$$Z_n = \frac{X_n - m}{\sigma_n}$$

où $m = \mathbb{E}(X_n)$ et σ_n l'écart-type de X_n.
Pour tous nombres réels a et b, $a < b$, on a :

$$\lim_{n\to+\infty} P(a \leqslant Z_n \leqslant b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx.$$

Définition 4.6 (Loi normale centrée réduite)

Une variable aléatoire X à densité f sur I suit une **loi normale centrée réduite** si $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ et $I = [a;b]$, a et b étant deux réels, $a < b$.
On note alors :

$$X \hookrightarrow \mathscr{N}(0;1).$$

Proposition 4.4

Soit $X \hookrightarrow \mathscr{N}(0;1)$. Alors,

1. $\forall u \in \mathbb{R}^+,\ P(X \leqslant -u) = P(X \geqslant u)$,
2. $P(X \leqslant 0) = P(X \geqslant 0) = \frac{1}{2}$.

Démonstration

1. $f(-x) = f(x)$ donc $\displaystyle\int_{-\infty}^{-u} f(x)\, dx = \int_u^{+\infty} f(x)\, dx$, d'où le résultat.
2. En prenant $u = 0$ dans le résultat précédent, on a $P(X \leqslant 0) = P(X \geqslant 0)$ et comme $P(X \leqslant 0) + P(X \geqslant 0) = 1$, on arrive au résultat souhaité. ■

Proposition 4.5 (Espérance et écart-type de la loi normale centrée réduite)

Soit $X \hookrightarrow \mathcal{N}(0;1)$.. Alors,

$$\mathbb{E}(X) = 0 \qquad \text{et} \qquad \sigma(X) = 1.$$

Démonstration

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} \frac{x}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx.$$

Or, la fonction $f : x \mapsto \frac{x}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ est impaire, ce qui signifie que :

$$\int_{-x}^{0} f(x)\, dx = -\int_{0}^{x} f(x)\, dx$$

et donc $\mathbb{E}(X) = 0$.

La démonstration de la propriété sur l'écart-type est hors-programme de Terminale. Cependant, voyons les points essentiels de la démonstration.

La formule de König-Huygens nous dit que $\mathbb{V}(X) = \mathbb{E}\left(X^2\right) - (\mathbb{E}(X))^2 = \mathbb{E}\left(X^2\right)$ car $\mathbb{E}(X) = 0$.

$$\mathbb{E}\left(X^2\right) = \int_{-\infty}^{+\infty} \frac{x^2}{\sqrt{2\pi}} e^{-\frac{x^4}{2}}\, dx = 2\int_{-\infty}^{0} \frac{x^2}{\sqrt{2\pi}} e^{-\frac{x^4}{2}}\, dx$$

À l'aide d'une intégration par parties, en posant $u(x) = \frac{x^2}{\sqrt{2\pi}}$ et $v(x) = e^{-\frac{x^4}{2}}$, on a :

$$\int_{-\infty}^{0} \frac{x^2}{\sqrt{2\pi}} e^{-\frac{x^4}{2}}\, dx = \underbrace{\left[-\frac{x}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\right]_{-\infty}^{0}}_{=0} + \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx = \frac{1}{2}.$$

On arrive donc bien à $\mathbb{V}(X) = 1$. Or, $\sigma(X) = \sqrt{\mathbb{V}(X)}$, donc $\sigma(X) = 1$. ■

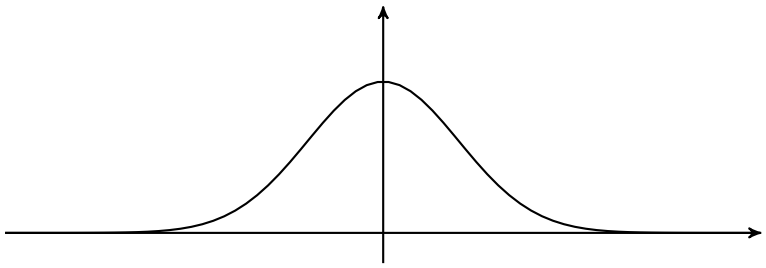
Théorème 4.2 (Répartition des valeurs d'une loi normale centrée réduite)

Soit X une variable aléatoire suivante la loi normale centrée réduite $\mathcal{N}(0;1)$.
Alors,

$$\forall\, \alpha \in]0;1[,\ \exists!\ u_\alpha \in \mathbb{R}^+ \mid P(-u_\alpha \leqslant X \leqslant u_\alpha) = 1 - \alpha.$$

Démonstration

La fonction de densité de la loi normale est représentée par la courbe suivante :



Par sa symétrie, on a :

$$\forall\, u \in \mathbb{R}^+,\ \mathrm{P}(-u \leqslant \mathrm{X} \leqslant u) = 2\mathrm{P}(0 \leqslant \mathrm{X} \leqslant u) = 2\int_0^u f(x)\,\mathrm{d}x = 2F(u),$$

où F est une primitive de f sur $\mathbb{R}$ qui s'annule en 0.
La fonction F est alors continue et strictement croissante sur $\mathbb{R}$ ($F(u)$ représente l'aire du domaine délimité par la courbe, l'axe des abscisses, l'axe des ordonnées et la droite d'équation $x = u$).
De plus, $2F(0) = 0$ et $\lim\limits_{u\to+\infty} 2F(u) = 1$ (par définition d'une fonction de densité) donc, d'après le théorème de la valeur intermédiaire, il existe un unique nombre u_α dans $]0\,;+\infty[$ tel que $2F(u_\alpha) = 1-\alpha$ pour $\alpha \in]0\,;1[$. D'où le tableau de variations suivant :

u	0		u_α		$+\infty$
$2F(u)$	0	↗	$1-\alpha$	↗	1

Le résultat est alors démontré. ■

Remarque

En classe de Terminale, on demande aux élèves de retenir les **Valeurs particulières de u_α suivantes :**

