Statistiques

Filière informatique de gestion

Karim Saïd

Haute école de gestion Arc

Table des matières

Chapitre 1

Statistiques descriptives

1.1 Introduction

Statistique vient du mot latin *status* qui signifie état, situation. Les premières ébauches de la statistique remontent aux recensements qui furent mis sur pied dès les premiers siècles de notre ère. Ce n'est pourtant qu'au $18^{\text{ème}}$ siècle qu'elle se constitue comme une discipline scientifique autonome. Aujourd'hui, la statistique est une branche des mathématiques appliquées intervenant dans divers domaines de la pensée humaine tels que la démographie, l'économie, la médecine, l'agronomie ou encore l'industrie.

La démarche statistique peut se décomposer en cinq étapes. Premièrement, il s'agit d'identifier précisément la population et le (les) caractère(s) à étudier. Suite à cela, des données seront récoltées par recensement ou échantillonnage. Ensuite, il faudra regrouper, classifier et présenter les données (*statistique descriptive*). Il conviendra alors de comparer les résultats avec des modèles théoriques (*calcul des probabilités*). Enfin, il s'agira d'interpréter les résultats et d'établir des hypothèses plausibles en vue de prévisions (*statistique inférentielle*) concernant des circonstances analogues.

Ce chapitre se borne à une introduction à la statistique descriptive en présentant, sur la base de deux exemples illustratifs, les quelques mesures qui caractérisent un ensemble fini de données.

1.2 Définitions

Définition.

1. On appelle *population* l'ensemble de référence sur lequel vont porter les observations. Il est d'usage de désigner par la lettre N la taille d'une population.
2. On appelle *échantillon* une partie de la population que l'on détermine par sondage lorsque la population est trop nombreuse à étudier ou impossible à observer dans sa totalité.
3. On appelle *individu* tout élément de la population.
4. Lorsque l'on peut ainsi étudier une caractéristique que possède chacun des individus, on appelle cela une *variable statistique* ou *caractère*.
5. Les différentes valeurs que peut prendre une variable statistique sont les *modalités* de cette variable.
6. Le nombre d'individus vérifiant une modalité donnée est appelé *l'effectif*.
7. La *fréquence* d'une modalité est le rapport entre l'effectif et le nombre d'observations. On l'exprime souvent en pour cent.

Notation. On note une variable statistique par une lettre majuscule X, Y, ... et ses modalités par la même lettre minuscule affectée d'indices : x_1, x_2, ... pour la variable X ou y_1, y_2, ... pour la variable Y.

Exemple. On fait une étude auprès des étudiants de la HE Arc. On aimerait connaître le sexe, l'âge au 1er janvier, la taille et le taux de satisfaction par rapport aux études (satisfait (S), insatisfait (I) et sans réponse (S)) de chaque étudiant.

La population considérée est "les étudiants de la HE Arc". Un *échantillon* est par exemple l'ensemble des étudiants inscrits en informatique de gestion. Tout étudiant inscrit dans cette filière est un *individu*.

Variable statistique	Modalités
X : sexe	$x_1 =$ homme, $x_2 =$ femme
Y : âge	$y_1 = 18$, $y_2 = 19$, ...
Z : taille	$z_i \in [150; 200]$
U : taux de satisfaction	$u_1 = T$, $u_2 = I$, $u_3 = S$

Définition.

1. Une variable statistique est dite *qualitative*, respectivement *quantitative* si ses valeurs peuvent être comptées, respectivement qu'elles ne peuvent pas l'être.
2. On dit d'une variable statistique qualitative X qu'elle est *nominale*, respectivement *ordinale*, si ses valeurs ne possèdent pas d'ordre, respectivement si ses valeurs possèdent une certain ordre.
3. On dit d'une variable statiatique quantitative X qu'elle est *discrète*, respectivement *continue*, si l'ensemble des valeurs de celle-ci est fini ou infini dénombrable, respectivement infini non dénombrable.

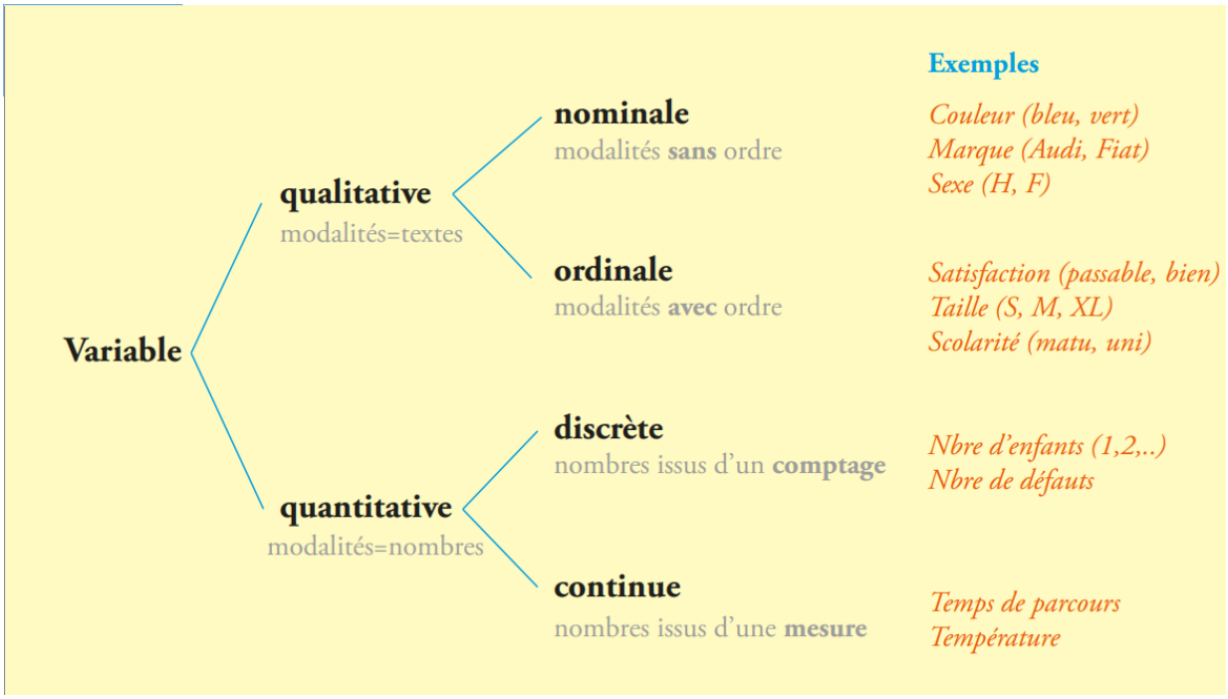


FIGURE 1.1 – Caractérisation des différentes variables statistiques.

Exemple. Dans notre exemple précédent, X est une variable statistique qualitative nominale, Y est une une variable statistique quantitative discrète, Z est une une variable statistique quantitative continue et U est une une variable statistique qualitative ordinale

1.3 Traitement des données

1.3.1 Regroupement des données par modalités

Exemple. On étudie l'état civil des 40 employés d'une compagnie.

Dans un premier temps, il s'agit de collecter l'information, dans ce cas l'état civil de chacun des individus de la population (les 40 employés de la compagnie) : les données brutes. La variable statistique est l'état civil. Elle est qualitative nominale et les modalités sont : marié(e), célibataire, divorcé(e) et veuf(ve).

On donne l'état civil des employés identifiés par un numéro :

1	Marié	11	Veuf	21	Célibataire	31	Célibataire
2	Mariée	12	Marié	22	Mariée	32	Divorcée
3	Célibataire	13	Célibataire	23	Marié	33	Divorcé
4	Divorcé	14	Célibataire	24	Marié	34	Marié
5	Marié	15	Mariée	25	Divorcée	35	Mariée
6	Célibataire	16	Célibataire	26	Mariée	36	Marié
7	Célibataire	17	Marié	27	Célibataire	37	Marié
8	Mariée	18	Veuve	28	Célibataire	38	Mariée
9	Marié	19	Marié	29	Marié	39	Célibataire
10	Divorcée	20	Divorcé	30	Veuf	40	Mariée

On obtient avec ces données brutes une information personnalisée. On va sacrifier le caractère individuel de l'information pour obtenir un portrait d'ensemble. On calcule pour chaque modalité le nombre d'individus ayant cette modalité : *l'effectif de la modalité* :

20 individus mariés
11 individus célibataires
6 individus divorcés
3 individus veufs

Il est d'usage de présenter la distribution des effectifs sous la forme d'un tableau :

Modalités	Effectifs	Fréquences
Mariés	20	50%
Célibataires	11	$27,5\%$
Divorcés	6	15%
Veufs	3	$7,5\%$
Total	40	100%

Pour trouver qu'il y a $27,5\%$ de célibataires, il suffit de calculer

$$\frac{11}{40} = 0,275 = 27,5\%.$$

Exemple. Dans un quartier composé de 50 ménages, on étudie le nombre de personnes par ménage.

Dans un premier temps, il s'agit de collecter les données brutes de chacun des individus de la population (les 50 ménages). La variable statistique est le nombre de personnes par ménage. Elle est quantitative discrète et les modalités sont : 1, 2, 3, 4, 5, 6 et 8.

Les données brutes sont :

1	1	1	1	1	2	2	2	2	2
2	2	2	2	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	4
4	4	4	4	4	4	4	4	4	5
5	5	5	5	5	6	6	6	8	8

On obtient avec ces données brutes une information personnalisée. Cette liste n'étant pas commode à lire, il convient à nouveau de sacrifier le caractère individuel de l'information pour obtenir un portrait d'ensemble. On va donc déterminer pour chaque modalité le nombre d'individus ayant cette modalité : *l'effectif de la modalité.*

Modalités x_i	Effectifs n_i	Fréquences f_i
1	5	10%
2	9	18%
3	15	30%
4	10	20%
5	6	12%
6	3	6%
8	2	4%

Dans le tableau ci-dessus, on a $x_1 = 1$, $x_2 = 2$, etc. Les x_i représentent le nombre de personnes par ménage. On a de plus $n_1 = 5$, $n_2 = 9$, etc. Les n_i indiquent le nombre de ménages comportant x_i personnes. Ainsi, on a par exemple 10 ménages comportant 4 personnes.

1.3.2 Représentation des données à l'intérieur des classes

Souvent, lors d'une étude statistique portant sur une variable statistique quantitative discrète ou continue, les données recueillies diffèrent à peu près toutes les unes des autres et sont étalées sur un large intervalle de valeurs. L'objectif de la statistique descriptive étant de résumer de la façon la plus adéquate possible cet ensemble de données, nous procédons alors à un regroupement de ces dernières à l'intérieur de *classes*, c'est-à-dire de sous-intervalles de valeurs. Les règles suivantes permettent de choisir judicieusement ces classes :

— On fixe un nombre de classes entre 5 et 15. Le nombre de classes dépend de la taille de la population et il faut éviter, si possible, des fréquences de classes trop petites.

— Les intervalles sont du type $[b_{i-1}; b_i[$ ou $]b_{i-1}; b_i]$.

b_{i-1} est la *borne inférieure* de la classe i ;

b_i est la *borne supérieure* de la classe i ;

$c_i = \dfrac{b_{i-1} + b_1}{2}$ est le *centre* de la classe i ;

$L_i = b_i - b_{i-1}$ est la *largeur* (ou *étendue* ou *amplitude*) de la classe i.

— En principe, on fixe les bornes des intervalles de telle sorte que ces derniers soient d'égales largeurs. Les bornes doivent permettre des calculs simples.

— Si on doit vraiment utiliser des classes de largeurs inégales, on place les classes de largeur égale au centre de la distribution.

Exemple. Dans une région française, on étudie la superficie de chacune des 500 exploitations agricoles exprimées en hectares.

Dans cet exemple, la population est l'ensemble des exploitations agricoles d'une région française, tandis qu'un individu est ici une exploitation agricole donnée. La population étant définie, elle est observée selon certains critères. Le critère retenu, c'est-à-dire la variable statistique est ici la superficie. Elle est de nature quantitative continue et les modalités sont des nombres représentant des superficies compris entre 0 ha et 40 ha.

Les données brutes que l'on receuille sur cette population sont inutilisables telles quelles. En vue de synthétiser l'information, on procède à des regroupements, à des classements et à l'établissement de tableaux statistiques. Le tableau ci-dessous constitue déjà une première simplification de l'information complète contenue dans un registre officiel comportant une ligne pour chacune des 500 exploitations.

Superfice en ha	Nombre d'exploitations	Fréquences en %
$]0; 10]$	48	$9,6$
$]10; 15]$	62	$12,4$
$]15; 20]$	107	$21,4$
$]20; 25]$	133	$26,6$
$]25; 30]$	84	$16,8$
$]30; 40]$	66	$13,2$

Les individus étant rassemblés par *classes*, on dit qu'on a affaire à un ensemble de données *groupées*. Ce qu'on gagne en simplicité par ce regroupement, on le perd en information. On sait par exemple que la classe $]20; 25]$ comporte 133 exploitations dont les superficies sont comprises entre 20 et 25. Mais on ne connaît rien de la répartition de ces 133 individus à l'intérieur de leur classe. Il est alors commode de formuler l'hypothèse d'une répartition uniforme au sein de chaque classe. On assigne ainsi à l'individu occupant la place x sur 133 dans la classe $]20; 25]$ (d'*étendue* 5), la valeur $20 + \dfrac{x}{133} \cdot 5$. Avec cette convention, le dernier individu (le $133^{\text{ème}}$) est affecté de la valeur 25, borne supérieure de l'intervalle.

1.4 Représentations graphiques

1.4.1 Diagramme en secteurs

La répartition d'une population et sa distribution de fréquences sont parfois plus expressives sur le plan visuel lorsqu'on les représente à l'aide d'un *diagramme circulaire*. Un diagramme circulaire consiste à représenter la population totale par un cercle et à la diviser en tranches, de façon proportionnelle aux effectifs de la variable statistique considérée. On obtient ainsi une représentation graphique de la répartition relative de la population, autrement dit de la distribution de fréquences.

Exemple. Reprenons notre exemple des exploitations agricoles. Ce qui caractérise "la taille d'une tranche" est l'angle au centre. Pour le trouver, il suffit de faire une règle de trois avec la relation 360° correspond à une fréquence de 100% ou, de manière équivalente, à un effectif de 500.

Superfice en ha	Effectifs	Fréquences en %	Angles en °
$]0;10]$	48	$9,6$	$34,56$
$]10;15]$	62	$12,4$	$44,64$
$]15;20]$	107	$21,4$	$77,07$
$]20;25]$	133	$26,6$	$95,76$
$]25;30]$	84	$16,8$	$60,48$
$]30;40]$	66	$13,2$	$47,52$

FIGURE 1.2 – Données avec angles.

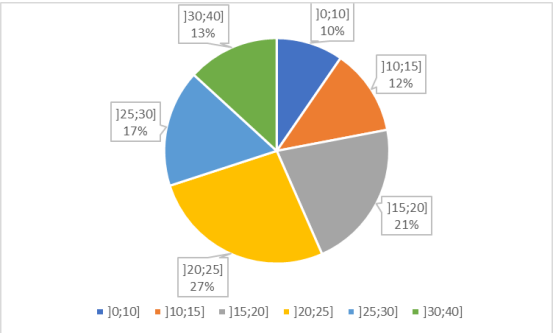


FIGURE 1.3 – Diagramme en secteurs.

Exemple. Reprenons notre exemple relatif à l'état civil des employés d'une compagnie. Pour représenter le diagramme en secteurs, il convient de déterminer l'angle de chaque tranche.

Etats civils	Effectifs	Fréquences en %	Angles en °
Mariés	20	50	180
Célibataires	11	$27,5$	99
Divorcés	6	15	54
Veufs	3	$7,5$	27

FIGURE 1.4 – Données avec angles.

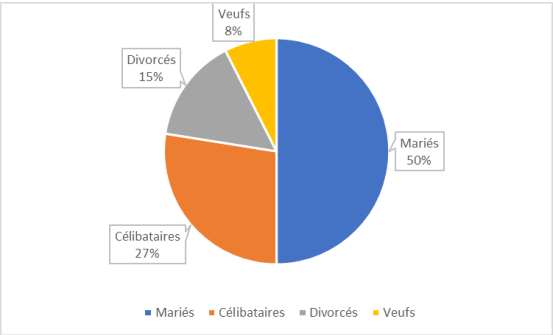


FIGURE 1.5 – Diagramme en secteurs.

1.4.2 Diagramme en bâtons

Lorque la variable statistique est quantitative discrète, la distribution des effectifs peut être représentée visuellement par un *diagramme en bâtons*. Il s'agit d'un diagramme dans lequel les modalités se trouvent sur l'axe horizontal et chaque bâton monte jusqu'à hauteur de l'effectif (ou de la fréquence) correspondant(e).

Exemple. Reprenons notre exemple relatif à l'état civil des employés d'une compagnie. Le diagramme en bâtons de cette distribution est représenté ci-dessous.

Modalités	Effectifs
Mariés	20
Célibataires	11
Divorcés	6
Veufs	3
Total	40

FIGURE 1.6 – Données.

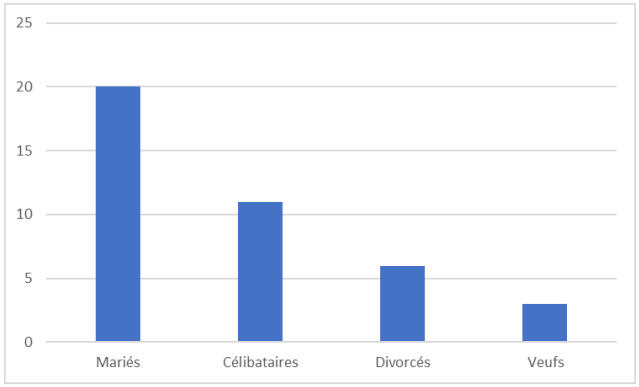


FIGURE 1.7 – Diagramme en bâtons.

Exemple. Reprenons notre exemple relatif au nombre de personnes par ménage.

Modalités	Effectifs
1	5
2	9
3	15
4	10
5	6
6	3
7	0
8	2

FIGURE 1.8 – Données.

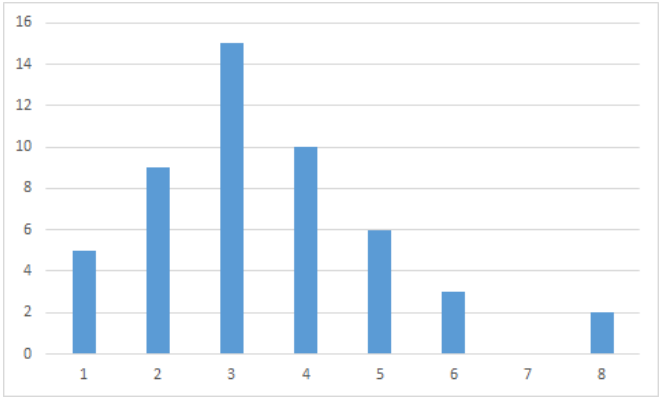


FIGURE 1.9 – Diagramme en bâtons.

1.4.3 Histogramme

Lorsque la variable statistique est quantitative continue ou discrète, mais que les données sont regroupées en classes, la distribution peut être représentée visuellement par un *histogramme*, qui est un diagramme en colonnes où les rectangles sont juxtaposés. En effet, les modalités sont ici remplacées par des classes et celles-ci sont formées d'intervalles successifs de sorte qu'il n'y a plus lieu de séparer ces rectangles.

Exemple. Dans notre exemple, les classes de superficie n'ont pas toutes la même *amplitude*. Certaines classes ont une amplitude de 10 ha, d'autres 5 ha. Pour être fidèle, une représentation graphique doit tenir compte de ces différences. Si, dans un *histogramme*, on représente les classes par des rectangles, alors, la surface totale représentant l'ensemble de la population, il faut que chaque rectangle ait une aire proportionnelle à l'effectif de la classe que ce dernier représente.

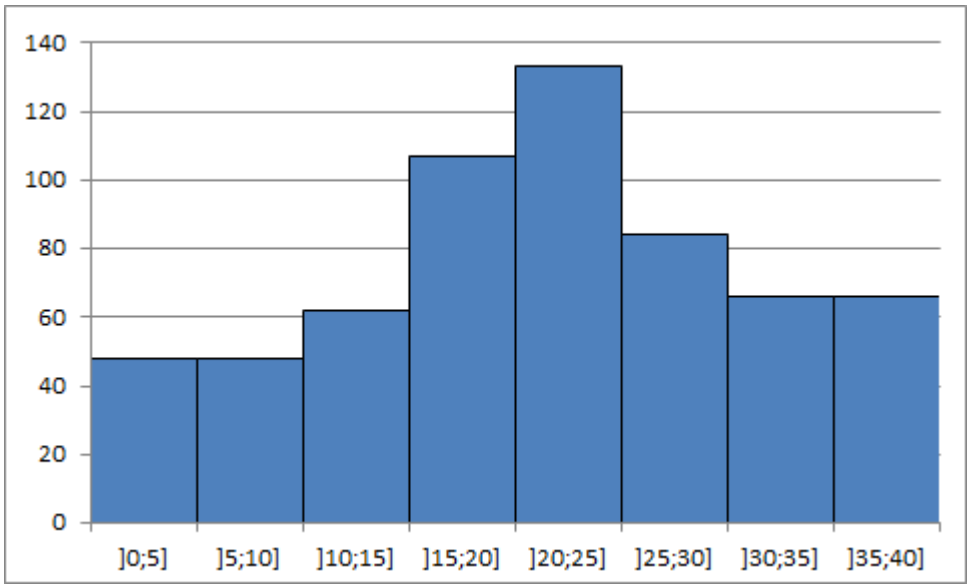


FIGURE 1.10 – Histogramme trompeur.

L'histogramme représenté ci-dessus est trompeur dans la mesure où il donne l'impression erronée que la classe initiale $[0; 10[$ contient 96 exploitations : 48 d'une surface de 0 à 5 ha et le même nombre d'une surface de 5 à 10 ha. Pour éviter cette déformation, il y a lieu de choisir une amplitude de référence (par exemple 5 ha) et de procéder à une correction des effectifs.

Avec cette correction, on obtient alors le tableau et l'histogramme correspondant suivants.

Superfice en ha	Nombre d'exploitations
$]0; 5]$	24
$]5; 10]$	24
$]10; 15]$	62
$]15; 20]$	107
$]20; 25]$	133
$]25; 30]$	84
$]30; 35]$	33
$]35; 40]$	33

FIGURE 1.11 – Effectifs corrigés.

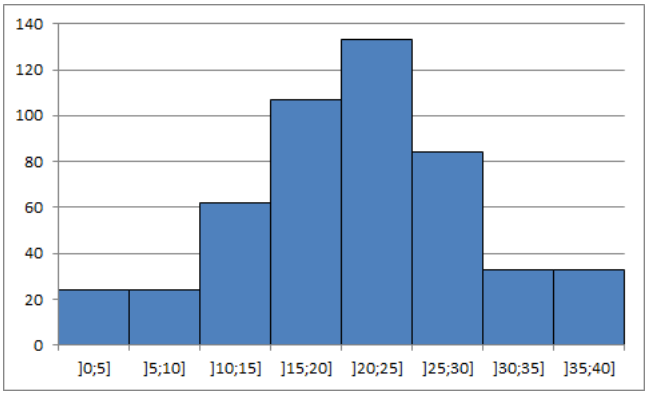


FIGURE 1.12 – Histogramme correct.

Algorithme de correction des effectifs

1. On choisit une classe de référence de largeur l (en général la plus fréquente).
2. Pour une classe quelconque de largeur L et d'effectif E, on calcule le rapport $x = \dfrac{E}{L}$.
3. On attribue alors à cette classe l'effectif corrigé $c = x \cdot l = \dfrac{E}{L} \cdot l$. Notons que cet effectif n'est pas forcément un nombre entier.

Exemple. Dans notre exemple, la classe de référence ayant pour largeur $l = 5$, la classe $]0; 10]$ a pour largeur $L = 10$, on calcule $x = \dfrac{E}{L} = \dfrac{48}{10} = 4,8$, ce qui conduit à l'effectif corrigé $c = x \cdot l = 4,8 \cdot 5 = 24$.

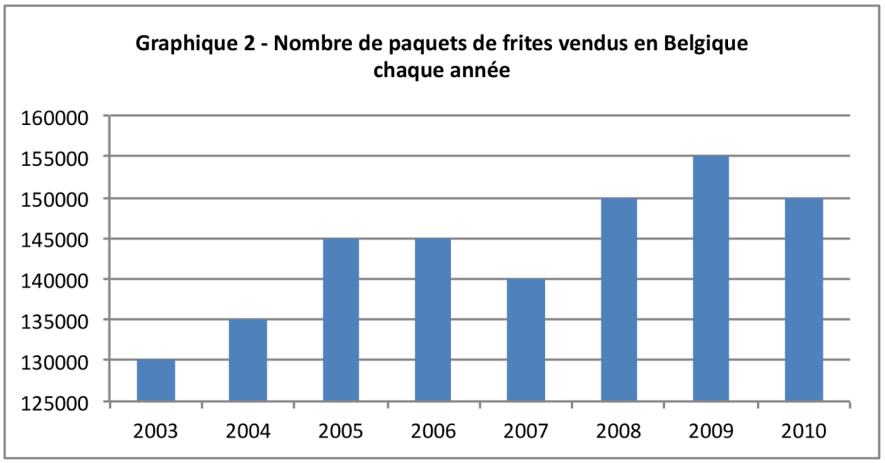
1.4.4 Diagrammes trompeurs ou faux

Dans la presse, à la télévision ou dans des tracts à caractère politique, il n'est pas rare d'y découvrir des diagrammes ou des graphes déformant la réalité, voir complètement faux. Le but de cette section consiste à donner quelques exemples et de mettre en avant les techniques utilisées pour déformer la réalité.

Exemple.

1. Dans cet exemple relatif à l'évolution du nombre de paquets de frites vendus en Belgique, nous allons voir comment présenter les données pour donner trois messages radicalement différents.

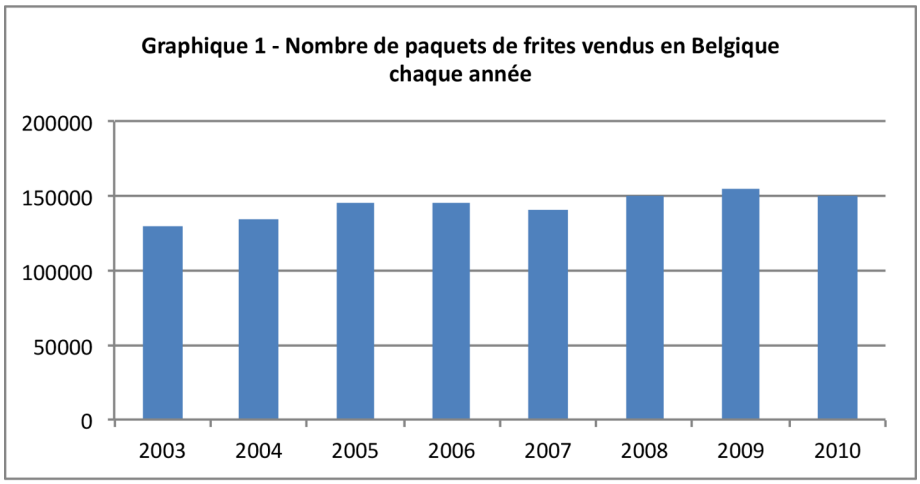
   a) En dépit des chiffres de 2007, le diagramme en bâtons ci-dessous semble indiquer une augmentation des ventes du nombre de paquets de frites.



   Cependant, on y regardant de plus près, on observe que l'axe des y ne part pas à 0, mais à 125'000 ! Sur ce diagramme, on peut y lire que 130'000 paquets de frites ont été vendus en 2003, contre 135'000 en 2004, ce qui correspond à une augmentation de 5'000 paquets en une année, soit d'environ $3,85\%$. Or, l'effet visuel du diagramme laisse supposer au premier abord que les ventes ont doublé durant cette période, c'est-à-dire qu'elles ont augmenté de 100% ! Remarquons enfin que le diagramme donne

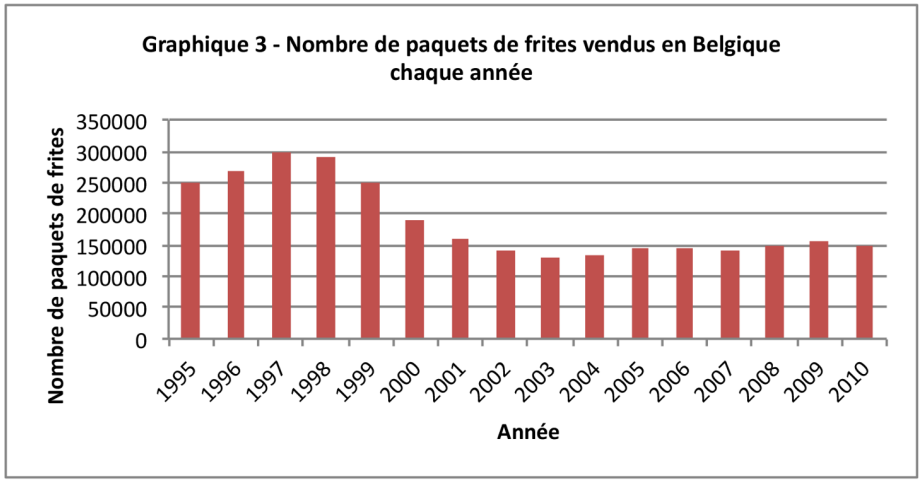
l'impression que les ventes ont été multipliées par 6 entre 2003 et 2009, alors qu'elles sont passées de 130'000 à 155'000, ce qui fait 25'000 de plus, soit une augmentation de presque 20% !

b) En présentant les mêmes données, mais en faisant partir l'axe des y de l'origine, le diagramme en bâtons ci-dessous semble indiquer une tendance de la vente des paquets de frites plutôt stable.