- $u_{0,05} \approx 1,96$ d'où $\mathrm{P}(-1,96 \leqslant \mathrm{X} \leqslant 1,96) \approx 0,95$.
- $u_{0,01} \approx 2,58$ d'où $\mathrm{P}(-2,58 \leqslant \mathrm{X} \leqslant 2,58) \approx 0,99$.

Cela signifie que si une variable aléatoire X suit la loi $\mathcal{N}(0\,;1)$, alors 95 % de ses valeurs se trouvent entre $-1,96$ et $1,96$, et 99 % entre $-2,58$ et $2,58$.

Définition 4.7 (Loi normale)

Une variable aléatoire X suit une loi normale $\mathcal{N}\left(\mu\,;\sigma^2\right)$, d'espérance μ et d'écart-type σ, si la variable aléatoire $\mathrm{Z} = \dfrac{\mathrm{X}-\mu}{\sigma}$ suit la loi normale centrée réduite $\mathcal{N}(0\,;1)$.

Proposition 4.6

Soit $X \hookrightarrow \mathcal{N}\left(\mu;\sigma^2\right)$.
Alors,

$$\begin{aligned} P(\mu-\sigma \leqslant X \leqslant \mu+\sigma) &\approx 0,68 && (\text{à } 10^{-2} \text{ près}) \\ P(\mu-2\sigma \leqslant X \leqslant \mu+2\sigma) &\approx 0,95 && (\text{à } 10^{-2} \text{ près}) \\ P(\mu-3\sigma \leqslant X \leqslant \mu+3\sigma) &\approx 0,997 && (\text{à } 10^{-3} \text{ près}) \end{aligned}$$

Démonstration

On a :

$$\begin{aligned} \mu-\sigma \leqslant X \leqslant \mu+\sigma &\Leftrightarrow -\sigma \leqslant X-\mu \leqslant \sigma \\ &\Leftrightarrow -1 \leqslant \frac{X-\mu}{\sigma} \leqslant 1 \end{aligned}$$

donc :

$$\begin{aligned} P(\mu-\sigma \leqslant X \leqslant \mu+\sigma) &= P\left(-1 \leqslant \frac{X-\mu}{\sigma} \leqslant 1\right) \\ &= P(-1 \leqslant Z \leqslant 1) \\ &\approx 0,68 \text{ (À l'aide de la calculatrice ou d'un logiciel de calcul)} \end{aligned}$$

Nous faisons de même pour les autres résultats. ■

Ce dernier résultat nous dit que si X est une variable aléatoire suivant la loi normale $\mathcal{N}\left(\mu;\sigma^2\right)$, alors :

- 68 % de ses valeurs se trouvent dans $[\mu-\sigma\,;\mu+\sigma]$;
- 95 % de ses valeurs se trouvent dans $[\mu-2\sigma\,;\mu+2\sigma]$;
- 99,7 % de ses valeurs se trouvent dans $[\mu-3\sigma\,;\mu+3\sigma]$.

Théorème 4.3 (Intervalles de fluctuation asymptotiques)

Soit $p \in]0;1[$. On pose alors X_n une suite de variables aléatoires suivant la loi binomiale $\mathcal{B}(n;p)$. Alors,

$$\forall \alpha \in]0;1[,\ \lim_{n\to+\infty} P\left(\frac{X_n}{n} \in I_n\right) = 1-\alpha$$

où

$$I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\,;p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

et où u_α est l'unique réel tel que $P(-u_\alpha \leqslant Z \leqslant u_\alpha) = 1-\alpha$, Z suivant la loi $\mathcal{N}(0;1)$.

Démonstration

$X_n \hookrightarrow \mathscr{B}(n;p)$ donc $\mathbb{E}(X_n) = np$ et $\sigma(X_n) = \sqrt{np(1-p)}$.
Ainsi,

$$Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}.$$

On a :

$$\begin{aligned} P(a \leqslant Z_n \leqslant b) &= P\left(a \leqslant \frac{X_n - np}{\sqrt{np(1-p)}} \leqslant b\right) \\ &= P(a\sqrt{np(1-p)} \leqslant X_n - np \leqslant b\sqrt{np(1-p)}) \\ &= P(a\sqrt{np(1-p)} + np \leqslant X_n \leqslant b\sqrt{np(1-p)} + np) \\ &= P\left(a\frac{\sqrt{np(1-p)}}{n} + p \leqslant \frac{X_n}{n} \leqslant b\frac{\sqrt{np(1-p)}}{n} + p\right). \end{aligned}$$

D'après le théorème de Moivre-Laplace,

$$\forall (a;b) \in \mathbb{R} \times \mathbb{R},\ \lim_{n\to+\infty} P(a \leqslant Z_n \leqslant b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx.$$

On en déduit alors :

$$\lim_{n\to+\infty} P\left(a\frac{\sqrt{np(1-p)}}{n} + p \leqslant \frac{X_n}{n} \leqslant b\frac{\sqrt{np(1-p)}}{n} + p\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx.$$