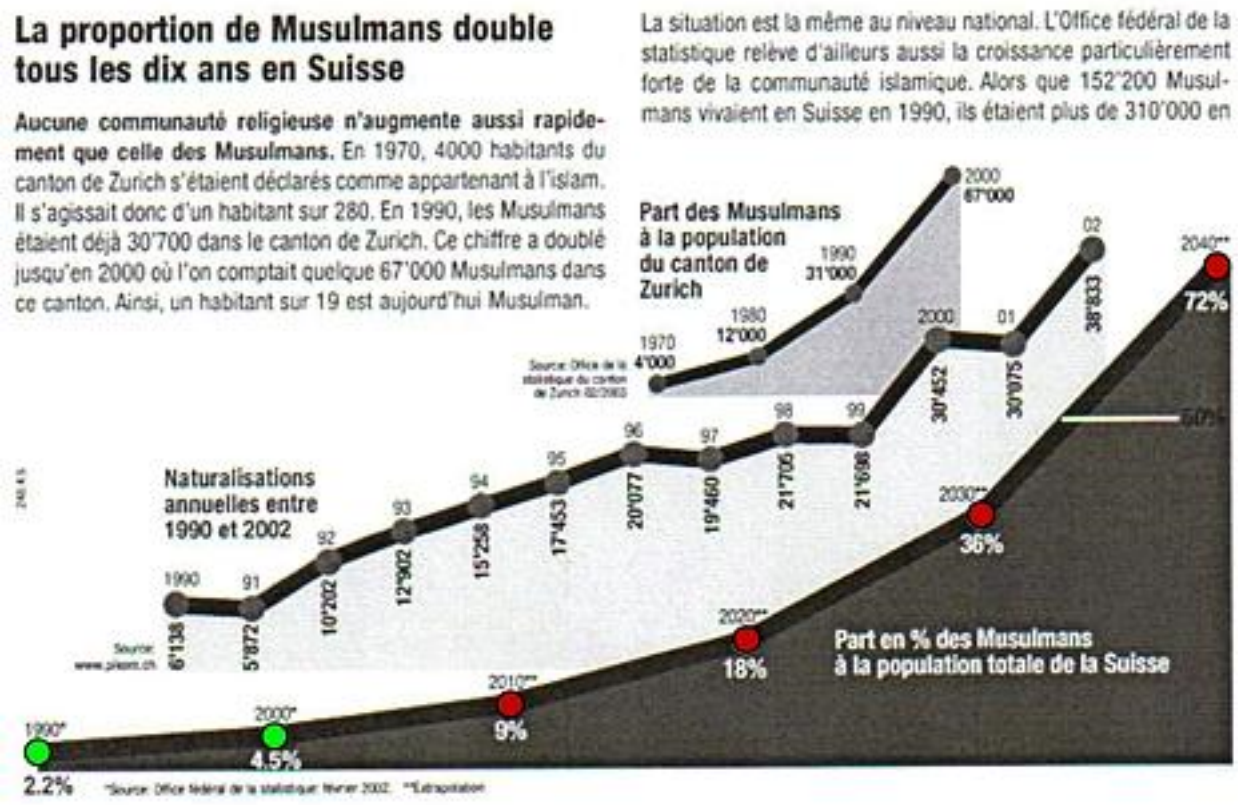


c) Les deux diagrammes ci-dessus contiennent uniquement les chiffres des ventes entre 2003 et 2010. Qu'en est-il si on considère les chiffres des années précédent 2003 ?

Le diagramme ci-dessous présente l'évolution du nombre de paquets de frites entre 1995 et 2010. En tenant compte de ces chiffres, il semble que les ventes de paquets de frites ont tendance à diminuer !



2. En vue des votations du 26 septembre 2004, un parti politique publie le document suivant.



Le texte explique qu'aucune communauté religieuse n'augmente autant rapidement que les musulmans. La courbe ci-dessus semble en effet indiquer que la croissance du nombre de musulmans en Suisse est exponentielle. Or, en y regardant de plus près, on observe que les chiffres de 1990 et de 2000 ($2,2\%$ et $4,5\%$) sont munis d'une étoile indiquant qu'ils proviennent de l'Office fédéral de la statistique. Les chiffres suivants (à partir de 2010) sont quant à eux munis de deux étoiles, pour préciser qu'il s'agit d'une extrapolation.

Mais comment arriver à un tel pronostic ? On observe que $4,5\%$ représente à peu près le double de $2,2\%$. Avec ces seuls deux chiffres, on en conclut que le pourcentage de la communauté musulmane de Suisse double tous les 10 ans pour atteindre ainsi 72% en 2040, soit le dernier point représenté sur le graphe. On comprend mieux pourquoi le graphe s'arrête à ce point. En effet, le suivant indiquerait que le taux de musulmans s'élèverait à 144% en 2050 ! Cette projection est donc basée sur un doublement arbitraire, mais renforcée par les chiffres zurichois, qui indiquent une forte progression entre 1970 et 2000. Cette extrapolation basée sur un seul canton ne donne aucune raison de penser que ces chiffres sont transposables à toute la Suisse. Notons enfin, que selon l'OFS, il y avait $4,9\%$ de musulmans en Suisse en 2011 et $5,3\%$ en 2018. Soit des valeurs bien différentes des 9% et 18% prédites par les auteurs du document ci-dessus !

3. Quant à l'affiche ci-dessous, elle contient un certain nombre d'éléments forts discutables. Jörg Mäder, conseiller national zurichois depuis 2019, décortique les nombreux éléments controversés de cette affiche sur cette video.

4. Le graphique publié par un quotidien en août 2008 (ci-dessous, à gauche) semble montrer que la consommation de viande s'est stabilisée ces dernières années. Cependant, on y regardant de plus près, on observe que que l'axe horizontal du graphique n'est pas linéaire : la moitié du graphique représente 50 ans, alors que l'autre moitié (la partie stable) ne concerne que 7 ans, donnant ainsi une impression erronée de la situation. Le graphique de droite montre les même données de façon correcte, et donne une impression différente.

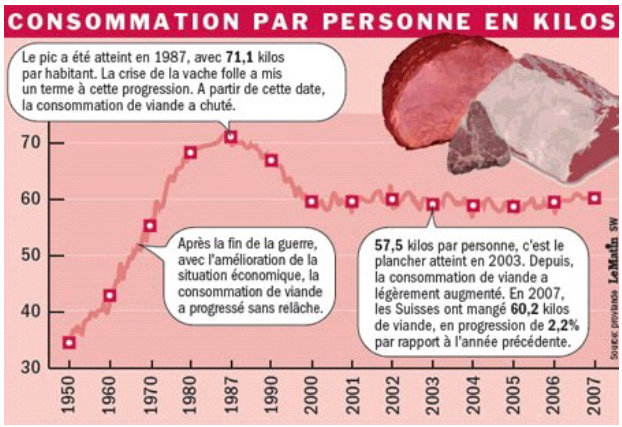


FIGURE 1.13 – Graphe faux.

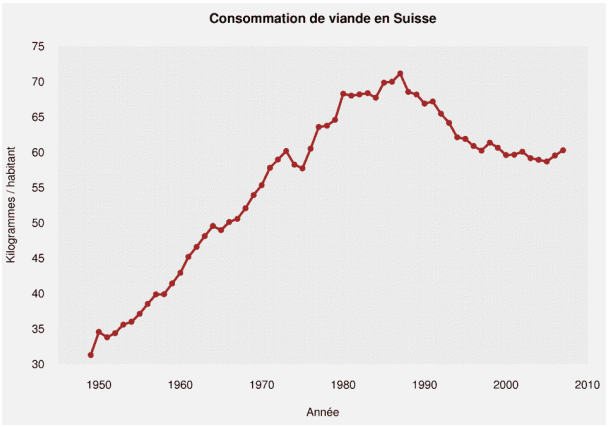
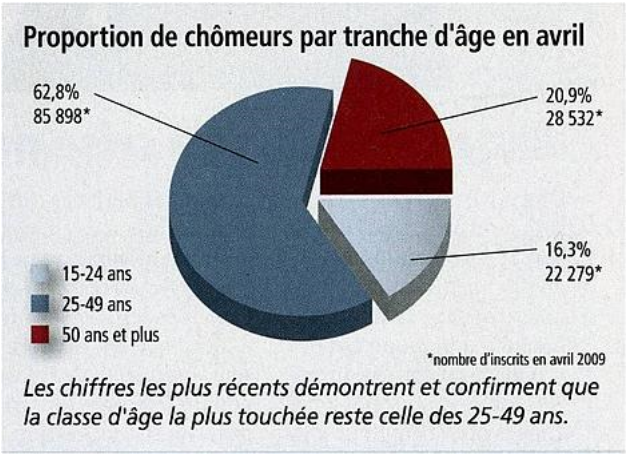


FIGURE 1.14 – Graphe correct.

On peut voir une autre différence entre les deux graphiques : celui de gauche indique des variations à l'intérieur des années. En fait, il apparaît que ces variations ont été ajoutées pour éviter que le graphique ne soit trop lisse. On peut s'étonner que de telles considérations purement esthétiques prennent le pas sur le traitement correct de l'information.

5. Le diagramme circulaire ci-dessous représente la proportion de chômeurs par tranche d'âge.



La légende conclut que la classe la plus touchée est celle des 25 à 49 ans. Les trois classes étant d'amplitudes différentes, il est difficile d'établir des comparaisons. Il n'est en effet pas surprenant que le plus grand nombre de chômeurs se trouve dans la classe la plus peuplée !

En fait, la valeur intéressante n'est pas la valeur absolue, mais le pourcentage de chômeurs à l'intérieur de chaque classe. Selon le rapport du SECO utilisé pour créer le graphique, ces taux sont de 4% pour les 15-24 ans, de 3.6% pour les 25-49 ans, et en-dessous de 3% pour les 50-65 ans. La classe la plus touchée est donc celle des jeunes, contrairement à ce qu'en dit l'auteur du diagramme.

6. Une chaîne de télévision a présenté le diagramme ci-dessous en 2011. Celui-ci rend compte du taux de dépenses publiques en 2011 en % du PIB de trois pays et de l'Union européenne.

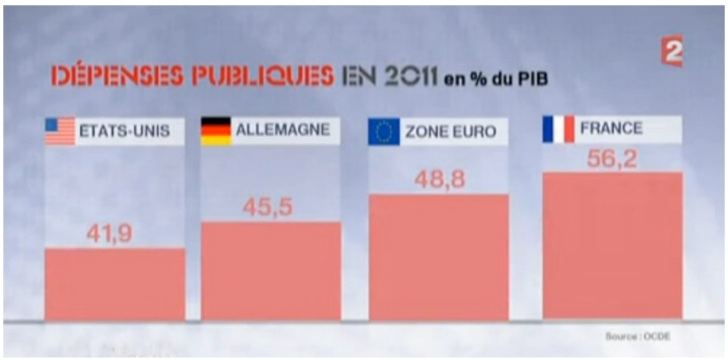


FIGURE 1.15 – Diagramme faux.

Si le 100% correspond à la zone comprise entre l'horizontale et à la parallèle passant juste en dessous du nom des pays, les 41,9% de dépenses publiques de Etats-Unis semblent être correctement représentées. Cela n'est pas le cas pour les autres pays ! En effet, les 56,2% de la France sont beaucoup trop hauts. Le diagramme donne l'impression que les dépenses publiques de la France sont de l'ordre de 80% ! Cette technique a pour objectif de susciter une émotion auprès de la population, en amplifiant la différence du taux de dépenses publiques par rapport à d'autres pays.

Ci-dessous, figure le diagramme correct, tel qu'il aurait dû être présenté aux téléspectateurs.

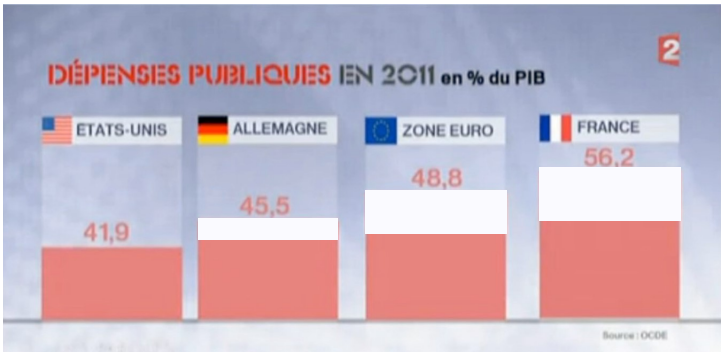
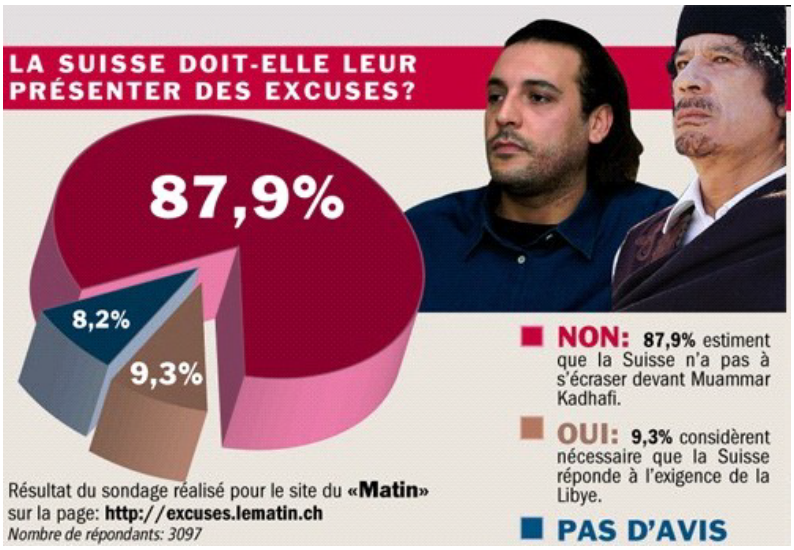


FIGURE 1.16 – Diagramme correct.

7. La figure ci-dessous illustre le fait que le prix des montres a augmenté de 40% sur 7 ans.

Le graphiste a voulu associer cette augmentation au diamètre des montres. On peut vérifier qu'ils augmentent bien de 40%. Le lecteur voit l'augmentation de la surface des horloges qui, elle, n'est pas de 40%, mais proche de 100% ! Enfin, les aiguilles ont été ajoutées dans un but purement esthétique, mais peuvent induire en erreur en donnant l'impression qu'elles contiennent de l'information.

8. Dans le diagramme circulaire ci-dessous, la somme des parties fait $105,4\%$! Le $8,2\%$ était probablement un $2,8\%$ à l'origine, ce qui donnerait la somme attendue de 100%.



1.4.5 Polygone des effectifs

A l'histogramme, on associe souvent le *polygone des effectifs*. Il s'agit d'une courbe polygonale telle que la surface comprise entre cette courbe et l'axe des abscisses soit égale à la surface de l'histogramme. Elle est obtenue en joignant les milieux des sommets des rectangles de l'histogramme. Pour la première et la dernière classe, on crée à cet effet deux classes fictives d'effectifs nuls.

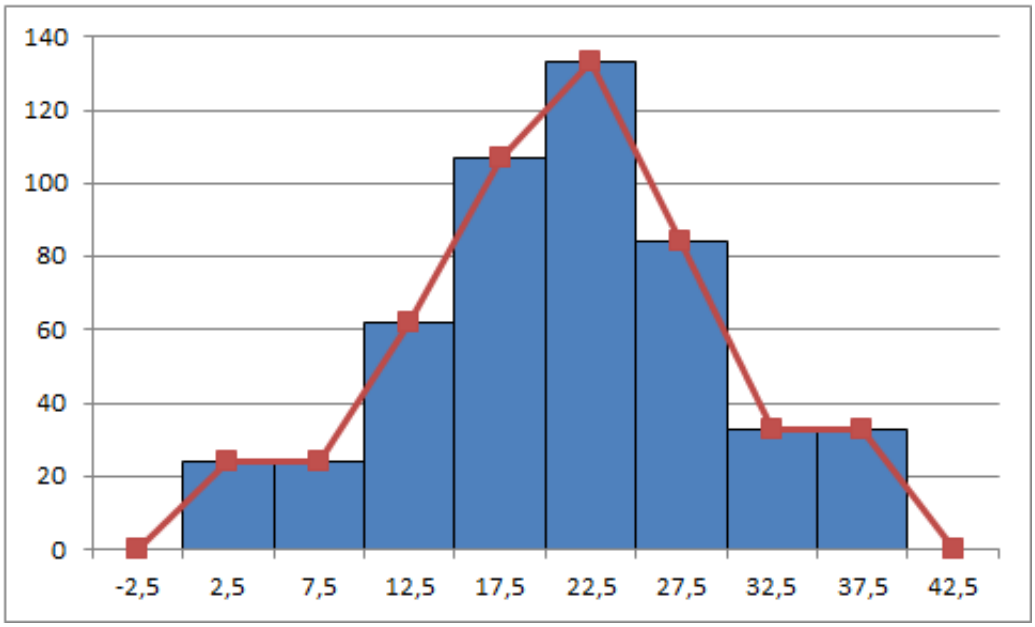


FIGURE 1.17 – Polygone des effectifs.

1.4.6 Polygone des effectifs cumulés

Aux données de départ, on associe le tableau des effectifs cumulés croissants et cumulés décroissants. On interprète les données de ce tableau comme suit. On peut affirmer, par exemple, que 350 exploitations agricoles ont une superficie d'au plus 25 ha. Par ailleurs, 283 exploitations ont une superficie supérieure à 20 ha.

Classes	Effectifs	Effectifs cumulés croissants	Effectifs cumulés décroissants
$]0;10]$	48	48	500
$]10;15]$	62	110	452
$]15;20]$	107	217	390
$]20;25]$	133	350	283
$]25;30]$	84	434	150
$]30;40]$	66	500	66

Les données contenues dans ce tableau peuvent être représentées par deux courbes : *le polygone des effectifs cumulés croissants* et le *polygone des effectifs cumulés décroissants*. Dans la représentation de ces courbes, on ne se préoccupe pas des différences d'amplitude des classes. Notons qu'il est également possible de réaliser un polygone des effectifs à partir des fréquences.

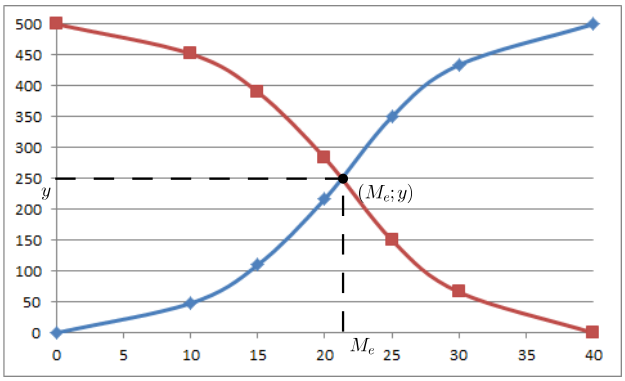


FIGURE 1.18 – Polygones des effectifs cumulés.

Remarque. L'intersection de ces deux polygones est un point $(M_e; y)$, dont la premiere coordonnee est appelee *mediane* M_e de la population. On observe, dans notre exemple, que $M_e \cong 21$. Cette valeur divise la population en deux parties d'effectifs egaux. En effet, soit y la seconde coordonnée du point d'intersection. Comme ce point est sur le polygone des effectifs cumulés croissants, on peut affirmer que y exploitations ont une superficie inferieure à M_e ; le reste, c'est-a-dire $500 - y$, ont une superficie superieure à M_e. Ce point étant également sur le polygone des effectifs cumulés décroissants, y decrit le nombre d'exploitations ayant une superficie superieure à M_e. On en deduit que $y = 500 - y$ et donc, que $y = \dfrac{500}{2} = 250$. Ainsi la moitié des exploitations ont une superficie supérieure (respectivement inférieure) à $M_e \cong 21$ ha.

Remarque. Dans une étude statistique, si on souhaite connaître la proportion de chaque valeur que peut prendre la variable statistique étudiée, on regarde sa fréquence f_i.

Si par contre on souhaite connaître la proportion des individus qui présentent des valeurs inférieures à une valeur fixée, on regarde la fréquence cumulée croissante F_i.

Pour visualiser la proportion des individus qui présentent des valeurs supérieures ou égales à une valeur fixée, on étudiera alors la fréquence cumulée décroissante F_i'.

1.5 Valeurs centrales

Une *valeur centrale* est une valeur caractéristique ou représentative d'un ensemble de données. Si cette valeur caractéristique a tendance à se situer au milieu d'un ensemble de données rangées par ordre de grandeur croissant, alors on dit qu'elle est une *mesure de tendance centrale* ou une *valeur centrale*.

1.5.1 Moyenne arithmétique

Cas discret

Définition. La *moyenne arithmétique* $\overline{x}$ est la valeur centrale la plus connue. Elle est égale au quotient de la somme de toutes les valeurs observées du caractère par l'effectif total. Ainsi

$$\overline{x} = \frac{n_1 x_1 + n_2 x_2 + ... + n_N x_N}{N}.$$

Exemple. Reprenons notre exemple du nombre de personnes par ménage :

Modalités x_i	Effectifs n_i	Fréquences f_i
1	5	10%
2	9	18%
3	15	30%
4	10	20%
5	6	12%
6	3	6%
7	0	0%
8	2	4%

la moyenne arithmétique $\overline{x}$ est donnée par

$$\overline{x} = \frac{5 \cdot 1 + 9 \cdot 2 + 15 \cdot 3 + 10 \cdot 4 + 6 \cdot 5 + 3 \cdot 6 + 2 \cdot 8}{50} = 3,44.$$

Cas continu

Pour des séries de données groupées, se fondant sur une répartition uniforme au sein des classes, on convient d'affecter à tous les individus d'une classe $]b_{i-1}, b_i]$ le centre $c = \dfrac{b_{i-1} + b_i}{2}$.

Exemple. Pour notre exemple des exploitations agricoles, à l'aide du tableau suivant

Classes x_i	Centres c_i	Effectifs n_i
$]0; 10]$	5	48
$]10; 15]$	12, 5	62
$]15; 20]$	17, 5	107
$]20; 25]$	22, 5	133
$]25; 30]$	27, 5	84
$]30; 40]$	35	66
Total		500

on tire la moyenne arithmétique des superficies de ces 500 exploitations agricoles, en calculant

$$\overline{x} = \frac{5 \cdot 48 + 12,5 \cdot 62 + 17,5 \cdot 107 + 22,5 \cdot 133 + 27,5 \cdot 84 + 35 \cdot 66}{500} = \frac{10500}{500} = 21 \text{ ha.}$$

1.5.2 Mode

Exemple. On constate que, dans un village de 500 habitants, il y a 490 personnes avec des cheveux noirs et 10 avec des cheveux blonds. Comment résumer la couleur des cheveux "moyenne" des habitants de ce village ? On répondra sûrement "noir", en pensant que l'écrasante majorité des habitants a les cheveux noirs. En réfléchissant ainsi, on donne comme réponse la valeur qui apparaît le plus fréquemment. Il s'agit du *mode*.

Cas discret

Définition. Le *mode*, noté M_o, est la valeur du caractère qui correspond à l'effectif le plus grand ou à la fréquence la plus importante. Cette valeur centrale est simple à percevoir, mais elle ne tient pas compte de l'ensemble des valeurs du caractère étudié.

Exemple. Les nombres 3, 5, 7, 7, 7, 9, 9 ont pour mode $M_o = 7$. Remarquons que le mode peut ne pas être unique. Ainsi, l'ensemble 3, 5, 7, 7, 7, 9, 9, 9, qui a deux modes : 7 et 9, est dit *bimodal*.

Exemple. Reprenons notre exemple du nombre de personnes par ménage.

Modalités x_i	Effectifs n_i
1	5
2	9
3	15
4	10
5	6
6	3
7	0
8	2

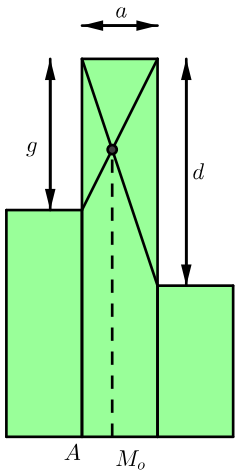
le mode est donné par $M_o = 3$, car 3 est la modalité au plus grand effectif.

Cas continu

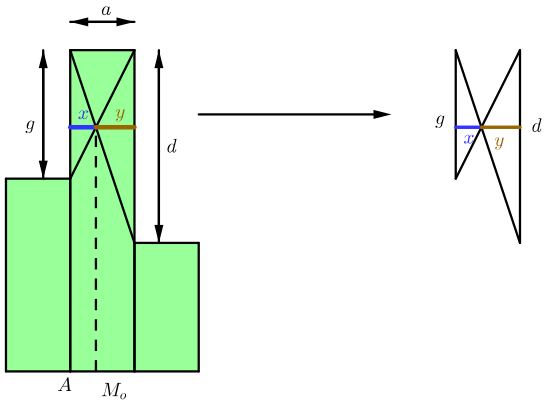
Pour des séries de données groupées par classes, la détermination du mode s'effectue comme suit :

1. On détermine les effectifs rectifiés.
2. On identifie la classe ayant le plus grand des effectifs rectifiés. Elle porte le nom de *classe modale* et peut ne pas être unique.
3. On convient que le mode est déporté à l'intérieur de la classe modale, à droite de sa borne inférieure A, en fonction des effectifs rectifiés des classes voisines. Le mode est alors défini par la formule

$$M_o = A + \frac{g}{g+d} \cdot a.$$



Preuve de la formule du mode :

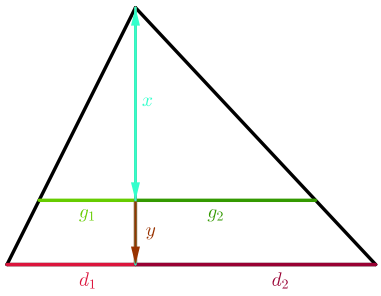


Il est clair que

— $x = M_o - A$.

— $y = a - x = a - (M_o - A) = a - M_o + A$.

Posons alors d_1, d_2, g_1 et g_2 comme le montre la figure ci-dessous.



Puisque l'on a deux triangles semblables, il est possible d'appliquer le théorème de Thalès :

— $\frac{y}{x} = \frac{d_1}{g_1} \Rightarrow y \cdot g_1 = x \cdot d_1$.

— $\frac{y}{x} = \frac{d_2}{g_2} \Rightarrow y \cdot g_2 = x \cdot d_2$.

On a alors

$$\begin{aligned} x \cdot d_1 + x \cdot d_2 &= y \cdot g_1 + y \cdot g_2 \\ x \cdot (d_1 + d_2) &= y \cdot (g_1 + g_2) \\ x \cdot d &= y \cdot g \\ x &= \frac{y \cdot g}{d}. \end{aligned}$$

Comme $x = M_o - A$ et $y = a - M_o + A$, cette dernière égalité s'écrit

$$\begin{aligned}
M_o - A &= \frac{g}{d} \cdot (a - M_o + A) \\
M_o - A &= \frac{g}{d} \cdot a - \frac{g}{d} \cdot M_o + \frac{g}{d} \cdot A \\
M_o + \frac{g}{d} \cdot M_o &= A + \frac{g}{d} \cdot a + \frac{g}{d} \cdot A \\
M_o \cdot \left(1 + \frac{g}{d}\right) &= A + \frac{g}{d} \cdot (a + A) \\
M_o \cdot \left(\frac{d}{d} + \frac{g}{d}\right) &= A + \frac{g}{d} \cdot (a + A) \\
M_o \cdot \left(\frac{d+g}{d}\right) &= A + \frac{g}{d} \cdot (a + A) \\
M_o &= \left(A + \frac{g}{d} \cdot (a + A)\right) \cdot \left(\frac{d}{d+g}\right) \\
M_o &= A \cdot \frac{d}{d+g} + \frac{g}{d} \cdot (a + A) \cdot \frac{d}{d+g} \\
M_o &= A \cdot \frac{d}{d+g} + (a + A) \cdot \frac{g}{d+g} \\
M_o &= \frac{Ad + ag + Ag}{d+g} \\
M_o &= \frac{Ad + Ag}{d+g} + \frac{ag}{d+g} \\
M_o &= \frac{A \cdot (d+g)}{d+g} + \frac{ag}{d+g} \\
M_o &= A \cdot \frac{d+g}{d+g} + \frac{g}{d+g} \cdot a \\
M_o &= A + \frac{g}{d+g} \cdot a.
\end{aligned}$$

Exemple. Dans notre exemple des exploitatations agricoles, après rectification des effectifs, on obtient le tableau suivant :

Superfice en ha	Nombre d'exploitations
$]0; 10]$	24
$]10; 15]$	62
$]15; 20]$	107
$]20; 25]$	133
$]25; 30]$	84
$]30; 40]$	33

La classe modale est donc la quatrième classe. Ainsi $A = 20$, $g = 133 - 107 = 26$, $d = 133 - 84 = 49$ et $a = 5$. Il s'ensuit que

$$M_o = 20 + \frac{26}{26 + 49} \cdot 5 \cong 21,73 \text{ ha}.$$

1.5.3 Médiane

Exemple. En 2016, un Suisse apprend par la presse que l'OFS estime le salaire brut moyen à 7491 francs. Il le compare avec son salaire qui se monte à 6942 francs et peste contre la pingrerie de son employeur chez qui il court réclamer une augmentation. Mais le salaire moyen est-il un indicateur pertinent dans ce cas ? Sûrement pas. Il est basé sur un grand nombre de personnes gagnant peu et un nombre restreint de managers gagnant des salaires indécents se montant à plusieurs millions, entraînant ainsi une distorsion vers le haut du salaire moyen. Il faudrait plutôt que notre individu se pose la question de savoir s'il gagne plus ou moins que la plupart de ses compatriotes. Pour répondre à cette interrogation, on va considérer la *médiane*. Cette indice coupe la population en deux parties égales. La médiane des salaires bruts en Suisse étant de 6502 francs en 2016 selon l'OFS, il est plutôt favorisé puisqu'il fait partie de la moitié de la population qui gagne le plus !

Cas discret

Définition. La *médiane*, notée M_e, est la valeur du caractère qui partage en deux l'effectif total. C'est la valeur du caractère qui correspond à une fréquence cumulée égale à 50%. Dans une population, il y a ainsi autant d'individus possédant une valeur du caractère inférieure au caractère médian que d'individus possédant une valeur du caractère supérieure à la médiane.

La *classe médiane* d'une variable continue est la première classe où la fréquence cumulée atteint ou dépasse 50%.

Exemple.

1. L'ensemble des nombres 3, 4, 4, 5, 6, 8, 8, 8, 10 a pour médiane $M_e = 6$.
2. L'effectif de l'ensemble 5, 5, 7, 9, 11, 12, 15, 18 étant pair, ce dernier a pour médiane $M_e = \dfrac{9+11}{2} = 10$.

Remarque. On constate que la médiane correspond à la valeur du caractère de l'individu occupant le rang

$$m = \frac{N+1}{2}.$$

— Si N est impair, il s'agit d'un individu réel occupant le rang entier m.