En prenant $a = -u_\alpha$ et $b = u_\alpha$, on obtient :

$$\lim_{n\to+\infty} P\left(-u_\alpha\frac{\sqrt{np(1-p)}}{n} + p \leqslant \frac{X_n}{n} \leqslant u_\alpha\frac{\sqrt{np(1-p)}}{n} + p\right) = \int_{-u_\alpha}^{u_\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx = 1 - \alpha.$$

Ainsi,

$$\lim_{n\to+\infty} P\left(\frac{X_n}{n} \in I_n\right) = 1 - \alpha.$$

Le théorème est ainsi démontré. ■

Remarque

L'intervalle $I_n = \left[p - u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ est appelé un **intervalle de fluctuation asymptotique** au seuil de $1-\alpha$ de la variable aléatoire $F_n = \frac{X_n}{n}$.

L'intervalle $\left[p - 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ est un intervalle de fluctuation asymptotique au seuil de 95 % (p désigne la proportion dans la population).

Pour que toutes les approximations précédentes soient réalisées, on prendra pour habitude d'avoir :

$$n \geqslant 30 \quad ; \quad np \geqslant 5 \quad ; \quad n(1-p) \geqslant 5.$$

On utilise les intervalles de fluctuation au seuil de 95 % pour déterminer si la fréquence f observée dans une population est compatible ou pas avec un modèle de Bernoulli.

Exemple 4.7 (d'après les ressources de l'E.N.)

Le responsable de la maintenance des machines à sous d'un casino doit vérifier qu'un certain type de machine est bien réglé sur une fréquence de succès de 0,06.
Il décide de régler chaque machine pour laquelle il aura observé, dans l'historique des jeux, une fréquence de succès se situant en dehors d'un intervalle de fluctuation au seuil de 95 %.
Lors du contrôle d'une machine, le technicien constate qu'elle a fourni 9 succès sur 85 jeux.

1. Déterminer la fréquence observée f de succès de cette machine.
2. Déterminer un intervalle de fluctuation asymptotique du cours au seuil de 95 %.
3. Le technicien va-t-il modifier le réglage de la machine ?
4. Quelle aurait été sa décision s'il y avait eu 21 succès sur 200 jeux ?

Solution :

1. La fréquence observée est : $f = \dfrac{8}{98} \approx 0,106$.
2. Ici, $n = 85 \geqslant 30$, $p = 0,06$ donc $np = 5,1 \geqslant 5$ et $n(1-p) = 79,9 \geqslant 5$.
   Un intervalle de fluctuation au seuil de 95 % est donc :
$$\mathrm{I} = \left[0,06 - 1,96\frac{\sqrt{0,06\times 0,94}}{\sqrt{85}} ; 0,06 + 1,96\frac{\sqrt{0,06\times 0,94}}{\sqrt{85}}\right] \approx [0,009 ; 0,111].$$
3. $f \in \mathrm{I}$ donc le réglage de la machine n'est pas à modifier.
4. Dans cette question, $f \approx 0,105$ et $\mathrm{I} \approx [0,027 ; 0,093]$ donc $f \notin \mathrm{I}$. Il faut donc régler la machine.

Remarque

Pour tout réel $p \in]0;1[$, la fonction $p \mapsto p(1-p)$ admet un maximum pour $p = \dfrac{1}{2}$ et vaut $\dfrac{1}{2}\left(1 - \dfrac{1}{2}\right) = \dfrac{1}{4}$. Donc,
$$p(1-p) \leqslant \frac{1}{4}.$$
On en déduit alors :
$$1,96\sqrt{p(1-p)} \leqslant 1,96\sqrt{\frac{1}{4}} \leqslant 1\,,$$
et donc :
$$1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \frac{1}{\sqrt{n}}.$$

Remarque (suite)

Ainsi,

$$\mathrm{I}_n = \left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\,;p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] \subseteq \left[p - \frac{1}{\sqrt{n}}\,;p - \frac{1}{\sqrt{n}}\right] = \mathrm{J}_n$$

et donc :

$$\mathrm{P}\left(\frac{\mathrm{X}_n}{n} \in \mathrm{I}_n\right) \leqslant \mathrm{P}\left(\frac{\mathrm{X}_n}{n} \in \mathrm{J}_n\right).$$

Ceci légitime l'utilisation de l'intervalle J_n en classe de Seconde.

Théorème 4.4

Soit $p \in]0\,;1[$ un réel. Soit alors (X_n) une suite de variables aléatoires suivant la loi binomiale $\mathscr{B}(n\,;p)$.

$$\exists\ n_0 \in \mathbb{N},\ n \geqslant n_0 \Rightarrow \mathrm{P}\left(\frac{\mathrm{X}_n}{n} \in \left[p - \frac{1}{\sqrt{n}}\,;p + \frac{1}{\sqrt{n}}\right]\right) \geqslant 0,95.$$

Démonstration

- Dans un premier temps, on démontre que :

$$\mathrm{P}\left(p - 2\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \frac{\mathrm{X}_n}{n} \leqslant p + 2\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) \leqslant \mathrm{P}\left(p - 2\frac{1}{\sqrt{n}} \leqslant \frac{\mathrm{X}_n}{n} \leqslant p + 2\frac{1}{\sqrt{n}}\right)$$

  en utilisant le même raisonnement que dans la remarque précédente.
- De plus,

$$\begin{aligned} & p - 2\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \frac{\mathrm{X}_n}{n} \leqslant p + 2\frac{\sqrt{p(1-p)}}{\sqrt{n}} \\ \Leftrightarrow\ & np - 2\sqrt{np(1-p)} \leqslant \mathrm{X}_n \leqslant np + 2\sqrt{np(1-p)} \\ \Leftrightarrow\ & -2 \leqslant \frac{\mathrm{X}_n - np}{\sqrt{np(1-p)}} \leqslant 2. \end{aligned}$$

  Donc,

$$\mathrm{P}\left(p - 2\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \frac{\mathrm{X}_n}{n} \leqslant p + 2\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = \mathrm{P}\left(-2 \leqslant \frac{\mathrm{X}_n - np}{\sqrt{np(1-p)}} \leqslant 2\right).$$

  Posons :

$$a_n = \mathrm{P}\left(-2 \leqslant \frac{\mathrm{X}_n - np}{\sqrt{np(1-p)}} \leqslant 2\right).$$

  D'après le théorème de Moivre-Laplace,

$$\lim_{n\to+\infty} a_n = \int_{-2}^{2} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x \approx 0,9545. \quad \text{(calculatrice)}$$

  Ainsi, (a_n) converge vers un nombre fini supérieur à 0,95.
  Il existe donc bien un entier n_0 à partir duquel $a_n \geqslant 0,95$. D'où le théorème. ■

Remarque

L'entier n_0 dépend de p. Son maximum est atteint pour $p = \frac{1}{2}$ et vaut $n_0 = 529$. (Résultat obtenu à l'aide d'un algorithme)

Proposition 4.7 (Intervalle de confiance)

Soit $p \in]0;1[$ et soit (X_n) une suite de variables aléatoires suivant la loi binomiale $\mathscr{B}(n;p)$. On pose alors $F_n = \frac{X_n}{n}$ la variable aléatoire représentant la fréquence du nombre de « succès ».
Alors, pour n assez grand, $p \in \left[F_n - \frac{1}{\sqrt{n}} ; F_n + \frac{1}{\sqrt{n}}\right]$ avec une probabilité de 95 %.

Démonstration

D'après le théorème précédent,

$$\exists\, n_0 \in \mathbb{N},\ n \geqslant n_0 \Rightarrow P\left(\frac{X_n}{n} \in \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]\right) \geqslant 0,95.$$

Or,

$$\frac{X_n}{n} \in \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right] \Leftrightarrow F_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \frac{1}{\sqrt{n}}.$$

D'où le résultat souhaité. ■

Remarque (Intervalle aléatoire)

L'intervalle $\left[F_n - \frac{1}{\sqrt{n}} ; F_n + \frac{1}{\sqrt{n}}\right]$ est appelé un **intervalle aléatoire** (intervalle dont les bornes sont des variables aléatoires).

La précision de l'estimation est alors donnée par l'amplitude de l'intervalle $\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}\right]$, soit $\frac{2}{\sqrt{n}}$. En notant ε la précision souhaitée, on peut donc calculer la taille n de l'échantillon pour que $\frac{2}{\sqrt{n}} \leqslant \varepsilon$. On a alors $n \geqslant \frac{4}{\varepsilon^2}$.

Par exemple, pour avoir une précision de 10^{-2}, on prendra un échantillon de taille $n = 40\,000$.

Exemple 4.8 (Sondages et élections)

Pour le 2e tour d'une élection présidentielle, l'institut de sondages A réalise un sondage qui montre que 52,5 % des intentions de votes sont pour le candidat 1, et donc 47,5 % pour le candidat 2 (en ne tenant pas compte ni de l'abstention ni des votes blancs).
L'institut B réalise son propre sondage et constate que 50,5 % des intentions de votes sont pour le candidat 1.
Les deux sondages ont été faits sur un échantillon de 1 112 personnes.
L'intervalle de confiance pour l'institut A est $\left[0,525-\frac{1}{\sqrt{1\,112}};0,525-\frac{1}{\sqrt{1\,112}}\right]$, soit $[0,495;0,555]$ pour le candidat 1.
L'intervalle de confiance pour l'institut B est $\left[0,505-\frac{1}{\sqrt{1\,112}};0,505-\frac{1}{\sqrt{1\,112}}\right]$, soit $[0,475;0,535]$ pour le candidat 1.
L'intersection des deux intervalles n'étant pas vide, les résultats des deux sondages ne sont pas en contradiction.
De plus, ces intervalles contiennent des valeurs inférieures à 0,5 ; ainsi, le candidat 1 n'est pas totalement sûr de gagner les élections.

Remarque

1. Remarquez l'importance de connaître la taille de l'échantillon sur lequel le sondage est effectué. Sans ce nombre, impossible de faire des conclusions probantes.
2. Comme pour les intervalles de fluctuation, il existe différents intervalles de confiance. Par exemple, l'intervalle :

$$\left[f-1,96\frac{\sqrt{f(1-f)}}{\sqrt{n}};f+1,96\frac{\sqrt{f(1-f)}}{\sqrt{n}}\right].$$

Exemple 4.9 (Contrôle du Loto)

Le site http://www.lototest.com offre certaines statistiques concernant le Loto, et notamment concernant le nouveau Loto (depuis 2008). On y trouve, au 16 août 2013, que sur 763 tirages, que la fréquence maximale enregistrée pour un numéro est de 2,49 % (le numéro 23) et la fréquence minimale est de 1,6 % (les numéros 42 et 19).
La question que l'on est en droit de se poser est si ces fréquences sont acceptables au seuil de 95 % ?