— Si N est pair, il s'agit d'un individu virtuel placé entre les rangs $N/2$ et $N/2+1$.

Exemple. Reprenons notre exemple du nombre de personnes par ménage. Pour déterminer la valeur de la médiane, il convient de calculer les effectifs cumulés croissants.

Modalités	Effectifs	Effectifs cumulés
1	5	5
2	9	14
3	15	29
4	10	39
5	6	45
6	3	48
8	2	50

La médiane est comprise entre les rangs 25 et 26. Donc, $M_e = 3$.

Cas continu

Exemple. Reprenons notre exemple des exploitations agricoles.

Classes	Effectifs	Effectifs cumulés croissants
$]0;10]$	48	48
$]10;15]$	62	110
$]15;20]$	107	217
$]20;25]$	133	350
$]25;30]$	84	434
$]30;40]$	66	500

La superficie médiane est comprise entre celles des $250^{\text{ème}}$ et $251^{\text{ème}}$ individus. Ces deux exploitations appartiennent à la classe $]20;25]$ d'effectif 133. Comme ces derniers occupent respectivement les $33^{\text{ème}}$ $(= 250 - 217)$ et $34^{\text{ème}}$ $(= 251 - 217)$ rangs, la médiane sera donc égale à

$$M_e = 20 + \frac{33,5}{133} \cdot 5 \cong 21,26 \text{ ha.}$$

Ce calcul repose sur l'hypothèse d'une répartition uniforme des 133 exploitations à l'intérieur de leur classe $]20;25]$.

Graphiquement, la médiane est la première coordonnée du point d'intersection des polygones des effectifs cumulés croissants et cumulés décroissants. Dans notre exemple le résultat graphique concorde bien avec le calcul précédent.

Remarque. Dans le cas d'une population très nombreuse, il est loisible de convertir tous les effectifs en fréquences et d'attribuer à la médiane la valeur du caractère correspondant à la fréquence de 50% des effectifs cumulés. La valeur obtenue est une approximation d'autant meilleure que l'effectif est important. Dans notre cas, cette convention conduit aux calculs suivants.

Classes	Effectifs	Effectifs cumulés croissants	Fréquence en %	Fréquence cumulés croissants en %
$]0;10]$	48	48	9,6	9,6
$]10;15]$	62	110	12,4	22
$]15;20]$	107	217	21,4	43,4
$]20;25]$	133	350	26,6	70
$]25;30]$	84	434	16,8	86,8
$]30;40]$	66	500	13,2	100

Desquels, il ressort que la médiane est donnée par

$$M_e = 20 + \frac{50 - 43,4}{26,6} \cdot 5 \cong 21,24 \text{ ha.}$$

Approximation très proche de la valeur exacte 21,26 ha.

1.5.4 Comparaison entre les valeurs centrales

Nous pouvons maintenant faire quelques comparaisons sommaires entre les trois mesures de tendance centrale.

La moyenne arithmétique

1. Elle est sans doute la mesure de tendance centrale la plus familière.
2. Elle demande plus de calculs, mais s'exprime algébriquement d'une manière simple.
3. Elle tient compte de toutes les données et est donc influencée par les données extrêmes de la distribution. Dans le cas où une distribution est fortement dissymétrique, ceci devient un inconvénient qui justifie l'usage de la médiane au lieu de la moyenne.
4. Elle est peu influencée par le choix des classes, mais ne peut cependant pas être calculée s'il y a une classe ouverte (par exemple une classe du type "80 ans et plus").
5. Elle se prête facilement aux manipulations algébriques à cause de son expression mathématique simple.
6. Sa valeur est stable, c'est-à-dire qu'elle varie peu d'un échantillon à l'autre, du fait qu'elle tient compte de toutes les données. C'est sa plus grande qualité pour faire de l'inférence statistique.
7. Il s'agit de la mesure de tendance centrale la plus utilisée.

Le mode

1. Il peut y en avoir plusieurs dans une distribution. Le cas échéant, la présence de plusieurs modes peut être une indication que la population étudiée se compose de sous-groupes distincts. Selon l'étude désirée, cela pourrait inviter à scinder la population.
2. Il est facile à déterminer,
3. Il ne tient pas compte de toutes les données : Par contre, il n'est pas influencé par les données extrêmes de la distribution.
4. Il peut être grandement influencé par le choix des classes.
5. Il n'est vraiment significatif que si l'effectif correspondant est nettement supérieur aux autres.
6. Sa valeur n'est pas stable, c'est-à-dire qu'elle varie beaucoup d'un échantillon à l'autre choisi dans la même population.
7. Il s'agit de la mesure de tendance centrale la moins utilisée.

La médiane

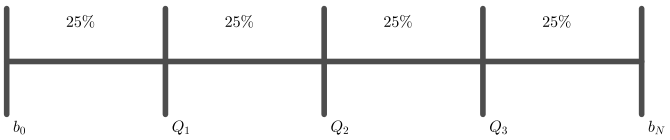
1. Elle provient d'une conception simple de la notion de centre.
2. Elle n'est pas très difficile à calculer, mais elle est plus difficile à exprimer algébriquement que la moyenne arithmétique.
3. Elle ne tient pas compte de toutes les données, mais uniquement de la position des données. Elle n'est donc pas influencée par les données extrêmes de la distribution.
4. Elle peut être influencée par le choix des classes.
5. Elle est surtout utilisée lorsque la distribution des effectifs est fortement dissymétrique ou lorsqu'elle contient des classes ouvertes.
6. Sa valeur est moins stable que celle de la moyenne. Ceci s'explique par le fait que cette valeur ne dépend que de quelques données parmi celles choisies dans un échantillon.
7. Elle est plus utilisée que le mode, mais moins que la moyenne arithmétique.

1.6 Quartiles et boîte à moustaches

1.6.1 Quartiles

Cas discret

Définition. On appelle *quartiles* les valeurs du caractère qui partagent l'effectif total de la série en 4 groupes d'effectifs égaux. On les note Q_1, Q_2 et Q_3. Un quart de l'effectif total possède donc un caractère inférieur à Q_1. Le deuxième quartile $Q_2 = M_e$ n'est autre que la médiane. Enfin, les trois quarts de la population se trouvent en dessous de la valeur définie par le troisième quartile Q_3.



Remarque. Il faut être attentif au fait qu'il existe de nombreuses méthodes différentes pour déterminer les quartiles, qui ne consuisent pas aux mêmes résultats. Dans ce cours, nous calculerons les quartiles selon la méthode établie par John Tukey en 1983.

On classe les N données dans l'ordre croissant et on les coupe en deux ensembles sur lesquels on calcule la médiane.

— Si N est impair, il y a une valeur centrale (la médiane), et on coupe les données en deux sous-ensembles en mettant la médiane dans chacun des deux ensembles. Le quartile Q_1 est alors la médiane du premier sous-ensemble ; le quartile Q_3 est la médiane du deuxième sous-ensemble.

— Si N est pair, il y a deux valeurs centrales (la médiane est la moyenne arithmétique de ces deux valeurs), et on coupe en deux sous-ensembles en mettant dans chaque sous-ensemble la valeur centrale correspondante. Le quartile Q_1 est alors la médiane du premier sous-ensemble ; le quartile Q_3 est la médiane du deuxième sous-ensemble.

Exemple.

N	Mesures	Sous-ensembles	Q_1	Q_3	Rang Q_1	Rang Q_3
4	1 3 4 5	$\{1;3\}$ et $\{4;5\}$	2	4,5	1,5	3,5
5	1 3 5 5 7	$\{1;3;5\}$ et $\{5;5;7\}$	3	5	2	4
6	1 3 4 6 7 9	$\{1;3;4\}$ et $\{6;7;9\}$	3	7	2	5
7	1 3 5 6 7 9 15	$\{1;3;5;6\}$ et $\{6;7;9;15\}$	4	8	2,5	5,5

Théorème.

— *Si N est pair, le rang du quartile Q_1 est donné par* $\dfrac{N+2}{4}$ *et celui de Q_3 par* $\dfrac{3N+2}{4}$.

— *Si N est impair, le rang du quartile Q_1 est donné par* $\dfrac{N+3}{4}$ *et celui de Q_3 par* $\dfrac{3N+1}{4}$.

Exemple. Reprenons notre exemple du nombre de personnes par ménage.

Modalités	Effectifs	Effectifs cumulés	Fréquences (en %)	Fréquences cumulées
1	5	5	10	10
2	9	14	18	28
3	15	29	30	58
4	10	39	20	78
5	6	45	12	90
6	3	48	6	96
8	2	50	4	100

On connaît déjà $Q_2 = M_e = 3$. Le quartile Q_1 est la valeur de l'observation de rang $\dfrac{50+2}{4} = 13$. Donc $Q_1 = 2$. Quant au quartile Q_3, il est égal à la valeur de l'observation de rang $\dfrac{3 \cdot 50 + 2}{4} = 38$. Donc $Q_3 = 4$.

Remarque. La plupart du temps, lorsqu'il s'agit par exemple de définir les quartiles, il n'est pas possible de trouver des rangs qui divisent la population en quatre classes d'effectif égal. Dans ce cas, on convertit les effectifs en fréquences et on définit les quartiles Q_1, M_e et Q_3 par les valeurs du caractère associées aux fréquences cumulées 25%, 50% et 75%.

Remarque. Il est possible de généraliser la notion de quartile à celle de *quantile d'ordre n*. Les autres quantiles les plus souvent utilisés sont :

— *Les déciles* D_1, D_2, ..., D_9 partagent l'effectif total en dix groupes égaux. Un dixième de la population a un caractère inférieur à D_1, et neuf dixièmes ont un caractère supérieur à D_1, . . ., et ainsi de suite. Le décile D_5 est égal à la médiane.

— *Les centiles* C_1, C_2,. . . , C_{99} partagent la population en 100 groupes d'effectifs égaux.

Cas continu

Exemple. Calculons les quartiles pour notre exemple des exploitations agricoles.

Classes	Effectifs	Effectifs cumulés croissants
$]0;10]$	48	48
$]10;15]$	62	110
$]15;20]$	107	217
$]20;25]$	133	350
$]25;30]$	84	434
$]30;40]$	66	500

On connait déjà la mediane $Q_2 = M_e \cong 21,26$.

Le premier quartile Q_1 est la valeur de la superficie de l'exploitation de rang $\frac{500+2}{4} = 125,5$. Comme elle se trouve dans la troisieme classe, d'effectif 107, on a

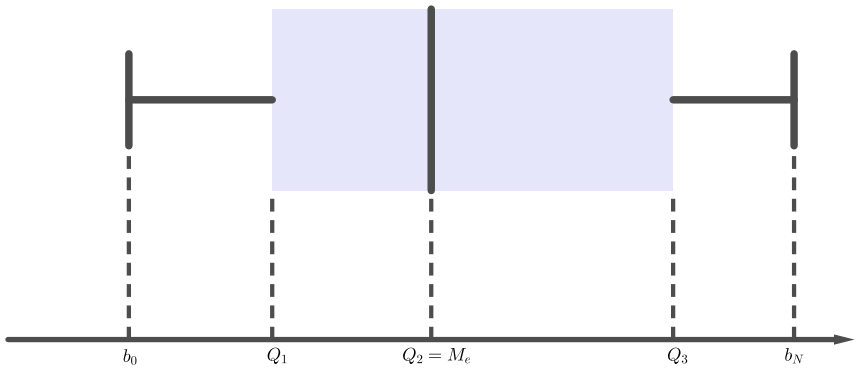
$$Q_1 = 15 + 5 \cdot \frac{125,5 - 110}{107} \cong 15,72 \text{ ha.}$$

Le troisième quartile Q_3 est défini par est la valeur de la superficie de l'exploitation de rang $\frac{3 \cdot 500 + 2}{4} = 375,5$. Celle-ci se trouvant en position 25,5 dans la classe $]25;30]$, d'effectif 84, on a

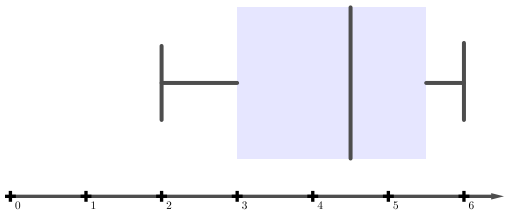
$$Q_3 = 25 + 5 \cdot \frac{25,5}{84} \cong 26,52 \text{ ha.}$$

1.6.2 Boîte à moustaches

Définition. Le *diagramme de Tukey*, plus communément appelé *boîte à moustaches* ou *box plot*, est une représentation codifiée des quantiles Q_1, M_e, Q_3 et des valeurs extrêmes b_0 et b_N de la distribution qui donne une information graphique concernant la symétrie de la distribution.



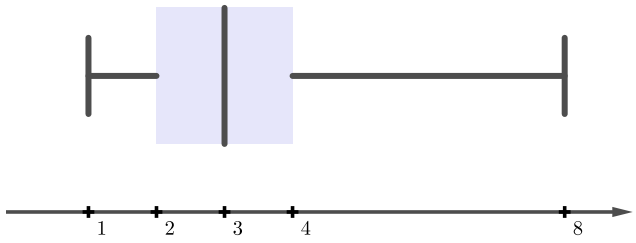
Exemple. Les notes d'une classe ont été représentées à l'aide de la boîte à moustache ci-dessous



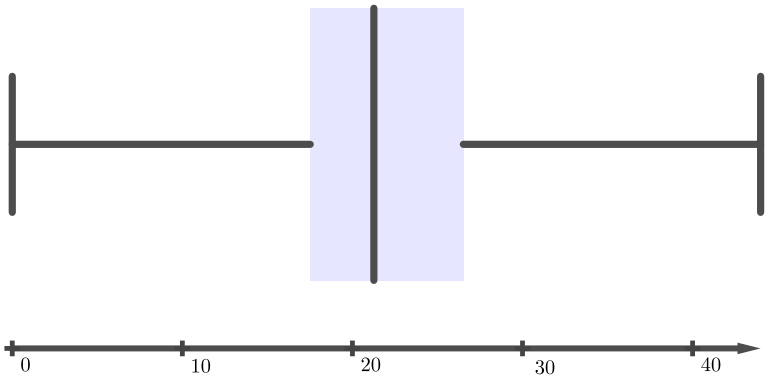
Cette boîte à moustaches fournit les informations suivantes :

- La moins bonne note est 2 et la meilleure 6.
- 25% des élèves ont fait une note égale ou inférieure à 3.
- La moitié des élèves ont fait 4,5 au moins (et au plus !).
- 75% des élèves ont fait une note inférieure ou égale à 5,5.
- 50% se tiennent dans un écart de 2,5.

Exemple. Dans notre exemple du nombre de personnes par ménage, la boîte à moustaches est donnée par



Exemple. Dans notre exemple des exploitations agricoles, la distribution étant presque symétrique, la boîte à moustaches s'étale symétriquement sur l'intervalle $[0; 40]$.



1.7 Mesures de dispersion

Si les valeurs centrales sont généralement nécessaires pour caractériser une série, elles ne sont toutefois pas suffisantes. Deux populations différentes peuvent avoir les mêmes valeurs centrales et différer notablement quant à la dispersion des individus autour de ces valeurs centrales.

Les deux ensembles $A = \{6; 8; 10; 12; 14\}$ et $B = \{2; 6; 10; 14; 18\}$ ont, par exemple, la même moyenne arithmétique et la même médiane, à savoir 10. Pourtant, les individus qui les composent ne sont pas répartis de la même manière autour de cette valeur centrale. L'ensemble B est moins régulier ou plus dispersé que l'ensemble A. On dit que A et B n'ont pas la même dispersion.

Pour comparer deux populations, on considère, outre leurs valeurs centrales, des mesures de leur dispersion. Les mesures classiques de dispersion sont les suivantes : *l'étendue, la variance et l'écart-type.*

1.7.1 Etendue de la série

C'est la valeur de dispersion la plus simple.

Définition. Aussi appelée *intervalle de variation*, *amplitude de la série* ou *intervalle maximal*, *l'étendue* E est la différence des valeurs extrêmes de la série.

Exemple. Dans notre exemple des exploitations agricoles, l'étendue vaut $E = 40 - 0 = 40$ ha.

Remarque. Simple à calculer, cette mesure de dispersion n'est pas très fiable puisqu'elle ne tient compte que de deux observations marginales et néglige toutes les autres.

1.7.2 Ecart interquartile

Définition. *L'écart interquartile* I_Q est défini par la différence des quartiles extrêmes. Autrement dit, on a

$$I_Q = Q_3 - Q_1.$$

Remarque. Cette mesure est plus fiable que l'étendue puisqu'elle exclut les 50% des valeurs marginales inférieures et supérieures.

Définition. *L'écart semi-interquartile* Q est défini par la moyenne arithmétique des écarts entre les quartiles et la médiane. Autrement dit, on a

$$Q = \frac{(Q_3 - M_e) + (M_e - Q_1)}{2} = \frac{Q_3 - Q_1}{2} = \frac{I_Q}{2}.$$

Exemple. Dans notre exemple du nombre de personnes par ménage, on a

$$I_Q = 4 - 2 = 2$$

et

$$Q = \frac{4 - 2}{2} = 1.$$

Exemple. Dans notre exemple des exploitations agricoles, on a donc

$$I_Q \cong 26,52 - 15,72 = 10,8 \text{ ha}$$

et

$$Q = \frac{26,52 - 15,72}{2} = 5,4 \text{ ha}.$$

Remarque. Il est également possible de définir *l'écart interdécile* I_D par

$$I_D = D_9 - D_1.$$

Cela définit un intervalle comprenant les 80% de la population.

1.7.3 Variance écart-type

Cas discret

Exemple. Deux classes de 20 élèves ont effectué un travail écrit de mathématiques, dont les résultats de ces travaux écrits sont présentés dans les tableaux ci-dessous.

Note x_i	Nombre d'élèves n_i
1	0
1,5	0
2	0
2,5	0
3	0
3,5	3
4	7
4,5	8
5	1
5,5	1
6	0

TABLE 1.1 – Notes de la première classe.

Note y_i	Nombre d'élèves n_i
1	0
1,5	1
2	0
2,5	2
3	4
3,5	0
4	0
4,5	3
5	6
5,5	2
6	2

TABLE 1.2 – Notes de la deuxième classe.

Au vu des résultats ci-dessus, il est naturel de se poser la question suivante.

Question : Laquelle de ces deux classes a été la plus performante lors de ce travail écrit ?

Un moyen de répondre à cette question consiste à calculer la moyenne arithmétique de chacune des deux classes :

$$\overline{x} = \frac{3,5 \cdot 3 + 4 \cdot 7 + 4,5 \cdot 8 + 5 \cdot 1 + 5,5 \cdot 1}{20} = 4,25$$

$$\overline{y} = \frac{1,5 \cdot 1 + 2,5 \cdot 2 + 3 \cdot 4 + 4,5 \cdot 3 + 5 \cdot 6 + 5,5 \cdot 2 + 6 \cdot 2}{20} = 4,25$$

Ces deux moyennes $\overline{x}$ et $\overline{y}$ sont égales alors que les résultats sont très différents !

La moyenne arithmétique ne donne pas d'informations sur la *dispersion des résultats* autour de la moyenne. Pour l'estimer, on essaie de quantifier la manière dont les notes sont réparties autour de la moyenne.

On obtient :

$x_i - \overline{x}$	n_i
$-3,25$	0
$-2,75$	0
$-2,25$	0
$-1,75$	0
$-1,25$	0
$-0,75$	3
$-0,25$	7
$0,25$	8
$0,75$	1
$1,25$	1
$1,75$	0

$y_i - \overline{y}$	n_i
$-3,25$	0
$-2,75$	1
$-2,25$	0
$-1,75$	2
$-1,25$	4
$-0,75$	0
$-0,25$	0
$0,25$	3
$0,75$	6
$1,25$	2
$1,75$	2

Le calcul de la moyenne de ces écarts est nul, car les écarts négatifs sont exactement compensés par les écarts positifs, ce qui n'amène aucun renseignement sur la dispersion. On choisit alors de calculer le carré des écarts à la moyenne.

On obtient alors les distributions suivantes :

$(x_i - \overline{x})^2$	n_i
10,5625	0
7,5625	0
5,0625	0
3,0625	0
1,5625	0
0,5625	3
0,0625	7
0,0625	8
0,5625	1
1,5625	1
3,0625	0

$(y_i - \overline{y})^2$	n_i
10,5625	0
7,5625	1
5,0625	0
3,0625	2
1,5625	4
0,5625	0
0,0625	0
0,0625	3
0,5625	6
1,5625	2
3,0625	2

Calculons alors la moyenne arithmétique de $(x_i - \overline{x})^2$ et $(y_i - \overline{y})^2$:

$$\begin{aligned}\overline{(x_i - \overline{x})^2} &= \frac{0,5625 \cdot 3 + 0,0625 \cdot 7 + 0,0625 \cdot 8 + 0,5625 \cdot 1 + 1,5625 \cdot 1}{20} \\ &= 0,2375\end{aligned}$$

$$\begin{aligned}\overline{(y_i - \overline{y})^2} &= \frac{7,5625 \cdot 1 + 3,0625 \cdot 2 + 1,5625 \cdot 4 + 0,0625 \cdot 3 + 0,5625 \cdot 6 + 1,5625 \cdot 2 + 3,0625 \cdot 2}{20} \\ &= 1,6375.\end{aligned}$$

Ces nombres ainsi trouvés sont une *mesure de la dispersion* des notes autour de la moyenne arithmétique. On voit ainsi que les notes de la première classe sont plus proches de la moyenne que celles de la deuxième classe.

Définition. On appelle *variance* V d'une série statistique la moyenne des carrés des écarts entre toutes les données et leur moyenne arithmétique. On a ainsi

$$V = \overline{(x_i - \overline{x})^2} = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + ... + (x_N - \overline{x})^2}{N}.$$

Définition. On appelle *ecart-type* σ, la racine carrée de la variance. Autrement dit, on a

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

Remarque. L'écart-type est une mesure de la dispersion plus significative que la variance. En effet, si les données x_i représentant une distance exprimée en mètres, V est en m^2 tandis que l'écart-type est exprimé en mètres.

Exemple. Soient les nombres -4, 3, 9, 11 et 17.

La moyenne arithmétique de ces nombres vaut

$$\overline{x} = \frac{-4 + 3 + 9 + 11 + 17}{5} = 7,2.$$

Du tableau suivant

	x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
	-4	$-11,2$	$125,44$
	3	$-4,2$	$17,64$
	9	$1,8$	$3,24$
	11	$3,8$	$14,44$
	17	$9,8$	$96,04$
Total	36	0	$256,8$

on en tire la variance

$$V = \frac{256,8}{5} = 51,36$$

et l'écart-type

$$\sigma = \sqrt{51,36} \cong 7,167.$$

Exemple. Calculons la variance et l'écart-type de notre exemple du nombre de personnes par ménage. A cet effet, il convient de dresser le tableau ci-dessous.

Modalités x_i	Effectifs n_i	Ecarts $x_i - \overline{x}$	Carrés des écarts $(x_i - \overline{x})^2$	Produits $n_i \cdot (x_i - \overline{x})^2$
1	5	$-2,44$	$5,9536$	$29,768$
2	9	$-1,44$	$2,0736$	$18,6624$
3	15	$-0,44$	$0,1936$	$2,904$
4	10	$0,56$	$0,3136$	$3,136$
5	6	$1,56$	$2,4336$	$14,6016$
6	3	$2,56$	$6,5536$	$19,6608$
8	2	$4,56$	$20,7936$	$41,5872$
Total	50			130,32

On en tire la variance

$$V = \frac{130,32}{50} = 2,6064$$

et donc

$$\sigma = \sqrt{V} \cong 1,614.$$

Cas continu

Comme pour le calcul de la moyenne arithmétique, on affecte à tous les individus d'une classe $]b_{i-1}; b_i]$ la valeur centrale $c = \dfrac{b_{i-1} + b_i}{2}$.

Exemple. Dans notre exemple des exploitations agricoles, cet écart se calcule à l'aide du tableau suivant.

Classes x_i	Centres c_i	Effectifs n_i	Carrés des écarts $(c_i - \overline{x})^2$	Produits $n_i \cdot (c_i - \overline{x})^2$
$]0; 10]$	5	48	256	12288
$]10; 15]$	12,5	62	72,25	4479,5
$]15; 20]$	17,5	107	12,25	1310,75
$]20; 25]$	22,5	133	2,25	299,25
$]25; 30]$	27,5	84	42,25	3549
$]30; 40]$	35	66	196	12936
Total		500	196	34862,5

La variance est donc égale à $V = \dfrac{34862,5}{500} = 69,725 \text{ hm}^2$ et l'écart-type est donné par $\sigma = \sqrt{69,725 \text{ hm}^2} \cong 8,35 \text{ hm}$.

Autre méthode de calcul

Le calcul de la variance (et donc de l'écart-type) n'est pas toujours commode. En particulier lorsque la moyenne est un nombre dont on ne donne qu'une approximation avec un développement décimal limité. Les calculs peuvent toutefois être simplifiés de la manière suivante.

Théorème. *La variance V peut être obtenue en calculant la différence entre la moyenne $\overline{x^2}$ des carrés des données x_i et le carré de leur moyenne $\overline{x}^2$. Ainsi, on a*

$$V = \overline{x^2} - \overline{x}^2.$$

Preuve. On a

$$\begin{aligned}
V &= \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + ... + (x_N - \overline{x})^2}{N} \\
&= \frac{(x_1^2 - 2x_1\overline{x} + \overline{x}^2) + (x_2^2 - 2x_2\overline{x} + \overline{x}^2) + ... + (x_N^2 - 2x_N\overline{x} + \overline{x}^2)}{N} \\
&= \frac{x_1^2 + x_2^2 + ... + x_N^2}{N} - \frac{2\overline{x}(x_1 + x_2 + ... + x_N)}{N} + \frac{\overline{x}^2 + \overline{x}^2 + ... + \overline{x}^2}{N} \\
&= \overline{x^2} - 2\overline{x} \cdot \overline{x} + \overline{x}^2 \\
&= \overline{x^2} - 2\overline{x}^2 + \overline{x}^2 \\
&= \overline{x^2} - \overline{x}^2.
\end{aligned}$$

□

Remarque. Cette seconde formulation sera préférée à la première chaque fois que les termes $x_i - \overline{x}$ sont plus compliqués que les termes x_i. Ce cas se présente fréquemment lorsque la moyenne n'est pas un nombre entier.

Exemple. Reprenons l'exemple des nombres -4, 3, 9, 11 et 17 de moyenne arithmétique $7,2$. Du tableau suivant

	x_i	x_i^2
	-4	16
	3	9
	9	81
	11	121
	17	289
Total	36	516

on en tire la variance $V = \dfrac{516}{5} - 7,2^2 = 51,36$ et l'écart-type $\sigma = \sqrt{51,36} \cong 7,167$. On retrouve bien les résultats obtenus plus haut.

Exemple. Reprenons l'exemple des exploitations agricoles.

Classes x_i	Centres c_i	Effectifs n_i	Carrés des centres c_i^2	Produits $n_i \cdot c_i^2$
$]0;10]$	5	48	25	1200
$]10;15]$	$12,5$	62	$156,25$	$9687,5$
$]15;20]$	$17,5$	107	$306,25$	$32768,75$
$]20;25]$	$22,5$	133	$506,25$	$67331,25$
$]25;30]$	$27,5$	84	$756,25$	63525
$]30;40]$	35	66	1225	80850
Total		500		$255362,5$

On en déduit que $\overline{x^2} = \dfrac{255362,5}{500} = 510,725$. Comme $\overline{x} = 21$, $\overline{x}^2 = 441$, il suit que

$$V = \overline{x^2} - \overline{x}^2 = 510,725 - 441 = 69,725 \text{ hm}^2$$

et l'écart-type est donc donné par $\sigma = \sqrt{69,725 \text{ hm}^2} \cong 8,35$ hm.

1.7.4 Ecart absolu moyen

On rappelle que la variance V d'une série statistique est la moyenne des carrés des écarts entre toutes les données et leur moyenne arithmétique. On a élevé ces écarts au carré pour faire disparaitre les nombre négatifs et que la moyenne des nombres ainsi obtenus ne sois pas nulle. Au lieu de prendre les carrés, il eût été possible de prendre les valeurs absolues. D'où la définition suivante :

Définition. On appelle *écart absolu moyen* e_a d'une série statistique la moyenne des valeurs absolues des écarts entre toutes les données et leur moyenne arithmétique. On a ainsi

$$e_a = \overline{|x_i - \overline{x}|} = \frac{|x_1 - \overline{x}| + |x_2 - \overline{x}| + ... + |x_N - \overline{x}|}{N}.$$

Exemple. Reprenons les nombres -4, 3, 9, 11 et 17 de moyenne arithmétique $\overline{x} = 7,2$.
Du tableau suivant

	x_i	$x_i - \overline{x}$	$\lvert x_i - \overline{x}\rvert$
	-4	$-11,2$	$11,2$
	3	$-4,2$	$4,2$
	9	$1,8$	$1,8$
	11	$3,8$	$3,8$
	17	$9,8$	$9,8$
Total	36	0	$30,8$

on en tire l'écart absolu moyen

$$e_a = \frac{30,8}{5} = 6,16.$$

1.7.5 Coefficient de variation

Pour caractériser une distribution, on utilise généralement une mesure de tendance centrale et une mesure de dispersion. Par exemple, on peut donner sa médiane et son intervalle semi-interquartille. Cependant, dans la grande majorité des cas, on décrit une distribution par sa moyenne et son écart-type. La moyenne indique autour de quelle valeur sont situées les données, alors que l'écart-type donne une idée de la dispersion. Cette idée de dispersion doit cependant être située dans son contecte. Si l'écart-type d'une distribution est égal à 10, peut-on dire que cette distribution est très dispersée ? Bien sûr, cela dépend de lôrdre de grandeur des données. En effet, si les données traitées sont de l'ordre de 2000 par exemple, cet écart-type est vraiment petit et les données sont sûrement très concentrées. Par contre, si les données sont de l'ordre de 12, par exemple, l'écart-type est grand et les données sont relativement dispersées. Il est donc utile de mesurer la *dispersion relative*.

Définition. Le *coefficient de variation* C d'une variable statistique est le rapport entre l'écart-type et la moyenne exprimé sous la forme d'un pourcentage :

$$C = \frac{\sigma}{\overline{x}}$$

Remarque. Si l'on souhaite porter un jugement sur la dispersion d'une série, la qualification suivante est généralement admise :

Coefficient de variation	Dispersion
0 à 10%	Faible
10 à 20%	Moyenne
Plus de 20%	Elevée

Exemple. Dans notre exemple des exploitations agricoles, ce coefficient vaut

$$C = \frac{\sigma}{\overline{x}} \cong \frac{8,35}{21} \cong 0,398 = 39,8\%.$$

Ainsi, la dispersion des données est élevée.

1.7.6 Comparaison des mesures de dispersion

L'étendue

1. Elle est très simple à calculer et à interpréter.
2. Elle ne tient pas compte de toutes les données et n'implique que les valeurs extrêmes.
3. Elle est utilisée pour donner une idée sommaire et rapide de la dispersion et pour déterminer les largeurs de classes lorsqu'on fait un regroupement en classes.
4. Sa valeur n'est pas stable, c'est-à-dire qu'elle varie beaucoup d'un échantillon à l'autre choisi dans une même population.
5. Elle est très peu utilisée.

L'écart semi-interquartille

1. Il est simple à calculer et à interpréter.
2. Il ne tient pas compte de toutes les données et n'est donc pas influencé par les données extrêmes.
3. Il est utilisé lorsque la distribution des effectifs est fortement dissymétrique. Dans ce cas, on utilise la médiane comme mesure de tendance centrale.
4. Sa valeur est moins stable que celle de la variance ou de l'écart-type.
5. Il est peu utilisé en général.

L'écart interquartille

1. Il présente les mêmes caractéristiques que l'écart semi-interquartille.

L'écart-type

1. Son calcul est plus long et son interprétation est moins immédiate.
2. Il tient compte de toutes les données.
3. Il se prête assez bien aux manipulations algébriques. On le retrouve ainsi dans plusieurs calculs en statistiques inférentielles.
4. Sa valeur est stable d'un échantillon à l'autre.
5. Il est, avec la variance, la mesure de dispersion la plus utilisée.

La variance

1. La variance a les mêmes caractéristiques que l'écart-type.
2. La présence de carrés accorde plus de poids aux grands écarts. Elle est ainsi fortement influencée par les données extrêmes.

L'écart absolu moyen

1. Il n'est pas difficile à calculer et son interprétation est naturelle.
2. Il tient compte de toutes les données et il accorde le même poids à chacun des écarts. Il est donc moins influencé que la variance par les données extrêmes.
3. Il se prête mal aux manipulations algébriques à cause des valeurs absolues. De ce fait, il est difficile d'effectuer certains calculs à l'aide de l'écart absolu moyen. Il s'agit d'un inconvénient sérieux.
4. Sa valeur est stable d'un échantillon à l'autre.
5. Son utilisation est très limitée.

Remarque. Le choix de la mesure de tendance centrale implique le choix de la mesure de dispersion :

mode	$\leftrightarrow$	étendue
médiane	$\leftrightarrow$	écart semi-interquartile
moyenne	$\leftrightarrow$	écart-type ou écart absolu moyen

Chapitre 2

Ajustement et régression

2.1 Introduction

Il arrive fréquemment que l'étude statistique d'une population porte simultanément sur deux ou plusieurs variables statistiques quantitatives. La question qui peut se poser alors est de rechercher et de déterminer une éventuelle liaison entre ces variables statistiques. Établir cette liaison, c'est effectuer un ajustement.

2.2 Les différents types de relations

Selon les circonstances, les trois types suivants de relations peuvent lier les caractères étudiés :

— la *relation totale* ou *relation fonctionnelle*;

— l'*absence de liaison*;

— la *liaison statistique*.

2.2.1 La liaison totale ou relation fonctionnelle

Définition. On dit qu'il y a *relation totale* ou *fonctionnelle* entre deux variables statistiques, lorsque la connaissance des valeurs prises par l'un d'eux permet de déterminer les valeurs prises par l'autre.

Il existe ainsi une liaison rigide entre les deux variables statistiques.

Exemple. Un examen de statistique porte sur 10 points. Etudions alors la relation entre le nombre de points obtenus X et la note Y dans un échantillon de 10 individus.

Points X	3	4	4,5	5	7	7,5	8	9	9,5	10
Note Y	2,5	3	3,25	3,5	4,5	4,75	5	5,5	5,75	6

Représentons le nuage de points correspondants. On observe que tous les points sont alignés.

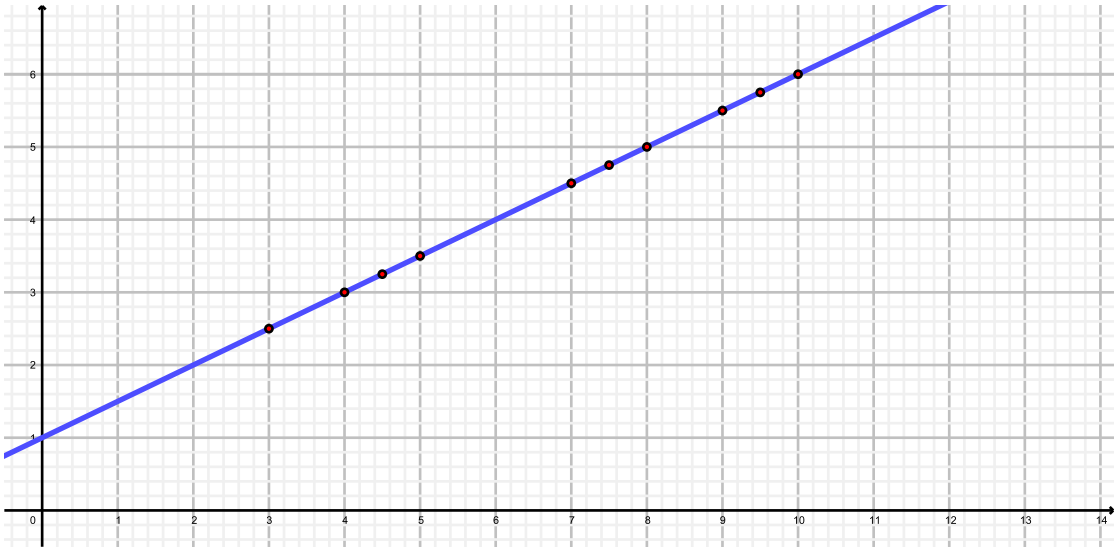


FIGURE 2.1 – Liaison totale.

Remarque. Dans le cas d'une relation fonctionnelle, les points dont les coordonnées $(x; y)$ sont définies par les valeurs prises par les variables statistiques X, Y sur les individus de la population sont tous exactement situés sur la même droite.

2.2.2 L'absence de liaison

Supposons qu'il n'y ait aucune relation entre les deux variables statistiques de la population étudiée.

Exemple. Admettons que X désigne la note obtenue à un examen de statistique par des étudiants et Y, le poids de ces derniers (en kg) dans un échantillon de 20 individus :

Note X	2,5	2,5	3	3	3	4	4	4	4	4
Poids Y	67	82	63	78	94	56	64	76	87	91

Note X	4,5	4,5	4,5	5	5	5	5	5,5	5,5	6
Poids Y	51	79	98	68	75	81	88	72	86	77

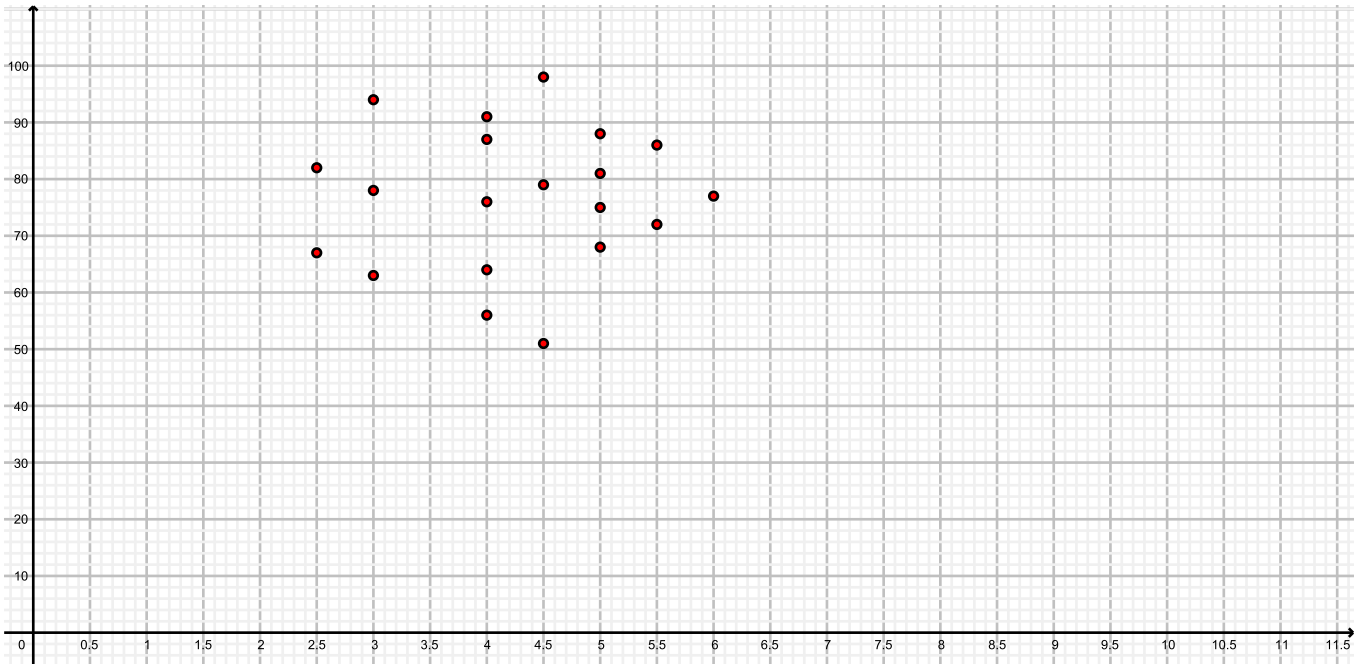


FIGURE 2.2 – Absence de liaison.

Dans ce cas, si on reporte dans le plan l'ensemble des points dont les coordonnées $(x; y)$ sont les valeurs des deux variables statistiques $(X; Y)$ pour chaque individu, alors ces derniers forment un nuage de points répartis de manière quelconque.

Remarque. Il est alors vain de chercher un ajustement entre deux caractères n'ayant aucun lien entre eux.

2.2.3 La liaison statistique

Définition. On dit qu'il existe une *liaison statistique* entre deux variables statistiques lorsque les variations de l'une des variables statistiques expliquent en partie les variations de l'autre.

Il existe donc une certaine dépendance entre ces deux caractères. La liaison statistique constitue ainsi une situation intermédiare entre l'absence de liaison et la relation fonctionnelle.

Exemple. Étudions la relation entre le poids (en kg) X et la taille (en cm) Y dans un échantillon de 20 individus :

Taille X	155	158	158	159	163	163	165	168	170	172
Poids Y	67,1	60,7	54,9	58,8	64,7	60,4	63	62,5	71,5	70,8

Taille X	173	175	176	178	178	180	182	186	189	196
Poids Y	63,1	74,8	71,1	73,1	63,5	69,4	70	82	76,5	84,6

On représente alors le *nuage de points* :

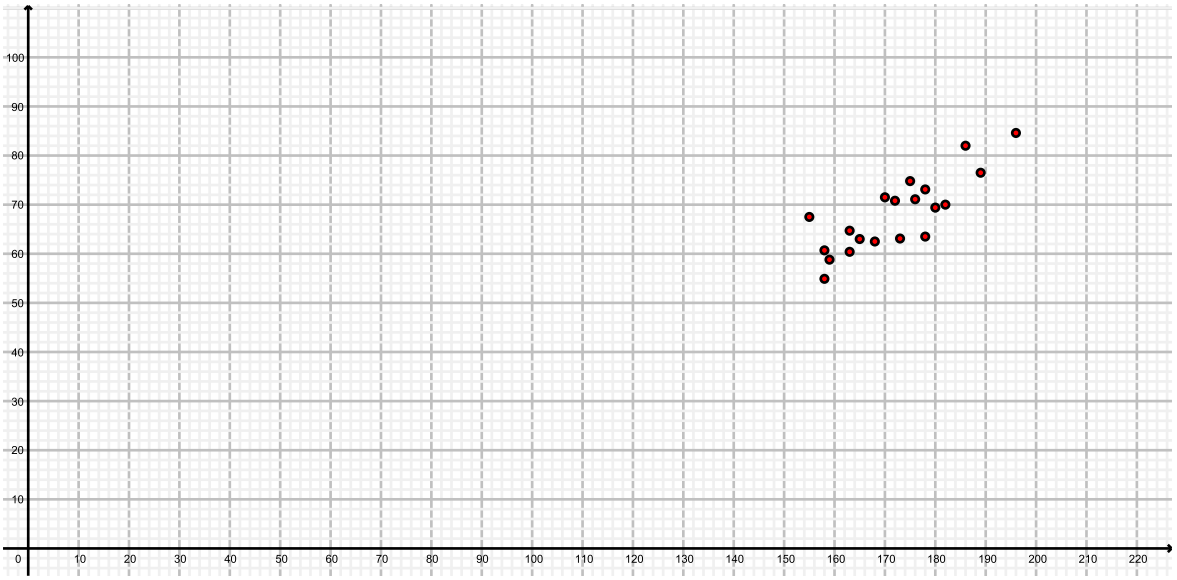


FIGURE 2.3 – Nuage de points.

Lorsqu'on représente graphiquement les points déterminés par les valeurs des deux variables statistiques, on obtient un nuage de points qui permet de suggérer une relation. Lorsque celle-ci a la forme d'une relation affine $y = px + h$, on procède à un *ajustement linéaire* (ou on effectue une *régression linéaire*) en cherchant la droite qui donne la meilleure représentation du nuage.

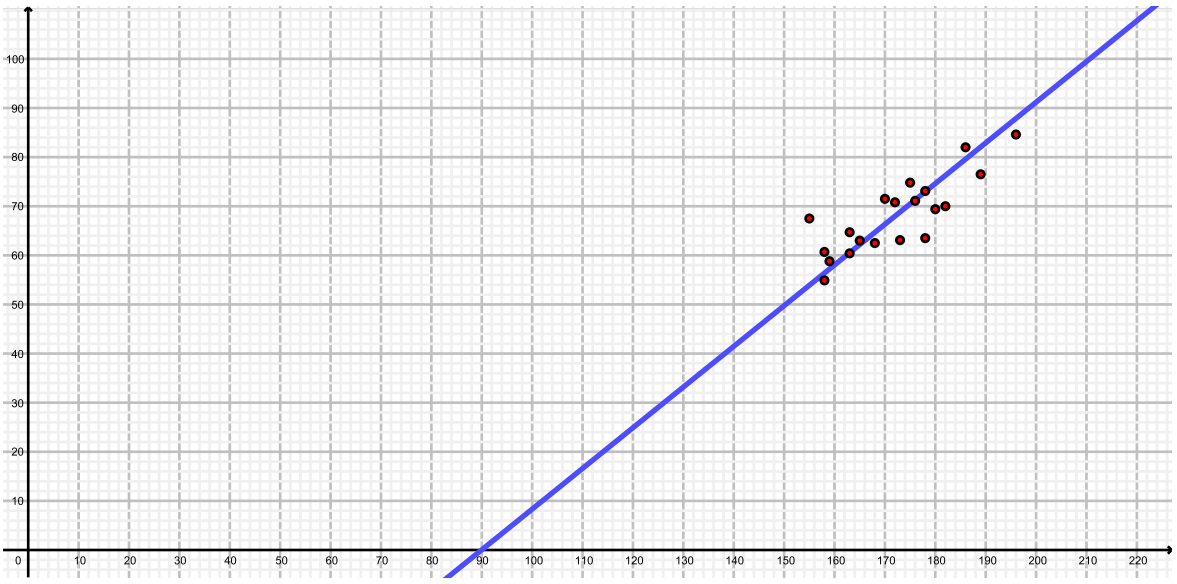


FIGURE 2.4 – Liaison statistique.

Remarque. Notons que le modèle de relation obtenue entre X et Y n'est valable que dans le voisinage des points donnés. En dehors de ce domaine, la relation n'est pas fiable et n'aurait même pas de sens.

2.3 Covariance

Définition. On appelle *covariance* de deux variables statistiques X, Y définis sur une population de taille N, le nombre

$$\mathrm{Cov}(X,Y) = \frac{\sum_{i=1}^{N}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{N} = \overline{(x_i - \overline{x}) \cdot (y_i - \overline{y})}.$$

Remarque.

1. Pour alléger la notation, on notera désormais $\sum$ au lieu de $\sum_{i=1}^{N}$.
2. Il découle de cette définition que la covariance de X et X n'est autre que la variance de X. Autrement dit,

$$\mathrm{Cov}(X,X) = \mathrm{Var}(X).$$

3. Intuitivement, la covariance est une mesure de la variation simultanée de deux variables statistiques. Elle indique le degré de variation conjointe de deux variables statistiques par rapport à leurs moyennes, et dans quelle mesure, les valeurs d'une variable statistique augmentent ou diminuent avec les valeurs croissantes de l'autre. Elle devient plus positive pour chaque couple de valeurs qui diffèrent dans le même sens, et plus négative pour chaque couple de valeurs qui diffèrent de leur moyenne dans le sens opposé.

Théorème. *La covariance peut être obtenue en calculant*

$$Cov(X;Y) = \overline{xy} - \overline{x} \cdot \overline{y}.$$

Preuve.

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \frac{1}{N} \cdot \sum (x_i - \overline{x})(y_i - \overline{y}) \\
&= \frac{1}{N} \cdot \sum (x_i \cdot y_i - x_i \cdot \overline{y} - \overline{x} \cdot y_i + \overline{x} \cdot \overline{y}) \\
&= \frac{1}{N} \cdot \sum (x_i \cdot y_i) - \frac{1}{N} \cdot \sum (x_i \cdot \overline{y}) - \frac{1}{N} \cdot \sum (\overline{x} \cdot y_i) + \frac{1}{N} \cdot \sum (\overline{x} \cdot \overline{y}) \\
&= \frac{1}{N} \cdot \sum (x_i \cdot y_i) - \overline{y} \cdot \frac{1}{N} \cdot \sum x_i - \overline{x} \cdot \frac{1}{N} \cdot \sum y_i + \frac{1}{N} \cdot N \cdot \overline{x} \cdot \overline{y} \\
&= \frac{1}{N} \cdot \sum (x_i \cdot y_i) - \overline{y} \cdot \overline{x} - \overline{x} \cdot \overline{y} + \overline{x} \cdot \overline{y} \\
&= \frac{1}{N} \cdot \sum (x_i \cdot y_i) - \overline{x} \cdot \overline{y} \\
&= \overline{xy} - \overline{x} \cdot \overline{y}.
\end{aligned}$$

□

Exemple. Reprenons les cinq premières valeurs de notre exemple précédent relatif à la relation entre la taille X et le poids Y.

Pour calculer la covariance de X et Y, il convient de dresser le tableau ci-dessous :

X	Y	$X \cdot Y$
155	67,1	10400,5
158	60,7	9590,6
158	54,9	8674,2
159	58,8	9349,2
163	64,7	10546,1
793	306,2	48560,6

On en tire que

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \overline{xy} - \overline{x} \cdot \overline{y} \\ &= \frac{48560,6}{5} - \frac{793}{5} \cdot \frac{306,2}{5} \\ &= -0,544. \end{aligned}$$

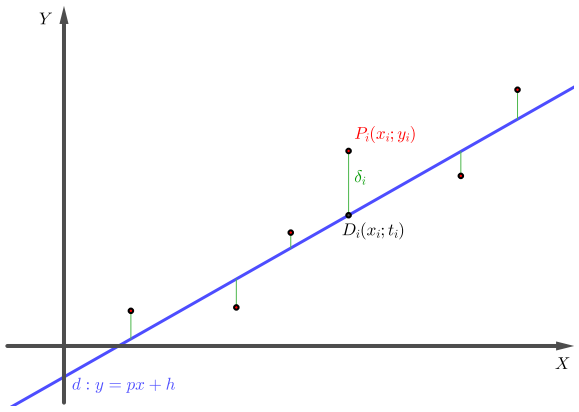
2.4 Ajustement linéaire

2.4.1 Position du problème

Un nuage de N points $P_i(x_i; y_i)$ du plan $\mathbb{R}^2$ devrait, en théorie, se situer exactement sur une droite. Les coordonnées y_i étant entachées d'erreur (elles représentent par exemple les mesures d'une expérience), les N points ne sont pas alignés. On cherche la droite d d'équation cartésienne $y = px+h$ qui minimise la somme des carrés des distances verticales entre les points mesurés $P_i(x_i; y_i)$ et les points théoriques $D_i(x_i; px_i + h)$ (cf. dessin ci-après), c'est-à-dire, la somme

$$\sum \delta_i^2 = \sum (y_i - t_i)^2 = \sum (t_i - y_i)^2 = \sum (px_i + h - y_i)^2.$$

Cette droite est appelée *droite des moindres carrés.*



2.4.2 Equation de la droite des moindres carrés

Théorème. *La droite des moindres carrés a pour équation* $y = px + h$ *avec*

$$p = \frac{Cov(X,Y)}{Var(X)} = \frac{\sum(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$$

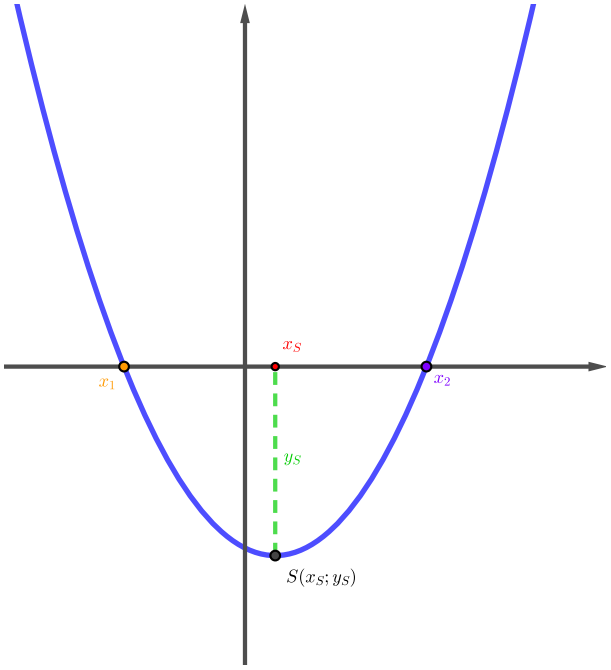
et

$$h = \overline{y} - p \cdot \overline{x}.$$

Preuve. On rappelle que la parabole d'équation $y = ax^2 + bx + c$ a pour coordonnées

$$S\left(-\frac{b}{2a}; f\left(-\frac{b}{2a}\right)\right).$$

En effet, le sommet de la parabole est situé au milieu des éventuels zéros x_1 et x_2.



Son abscisse x_S est donc donnée par

$$\begin{aligned} x_S &= \frac{x_1 + x_2}{2} = \frac{\frac{-b - \sqrt{\Delta}}{2a} + \frac{-b + \sqrt{\Delta}}{2a}}{2} \\ &= -\frac{2b}{2a} \cdot \frac{1}{2} = -\frac{b}{2a}. \end{aligned}$$

L'abscisse x_S étant alors connue, on en tire l'ordonnée y_S par

$$y_S = f\left(-\frac{b}{2a}\right).$$

La droite optimale d ayant pour équation $y = px + h$, toute autre droite conduit à une somme $\sum \delta_i^2$ supérieure à la valeur correspondante à d. Prouvons maintenant que la droite d contient le centre de gravité $G(\overline{x};\overline{y})$. Il se trouve que la droite de pente p a une ordonnée à l'origine h qui rend minimale la somme

$$\begin{aligned} S(h) &= \sum \delta_i^2 \\ &= \sum (px_i + h - y_i)^2 \\ &= \sum (h + px_i - y_i)^2 \\ &= \sum [h^2 + 2(px_i - y_i)h + (px_i - y_i)^2] \\ &= \sum h^2 + \sum [2(px_i - y_i)h] + \sum (px_i - y_i)^2 \\ &= N \cdot h^2 + 2\sum (px_i - y_i)h + \sum (px_i - y_i)^2. \end{aligned}$$

Cette expression quadratique avec $a = N$, $b = 2\sum (px_i - y_i)$ et $c = \sum (px_i - y_i)^2$ est minimale pour

$$\begin{aligned} h &= -\frac{b}{2a} & &= -\frac{2\sum (px_i - y_i)}{2N} \\ &= \frac{\sum (y_i - px_i)}{N} & &= \frac{\sum y_i}{N} - \frac{\sum px_i}{N} \\ &= \frac{1}{N}\sum y_i - p \cdot \frac{1}{N}\sum x_i & &= \overline{y} - p \cdot \overline{x}. \end{aligned}$$

Ainsi, l'équation de la droite s'écrit

$$y = p \cdot x + \overline{y} - p \cdot \overline{x}.$$

Cela prouve que la droite passe par le barycentre $G(\overline{x};\overline{y})$.

La droite optimale ayant pour équation $y = p \cdot x + \overline{y} - p \cdot \overline{x}$, sa pente p rend forcément minimale l'expression quadratique

$$\begin{aligned} Q(p) &= \sum \delta_i^2 \\ &= \sum (px_i + h - y_i)^2 \\ &= \sum (px_i + \overline{y} - p\overline{x} - y_i)^2 \\ &= \sum [(x_i - \overline{x})p - (y_i - \overline{y})]^2 \\ &= \sum [(x_i - \overline{x})^2 p^2 - 2(x_i - \overline{x})p(y_i - \overline{y}) + (y_i - \overline{y})^2] \\ &= \sum [(x_i - \overline{x})^2 p^2] - 2\sum [(x_i - \overline{x})p(y_i - \overline{y})] + \sum [(y_i - \overline{y})^2] \\ &= \sum (x_i - \overline{x})^2 p^2 + (-2)\sum (x_i - \overline{x})(y_i - \overline{y})p + \sum [(y_i - \overline{y})^2]. \end{aligned}$$

Cette expression quadratique avec $a = \sum(x_i - \overline{x})^2$ et $b = (-2)\sum(x_i - \overline{x})(y_i - \overline{y})$ est minimale pour

$$p = \frac{-b}{2a} = \frac{2\sum(x_i - \overline{x})(y_i - \overline{y})}{2\sum(x_i - \overline{x})^2} = \frac{\frac{1}{N} \cdot \sum(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{N} \cdot \sum(x_i - \overline{x})^2} = \frac{\mathrm{Cov}(X;Y)}{\mathrm{Var}(X)}.$$

□

Exemple. En 1969, dans cinq villes des États-Unis, des chercheurs ont étudié le taux d'absentéisme (féminin) en fonction de la pollution de l'air par les poussières de soufre (en $\mu g/m^3$). Le tableau suivant fait état des résultats (seules les absences de plus de sept jours ont été comptabilisées).