La probabilité d'un numéro quelconque est : $p=\frac{1}{49}\approx 2,04\,\%$.
Un intervalle de fluctuation au seuil de 95 % est :

$$\left[0,0204-1,96\frac{\sqrt{0,0204\times(1-0,0204)}}{\sqrt{763}};0,0204-1,96\frac{\sqrt{0,0204\times(1-0,0204)}}{\sqrt{763}}\right]\approx[0,0104;0,031].$$

On constate donc que la fréquence minimale observée ainsi que la fréquence maximale observée sont dans cet intervalle.
On peut donc conclure qu'il n'y a aucune anomalie concernant les tirages observés.

Deuxième partie

Statistiques à 2 variables

5 Introduction

Dans cette partie, nous allons considérer une série statistique $\{(x_i;y_i)\}_{1\leqslant i\leqslant n}$, où les x_i et y_i sont deux caractères que l'on observe en même temps.

L'objectif principal des séries statistiques à deux variables est de voir s'il y a corrélation entre les deux caractères, c'est-à-dire si l'un varie en fonction de l'autre.

Représentation graphique

5.1.1 Nuage de points

Définition 5.1

Soit la série statistique $\{(x_i;y_i)\}_{1\leqslant i\leqslant n}$. On appelle **nuage de points** associé à cette série l'ensemble des points M_i de coordonnées $(x_i;y_i)$, pour $1 \leqslant i \leqslant n$.

5.1.2 Point moyen

Définition 5.2

Soit la série statistique $\{(x_i;y_i)\}_{1\leqslant i\leqslant n}$. On appelle **point moyen** le point de coordonnées $(\overline{x};\overline{y})$, où $\overline{x}$ représente la moyenne des x_i et $\overline{y}$ celle des y_i :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad ; \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

5.1.3 Graphique

On représente le nuage de points ainsi que le point moyen dans un repère orthogonal.

Exemple 5.1

Un particulier observe sa consommation de fuel en fonction de la température extérieure. Les résultats sont reportés dans le tableau suivant :

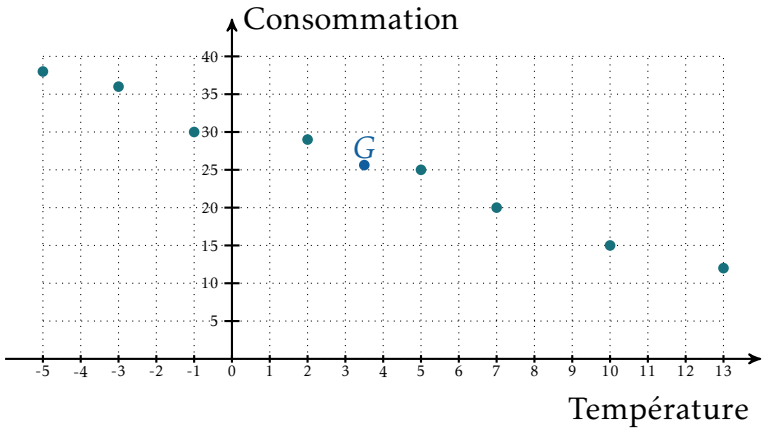
Température (en ° C)	−5	−3	−1	2	5	7	10	13
Consommation (en L)	38	36	30	29	25	20	15	12

Exemple 5.1 (suite)

Le point moyen G est tel que :

$$\overline{x} = \frac{-5-3-1+2+5+7+10+13}{8} = 3,5$$
$$\overline{y} = \frac{38+36+30+29+25+20+15+12}{8} = 25,625$$

Le nuage de points correspondant à cette série statistique est :



6 Méthodes d'ajustement

Définition 6.1 (Ajustement)

Un **ajustement** est une représentation graphique qui s'approche le plus possible du nuage de point d'une série statistique à deux variables.

La méthode de Mayer

Définition 6.2

La méthode de Mayer consiste à partager la série en deux séries de même effectif afin de trouver leur point moyen G_1 et G_2. L'ajustement sera alors la droite (G_1G_2). C'est un **ajustement affine**.

Exemple 6.1

Reprenons l'exemple du chapitre précédent : on peut partager la série en deux séries :

$$S_1 = \{(-5;38),(-3;36),(-1;30),(2;29)\}$$

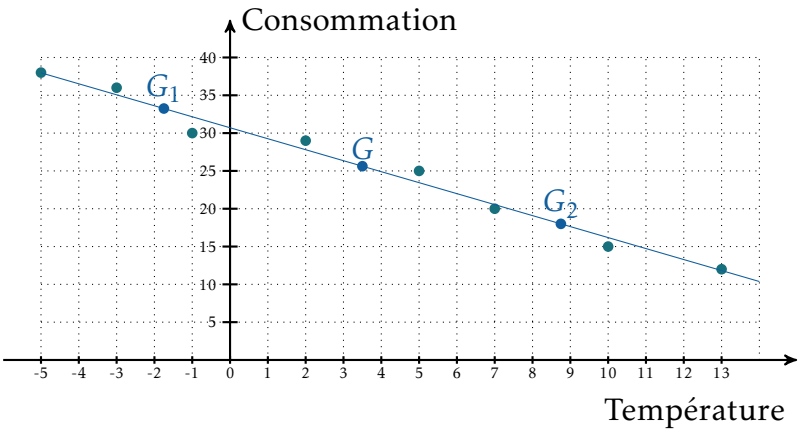
et

$$S_2 = \{(5;25),(7;20),(10;15),(13;12)\}$$

dont les points moyens sont :

$$G_1(-1,75;33,25) \qquad \text{et} \qquad G_2(8,75;18)$$

On a alors l'ajustement page suivante :



Nous observons que l'ajustement obtenu n'est pas trop éloigné du nuage de points. Nous pouvons alors considéré qu'il est judicieux.