Villes	$\mu g/m^3$	Nombres d'absences pour mille employées
Cincinnatti	7	19
Indianapolis	13	44
Woodbridge	14	53
Camden	17	61
Harrison	20	88

Déterminons l'équation de la droite des moindres carrés $d : y = px + h$, qui ajuste les points $P_i(x_i; y_i)$ (avec X = taux de pollution, Y = nombre d'absences). On construit le tableau des données

X	Y	X^2	XY
7	19	49	133
13	44	169	572
14	53	196	742
17	61	289	1037
20	88	400	1760
71	265	1103	4244

On en tire les moyennes $\overline{x} = \frac{71}{5} = 14,2$, $\overline{y} = 53$, $\overline{x^2} = \frac{1103}{5} = 220,6$ et $\overline{xy} = \frac{4244}{5} = 848,8$. Il s'ensuit que

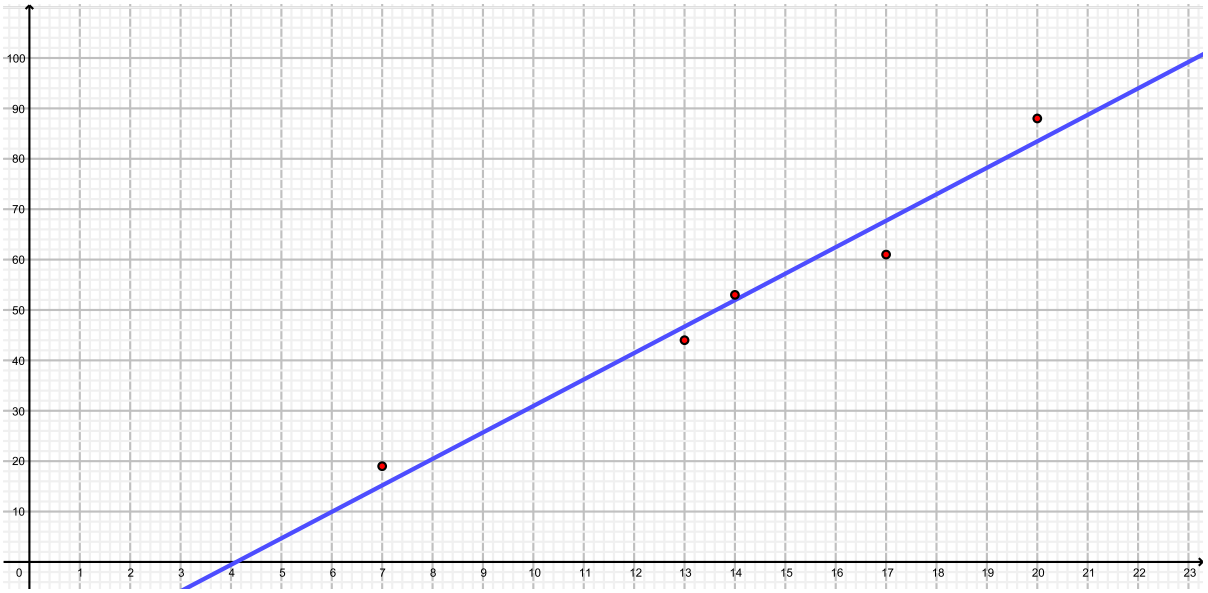
$$p = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\frac{4244}{5} - \frac{71}{5} \cdot 53}{\frac{1103}{5} - \left(\frac{71}{5}\right)^2} = \frac{2405}{474} \cong 5,074$$

et

$$h = \overline{y} - p\overline{x} = 53 - \frac{2405}{474} \cdot \frac{71}{5} = -\frac{9029}{474} \cong -19,049.$$

La droite des moindres carrés (représentée ci-dessous) admet donc l'équation

$$d : y = \frac{2405}{474}x - \frac{9029}{474}.$$



2.4.3 Equations normales

Théorème. *La pente p et l'ordonnée à l'origine h de la droite des moindres carrés satisfont aux équations, dites équations normales, du système linéaire suivant.*

$$\left\{ \begin{array}{rcl} p \cdot \sum x_i + N \cdot h & = & \sum y_i \\ p \cdot \sum x_i^2 + h \cdot \sum x_i & = & \sum x_i y_i \end{array} \right. .$$

Preuve. On peut retrouver ces relations par une démarche purement algébrique [1]

Nous avons établi que

$$h = \overline{y} - p \cdot \overline{x}.$$

Multipliée par N, cette relation s'écrit aussi

$$\begin{array}{rcl} N \cdot h & = & N \cdot \overline{y} - p \cdot N \cdot \overline{x} \\ N \cdot h & = & N \cdot \frac{1}{N} \sum y_i - p \cdot N \cdot \frac{1}{N} \sum x_i \\ N \cdot h & = & \sum y_i - p \cdot \sum x_i \end{array}$$

ou encore

$$p \cdot \sum x_i + N \cdot h = \sum y_i$$

1. On obtient d'ordinaire la pente p et l'ordonnée h comme étant le couple $(p; h)$ correspondant au minimum de la fonction de deux variables

$$E(p; h) = \sum \delta_i^2 = \sum (px_i + h - y_i)^2,$$

c'est-à-dire le point pour lequel les dérivées partielles $\frac{\partial E}{\partial h}$ et $\frac{\partial E}{\partial p}$ s'annulent.

qui n'est autre que la première équation normale.
Quant à la relation $p \cdot \mathrm{Var}(X) = \mathrm{Cov}(X,Y)$, multipliée par N, elle s'écrit aussi

$$\begin{aligned}
p \cdot \sum (x_i - \overline{x})^2 &= \sum (x_i - \overline{x})(y_i - \overline{y}) \\
p \cdot \sum (x_i^2 - 2x_i\overline{x} + \overline{x}^2) &= \sum (x_i y_i - x_i\overline{y} - \overline{x}y_i + \overline{x} \cdot \overline{y}) \\
p \cdot \sum x_i^2 - 2p\overline{x} \cdot \sum x_i + Np\overline{x}^2 &= \sum x_i y_i - \overline{y} \cdot \sum x_i - \overline{x} \cdot \sum y_i + N\overline{x} \cdot \overline{y} \\
p \cdot N \cdot \tfrac{1}{N} \sum x_i^2 - 2p\overline{x} \cdot \sum x_i + Np\overline{x}^2 &= \sum x_i y_i - \overline{y} \cdot \sum x_i - \overline{x} \cdot \sum y_i + N\overline{x} \cdot \overline{y} \\
p \cdot \sum x_i^2 - 2pN\overline{x}^2 + Np\overline{x}^2 &= \sum x_i y_i - \overline{y}N\overline{x} - \overline{x}N\overline{y} + N\overline{x} \cdot \overline{y} \\
p \cdot \sum x_i^2 - Np\overline{x}^2 &= \sum x_i y_i - N\overline{x} \cdot \overline{y} \\
p \cdot \sum x_i^2 + N\overline{x} \cdot \overline{y} - Np\overline{x}^2 &= \sum x_i y_i \\
p \cdot \sum x_i^2 + N\overline{x}(\overline{y} - p\overline{x}) &= \sum x_i y_i \\
p \cdot \sum x_i^2 + N \cdot \tfrac{1}{N} \sum x_i (\overline{y} - p\overline{x}) &= \sum x_i y_i \\
p \cdot \sum x_i^2 + \sum x_i (\overline{y} - p\overline{x}) &= \sum x_i y_i \\
p \cdot \sum x_i^2 + (\overline{y} - p\overline{x}) \sum x_i &= \sum x_i y_i
\end{aligned}$$

ou encore

$$p \sum x_i^2 + h \cdot \sum x_i = \sum x_i y_i$$

qui n'est rien d'autre que la seconde équation normale. □

Exemple. Soient les 4 points de coordonnées $(-4;-3)$, $(-1;0)$, $(2;4)$, $(5;7)$. Du tableau suivant

x_i	y_i	x_i^2	$x_i \cdot y_i$
-4	-3	16	12
-1	0	1	0
2	4	4	8
5	7	25	35
2	8	46	55

on déduit le système des équations normales

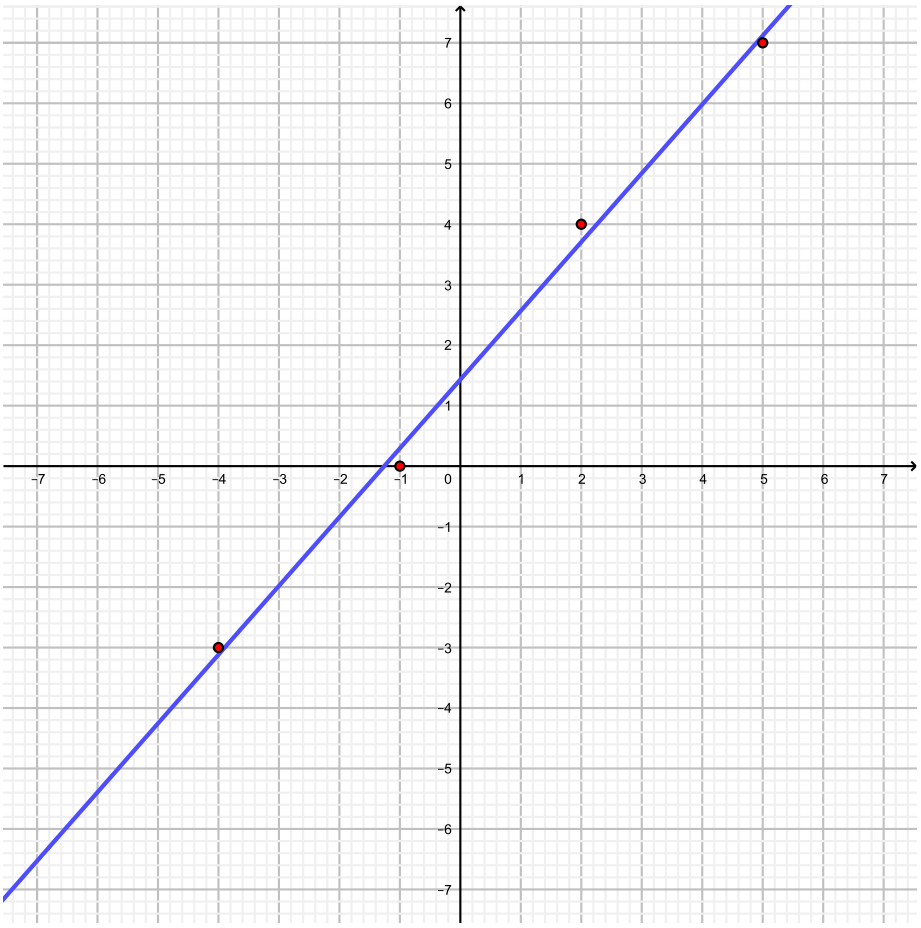
$$\begin{cases} 2p + 4h &= 8 \\ 46p + 2h &= 55 \end{cases}.$$

De la première équation, on tire $2p = 8 - 4h$, c'est-à-dire $p = 4 - 2h$.
On injecte dans la seconde $46(4 - 2h) + 2h = 55$ et on trouve

$$h = \frac{43}{30} \cong 1,433 \text{ et } p = \frac{17}{15} \cong 1,133$$

puis l'équation

$$y = \frac{17}{15}x + \frac{43}{30}.$$



2.5 Corrélation

Suivant leurs dispositions dans le plan, les points forment un nuage plus ou moins fortement dispersé autour de la droite de régression déterminée par la méthode des moindres carrés. Si les points sont globalement très voisins de la droite, on dira que la *corrélation linéaire* entre X et Y est forte. Dans le cas contraire, c'est-à-dire si le nuage de points est très dispersé, la corrélation sera dite faible.

2.5.1 Droite de régression de x en y

Jusqu'ici, on a toujours cherché une droite d'ajustement $d : y = px + h$ exprimant y en fonction de x. Géométriquement, cette méthode revient à minimiser la somme des carrés des écarts verticaux entre les points mesurés et les points de mêmes abscisses situés sur la droite. Cette droite est appelée *droite de régression de y en x*. Rien n'empêche d'inverser les rôles respectifs de x et y et de chercher la droite d' qui va minimiser la somme des carrés des écarts horizontaux séparant les points mesurés de ceux se trouvant sur la droite. Une telle droite est appelée *droite de régression de x en y*. On détermine son équation

$$d' : x = p'y + h'$$

de la même manière que celle appliquée pour la première droite mais en échangeant les rôles de x et y.

2.5.2 Le coefficient de la corrélation

Si les points sont parfaitement alignés, alors les droites d et d' sont confondues. Dans ce cas, le produit de leurs pentes p et $p' = \dfrac{1}{p}$ est donné par $p \cdot p' = 1$.

En effet, si d a pour équation $d : y = px + h$, il suffit d'isoler x de cette dernière équation pour trouver l'équation de d' :

$$\begin{array}{rcl} y & = & px + h \\ y - h & = & px \\ \frac{y}{p} - \frac{h}{p} & = & x \\ x & = & \frac{1}{p}y - \frac{h}{p}. \end{array}$$

d' a donc bien $p' = \frac{1}{p}$ comme pente.

Il s'ensuit que $p \cdot p' = p \cdot \frac{1}{p} = 1$.

En général, les points mesurés ne sont toutefois pas alignés ; la *corrélation* sera alors d'autant plus forte que les deux droites de régression sont proches, c'est-à-dire que le produit de leurs pentes sera proche de 1.

Définition. Pour mesurer la corrélation, on introduit le *coefficient de corrélation linéaire* r défini comme étant la moyenne géométrique des pentes p et p'

$$r = \sqrt{p \cdot p'}.$$

Théorème. *Le coefficient de corrélation peut s'obtenir à l'aide de l'expression suivante*

$$r = \frac{Cov(X,Y)}{\sigma(X) \cdot \sigma(Y)}$$

dans laquelle $\sigma(X) = \sqrt{Var(X)}$ *et* $\sigma(Y) = \sqrt{Var(Y)}$ *désignent les écarts-types des variables* X *et* Y.

Preuve. Comme $p = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$ et $p' = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}$, on a donc

$$r^2 = \frac{\mathrm{Cov}(X,Y)^2}{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}.$$

En prenant la racine carrée des deux côtés, on en déduit l'expression suivante pour le coefficient de corrélation

$$r = \frac{\mathrm{Cov}(X,Y)}{\sigma(X) \cdot \sigma(Y)}.$$

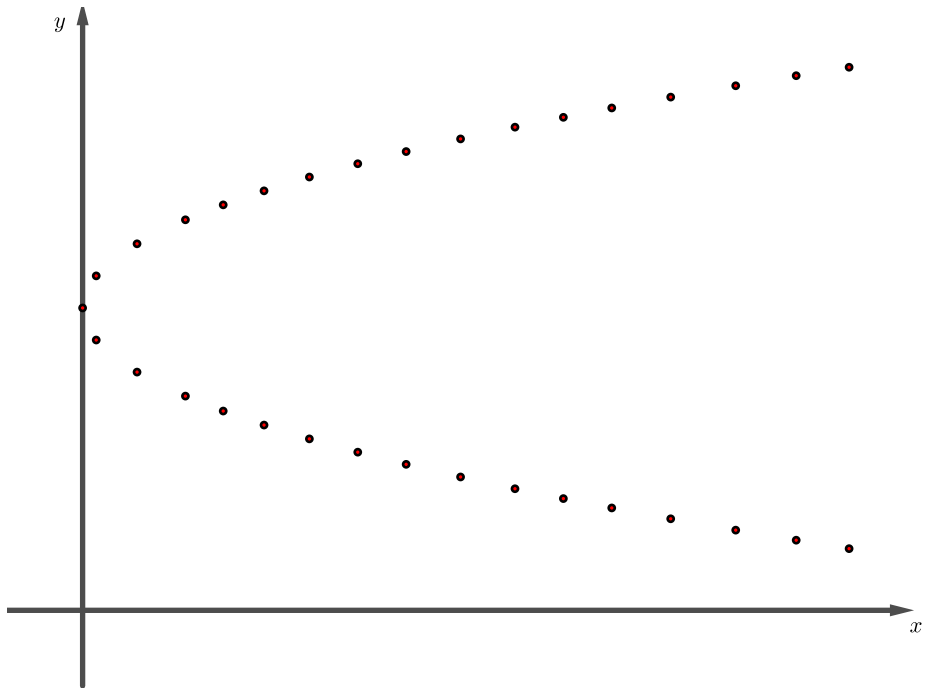
□

Remarque.

1. S'agissant de la corrélation, l'interprétation suivante est généralement admise :

Corrélation	Interprétation
1	Corrélation positive parfaite
0,8 à 1	Bonne à très bonne corrélation positive
0,2 à 0,8	Corrélation positive moyenne
0 à 0,2	Corrélation positive très faible
0	Corrélation nulle
$-0,2$ à 0	Corrélation négative très faible
$-0,8$ à $-0,2$	Corrélation négative moyenne
-1 à $-0,8$	Bonne à très bonne corrélation négative
-1	Corrélation négative parfaite

2. Toutefois, $r = 0$ ou $r \cong 0$ n'implique pas nécessairement l'indépendance des variables X et Y. Il indique seulement que les droites de régressions sont presque parallèles aux axes de coordonnées. Dans l'exemple ci-dessous, il existe une relation fonctionnelle quadratique entre X et Y , alors que la corrélation linéaire est presque nulle. Cet exemple montre bien que le coefficient de corrélation linéaire ne doit être utilisé pour caractériser l'intensité de la corrélation que dans le cas où celle-ci est approximativement linéaire.



Exemple. Reprenons l'exemple, traité précédemment, du taux d'absentéisme dans des villes américaines.

X	Y	X^2	$X \cdot Y$	Y^2
7	19	49	133	361
13	44	169	572	1936
14	53	196	742	2809
17	61	289	1037	3721
20	88	400	1760	7744
71	265	1103	4244	16571

On a alors $\overline{x} = \dfrac{71}{5} = 14,2$, $\overline{y} = \dfrac{265}{5} = 53$, $\overline{x^2} = \dfrac{1103}{5} = 220,6$, $\overline{y^2} = \dfrac{16571}{5} = 3314,2$ et $\overline{xy} = \dfrac{4244}{5} = 848,8$.

Il s'ensuit que

$$r = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{\overline{x^2} - \overline{x}^2} \cdot \sqrt{\overline{y^2} - \overline{y}^2}} = \frac{848,8 - 14,2 \cdot 53}{\sqrt{220,6 - 14,22^2} \cdot \sqrt{3314,2 - 53^2}} \cong 0,983.$$

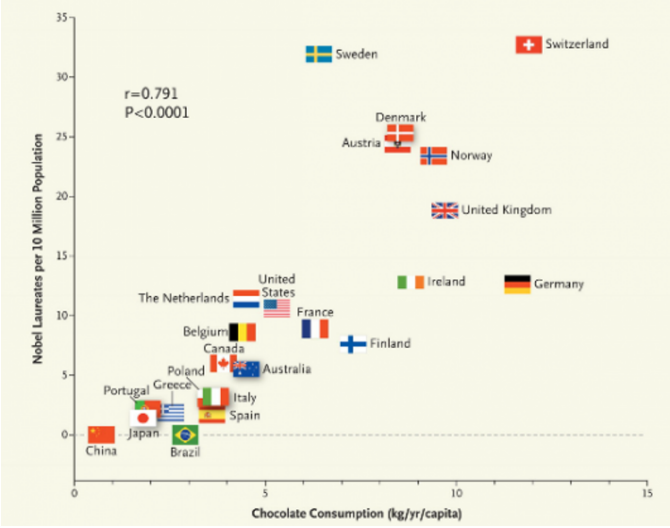
Il dénote une corrélation linéaire forte.

2.5.3 Corrélation et relation causale

Il ne faut pas confondre *corrélation* et *relation causale*. Une bonne corrélation entre deux grandeurs peut révéler une relation de cause à effet entre elles, mais pas nécessairement.

Exemple.

1. Le diagramme ci-dessous présente pour chacun des pays, la consommation de chocolat sur l'axe horizontal et le nombre de prix Nobel par tranche de 10'000 habitants.



Le coefficient de corrélation égal à 0,791, rentre presque dans la catégorie "bonne à très bonne corrélation" ! Il serait cependant hâtif de conclure que le fait de manger du chocolat rend intelligent et augmente des chances de gagner un prix Nobel. Cette corrélation n'est pas dûe à un lien de causalité, mais à un facteur externe. En effet, le chocolat est plutôt un produit de luxe et les pays développés sont pourvus d'écoles performantes. Ainsi, plus un pays est développé, plus ses habitants auront les moyens de consommer du chocolat et auront davantage de chance d'acquérir une formation pour devenir de brillants chercheurs susceptibles de remporter des prix Nobel.

2. Depuis une dizaine d'années, la taille d'une personne née en 2001, est très bien corrélée avec la puissance de calcul des ordinateurs personnels. Cette excellente corrélation ne révèle bien évidemment aucune relation de cause à effet, ni de cause commune.
3. Dans un article de la revue *Science et Avenir*, une étude statistique montrait une corrélation positive entre l'utilisation de crème solaire et le cancer de la peau. Des journalistes pressés en avaient conclu un peu vite à la nocivité de la crème solaire. En fait, "*Utilisation de crème solaire*" et "*Cancer de la peau*" n'étaient que la conséquence d'une même cause : l'exposition au soleil. Plus on s'expose au soleil plus on risque le cancer de la peau, mais plus aussi on utilise de crème solaire.
4. La taille moyenne des Japonais a augmenté de 15 cm depuis la fin de la deuxième guerre mondiale alors que la distance entre le Japon et les États-Unis augmente de 2 ou 3 cm par an à cause de la dérive des continents. Il y a corrélation parce que les deux phénomènes augmentent avec le temps, mais il n'y a pas bien évidemment la moindre causalité.

Remarque. L'existence d'une corrélation, aussi bonne soit-elle, n'est jamais la preuve d'une relation de cause à effet.

Chapitre 3

Probabilités

3.1 Introduction

Le *calcul des probabilités* a pour objet l'étude des *phénomènes aléatoires* (dus au hasard). Il trouve son origine dans les jeux de hasard pratiqués dès l'Antiquité, mais surtout très en vogue aux XVI^{e} et XVII^{e} siècles. De grands mathématiciens et éminents savants, tels Pascal [1], Fermat [2], Huygens [3], Jacques Bernouilli [4], Moivre [5], Laplace [6] et Gauss [7], ont contribué à son essor. Au XIX^{e} siècle, le calcul des probabilités s'est enrichi de nombreuses applications et développements, notamment en mécanique statistique. En 1933, Kolmogorov [8] publie dans *Fondements de la théorie des probabilités* une présentation axiomatisée du calcul des probabilités.

Aujourd'hui, le calcul des probabilités est très proche d'autres branches des mathématiques pures. Sa maîtrise est devenue importante dans de très nombreux domaines de la pensée humaine comme la physique, la chimie, la biologie, la médecine, la psychologie, la sociologie, l'économie et les sciences politiques.

3.2 Notion intuitive

Définition. Une *expérience aléatoire* est une expérience qui possède les deux propriétés suivantes :

1. On ne peut prédire avec certitude le résultat de l'expérience.
2. On peut décrire, avant l'expérience, l'ensemble des résultats possibles.

1. Blaise Pascal (1523-1663), mathématicien, physicien, inventeur, philosophe, moraliste et théologien français. .
2. Pierre de Fermat (première décennie du XVII^{e} siècle-1665), juriste, polymathe et mathématicien français.
3. Christian Huygens (1629-1695), mathématicien, astronome et physicien néerlandais.
4. Jacques Bernouilli (1654-1705), mathématicien et physicien suisse.
5. Abraham de Moivre (1667-1754), mathématicien français.
6. Pierre-Simon de Laplace (1749-1827), mathématicien, astronome et physicien français.
7. Johann Carl-Friedrich Gauss (1777-1855), mathématicien, astronome et physicien allemand.
8. Andreï Nikolaïevitch Kolmogorov (1903-1987), mathématicien russe.

Exemple.

1. Jeter une pièce de monnaie.
2. Jeter un dé.
3. Tirer une carte dans un jeu de 52 cartes.
4. Jouer à la loterie.

Exemple. Quelle est la probabilité ("combien de chances sur combien a-t-on") d'obtenir pile lorsque l'on jette une pièce de monnaie ?

Exemple. On jette un dé. Quelle est la probabilité d'obtenir un résultat

a) Impair ?

b) Multiple de 3 ?

c) Strictement inférieur à 5 ?

d) Supérieur ou égal à 1 ?

e) Strictement supérieur à 6 ?

Les exercices ci-dessus montrent qu'il parait naturel l'admettre que la probabilité qu'un événement se réalise est le quotient du nombre de possibilités qu'il a de se produire et du nombre total de possibilités.

On calcule donc la probabilité $P(A)$ d'un événement A à l'aide de la *formule de Laplace.*

$$P(A) = \frac{\text{Nombre de cas favorables}}{\text{Nombre de cas possibles}}.$$

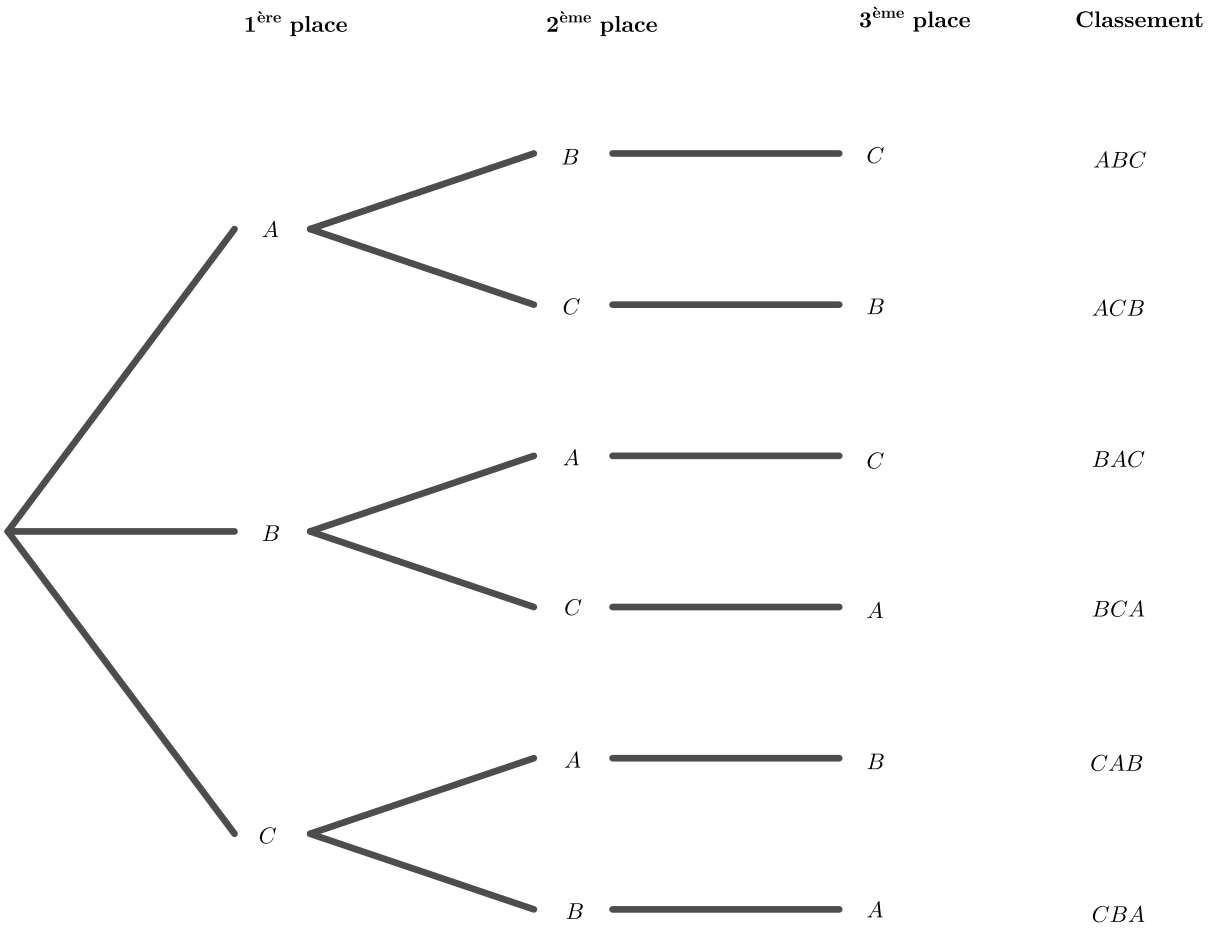
Remarque. La formule de Laplace n'est valable que lorsque les cas possibles ont tous la même probabilité de se réaliser, On dit qu'ils sont *équiprobables.*

3.3 Analyse combinatoire

3.3.1 Introduction

La principale difficulté de l'application de la formule de Laplace réside dans le calcul du nombre de cas possibles de l'expérience, ainsi que dans le nombre $n(A)$ de cas favorables à un événement A donné. La présente section a donc pour objectif de dénombrer des objets en grand nombre.

Exemple. Supposons que trois équipes participent à un tournoi dans lequel sont déterminées une première, une deuxième et une troisième place. Pour faciliter l'identification des équipes, nous allons les désigner par les lettres A, B, C. Cherchons le nombre de manières différentes permettant d'attribuer le classement de ces 3 équipes. On peut illustrer ce raisonnement par un diagramme en arbre.



On remarque que le nombre de possibilités de classement (6) est le produit du nombre de possibilités (3) d'attribuer la première place, par le nombre de possibilités (2) d'attribuer la deuxième place (après que la première place a été attribuée), par le nombre de possibilités (1) d'attribuer la troisième place (les deux premières étant déjà fixées).

Le raisonnement ci-dessus illustre la règle générale suivante, que nous utiliserons comme axiome fondamental :

Théorème. *Si une épreuve est composée de deux opérations successives, la première pouvant mener à n_1 issues différentes et le deuxième à n_2 issues différentes, alors l'épreuve peut se réaliser de $n_1 \cdot n_2$ manières différentes.*

Remarque. L'analyse combinatoire ne consiste pas en l'énumération de toutes les possibilités (souvent long et fastidieux) mais bien le dénombrement de celle-ci par un calcul.

Exemple. Une classe se compose de 12 filles et 9 garçons. De combien de façons peuvent être choisis un président de classe, un vice-président, un trésorier et un secrétaire, si le trésorier doit être une fille, le secrétaire un garçon, et si un étudiant ne peut exercer plus d'une charge ?

Il y a donc $12 \cdot 9 \cdot 19 \cdot 18 = 36'936$ choix possibles.

Exemple. Combien peut-on former de nombres entiers de quatre chiffres, si ces nombres doivent être des multiples de 5 ?

Il existe donc $9 \cdot 10 \cdot 10 \cdot 2 = 1'800$ tels nombres.

3.3.2 Permutations simples

Définition. Soit $n \in \mathbb{N}$. On définit l'entier n *factorielle*, noté $n!$, comme suit :

$$\begin{aligned} 0! &= 1; \\ n! &= n \cdot (n-1) \cdots 2 \cdot 1. \end{aligned}$$

Définition. On appelle *permutation simple* de n éléments tout classement de ces n éléments distincts dans un ordre particulier.

Remarque. Deux permutations ne diffèrent que par l'ordre des objets.

Exemple. De combien de manières différentes peut-on disposer 5 personnes sur une rangée de 5 chaises ?

Il est possible de le faire de $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$ manières différentes.

Théorème. *Le nombre P_n de permutations simples de n éléments est donné par*

$$P_n = n!$$

3.3.3 Permutations avec répétitions

Exemple. Combien de mots différents peut-on écrire avec toutes les lettres des mots LUNE, SEVERES ?

Pour le mot LUNE, le cas est clair : il y a $4! = 24$ mots différents puisque les 4 lettres sont distinctes. Pour le mot SEVERE, la situation est plus complexe puisque trois lettres sont identiques. Supposons que les trois E soient de couleurs différentes. Il y a alors $6! = 720$ mots colorés différents. Notons qu'un groupe de $3! = 6$ mots colorés conduit au même mot unicolore puisqu'il y a $3! = 6$ façons de placer les E sans changer le "sens" du mot. On peut donc composer $\frac{6!}{3!} = 120$ mots différents.

	SEVERE SEVERE SEVERE SEVERE SEVERE SEVERE	SEVERE	
720 mots	⋮	⋮	120 mots
	EEESVR EEESVR EEESVR EEESVR EEESVR EEESVR	EEESVR	

Comme le mot au pluriel SEVERES compte 3 E et 2 S, le nombre de mots de 7 lettres qu'on peut composer sera donc $\frac{7!}{3! \cdot 2!} = 420$.

3.3.4 Arrangements simples

Définition. On appelle *arrangement simple* de k éléments distincts parmi n tout choix de ces k éléments en les classant dans un ordre particulier.

Exemple. Huit athlètes participent à la finale du championnat du monde d'une course (100 mètres). Combien de podiums différents sont-ils possibles ?

Il existe $8 \cdot 7 \cdot 6 = 336$ podiums possibles.

Théorème. *Le nombre A_n^k d'arrangements simples, $k \leq n$, est donné par*

$$A_n^k = \frac{n!}{(n-k)!}.$$

3.3.5 Arrangements avec répétitions

Définition. On appelle *arrangement avec répétition* de k éléments choisis parmi n tout choix de k éléments distincts ou non (on peut choisir plusieurs fois le même) en les classant dans un ordre particulier.

Exemple. Lors d'une consultation populaire portant sur quatre objets, les électeurs peuvent répondre à chacune des questions posées par *Oui*, *Non* ou alors voter *blanc*. Les quatre réponses figurent sur une même feuille. Combien de piles différentes faut-il prévoir pour le dépouillement, si chaque pile ne doit comporter que des bulletins où les quatre réponses sont identiques ?

Il s'agit de prévoir $3 \cdot 3 \cdot 3 \cdot 3 = 3^4 = 81$ piles.

3.3.6 Combinaisons simples

Définition. On appelle *combinaison simple* de k éléments distincts parmi n tout choix de ces k éléments sans les classer dans un ordre particulier.