Obtenir un ajustement permet de faire des estimations. Par exemple, à l'aide de l'ajustement affine obtenu par la méthode de Mayer, on peut estimer graphiquement à 10 litres la consommation de fuel pour une température extérieure de 14° C.

La méthode des moindres carrés

Définition 6.3 (Covariance)

Soit la série statistique $\{(x_i;y_i)\}_{1\leqslant i\leqslant n}$. On appelle **covariance** de cette série le nombre :

$$\sigma_{xy}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$$

Proposition 6.1 (Formule de Koenig)

$$\sigma_{xy}=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\overline{x}\cdot\overline{y}$$

Démonstration

$$\begin{aligned}\sigma_{xy}&=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})\\&=\frac{1}{n}\sum_{i=1}^{n}(x_iy_i-x_i\overline{y}-y_i\overline{x}+\overline{x}\cdot\overline{y})\\&=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\overline{y}\cdot\frac{1}{n}\sum_{i=1}^{n}x_i-\overline{x}\cdot\frac{1}{n}\sum_{i=1}^{n}y_i+\overline{x}\cdot\overline{y}\cdot\frac{1}{n}\sum_{i=1}^{n}1\\&=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\overline{y}\cdot\overline{x}-\overline{x}\cdot\overline{y}+\overline{x}\cdot\overline{y}\\\sigma_{xy}&=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\overline{x}\cdot\overline{y}\end{aligned}$$

■

Proposition 6.2 (Méthode des moindres carrés)

Soient la série statistique $\{(x_i;y_i)\}_{1\leqslant i\leqslant n}$ et la droite $(d):y=ax+b$.
On pose alors $M_i(x_i;y_i)$, $P_i(x_i;ax_i+b)$ et :

$$e=\sum_{i=1}^{n}M_iP_i^2=\sum_{i=1}^{n}[y_i-(ax_i+b)]^2$$

Alors, e est minimum pour $a=\frac{\sigma_{xy}}{V_x}$ et $b=\overline{y}-a\overline{x}$, où V_x est la variance des x_i.
La droite (d) ainsi obtenue est appelée la droite de régression de y en x.

Démonstration

M_iP_i représente, pour $1 \leqslant i \leqslant n$, la différence des ordonnées de chacun des points M_i et P_i. En élevant au carré ces distances, on est certain d'avoir une quantité positive et en ajoutant ces carrés, on obtient e.
Ainsi, dire que e est minimal signifie que la droite (d) est la plus proche possible du nuage de points M_i.

Posons maintenant :

$$z_i = y_i - ax_i - b$$

Ainsi,

$$e = \sum_{i=1}^{n} z_i^2$$

Si l'on note V_z la variance des z_i, on a :

$$nV_z = e - \overline{z}^2$$

Donc

$$e = nV_z + \overline{z}^2$$

Ainsi, minimiser e revient à minimiser nV_z et $\overline{z}^2$ car ce sont deux quantités positives.

1 **Minimisation de nV_z.**

$$\begin{aligned} nV_z &= \sum_{i=1}^{n} (z_i - \overline{z})^2 \\ &= \sum_{i=1}^{n} \left[(y_i - \overline{y}) - a(x_i - \overline{x})\right]^2 \\ &= \sum_{i=1}^{n} \left[(y_i - \overline{y})^2 - 2a(x_i - \overline{x})(y_i - \overline{y}) + a^2(x_i - \overline{x})^2\right] \\ &= \sum_{i=1}^{n} (y_i - \overline{y})^2 - 2a\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) + a^2\sum_{i=1}^{n} (x_i - \overline{x})^2 \\ &= nV_y - 2na\sigma_{xy} + na^2V_x \end{aligned}$$

Ainsi, en simplifiant par n (qui est non nul), on a :

$$V_z = V_x a^2 - 2\sigma_{xy}a + V_y$$

polynôme du second degré en a.
Le coefficient de a^2 étant strictement positif, le minimum est atteint pour $a = -\dfrac{-2\sigma_{xy}}{2 \times V_x} = \dfrac{\sigma_{xy}}{V_x}$

2 **Minimisation de $\overline{z}$.**

$\overline{z}$ est minimal quand il est nul.

$$\begin{aligned} \overline{z} = 0 &\Leftrightarrow \overline{y} - a\overline{x} - b = 0 \\ &\Leftrightarrow b = \overline{y} - a\overline{x} \end{aligned}$$

■

Remarque

Cette dernière égalité implique que la droite (d) passe nécessairement par le point moyen G.

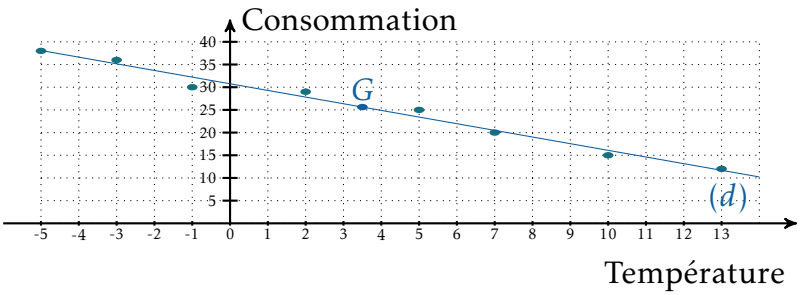
Exemple 6.2

Reprenons l'exemple de la consommation de fuel.
On trouve :

$$\sigma_{xy} \approx -52,062 \qquad \text{et} \qquad V_x \approx 35,5$$

D'où :

$$(d) : y = -1,467x + 30,758$$



Coefficient de corrélation linéaire

Définition 6.4

Le **coefficient de corrélation linéaire** est défini par l'égalité :

$$\sigma = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

Ce nombre permet de qualifier la corrélation entre les x_i et les y_i : si $r > 0,95$, on dira que la corrélation est forte et donc que l'ajustement linéaire est justifié.

Exemple 6.3

Dans l'exemple précédent, on a :

$$r = \frac{-52,062}{5,958 \times 8,817} \approx 0,99$$

L'ajustement affine est alors justifié.

Nous avons vu jusqu'à présent des nuages de points qui semblaient être autour d'une droite. Cependant, ce n'est pas toujours le cas.
Nous allons voir maintenant des ajustements qui ne sont pas affines.

Ajustement exponentiel

Définition 6.5

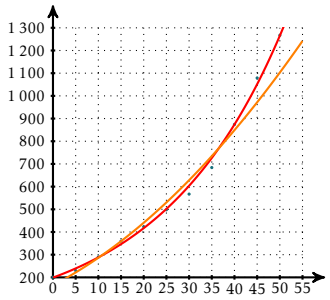
Soit la série statistique $\{(x_i;y_i)\}_{1\leqslant i\leqslant n}$. Un **ajustement exponentiel** est un ajustement obtenu en considérant la série statistique $\{(x_i;\ln y_i)\}_{1\leqslant i\leqslant n}$.

Exemple 6.4

Le tableau ci-dessous donne l'évolution du nombre de personnes âgées de plus de 85 ans, en France métropolitaine, de 1950 à 2000.

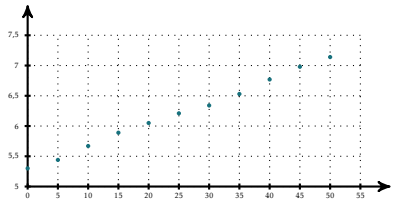
Année	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995	2000
Rang (x_i)	0	5	10	15	20	25	30	35	40	45	50
Effectif (y_i) en milliers	201	231	290	361	423	498	567	684	874	1 079	1 267

Le nuage de points correspondant est le suivant :



La forme du nuage de points ne nous permet pas de faire un ajustement affine car il semble suivre une courbe exponentielle (ou parabolique).
On considère alors les valeurs $z_i = \ln y_i$ pour $1 \leqslant i \leqslant 11$.

Rang (x_i)	0	5	10	15	20	25	30	35	40	45	50
$z_i = \ln y_i$	5,3	5,44	5,67	5,89	6,05	6,21	6,34	6,53	6,77	6,98	7,14



L'allure de ce nouveau nuage de points nous permet de faire un ajustement affine de z en x. À l'aide de la méthode des moindres carrés, on trouve alors :

$$z = 0,037x + 5,291.$$

Dans la mesure où $z = \ln y$, on peut dire que $y = \mathrm{e}^z$. On a alors :

$$y = \mathrm{e}^{0,037x+5,291} = 199\mathrm{e}^{0,037x}.$$

(La fonction $x \mapsto 199\mathrm{e}^{0,037x}$ est représentée en rouge)

Ajustement quadratique

Nous l'avons vu à travers l'exemple précédent, lorsque le nuage de points est courbé, on peut l'ajuster via un ajustement exponentiel, mais on peut aussi l'approcher avec une parabole.

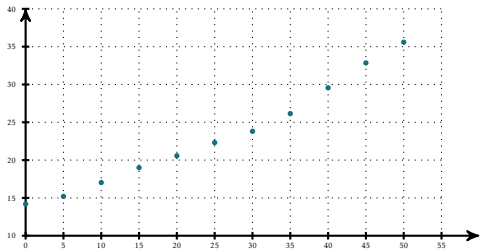
Définition 6.6

Soit la série statistique $\{(x_i;y_i)\}_{1\leqslant i\leqslant n}$, où $y_i \geqslant 0$. Un **ajustement quadratique** est un ajustement obtenu en considérant la série statistique $\left\{\left(x_i;\sqrt{y_i}\right)\right\}_{1\leqslant i\leqslant n}$.