Exemple. Dans un jeu de 36 cartes, on en tire 5 au sort. Combien y a-t-il de possibilités ?

Dans un premier temps, voyons ce qui se passe si l'on tient compte de l'ordre. Dans ce cas, il y a

$$36 \cdot 35 \cdot 34 \cdot 33 \cdot 32 = \frac{36 \cdot 35 \cdot 34 \cdot 33 \cdot 32 \cdot 31 \cdot 30 \cdots 1}{31 \cdot 30 \cdots 1} = \frac{36!}{31!} = 45'239'040 \text{ possibilités.}$$

Dans le décompte ci-dessus, deux mains contenant 5 cartes identiques mais classées dans un ordre différent sont considérées comme différentes. Puisqu'il existe $5!$ mains dans le dénombrement ci-dessous contenant 5 mêmes cartes données, on en déduit que le nombre de mains de 5 cartes d'un jeu de 36 cartes est donné par

$$\frac{45'239'040}{5!} = \frac{36!}{31! \cdot 5!} = 376'992 \text{ mains.}$$

Théorème. *Le nombre C_n^k (noté aussi $\binom{n}{k}$) de combinaisons simples, appelé coefficient binomial, est donné par*

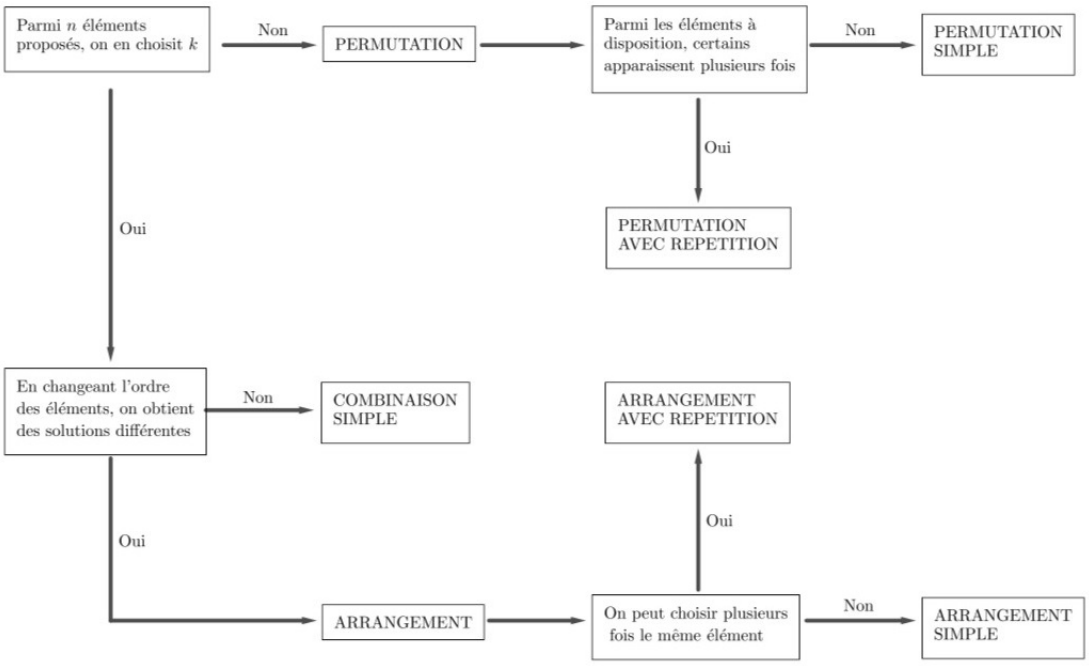
$$C_n^k = \frac{n!}{k! \cdot (n-k)!}.$$

Exemple. De combien de manières différentes peut-on former un comité de trois personnes à partir d'une classe de 24 élèves ?

Il existe $C_{24}^3 = \frac{24!}{(24-3)! \cdot 3!} = 2'024$ comités.

3.3.7 Resumé

La figure ci-dessous résume la méthode de déombrement à choisir en fonction d'une situation donnée.



3.4 Formule de Laplace

Définition. *L'univers* d'une expérience aléatoire est l'ensemble Ω de toutes les issues possibles que l'on peut obtenir au cours de cette expérience.

Exemple. Décrivons l'univers ainsi que le nombre d'issues possibles des expériences aléatoires proposées :

1. Lancer une pièce de monnaie : $\Omega = \{P; F\}$, 2 issues possibles.
2. Jeter un dé : $\Omega = \{1; 2; 3; 4; 5; 6\}$, 6 issues possibles.
3. Jeter deux fois de suite le même dé : $\Omega = \{(1; 1); (1; 2); (1; 3); ...; (6; 5); (6; 6)\}$, $6 \cdot 6 = 36$ issues possibles.

Définition. Soit Ω l'univers d'une expérience aléatoire.

Un *événement* est un sous-ensemble de l'univers Ω. On note les événements par des lettres majuscules.

Le sous-ensemble vide ϕ est *l'événement impossible* et l'univers Ω est *l'événement certain.*

Exemple. On tire au hasard un jeton parmi les 3 jetons suivants : 1 , 2 et 3 .

L'Univers Ω est donné par $\Omega = \{1; 2; 3\}$.

Les 8 événements possibles sont :

— $A =$ "Obtenir le jeton 1", $A = \{1\}$.

— $B =$ "Obtenir le jeton 2", $B = \{2\}$.

— $C =$ "Obtenir le jeton 3", $C = \{3\}$.

— $D =$ "Obtenir le jeton 1 ou 2", $D = \{1; 2\}$.

— $E =$ "Obtenir le jeton 1 ou 3", $E = \{1; 3\}$.

— $F =$ "Obtenir le jeton 2 ou 3", $F = \{2; 3\}$.

— $G =$ "Obtenir le jeton 1 ou 2 ou 3", $G = \{1; 2; 3\} = \Omega$.

— $H =$ "Obtenir le jeton 4", $H = \phi$.

On rappelle que dans le cas d'événements équiprobables, la probabilité $P(A)$ se produise se calcule à l'aide de ladite *Formule de Laplace* :

$$P(A) = \frac{\text{Nombre de cas favorables}}{\text{Nombre de cas possibles}}.$$

Exemple. Si on tire deux cartes d'un jeu de 36 cartes bien brassé et si le tirage se fait au hasard, sans tricher. L'univers sera constitué de tous les tirages possibles de 2 cartes parmi les 36. Sans les décrire, nous savons qu'il y en a

$$C_{36}^2 = \frac{36!}{34! \cdot 2!} = 630 \text{ possibilités.}$$

Si maintenant, on s'intéresse parmi ces possibilités à l'événement

$$A = \text{"Obtenir deux as"},$$

nous pouvons calculer le nombre de possibilités d'obtenir 2 as par

$$C_4^2 = \frac{4!}{2! \cdot 2!} = 6 \text{ possibilités.}$$

La probabilité d'obtenir 2 as en tirant au hasard 2 cartes dans un jeu de 36 cartes est donc :

$$P(A) = \frac{6}{630} \cong 0,00952 = 0,952\%.$$

On en déduit de plus la probabilité de l'événement complémentaire

$${}^cA = \text{" Ne pas obtenir deux as"}$$

par

$$P({}^cA) = 1 - P(A) \cong 1 - 0,00952 = 0,952\%.$$

Remarque.

— Cette définition est valable uniquement si tous les tirages ont la même chance de se réaliser. On dira alors que les résultats sont *équiprobables*. Par exemple, les résultats "on obtient pile" ou "on obtient face" en lançant une pièce de monnaie pourraient ne pas être équiprobables si la pièce est faussée. Dès lors, on ne pourrait plus utiliser la formule de Laplace.

— La probabilité d'un événement est un nombre réel compris entre 0 et 1. On l'exprime volontiers sous la forme d'un pourcentage.

— Dans la réalité, il est relativement rare qu'il soit possible de dénombrer les cas favorables et les cas possibles. Par exemple, les meilleurs météorologues ne savent pas chiffrer avec certitude la probabilité de l'événement "il fera beau demain".

3.5 Loi de probabilité

3.5.1 Définition

Définition. Si, à chacune des valeurs possibles d'une variable aléatoire X, on associe la probabilité de l'événement correspondant, on obtient *la loi de probabilité* (ou *distribution de probabilité*) de X.

Exemple. On lance deux fois une pièce de monnaie. Si on définit la variable aléatoire X par le nombre de faces obtenues, alors on obtient la loi de probabilité suivante.

Événement élémentaire	Variable aléatoire X	Probabilité $P(\{X\})$
PP	0	1/4
PF ou FP	1	1/2
FF	2	1/4

Exemple. On jette deux dés. Si on définit la variable X par la somme des résultats des deux dés, on obtient la loi de probabilité suivante :

Variable aléatoire X	Probabilité $P(\{X\})$
2	1/36
3	1/18
4	1/12
5	1/9
6	5/36
7	1/6
8	5/36
9	1/9
10	1/12
11	1/18
12	1/36

3.5.2 Espérance mathématique

Exemple. Cinq élèves se sont présentés à un examen de latin. Quatre élèves ont obtenu la note 5 et un a obtenu 6. On se demande quelle est la note moyenne d'un tel examen.

Il serait absurde de prétendre que la note moyenne est de $\frac{5+6}{2} = 5,5$. En effet, les résultats de cet examen montrent que les chances d'obtenir un 6 sont beaucoup plus faibles que d'obtenir un 5.

On peut calculer la moyenne μ de l'examen comme suit :

$$\mu = \frac{4 \cdot 5 + 1 \cdot 6}{5} = 5 \cdot \frac{4}{5} + 6 \cdot \frac{1}{5} \cdot 6 = 5,2.$$

Définition. Soit X une variable aléatoire discrète prenant les n valeurs x_1, x_2,..., x_n. avec les probabilités respectives p_1, p_2,..., p_n.

On appelle *espérance mathématique* de X le nombre $\mathbb{E}(X)$ défini par

$$E(X) = \sum_{k=1}^{n} p_k x_k = x_1 p_1 + x_2 p_2 + ... + x_n p_n.$$

Exemple. Dans l'exemple du jet des deux pièces de monnaie, l'espérance mathématique du nombre de faces obtenues est donc

$$E(X) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1.$$

Exemple. Dans l'exemple du jet de deux dés avec la variable aléatoire X égale à la somme des résultats, l'espérance mathématique de X est égale à

$$E(X) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{1}{18} + 4 \cdot \frac{1}{12} + 5 \cdot \frac{1}{9} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{1}{6} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{1}{9} + 10 \cdot \frac{1}{12} + 11 \cdot \frac{1}{18} + 12 \cdot \frac{1}{36} \cdot 12 = 7.$$

Remarque. Dans un jeu de hasard, l'espérance mathématique E du jeu correspond au gain qu'un joueur peut espérer retirer du jeu.

Définition. On dit d'un jeu qu'il est

$$\begin{array}{ll} \textit{-favorable} & \text{si } E > 0\,; \\ \textit{-défavorable} & \text{si } E < 0\,; \\ \textit{-équilibré} & \text{si } E = 0. \end{array}$$

3.5.3 Variance et écart-type

Définition. Soit X une variable aléatoire discrète prenant les valeurs x_1, x_2,..., x_n avec les probabilités p_1, p_2,..., p_n. On appelle *variance* de X le nombre réel noté $\mathrm{Var}(X)$ défini par

$$\mathrm{Var}(X) = \sum_{k=1}^{n} (x_k - E(X))^2 \cdot p_k.$$

Définition. Soit X une variable aléatoire discrète prenant les valeurs x_1, x_2,..., x_n avec les probabilités p_1, p_2,..., p_n. On appelle *écart-type* de X le nombre réel noté σ et défini par

$$\sigma = \sqrt{\mathrm{Var}(X)}.$$

Exemple. Dans l'exemple du jet des deux pièces de monnaie, on obtient

$$\mathrm{Var}(X) = (0-1)^2 \cdot \frac{1}{4} + (1-1)^2 \cdot \frac{1}{2} + (2-1)^2 \cdot \frac{1}{4} = 0,5$$

et

$$\sigma = \sqrt{0,5} \cong 0,707.$$

3.5.4 Loi binomiale

Considérons les problèmes suivants :

1. On lance une pièce de monnaie 20 fois et on cherche la probabilité d'obtenir 7 fois "Pile" exactement dans n'importe quel ordre.
2. On jette 15 fois un dé et on cherche la probabilité d'obtenir exactement 6 fois la face 3 dans n'importe quel ordre.
3. Un tireur touche sa cible avec la probabilité 90%. Quelle est la probabilité qu'il atteigne la cible exactement 11 fois en tirant 17 coups ?

4. On soigne 20 patients avec un traitement qui se révèle efficace dans 70% des cas. Quelle est la probabilité que 12 patients traités guérissent ?
5. Un étudiant répond au hasard à un QCM comprenant 25 questions. Pour chacune d'elles, 4 réponses sont proposées dont une seule est correcte. Quelle est la probabilité que l'étudiant réponde correctement à 6 questions ?
6. Quelle est la probabilité qu'une famille de 7 enfants compte exactement 2 garçons ?
7. Quelle est la probabilité de deviner les résultats de 8 matches d'une journée de championnat au cours de laquelle se déroulent 13 parties.

Dans chacune de ces situations, on répète un certain nombre de fois la même expérience débouchant sur deux issues : *succès* et *échec*, de probabilités complémentaires $P(\text{succès}) = p$ et $P(\text{échec}) = 1 - p$. Une telle expérience aléatoire est appelée *expérience de Bernoulli*. Si on définit la variable aléatoire X dénombrant le nombre total de succès réalisés sur n répétitions indépendantes de cette même expérience, alors les valeurs possibles de X sont $k = 0, 1, ..., n$ et les probabilités correspondantes sont données par

$$\begin{aligned} P(X = k) &= \text{Nombre de cas favorables} \cdot \text{Probabilité d'un cas favorable} \\ &= C_n^k \cdot p^k \cdot (1-p)^{n-k}. \end{aligned}$$

En effet, $\underbrace{SS...S}_{k}\underbrace{EE...E}_{n-k}$ est un cas favorable, de probabilité $p^k \cdot (1-p)^{n-k}$.

Or, il existe $\dfrac{n!}{k! \cdot (n-k)!} = C_n^k$ tels cas (le nombre de mots qu'on peut écrire avec k lettres S et $n-k$ lettres E).

On dit que la variable aléatoire X suit une *loi binomiale* de paramètres n et p. On note $X \sim \mathcal{B}(n; p)$ et on a

$$P(X = k) = C_n^k \cdot p^k \cdot (1-p)^{n-k}.$$

Exemple. Les problèmes posés ci-dessus ont donc les solutions suivantes.

$P(X = 7) = C_{20}^7 \cdot \left(\frac{1}{2}\right)^7 \cdot \left(\frac{1}{2}\right)^{13} \cong 7,39\%.$

$P(X = 6) = C_{15}^6 \cdot \left(\frac{1}{6}\right)^6 \cdot \left(\frac{5}{6}\right)^9 \cong 2,08\%.$

$P(X = 11) = C_{17}^{11} \cdot 0,9^{11} \cdot 0,1^6 \cong 0,338\%.$

$P(X = 12) = C_{20}^{12} \cdot 0,7^{12} \cdot 0,3^8 \cong 11,44\%.$

$P(X = 6) = C_{25}^6 \cdot \left(\frac{1}{4}\right)^6 \cdot \left(\frac{3}{4}\right)^{19} \cong 18,28\%.$

$P(X = 2) = C_7^2 \cdot \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^5 \cong 16,41\%.$

$P(X = 8) = C_{13}^8 \cdot \left(\frac{1}{3}\right)^8 \cdot \left(\frac{2}{3}\right)^5 \cong 2,58\%.$

Notons que la loi binomiale a pour espérance mathématique et pour variance

$$\begin{aligned} E(X) &= n \cdot p \\ \text{Var}(X) &= n \cdot p \cdot (1-p) \end{aligned}.$$

Exemple. Combien de fois faut-il lancer un dé pour que la probabilité qu'il retombe au moins une fois sur 6 soit de 99% ? On cherche n tel que $P(X \geq 1) = 99\%$:

$$\begin{aligned} P(X \geq 1) &= 99\% \\ 1 - P(X = 0) &= 99\% \\ 1 - C_n^0 \cdot \left(\tfrac{1}{6}\right)^0 \cdot \left(1 - \tfrac{1}{6}\right)^{n-0} &= 99\% \\ 1 - \left(\tfrac{5}{6}\right)^n &= 99\% \\ -\left(\tfrac{5}{6}\right)^n &= -1\% \\ \left(\tfrac{5}{6}\right)^n &= 0,01 \\ \log\left(\tfrac{5}{6}\right)^n &= \log 0,01 \\ n \cdot \log\left(\tfrac{5}{6}\right) &= \log 0,01 \\ n &= \tfrac{\log 0,01}{\log\left(\frac{5}{6}\right)} \\ n &\cong 25,26 \end{aligned}$$

Il faudra ainsi lancer le dé 26 fois.

3.6 Variables aléatoires continues

3.6.1 Introduction

On rappelle qu'une variable aléatoire X est dite *continue*, si l'ensemble des valeurs de celle-ci est infini non dénombrable. Par exemple, la variable aléatoire X décrivant la taille (en cm) des hommes en Suisse est continue. De toutes les lois usuelles des variables aléatoires continues, la *loi normale* est la plus fréquemment rencontrée. Lorsqu'une grandeur, qui se reproduit, est soumise à l'influence d'un grand nombre de facteurs de variations indépendants les uns des autres, chacun exerçant des actions individuelles de faible intensité dont les effets tendent à se compenser, on peut établir que la distribution des valeurs de cette grandeur suit une *loi de Laplace-Gauss* dite *loi normale*.

3.6.2 Cloche de Gauss

Définition. On appelle *Cloche de Gauss* ou *gaussienne* la courbe représentant une fonction f dont l'expression fonctionelle est de la forme

$$f(x) = C \cdot e^{-a \cdot (x-m)^2}.$$

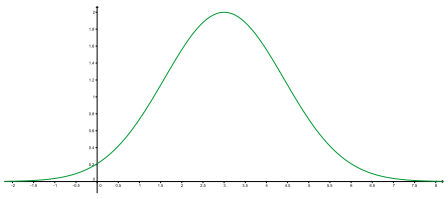
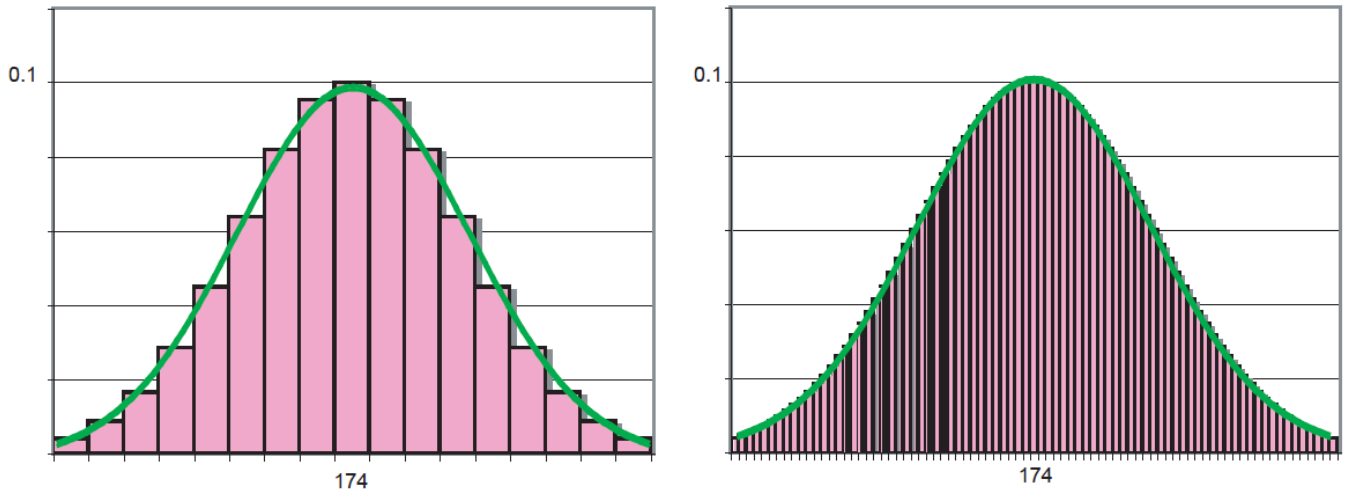


FIGURE 3.1 – Gaussienne de paramètres $C = 2$, $a = 0,25$ et $m = 3$.

3.6.3 Loi normale

Exemple. En 2001, les hommes français avaient une taille moyenne de 174 cm avec un écart-type de 7 cm. Si l'on représente l'histogramme des fréquences associé à une répartition en classes de taille de largeur 5 cm, on obtient une figure comportant 10 fois moins de rectangles verticaux que dans le cas où l'on divise la population en classes de largeur 0,5 cm.



Les frontières supérieures des histogrammes correspondants sont des courbes en escalier qui, à mesure que la largeur des classes est réduite et que leur nombre croît, se rapprochent de la gaussienne d'expression fonctionnelle

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}.$$

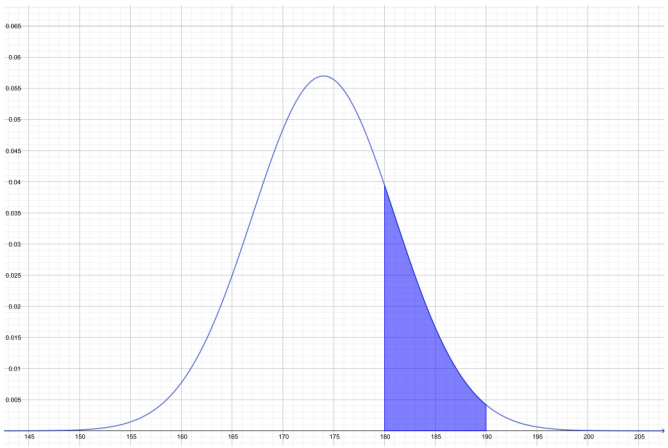
Cette courbe s'exprime en fonction de la moyenne μ (ici $\mu = 174$) et de l'écart-type σ (ici $\sigma = 7$) de la population.

Déterminons la probabilité qu'un individu donné ait une taille comprise entre 180 et 190 cm.

La probabilité cherchée n'est rien d'autre que l'aire du domaine compris [9] entre la courbe, l'axe horizontal et les verticales d'équations $x = 180$ et $x = 190$.

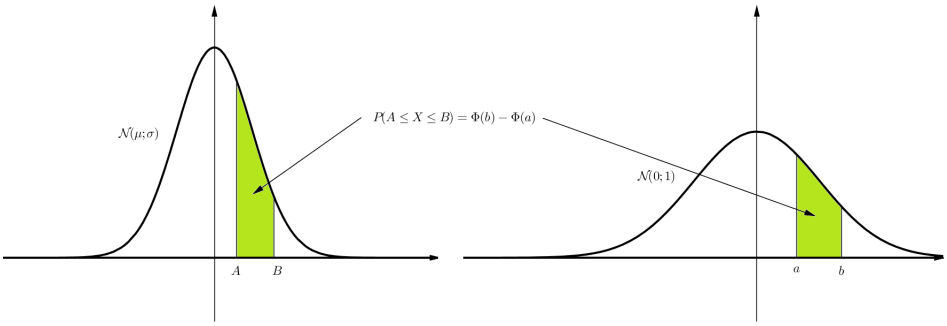
9. C'est-à-dire au moyen de l'intégrale définie

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \int_{180}^{190} e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}.$$



Malheureusement, cette aire ne se laisse pas calculer analytiquement (car $f(x)$ n'admet pas de primitive explicite). Le calcul de cette probabilité s'avère très complexe et requiert une approximation numérique. Pour éviter de devoir réaliser une lors de chaque calcul de probabilité de ce type, une solution consiterait à établir des tables numériques. Or il en faudrait une infinité, soit une par paire de valeurs de μ et σ.

On peut toutefois facilement démontrer[10] que l'aire $P(A \leq X \leq B)$ sous la courbe est égale à l'aire d'une autre surface située entre deux autres bornes a et b sous la *courbe normale centrée réduite* de moyenne $\mu = 0$ et d'écart-type $\sigma = 1$. Cela permet de calculer ce type de probabilité à l'aide d'une unique table numérique.



Comme $\mu = 174$ et $\sigma = 7$, les nouvelles bornes a et b sont définies par les formules suivantes

$$a = \frac{180 - \mu}{\sigma} \cong 0,86 \text{ et } b = \frac{190 - \mu}{\sigma} \cong 2,29.$$

L'aire sous la gaussienne considérée est donnée par une table (cf annexe) qui fournit les aires $\Phi(z)$ sous cette courbe à gauche des bornes supérieures z.

10. Pour calculer

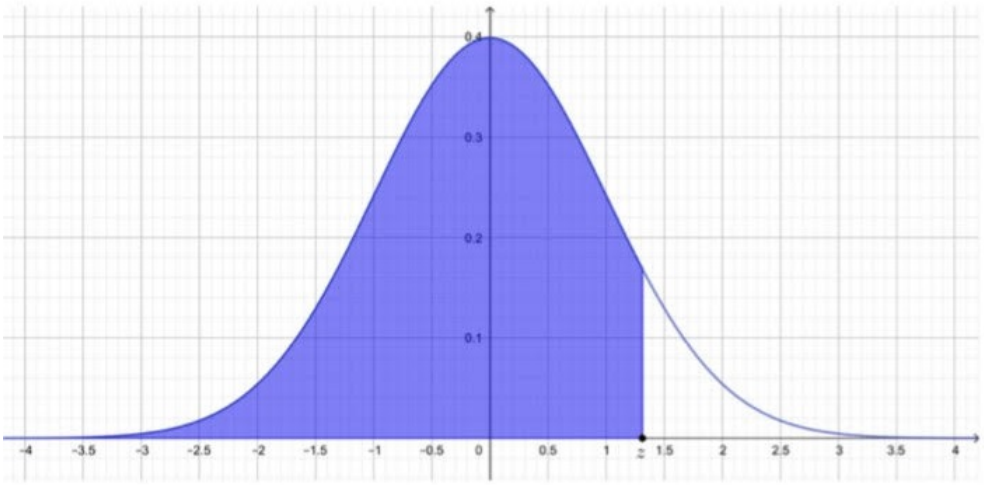
$$\frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \int_A^B e^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2} dx,$$

on effectue le changement de variable

$$z = \frac{x - \mu}{\sigma}.$$

On en déduit que

$$I = \frac{1}{\sqrt{2\pi}} \cdot \int_{(A-\mu)/\sigma}^{(B-\mu)/\sigma} e^{-\frac{1}{2} \cdot z^2} dz.$$



z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936

Dans la table, on y trouve $\Phi(2,29) = 0,9890$ et $\Phi(0,86) = 0,8051$. Ainsi, le pourcentage d'hommes dont la taille est comprise entre 180 et 190 cm est donné par

$$P(180 \leq X \leq 190) \cong \Phi(2,29) - \Phi(0,86) = 0,9890 - 0,8051 = 0,1839 = 18,39\%.$$

Définition. On dit que la variable aléatoire X suit une *loi normale* (ou une *loi de Laplace-Gauss*) de moyenne μ et d'écart-type σ, notée $\mathcal{N}(\mu;\sigma)$ si la probabilité que X varie entre a et b est donnée par

$$P(a \leq X \leq b) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \int_a^b e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}.$$

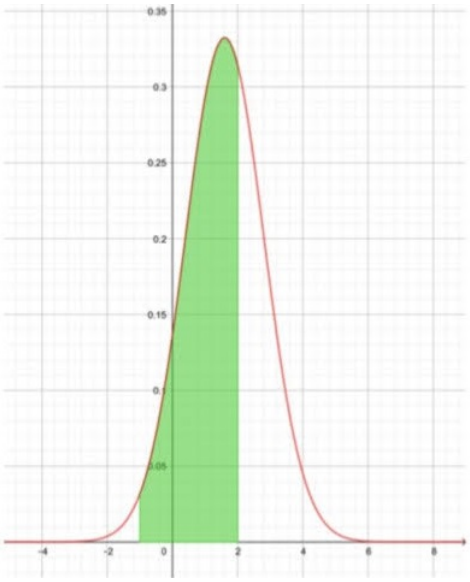


FIGURE 3.2 – $P(a \leq X \leq b)$ avec $\sigma = 1,2$, $\mu = 1,6$, $a = -1$ et $b = 2$.