Exemple 6.5

En considérant l'exemple précédent, on a :

Rang (x_i)	0	5	10	15	20	25	30	35	40	45	50
$z_i = \sqrt{y_i}$	14,18	15,2	17,03	19	20,57	22,32	23,81	26,15	29,56	32,85	35,59

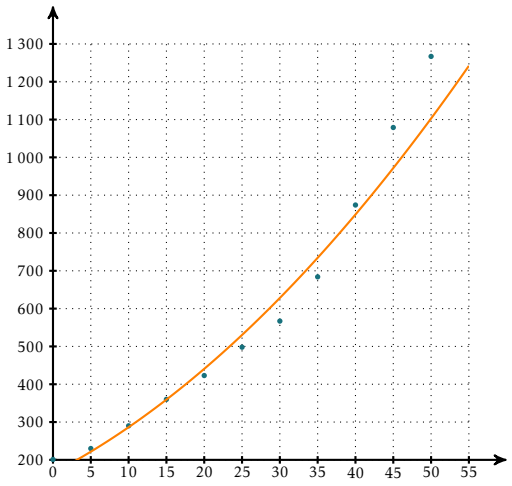


Ce dernier nuage de points nous permet de faire un ajustement affine de z en x. À l'aide de la méthode des moindres carrés, on a :

$$z = 0,407x + 12,852$$

Et donc :

$$y = z^2 = 0,166x^2 + 10,462x + 165,174$$



Nous voyons ici que cet ajustement « colle » moins au nuage de points, ce qui sous-entend que ce dernier ajustement est moins bon que l'ajustement exponentiel.

Ajustement logarithmique

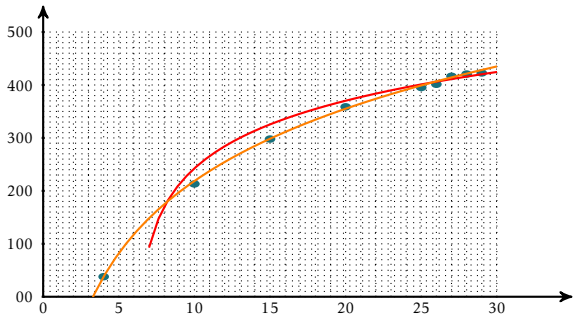
Définition 6.7

Soit la série statistique $\{(x_i;y_i)\}_{1\leqslant i\leqslant n}$. Un **ajustement logarithmique** est un ajustement obtenu en considérant la série statistique $\{(x_i;\exp y_i)\}_{1\leqslant i\leqslant n}$.

Exemple 6.6

Le tableau ci-dessous donne la production d'électricité d'origine nucléaire en France, exprimée en milliards de kWh, entre 1979 et 2004.
Les rangs des années sont calculés par rapport à l'année 1975.

Année	1979	1985	1990	1995	2000	2001	2002	2003	2004
Rang x_i	4	10	15	20	25	26	27	28	29
Production y_i	37,9	213,1	297,9	358,8	395,2	401,3	416,5	420,7	427,7



L'allure du nuage de point nous fait penser à une courbe logarithmique. On pourrait donc effectuer le changement de variable :

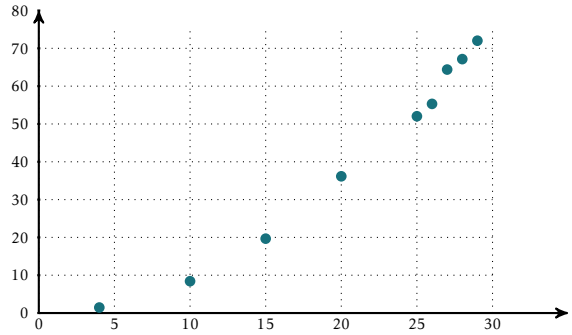
$$z_i = \mathrm{e}^{y_i} \quad , \quad 1 \leqslant i \leqslant 9$$

Mais pour des raisons de simplifications de calculs, je préfère le changement de variable suivant :

$$z_i = \mathrm{e}^{0{,}01y_i} \quad , \quad 1 \leqslant i \leqslant 9$$

On a alors :

Rang x_i	4	10	15	20	25	26	27	28	29
$z_i = \mathrm{e}^{0{,}1y_i}$	1,46	8,42	19,67	36,16	52,04	55,31	64,39	67,15	72,02



Exemple 6.6 (suite)

L'équation de la droite de régression de z en x obtenue par la méthode des moindres carrés est :

$$z = 2,925x - 17,946$$

D'où :

$$y = \ln z = \ln(2,925x - 17,946) \quad , \quad x > 6$$

La fonction $x \mapsto \ln(2,925x - 17,946)$ est tracée en rouge sur le premier graphique.

Remarque

On peut effectuer n'importe quel changement de variable qui nous paraît judicieux. En effet, si l'on effectue le changement de variable $z_i = f(y_i)$ et si l'on effectue un ajustement affine sur les z_i, alors $y_i = f^{-1}(z_i)$, où f^{-1} est la fonction réciproque de f (sous réserve qu'elle existe).

Plutôt que d'effectuer un changement de variable, on peut aussi considérer que la fonction dont la courbe représentative approchera le nuage de point est de la forme :

$$f(x) = a \ln x + b$$

À nous de déterminer les meilleures valeurs de a et b. On peut par exemple considérer que la courbe passe par le 1er et le dernier point du nuage. Ainsi, on a à résoudre le système :

$$\begin{cases} a \ln 4 & +b = 37,9 \\ a \ln 29 & +b = 427,7 \end{cases}$$

On peut alors prendre $a = 197$ et $b = -235$. On obtient alors la courbe orange dessinée sur le premier graphique et on constate que f est une bien meilleure approximation que celle obtenue par la méthode précédente.