Remarque. Cette courbe a les propriétés remarquables suivantes :

- Elle est symétrique par rapport à l'axe vertical $x = \mu$. Ce qui indique que la moyenne est aussi la médiane de la population. En particulier, la courbe centrée normale est symétrique par rapport à l'axe vertical.
- L'aire de la surface comprise entre l'axe horizontal et la courbe est égale à 1. En effet, elle est la même que celle de l'histogramme composé des rectangles représentant les fréquences de chaque classe. La somme des aires est donc la somme de toutes les fréquences, c'est-à-dire $100\% = 1$.
- Le pourcentage $P(A \leq X \leq B)$ d'individus dont la taille X est comprise entre deux bornes A et B (180 et 190, par exemple) est égale à l'aire sous la courbe limitée par les verticales $x = A$ et $x = B$. Il est clair que $P(A \leq X \leq B)$ est aussi la probabilité qu'un individu choisi au hasard ait une taille X comprise entre A et B. On dit alors que la variable aléatoire X suit une loi normale de paramètres μ et σ, ce qu'on note $X \sim \mathcal{N}(\mu;\sigma)$.
- Ainsi, si X est une variable aléatoire suivant une loi normale $\mathcal{N}(\mu,\sigma)$, on calcule la probabilité $P(A \leq X \leq B)$ comme suit :
  1. On définit les nouvelles bornes

$$a = \frac{A-\mu}{\sigma} \text{ et } b = \frac{B-\mu}{\sigma}.$$

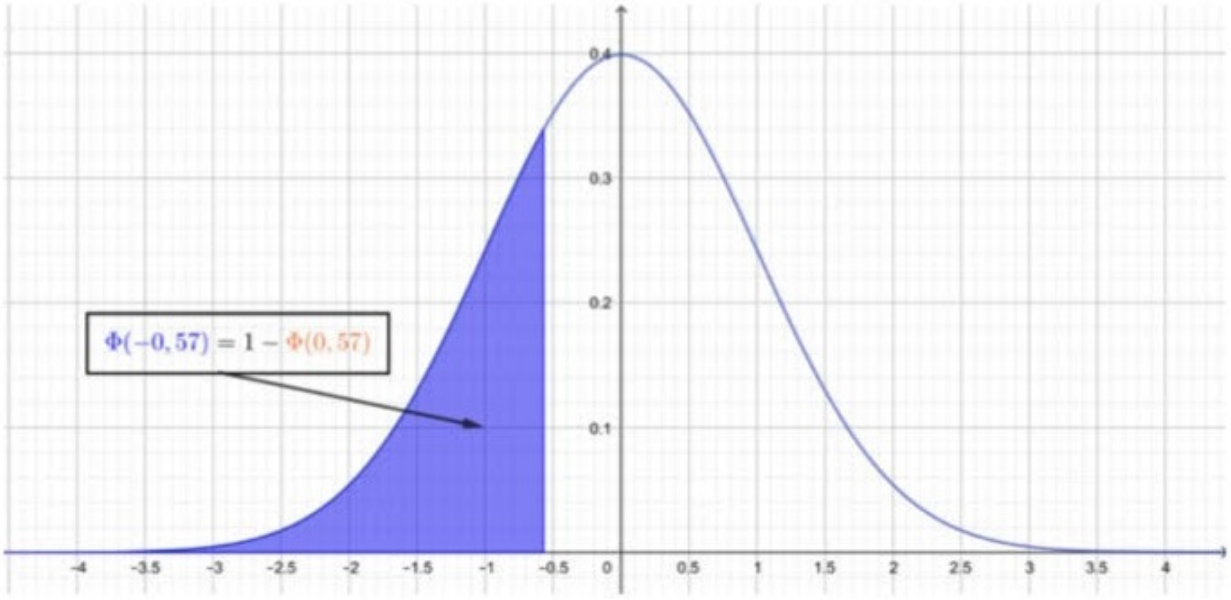
2. On détermine, à l'aide de la table, l'aire sous la gaussienne représentant $f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}$ entre les bornes a et b, c'est-à-dire $\Phi(b) - \Phi(a)$.
3. On en conclut que

$$P(A \leq X \leq B) \cong \Phi(b) - \Phi(a).$$

Exemple. Reprenons notre exemple relatif à la taille des hommes français en 2001. Déterminons maintenant le pourcentage d'hommes dont la taille est inférieure à 170 cm, c'est-à-dire la probabilité, pour la variable aléatoire taille X, d'être inférieure à $B = 170$. La nouvelle borne obtenue après conversion en unités centrées réduites est donc

$$b = \frac{B - \mu}{\sigma} = \frac{170 - 174}{7} \cong -0,57.$$

Il s'agit alors de calculer $\Phi(-0,57)$. Or, la table ne donne pas les valeurs de $\Phi(z)$ pour $z < 0$. Celles-ci s'obtiennent en exploitant la symétrie de la cloche de Gauss.



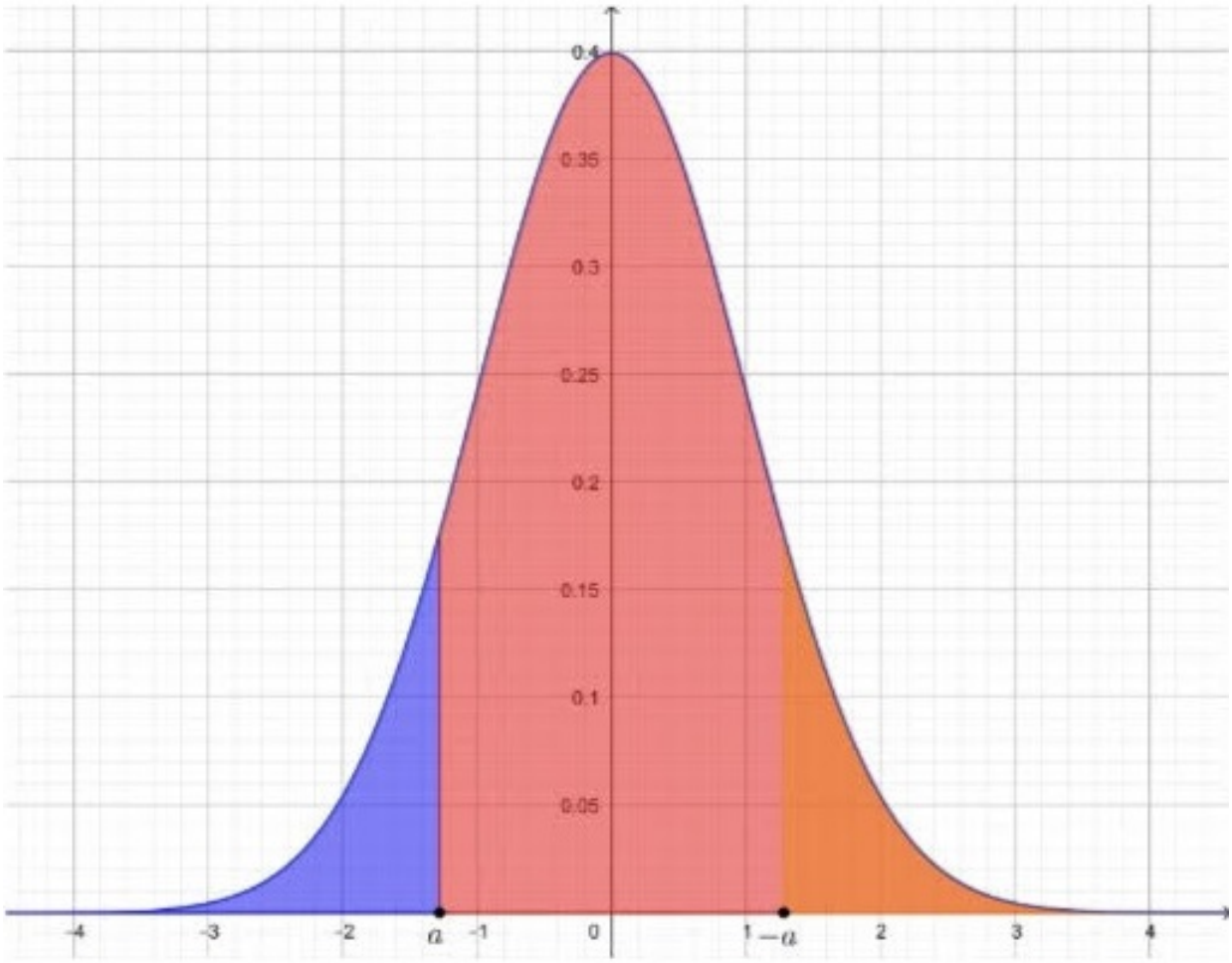
Il est clair en effet que

$$\Phi(-z) = 1 - \Phi(z).$$

On en déduit que

$$P(X \leq 170) = \Phi(-0,57) = 1 - \Phi(0,57) = 1 - 0,7157 = 0,2843 = 28,43\%.$$

Déterminons alors la taille au-dessus de laquelle on compte 90% des hommes français. On cherche A tel que $P(X \geq A) = 0,9$.



On pose donc

$$\begin{aligned} P(X \geq A) &= 0,9 \\ 1 - \Phi(a) &= 0,9 \\ \Phi(-a) &= 0,9 \end{aligned}$$

Sur la table, on lit que $-a = 1,28$, c'est-à-dire $a = -1,28$.

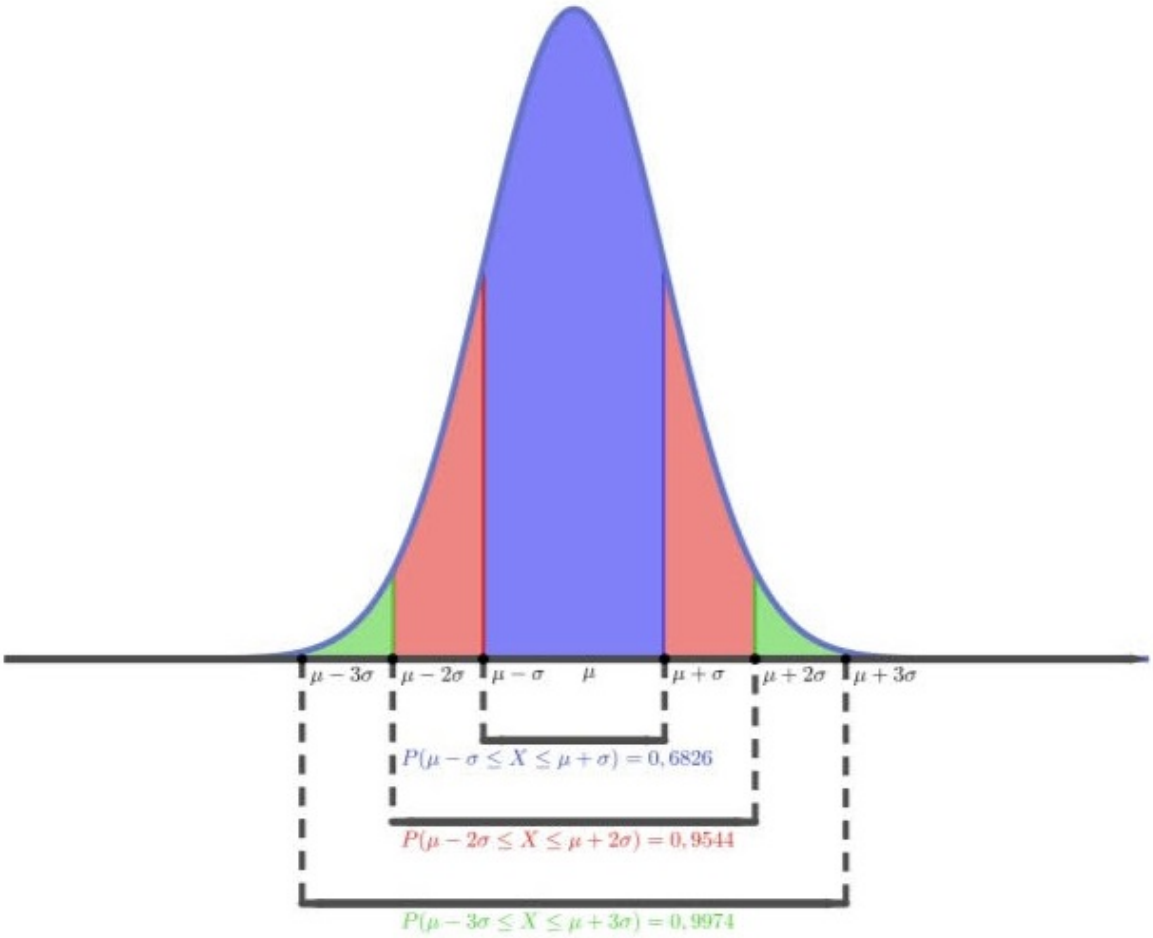
On pose alors

$$\begin{aligned} \frac{A-174}{7} &= -1,28 \\ A - 174 &= -8,96 \\ A &= 165,04. \end{aligned}$$

Ainsi, 90% des hommes français mesurent au moins 165,04 cm.

Théorème. *Pour une variable aléatoire X suivant une loi normale $\mathcal{N}(\mu, \sigma)$, on a*

- $P(\mu - \sigma \leq X \leq \mu + \sigma) = 0,6826$, *ce qui signifie que* $68,26\%$ *des réalisations de* X *seront comprises entre la moyenne plus ou moins 1 écart-type.*
- $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0,9544$, *ce qui signifie que* $95,44\%$ *des réalisations de* X *seront comprises entre la moyenne plus ou moins 2 écart-type.*
- $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0,9974$, *ce qui signifie que* $99,74\%$ *des réalisations de* X *seront comprises entre la moyenne plus ou moins 3 écart-type.*



Preuve.

—

$$\begin{aligned}
(\mu - \sigma \leq X \leq \mu + \sigma) &= \Phi\left(\tfrac{\mu+\sigma-\mu}{\sigma}\right) - \Phi\left(\tfrac{\mu-\sigma-\mu}{\sigma}\right) \\
&= \Phi\left(\tfrac{\sigma}{\sigma}\right) - \Phi\left(\tfrac{-\sigma}{\sigma}\right) \\
&= \Phi(1) - \Phi(-1) \\
&= \Phi(1) - (1 - \Phi(1)) \\
&= \Phi(1) - 1 + \Phi(1) \\
&= 2\Phi(1) - 1 \\
&= 2 \cdot 0,8413 \\
&= 0,6826.
\end{aligned}$$

—

$$\begin{aligned}(\mu-2\sigma\leq X\leq\mu+2\sigma) &= \Phi\left(\tfrac{\mu+2\sigma-\mu}{\sigma}\right)-\Phi\left(\tfrac{\mu-2\sigma-\mu}{\sigma}\right)\\ &= \Phi\left(\tfrac{2\sigma}{\sigma}\right)-\Phi\left(\tfrac{-2\sigma}{\sigma}\right)\\ &= \Phi(2)-\Phi(-2)\\ &= \Phi(2)-(1-\Phi(2))\\ &= \Phi(2)-1+\Phi(2)\\ &= 2\Phi(2)-1\\ &= 2\cdot 0,0,9772\\ &= 0,9544.\end{aligned}$$

—

$$\begin{aligned}(\mu-3\sigma\leq X\leq\mu+3\sigma) &= \Phi\left(\tfrac{\mu+3\sigma-\mu}{\sigma}\right)-\Phi\left(\tfrac{\mu-3\sigma-\mu}{\sigma}\right)\\ &= \Phi\left(\tfrac{3\sigma}{\sigma}\right)-\Phi\left(\tfrac{-3\sigma}{\sigma}\right)\\ &= \Phi(3)-\Phi(-3)\\ &= \Phi(3)-(1-\Phi(3))\\ &= \Phi(3)-1+\Phi(3)\\ &= 2\Phi(3)-1\\ &= 2\cdot 0,9987\\ &= 0,9974.\end{aligned}$$

□

Exemple. Dans notre exemple relatif à la taille des français, on a
$P(157\leq X\leq 181)=0,6826$, $P(160\leq X\leq 188)=0,9544$ et $P(152\leq X\leq 195)=0,9974$.

3.6.4 Approximation d'une loi binomiale

Exemple. Dans un cabinet médical, on teste l'allergie de patients à une certaine substance. On sait qu'un patient développera l'allergie avec une probabilité de 30%. Soit X la variable aléatoire donnant le nombre de patients allergiques sur une population de n patients testés. On sait que X suit une loi binomiale, c'est-à-dire

$$P(X=k)=C_n^k\cdot 0,3^k\cdot 0,7^{n-k},\text{ avec } k=0,1,...,n.$$

Les graphiques ci-dessous représentent les lois de probabilités pour les cas où $n=5$ et $n=10$.

k	$P(X=k)$
0	0,1681
1	0,3602
2	0,3087
3	0,1323
4	0,0284
5	0,0024

FIGURE 3.3 – Loi avec $n=5$.

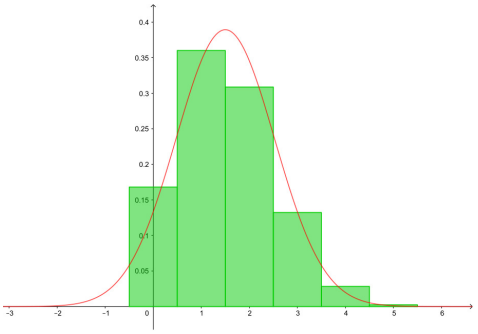


FIGURE 3.4 – Distribution avec $n=5$.

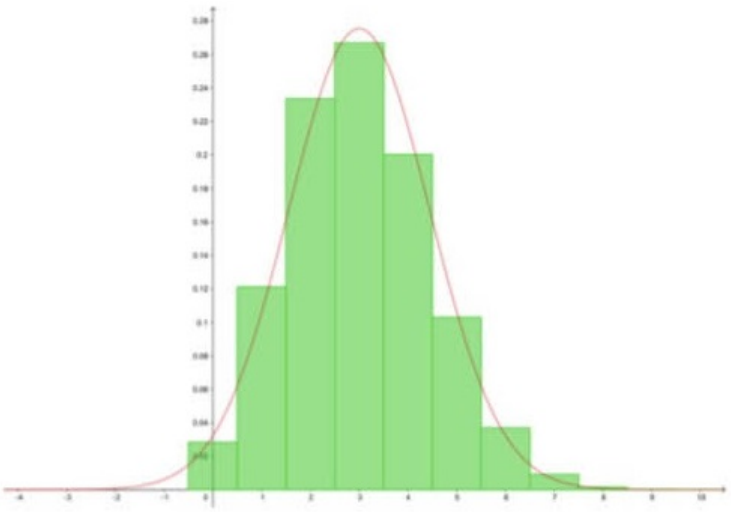


FIGURE 3.5 – Distribution avec $n = 10$.

La probabilité que X prenne la valeur 2, par exemple, est donnée par l'aire du rectangle dont la base, de longueur 1, est centrée en 2. Ainsi, l'aire de l'histogramme vaut 1 puisqu'elle correspond à la somme des probabilités de tous les cas possibles.

Sur ces deux figures, apparaissent également deux gaussiennes, définies par l'expression fonctionnelle

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

avec $\mu = E(X) = n \cdot p$ et $\sigma = \sqrt{\text{Var}(X)} = \sqrt{n \cdot p \cdot (1-p)}$.

Ainsi la gaussienne qui ajuste le mieux l'histogramme représentant une loi binomiale $\mathcal{B}(n; p)$ est donnée par l'expression fonctionnelle ci-dessus. Ceci constitue un résultat très important, découvert en 1733 par Abraham de Moivre.

Le graphique suivant montre la même approximation pour $n = 30$. On observe que l'ajustement est d'autant meilleur que n est grand.

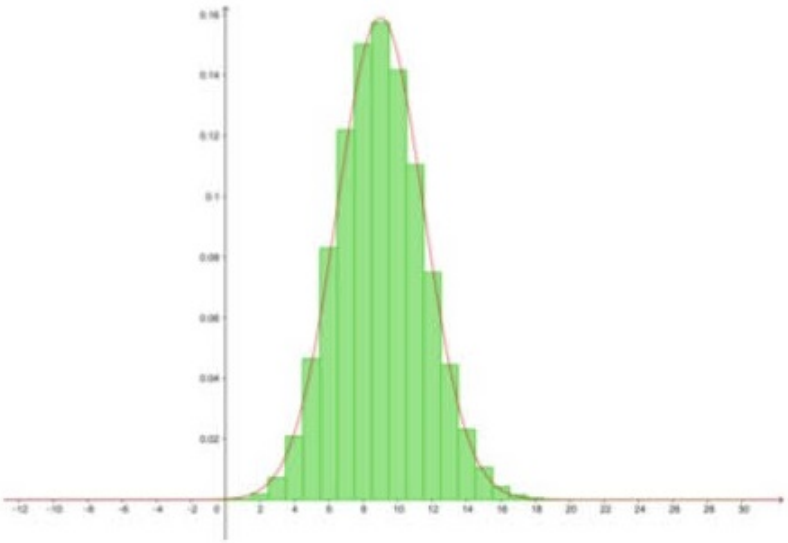


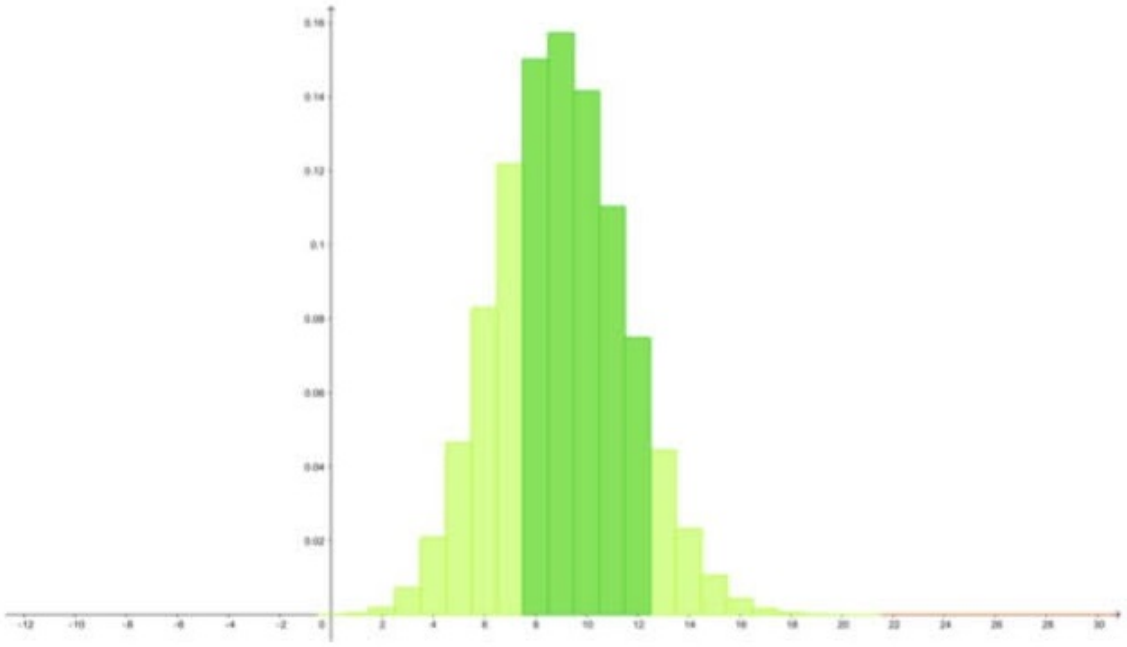
FIGURE 3.6 – Distribution avec $n = 30$.

Mais quel est au juste l'intérêt d'une telle approximation ? Reprenons notre problème initial et supposons que nous soit posé la question suivante : *Quelle est la probabilité que, sur 30 patients testés, le nombre de patients allergiques soit compris entre 8 et 12 ?* La loi binomiale fournit la réponse suivante :

$$P(8 \leq X \leq 12) = C_{30}^{8} \cdot 0,3^{8} \cdot 0,7^{22} + C_{30}^{9} \cdot 0,3^{9} \cdot 0,7^{21} + ... + C_{30}^{12} \cdot 0,3^{12} \cdot 0,7^{18}.$$

Le calcul effectif de d'expression ci-dessus s'avère fastidieux. Il le serait d'autant plus si on avait testé 500 patients plutôt que 30. Il convient donc de trouver une astuce.

En fait, la probabilité demandée correspond à la somme des aires de 5 rectangles dont les bases (de longueur 1) sont centrées en 8, 9, ..., 12.



Cette aire totale est très voisine de l'aire comprise sous la gaussienne entre les verticales $x = 7,5$ et $x = 12,5$. Autrement dit, il est possible d'approximer cette probabilité par une loi normale dont les bornes sont $A = 8 - 0,5 = 7,5$ et $B = 12 + 0,5 = 12,5$.

Ici, $\mu \overset{\text{déf}}{=} E(x) = n \cdot p = 30 \cdot 0,3 = 9$ et $\sigma = \sqrt{n \cdot p \cdot (1-p)} = \sqrt{30 \cdot 0,3 \cdot 0,7} = \sqrt{6,3} \cong 2,51$.

Les nouvelles bornes sont

$$a = \frac{7,5 - 9}{2,51} \cong -0,60 \text{ et } b = \frac{12,5 - 9}{2,51} \cong 1,39.$$

Sur la table, on y trouve

$$\Phi(1,39) \cong 0,9177 \text{ et } \Phi(-0,60) = 1 - \Phi(0,60) \cong 0,2743.$$

Finalement, la probabilité demandée vaut

$$P(8 \leq X \leq 12) \cong \Phi(1,39) - \Phi(0,60) \cong 0,9177 - 0,2743 = 64,34\%.$$

La valeur calculée avec la loi binomiale est donnée par $63,42\%$.

Ainsi, si X est une variable aléatoire suivant une loi binomiale $\mathcal{B}(n,p)$, on calcule la probabilité $P(A \leq X \leq B)$ comme suit :

1. On calcule la moyenne $\mu = n \cdot p$ et l'écart-type $\sigma = \sqrt{n \cdot p \cdot (1-p)}$.
2. On définit les nouvelles bornes
$$a = \frac{A - 0,5 - \mu}{\sigma} \text{ et } b = \frac{B + 0,5 - \mu}{\sigma}.$$
3. On détermine, à l'aide de la table, l'aire sous la gaussienne représentant $\frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}$ entre les bornes a et b, c'est-à-dire $\Phi(b) - \Phi(a)$.
4. On en conclut que
$$P(A \leq X \leq B) \cong \Phi(b) - \Phi(a).$$

Remarque. Dans la pratique, on considère comme satisfaisante l'approximation de la loi binomiale par une loi normale si $n \cdot p \geq 5$ et si $n \cdot (1-p) \geq 5$. Lorsque l'une ou l'autre de ces conditions n'est pas remplie, l'approximation peut se révéler assez grossière.

Exemple. On jette un dé 300 fois.

On a affaire à une loi binomiale de d'espérance $\mu = 300 \cdot \frac{1}{6} = 50$ et d'écart-type $\sigma = \sqrt{300 \cdot \frac{1}{6} \cdot \frac{5}{6}} \cong 7,07$.

Pour calculer la probabilité que le nombre de fois que le dé retombe sur 6 soit compris entre 35 et 45, on calcule les nouvelles bornes

$$a = \frac{3,5 - 0,5 - 50}{7,07} \cong -2,19 \text{ et } b = \frac{45 + 0,5 - 50}{7,07} \cong -0,78.$$

Sur la table, on y trouve $\Phi(-2,19) = 1 - \Phi(2,19) = 1 - 0,9857 = 0,0143$ et $\Phi(-0,78) = 1 - \Phi(0,78) = 1 - 0,7823 = 0,2177$. On en conclut que

$$P(35 \leq X \leq 45) = \Phi(-0,78) - \Phi(-2,19) = 0,2177 - 0,0143 = 20,34\%.$$

Déterminons alors la probabilité que le nombre de fois où le dé retombe sur 6 soit d'au moins 60 fois. On calcule la nouvelle borne

$$a = \frac{60 - 0,5 - 50}{7,07} \cong 1,34.$$

A l'aide de la table, on trouve $\Phi(1,34) = 0,9099$. On en déduit que

$$P(X \geq 60) = 1 - \Phi(1,34) = 1 - 0,9099 = 9,01\%.$$

Déterminons enfin jusqu'à combien de fois le dé doit retomber sur 6 pour que la probabilité soit de 95%. Autrement dit, on cherche B tel que $P(X \leq B) = 95\%$.

On pose

$$\begin{aligned} P(X \leq B) &= 0,95 \\ \Phi(b) &= 0,95 \\ b &= 1,64. \end{aligned}$$

On résout alors

$$\begin{aligned} \frac{B+0,5-50}{7,07} &= 1,64 \\ B - 49,5 &= 11,5948 \\ B &= 61,09. \end{aligned}$$

D'où $P(X \leq 62) \cong 95\%$.

Chapitre 4

Introduction à la statistique inférentielle

4.1 Introduction

Pour recueillir des informations sur une population statistique, on dispose de deux méthodes.

— La méthode exhaustive ou recensement : on examine chacun des individus de la population selon le ou les caractères étudiés (exemple : le recensement fédéral). On décrit alors les propriétés de la population avec les outils de la statistique descriptive.

— La méthode des sondages : on examine une fraction seulement, un échantillon de la population (exemple : dans une fabrique d'allumettes, on prélève à la sortie de fabrication des échantillons sur lesquels on contrôle la qualité du produit). On cherche alors à tirer des conclusions concernant la population-mère sur la base d'observations réduites à des échantillons. Les méthodes développées à cette fin ressortissent à la statistique inférentielle.

Il arrive fréquemment que l'on doive rejeter la méthode exhaustive, soit à cause de son coût ou de sa durée, soit en raison de ses effets destructeurs ; pour connaître la durée de vie d'une production d'ampoules ou savoir si des allumettes s'enflamment, il faut détruire les objets testés. Dans ces cas, il n'y a pas d'autres moyens que d'effectuer des sondages.

4.1.1 Échantillonnage et estimation

L'échantillonnage est un processus de déduction, qui consiste à passer d'une population totale, dont les caractéristiques sont connues, à un échantillon de cette population. L'estimation consiste à induire, à partir des résultats observés sur un échantillon particulier, des résultats concernant la population globale.

Exemple. (Échantillonnage) La production d'une machine comprend, en moyenne, une pièce défectueuse sur six. Si l'on prélève un échantillon de 10 pièces sur cette production, combien sont défectueuses ? On est ici en présence d'une loi binomiale

$$P(X = k) = C_{10}^{k} \cdot \left(\frac{1}{6}\right)^{k} \cdot \left(\frac{5}{6}\right)^{n-k}.$$

Parmi tous les échantillons qui peuvent être extraits, on sait que $P(X = 0) \cong 16\%$ ne contiendront aucune pièce défectueuse, $P(X = 1) \cong 32\%$ en compteront une, $P(X = 2) \cong 29\%$ en compteront 2, etc. . . Ainsi, si l'on connaît la composition de la population, on peut, sous certaines conditions, en déduire a priori la composition de l'échantillon extrait. Plus la taille de

l'échantillon est grande, plus sa structure a de chance d'être voisine de celle de la population globale.

Exemple. (Estimation) De la production d'une machine, on prélève un échantillon de 120 pièces et on observe que 23 d'entre elles sont défectueuses. Quelle proportion de pièces défectueuses cette machine produit-elle effectivement ? On cherche à répondre à cette question en trouvant un intervalle de confiance (une fourchette) dans lequel il est probable que se touve la vraie valeur de la proportion de pièces défectueuses. On pourra établir, par exemple, que la probabilité de trouver la proportion exacte dans l'intervalle $[10\%; 23\%]$ est égale à 95%.

4.1.2 Estimation

La valeur inconnue d'une population, à estimer à partir d'un échantillon, est appelée un paramètre. Les paramètres à estimer le plus souvent sont la moyenne, l'écart-type, la variance ou la fréquence (ou pourcentage). Pour distinguer les paramètres de la population de leurs estimations sur échantillons, on utilise des symboles différents.

	Paramètres de la population	Estimations sur l'échantillon
Moyenne	μ	$\overline{x}$
Écart-type	σ	s
Fréquence	π	p
Taille	N	n

La moyenne arithmétique $\overline{x}$, de même que l'écart-type s, fournissent un seul "point", une seule valeur numérique, comme estimation du paramètre μ, respectivement σ. Une telle estimation est dite *estimation ponctuelle* du paramètre de la population.

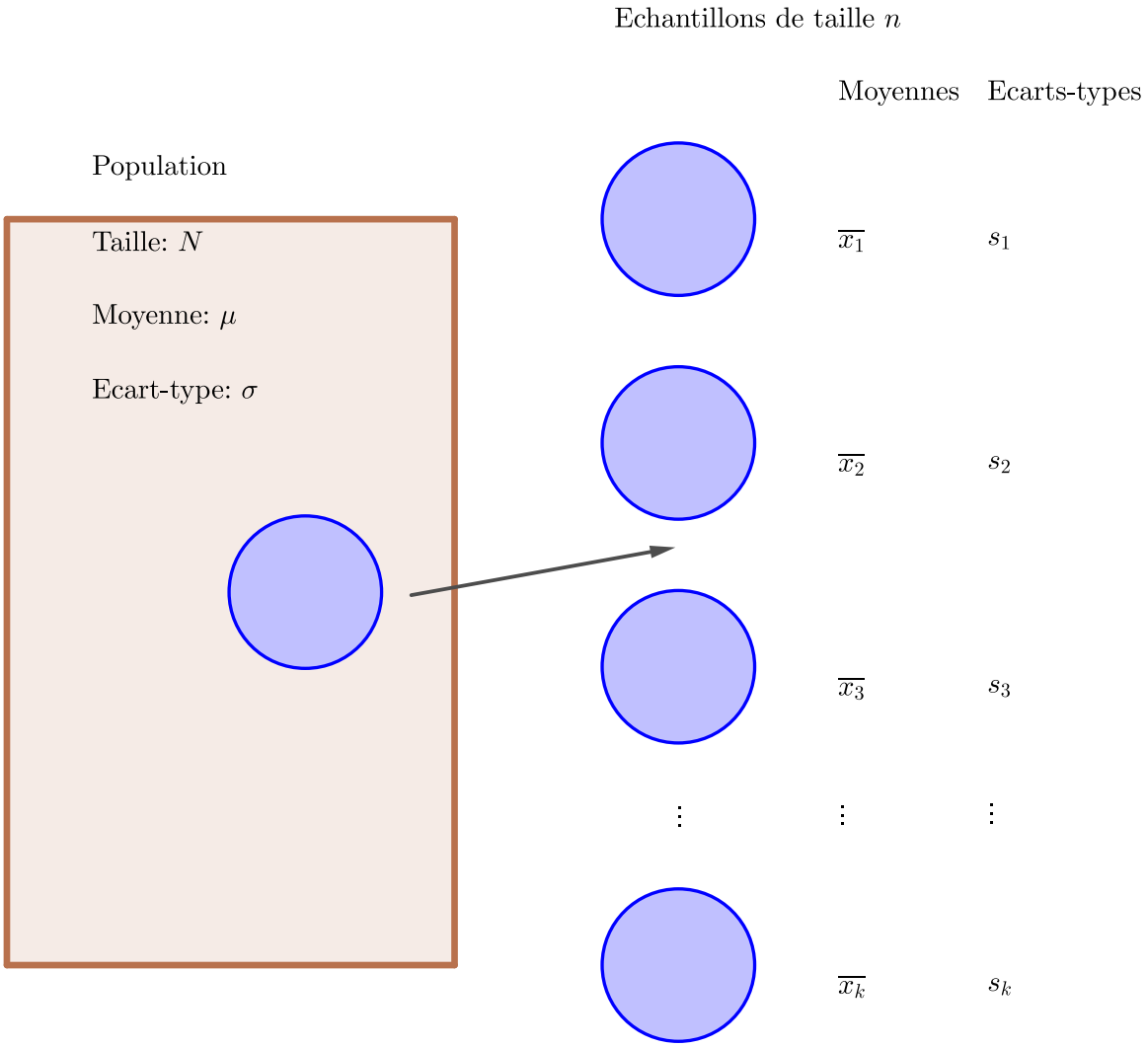
L'estimation ponctuelle d'un paramètre consiste donc en l'évaluation de la valeur du paramètre de la population à l'aide d'une valeur unique calculée sur un échantillon. Mais cette valeur peut varier suivant le choix de l'échantillon. Par exemple, si n et N désignent respectivement les tailles de l'échantillon et de la population, il y aura C_N^n échantillons possibles et donc, autant d'estimations du paramètre. Si la population est infinie, le nombre d'échantillons possibles est illimité.

4.2 Intervalle de confiance pour la moyenne

Supposons qu'on choisisse d'extraire d'une population donnée de moyenne μ, tous les échantillons de taille fixe n. Pour chacun d'eux, on calcule la valeur de la variable aléatoire

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n-1}}}.$$

dans laquelle $\overline{x}$ et s désignent respectivement la moyenne et l'écart-type de l'échantillon.



Lorsque la taille n des échantillons est suffisamment grande (on convient en général de la borne $n \geq 30$), la distribution d'échantillonnage de la variable t est approximativement une distribution normale, que la distribution de la population soit normale ou non.

De plus, lorsque la distribution de la population est normale, la distribution d'échantillonnage est une distribution normale.

Ce résultat permet de définir un intervalle de confiance pour la moyenne inconnue μ de la population à partir d'un seul échantillon. On sait en effet, par exemple, qu'il y a 95% de chances que

$$-1,96 < t < 1,96$$

c'est-à-dire que

$$-1,96 < \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n-1}}} < 1,96$$

ou encore que

$$\overline{x} - 1,96 \cdot \frac{s}{\sqrt{n-1}} < \mu < \overline{x} + 1,96 \cdot \frac{s}{\sqrt{n-1}}.$$

Ces deux dernières inégalités constituent un encadrement de la moyenne μ, laquelle va donc se trouver (avec 95% de chances) dans l'intervalle dont les bornes sont

$$\overline{x} \pm 1,96 \cdot \frac{s}{\sqrt{n-1}}.$$

Autrement dit, le risque de trouver la moyenne effective μ de la population hors de cet intervalle centré sur la moyenne de l'échantillon $\overline{x}$ n'est que de 5%. De manière générale, les bornes d'un intervalle de confiance de niveau α pour la moyenne sont données par

$$\overline{x} \pm z_\alpha \cdot \frac{s}{\sqrt{n-1}}$$

avec

α	99,73%	99%	98%	96%	95,45%	95%	90%	80%	68,27%	50%
z_α	3,00	2,58	2,33	2,05	2,00	1,96	1,645	1,28	1,00	0,6745

Exemple. On a prélevé, par sondage aléatoire, un échantillon de 10'000 ménages dans une région qui comptait 700'000 ménages lors du recensement. Sur cet échantillon, on a observé, pour un mois particulier, une consommation moyenne de 950 euros par ménage, avec un écart-type de 700 euros. Déterminer un intervalle de confiance à 95% se rapportant à la consommation moyenne des ménages dans cette région.

Ici $n = 10'000$, $N = 700'000$, $\overline{x} = 950$ et $s = 700$.

$$\frac{s}{\sqrt{n-1}} = \frac{700}{\sqrt{9'999}} \cong 7.$$

L'intervalle de confiance à 95% pour la moyenne μ est donc donné par

$$950 - 1,96 \cdot 7 < \mu < 950 + 1,96 \cdot 7,$$

donc

$$936,28 < \mu < 963,72.$$

4.3 Intervalle de confiance pour la fréquence

Soit une population (considérée comme infinie) au sein de laquelle un pourcentage π d'individus présente un certain caractère. À tout échantillon de taille n extrait de cette population, on associe la variable aléatoire

$$t = \frac{p - \pi}{\sqrt{\frac{p \cdot (1-p)}{n-1}}}$$

dont la valeur dépend de la fréquence p du caractère observée dans l'échantillon.

On peut démontrer que, pour autant que la taille de l'échantillon soit assez grande et que π ne soit trop voisin ni de 0, ni de 1, alors la variable aléatoire t suit une loi normale centrée réduite. Il s'ensuit que l'intervalle de confiance de niveau α pour la fréquence effective dans la population, déduit de l'observation de l'échantillon, aura pour bornes

$$p \pm z_\alpha \cdot \sqrt{\frac{p \cdot (1-p)}{n-1}}.$$

Exemple. Le daltonisme est une anomalie au niveau de l'une des parties constitutives de l'oeil qui donne naissance à de nombreux dysfonctionnements de la vision ou de la perception des couleurs. Cherchant à déterminer la fréquence d'hommes daltoniens dans la population, on extrait un échantillon de 626 hommes et on constate que 67 d'entre eux sont daltoniens.

a) Déterminer un intervalle de confiance de niveau 95% pour la fréquence du daltonisme dans la population.

Ici on a $n = 626$, $p = \frac{67}{626} = 0,10703$, $z_\alpha = 1,96$. Les bornes de l'intervalle de confiance sont données par

$$p \pm z_\alpha \cdot \sqrt{\frac{p \cdot (1-p)}{n-1}} = 0,10703 \pm 1,96 \cdot \sqrt{\frac{0,10703 \cdot 0,8930}{625}},$$

c'est-à-dire $0,08279$ et $0,13127$.

D'où l'intervalle de confiance à 95% : $[8,279\%; 13,127\%]$.

b) Quel est le niveau de confiance de l'intervalle $[7,822\%; 13,584\%]$?

On cherche z tel que

$$0,10703 + z \cdot \sqrt{\frac{0,10703 \cdot 0,8930}{625}} = 0,13584$$

et on trouve $z = 2,33$. L'intervalle a donc un niveau de confiance de 98%.

4.4 La décision statistique

Dans la pratique, on est souvent amené à prendre des décisions diverses au sujet d'une population à partir de la seule information fournie par un échantillon. De telles décisions sont appelées décisions statistiques. On voudra, par exemple, décider à partir d'un échantillon si un sérum est effectivement efficace pour guérir une maladie, si une méthode pédagogique est meilleure qu'une autre, si une machine est toujours bien réglée, si un dé est équilibré, etc.

4.4.1 Les hypothèses statistiques

Pour parvenir à une décision à partir d'un échantillon, il est indiqué de faire des hypothèses sur la population de laquelle il est extrait. De telles hypothèses statistiques, qui peuvent être vraies ou fausses, sont en général des affirmations relatives à la distribution de probabilité de la population.

Dans la plupart des cas, on formulera une hypothèse statistique dans le seul but de la rejeter. Si l'on veut, par exemple, établir qu'une pièce de monnaie est déséquilibrée, on posera l'hypothèse qu'elle est parfaite et que la probabilité qu'elle tombe sur pile est égale à $\frac{1}{2}$. De la même façon, si l'on veut prouver qu'un procédé de fabrication est meilleur qu'un autre, on supposera qu'il n'y a aucune différence entre eux (c'est-à-dire que toutes les différences observées sont uniquement dues à des fluctuations d'échantillonnage de la même population). Une telle hypothèse est une *hypothèse nulle*, on la désigne par le symbole H_0. Toute hypothèse qui diffère d'une hypothèse donnée est une *hypothèse alternative*. En regard de l'hypothèse nulle $p = \frac{1}{2}$, les hypothèses $p = \frac{3}{4}$, $p \neq \frac{1}{2}$ ou $p > \frac{1}{2}$ sont des hypothèses alternatives. On désignera par le symbole H_a toute hypothèse alternative de l'hypothèse nulle.

4.4.2 Tests d'hypothèses

Imaginons que, sous une hypothèse particulière supposée vraie, l'on ait trouvé que les observations d'un échantillon aléatoire diffèrent sensiblement des observations espérées quand seul le hasard est responsable des variations d'échantillonnage. On dira alors que les différences observées sont *significatives* et l'on sera enclin à rejeter l'hypothèse (ou au moins à ne pas l'accepter) du fait qu'il y a de fortes chances (par exemple 95%) qu'elle ne soit pas vraie. Si, par exemple, on lance 20 fois une pièce de monnaie qui tombe 16 fois sur pile, on aura tendance à rejeter l'hypothèse que la pièce est équilibrée, bien qu'il soit concevable, mais peut probable, que tel soit effectivement le cas.

On appelle *tests d'hypothèses*, *tests de signification* ou *règles de décision*, les procédés qui permettent de décider si des hypothèses sont vraies ou fausses ou de déterminer si des échantillons observés diffèrent significativement des résultats supposés.

4.4.3 Erreurs de 1ère et de 2e espèces - Niveau de signification

Si l'on rejette une hypothèse quand elle doit être acceptée, on dit qu'on commet une erreur de 1ère espèce. Par contre, si l'on accepte une hypothèse qui doit être rejetée, on dit qu'on fait une erreur de 2e espèce. Chacun de ces cas correspond à une décision erronée ou bien à une erreur de jugement. Pour qu'un test d'hypothèse soit efficace, celui-ci doit être conçu de manière à minimiser les erreurs de décision.

Lorsqu'on teste une hypothèse, la probabilité avec laquelle on est disposé à risquer une erreur de 1ère espèce est appelé seuil de signification du test. Cette probabilité α est en général spécifiée avant d'extraire tous les échantillons. Dans la pratique, un seuil de signification de 5% ou 1% est commode, bien que d'autres valeurs puissent également être utilisées. Si l'on choisit par exemple 5% comme seuil de signification en construisant un test d'hypothèse, il y a alors 5 chances sur 100 que l'on rejette l'hypothèse alors qu'elle doit être acceptée. Cela signifie que l'on est sûr à 95% d'avoir pris la bonne décision. On dit alors que l'on a rejeté l'hypothèse à un seuil de signification égal à 5%, ce qui signifie que l'on a tort avec une probabilité de 5%.

On peut représenter les erreurs de première espèce (de probabilité α) et de seconde espèce (de probabilité β) d'un test d'hypothèse dans le tableau suivant.

Situation	Accepter H_0	Rejeter H_0
H_0 vraie	$1-\alpha$	α
H_0 fausse	β	$1-\beta$

Afin d'illustrer ces différentes situations avec un exemple de la vie courante, supposons qu'une personne pose les hypothèses suivantes un matin avant de sortir pour se rendre à son travail :

Hypothèse nulle H_0 : il va pleuvoir ;
Hypothèse alternative H_a : le temps sera sec.

Supposons qu'en observant un instant le ciel, la personne rejette l'hypothèse de la pluie et qu'elle sorte sans parapluie. Si le temps s'avère sec par la suite, la personne aura pris la bonne décision. En revanche, s'il pleut, alors elle aura commis une erreur : celle d'avoir rejeté une hypothèse vraie. Il s'agit donc d'une erreur de première espèce.

Dans le cas contraire, si la personne accepte l'hypothèse selon laquelle il va pleuvoir et qu'elle emporte un parapluie, alors elle aura pris une bonne décision s'il pleut effectivement. Elle aura commis une erreur si le temps se révèle sec : celle d'avoir accepté une hypothèse

fausse. Il s'agit là d'une erreur de seconde espèce. Ces différents cas peuvent être repésentés dans le tableau suivant.

Situation	Accepter H_0 prendre le parapluie	Rejeter H_0 laisser le parapluie
H_0 vraie : il pleut	bonne décision : $1-\alpha$	erreur de la première espèce : α
H_0 fausse : temps sec	erreur de la seconde espèce : β	bonne décision : $1-\beta$

4.5 Tests d'hypothèses pour une moyenne

On présente dans la suite, au travers d'exemples, trois tests d'hypothèses relatifs à une moyenne : le test *unilatéral à droite*, le test *unilatéral à gauche* et le test *bilatéral*.

4.5.1 Test unilatéral à droite

Exemple. Une chaîne de montage de réfrigérateurs fonctionne de façon optimale si le temps de passage dans la chaîne n'excède pas 30 minutes. Un échantillon aléatoire de 100 réfrigérateurs a été choisi et le temps de passage a été mesuré pour chacun d'eux. On a constaté une durée moyenne de passage $\overline{x} = 31$ minutes et un écart-type s égal à 6 minutes. Le fonctionnement de cette chaîne est-il optimal ?

Un temps de passage inférieur à 30 minutes n'étant pas problématique, on pose les hypothèses de départ suivantes pour la moyenne effective des temps de passage

$$H_0 : \mu = 30, \qquad H_a : \mu > 30.$$

On choisit un seuil de signification de 5%. Avec l'hypothèse nulle, et comme la taille $n = 100$ de l'échantillon est plus grande que 30, il y a une probabilité de 5% que la variable aléatoire

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n-1}}}$$

pour un échantillon quelconque, dépasse le seuil critique $t_c = 1,645$ (dans la table, on trouve en effet $\Phi(1,645) = 0,95$. Pour l'échantillon extrait, on a

$$t = \frac{31-30}{\frac{6}{\sqrt{99}}} \cong 1,685 > t_c = 1,645.$$

Comme $t > t_c = 1,645$, on rejette l'hypothèse nulle et on en conclut que le temps de passage moyen est significativement supérieur à 30 minutes.

Remarque. Rejeter l'hypothèse nulle quand

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n-1}}} > t_c,$$

revient à décider de son rejet quand $\overline{x} > \mu + t_c \cdot \frac{s}{\sqrt{n-1}}$, c'est-à-dire quand la moyenne $\overline{x}$ de l'échantillon est plus grande que le seuil inférieur de la région de rejet : $\left]\mu + t_c \cdot \frac{s}{\sqrt{n-1}}; +\infty\right[$.

4.5.2 Test unilatéral à gauche

Exemple. Un producteur de parfum entend s'assurer que ses flacons ont bien une contenance de 40 ml. Un échantillon aléatoire de 50 flacons donne une moyenne de 39 ml avec un écart-type de 4 ml.

On va tester les hypothèses suivantes

$$H_0 : \mu = 40, \qquad H_a : \mu < 40.$$

en choisissant un seuil de signification de $\alpha = 1\%$. Comme $\Phi(-2,33) = 1\%$, y a une probabilité de 1% que la variable aléatoire

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n-1}}} < t_c = -2,33,$$

pour un échantillon quelconque, soit inférieure au seuil critique $t_c = 2,33$. Or

$$t = \frac{39 - 40}{\frac{4}{\sqrt{49}}} \cong -1,75.$$

Ainsi, comme $t > t_c = -2,33$, on ne rejette pas l'hypothèse nulle, puisqu'elle ne se trouve pas dans la région de rejet $]-\infty; -2,33[$. Ce qui s'exprime aussi par le fait que la moyenne observée, à savoir 39, est plus grande que le seuil critique de rejet

$$39 > \mu + t_c \cdot \frac{s}{\sqrt{n-1}} = 40 - 2,33 \cdot \frac{4}{7} = 38,67.$$

4.5.3 Test bilatéral

Exemple. Une entreprise fabrique des câbles en acier. Elle veut vérifier, pour un important lot de production, si le diamètre des câbles est conforme aux normes qui imposent un diamètre de 0,9 cm. A cette fin, un échantillon aléatoire de 100 câbles est extrait. Les mesures font apparaître un diamètre moyen de 0,93 cm avec un écart-type de 0,12 cm.

On va tester les hypothèses suivantes

$$H_0 : \mu = 0,9, \qquad H_a : \mu \neq 0,9.$$

en choisissant un seuil de signification $\alpha = 5\%$. Comme $1 - \Phi(1,96) = 2,5\%$, il y a une probabilité de 95% que la variable aléatoire

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n-1}}},$$

pour un échantillon quelconque, soit dans l'intervalle d'acceptation $]-1,96; 1,96[$. Autrement dit, il y a 5% de chances de trouver cette variable dans la région de rejet $|t| > 1,96$.

Or $t = \frac{0,93-0,9}{\frac{0,12}{\sqrt{99}}} \cong 2,48$. Ainsi, comme $|t| > t_c = 1,96$, on rejette l'hypothèse nulle, puisqu'elle se trouve dans la région de rejet. Ce qui s'exprime aussi par le fait que la moyenne observée, à savoir 0,93, est plus grande que le seuil critique de rejet

$$0,93 > \mu + t_c \cdot \frac{s}{\sqrt{n-1}} = 0,9 + 1,96 \cdot \frac{0,12}{\sqrt{99}} = 0,923.$$

4.5.4 Seuils critiques

Le tableau ci-dessous donne les valeurs critiques pour des tests unilatéraux ou bilatéraux et les niveaux de signification correspondants.

Niveaux de signification	10%	5%	1%	0,5%	0,2%
Valeurs critiques t_c pour tests unilatéraux	$-1,28$ ou 1,28	$-1,645$ ou 1,645	$-2,33$ ou 2,33	$-2,58$ ou 2,5	$-2,88$ ou 2,88
Valeurs critiques t_c pour tests bilatéraux	$-1,645$ et 1,645	$-1,96$ et 1,96	$-2,58$ et 2,58	$-2,81$ et 2,8	$-3,08$ et 3,08

4.6 Tests d'hypothèses pour une fréquence

Supposons que π désigne le pourcentage d'individus présentant un certain caractère dans une population. En théorie de l'échantillonnage, on peut démontrer que la variable aléatoire définie par l'expression

$$t = \frac{p - \pi}{\sqrt{\frac{\pi \cdot (1-\pi)}{n}}}$$

dans laquelle n désigne la taille (n assez grande, c'est-à-dire $n \geq 30$) de l'échantillon et p le pourcentage d'individus de l'échantillon présentant le caractère considéré, suit une loi normale centrée réduite. Ce résultat est valable pour autant que les conditions suivantes soient remplies :

$$n \geq 30,\ n \cdot \pi \geq 5 \text{ et } n \cdot (1-\pi) \geq 5.$$

Le test d'hypothèse pour une fréquence repose alors sur les mêmes principes que le test d'hypothèses pour une moyenne. Si π et p dénotent respectivement les fréquences d'un caractère dans une population et dans l'un de ses échantillons de taille n, alors on posera les hypothèses de départ suivantes :

$$\begin{array}{lll} H_0 : & \pi = \pi_0 : & \text{valeur présumée de la fréquence ;} \\ H_a : & \pi > \pi_0 : & \text{pour un test unilatéral à droite ;} \\ H_a : & \pi < \pi_0 : & \text{pour un test unilatéral à gauche ;} \\ H_a : & \pi \neq \pi_0 : & \text{pour un test bilatéral.} \end{array}$$

L'hypothèse nulle sera rejetée si la variable aléatoire

$$t = \frac{p - \pi}{\sqrt{\frac{\pi \cdot (1-\pi)}{n}}}$$

prend une valeur :

— supérieure au seuil critique t_c, pour une test unilatéral à droite ;

— inférieure au seuil critique t_c, pour une test unilatéral à gauche ;

— à l'extérieur de l'intervalle d'acceptation $]-t_c; t_c[$ pour un test bilatéral.

Le seuil critique est défini en fonction du niveau de signification du test.

Exemple. Sur la base d'un sondage effectué auprès de 200 citoyens, il ressort qu'un candidat aux élections reccueille 52% des intentions de vote. On veut alors tester l'hypothèse, au niveau de signification de 5%, que ce candidat sera effectivement élu. Pour ce faire, on pose les hypothèses

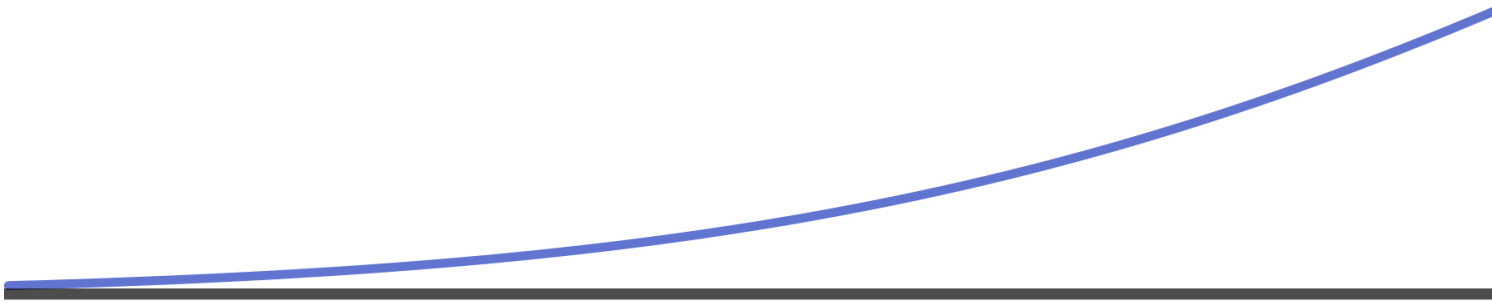
$$H_0 : \pi_0 = 0,5, \qquad\qquad H_a : \pi_0 > 0,5.$$

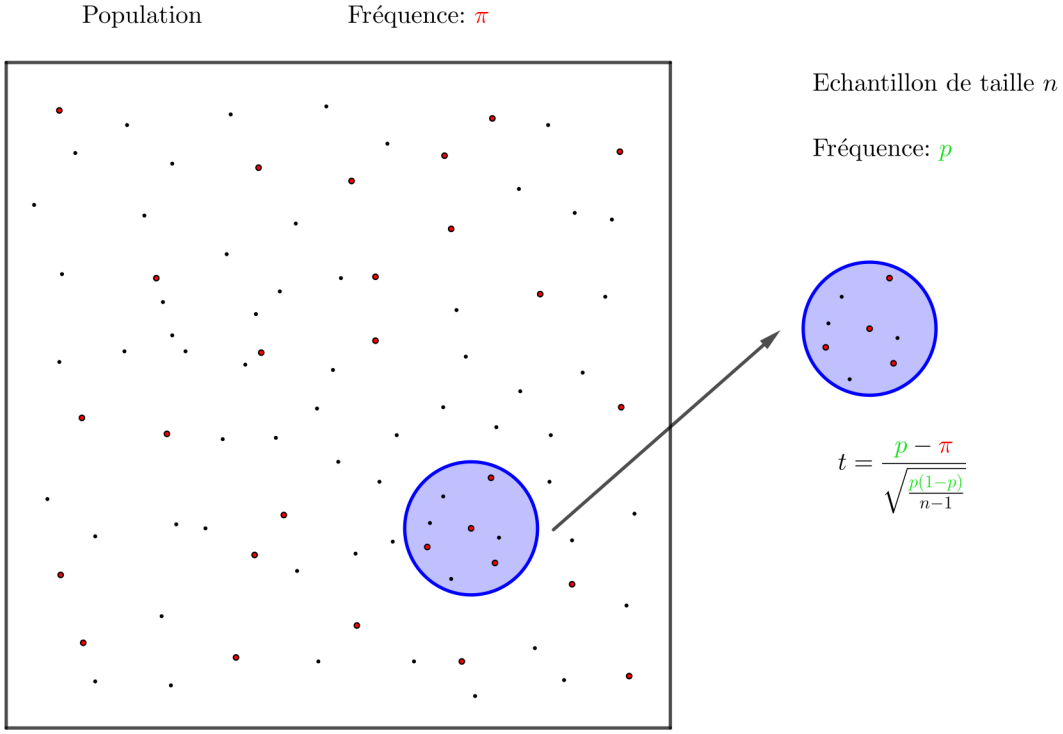
La valeur critique, correspondant à un test unilatéral à droite de niveau de signification $\alpha = 5\%$, est $t_c = 1,645$. Comme le ratio

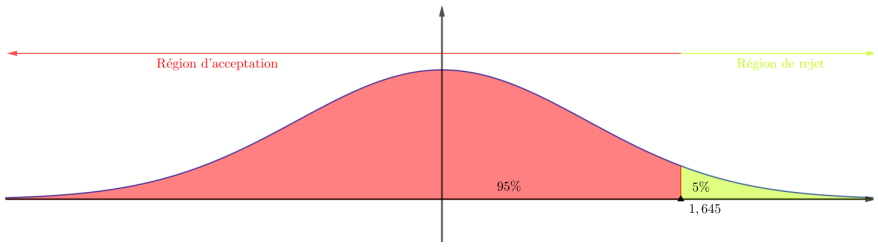
$$\frac{p - \pi}{\sqrt{\frac{\pi \cdot (1-\pi)}{n}}} = \frac{0,52 - 0,5}{\sqrt{\frac{0,5 \cdot 0,5}{200}}} = 0,565 < 1,645.$$

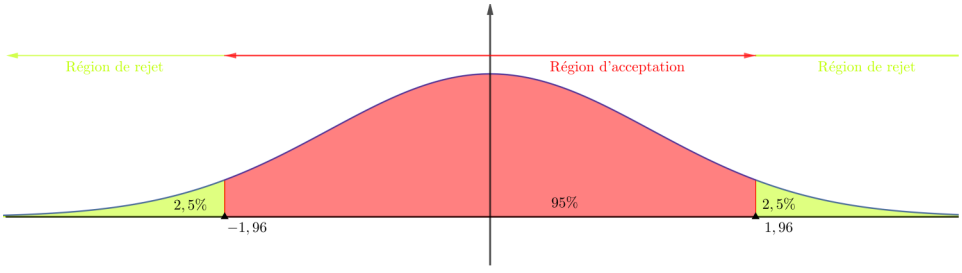
est inférieur au seuil de rejet, l'hypothèse nulle ne peut pas être rejetée. De ce fait, on ne peut pas conclure à la vraisemblance de l'élection du candidat.

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000




Population
Fréquence: π
Echantillon de taille n
Fréquence: p
$t = \frac{p - \pi}{\sqrt{\frac{p(1-p)}{n-1}}}$


Région d'acceptation
Région de rejet
95%
5%
1,645


Région de rejet
Région d'acceptation
Région de rejet
2,5%
95%
2,5%
−1,96
1,96

Bibliographie

[1] Christian Affolter, Sylvie Gonano, *Support de cours*, HEG Arc.

[2] Hubert Bovet, *Probabilités*, Editions Polymaths, 2005.

[3] Hubert Bovet, *Statistique*, Editions Polymaths, 2002.

[4] CRM, *Probabilités*, Editions du Tricorne, 2005.

[5] Olivier Dessibourg , *La vérité cachée des chiffres* , Le Temps, 2010.

[6] Yadolah Dodge, *Premiers pas en statistiques*, Springer, 1999.

[7] Yadolah Dodge, Valentin Rousson *Analyse de régression appliquée*, Dunod, 2004.

[8] Jean-Pierre Favre, *Mathématiques pour la maturité professionnelle*, Editions Digilex, 2016.

[9] Peter Frommenwiler, Kurt Studer, *Algèbre et analyse de données*, Cornelsen, 2014.

[10] Jean-Philippe Javet, *Supports de cours*, Gymnase de Morges.

[11] Dagris Musitelli, *Introduction aux statistiques*, CPLN-EPC.

[12] Stephane Perret, *Supports de cours*, Lycée Cantonal de Porrentruy.

[13] Frédéric Schütz, *Statistiques et fantaisies journalistiques*, CFJM, 2015.

[14] Maxime Zuber, *Supports de cours*, HEG Arc.

Index