Statistiques, informatique et modélisation

LP341

Année 2005-2006

# Statistiques

Philippe Tourrenc, Anne-Laure Melchior

Université Pierre et Marie Curie

# Table des matières

°°°°°°°°°°°°°°°°°°°°°°°°°°°°°°

# Avertissement

Ce polycopié s'adresse à vous, étudiants de la licence SVP. Ce ne veut pas être un texte abstrait, aseptisé et bien pensant, ni même un cours "bien léché" mais une incitation, une provocation à la réflexion.

Les méthodes statistiques constituent un outil scientifique puissant dont le champ d'application est si vaste que l'on n'en voit pas les limites. Nous pourrions enseigner le maniement de cet outil dans tel ou tel domaine spécialisé. Très vite vous vous endormiriez et nous aussi. Nous voulons au contraire traiter de ce que tout le monde connaît, ou tout au moins de ce dont tout le monde a entendu parler.

Les statistiques ne sont pas une création du huitième jour mais un ensemble d'outils qui furent fabriqués dans des buts précis (pas toujours recommandables), avec les moyens que mettaient à la disposition des hommes leur imagination, leur liberté et leur expérience. C'est bien moins le maniement des outils les plus sophistiqués que la nature, les buts et les potentialités des statistiques que nous voulons vous présenter : suivant Montaigne, nous parions sur une tête bien faite plutôt que bien pleine.

Nous espérons ainsi susciter l'admiration pour nos prédécesseurs tout en développant en vous le virus de la critique et du doute scientifique. Si, de surcroît, vous vous appropriez cette liberté de vos ancêtres et qu'un jour vous créez vos propres outils, même si, sans gloire, ils ne servent qu'à vous, alors nous aurons gagné notre pari. Si demain, jeune néophyte face à une situation bien connue, sauf de vous même, vous vous adaptez, si vous assimilez les méthodes à mettre en oeuvre et si vous en comprenez les raisons, alors nous aurons gagné notre pari. Si vous ne vous laissez pas trop manipuler par les affirmations officielles, les statistiques des uns et des autres, alors nous auront gagné notre pari.

Nous aurons gagné notre pari mais bien sûr nous ne le saurons pas et vous aurez sans doute oublié que c'était un pari...

# Introduction

*Statistique*(*) : le mot n'a pas trois siècles mais la pratique en est multi-millénaire. Tout a commencé par des études démographiques, il y a 5000 ans déjà chez les Sumériens(†).

Les Chinois, les Egyptiens et plus près de nous les Romains recensaient les personnes et les biens. On se souvient que Marie donna naissance à Jésus dans une étable de Bethlehem lors du voyage qu'elle entreprit avec Joseph vers la Judée pour qu'il s'y fasse recenser. C'était il y a 2000 ans environ.

Les Carolingiens voulaient connaître leur patrimoine et les Capétiens l'état du royaume. Nul ne doit échapper à l'impôt et tous les récidivistes, ceux qui par deux fois volent un pain, doivent finir aux galères : tolérance zéro. Alors, tout au long des siècles on compte, on classe, on dresse des registres, on ouvre un "*bureau de la statistique républicaine*". On connaît les richesses du pays. Il devient impossible d'échapper à la conscription, napoléonienne d'abord, républicaine ensuite.

Intimement liées, les statistiques et la taxonomie se développent pour le plus grand profit des sciences naturelles : c'est à partir des mêmes observations et des mêmes classifications, que l'évolutionisme de Darwin et le transformisme de Lamarck s'opposent ; ce sont les statistiques sur ses petits pois qui conduisent Mendel aux lois de l'hérédité.

Cette frénésie de mensuration et de classification n'épargne pas l'homme. On apprend à distinguer les brachycéphales des dolichocéphales, on mesure l'angle facial, la taille des oreilles et bien d'autres choses. Mais un jour tout dérape : on en tire des conclusions. L'outil scientifique, devenu outil de répression avec l'anthropométrie‡ de Bertillon, devient un outil d'exclusion§ et d'extermination.

Tout en se développant, les statistiques évoluent. Initialement descriptives, elles atteignent au $18^{ème}$ siècle un haut degré de sophistication dans ce domaine, principalement sous l'impulsion de l'école allemande. Dans le même temps le calcul des probabilités se développe : Jacques Bernoulli(¶) démontre la loi des grands nombres, Thomas Bayes (1702-1761) introduit les probabilités des causes.

Dès lors les statistiques deviennent un outil qui permet de prédire avant même de comprendre. C'est de cette époque que l'on peut dater la naissance des statistiques mathématiques et des techniques de sondage, encore que des méthodes de collecte, de

*Du latin moderne (*statisticus = relatif à l'état*), le terme passe à l'Allemand avec son acception moderne (*Statistik*), puis au Français vers la fin du $18^{ème}$ siècle.

†Nous serions heureux que la lecture de cette introduction donne l'envie d'ouvrir une encyclopédie, un dictionnaire ou d'aller sur la toile. Mais que l'on se rassure ! Nous ne prendrons pas le risque d'écrire un polycopié qui, du premier au dernier mot, devrait se lire un dictionnaire à la main.

‡Appliquée à l'homme, la biométrie devient anthropométrie.

§Les statistiques sont aujourd'hui encore un puissant outil de répression et d'exclusion couramment utilisé par les administrations. Attention aux dérapages !

¶Jacques Bernoulli (1654-1705) est l'oncle de Daniel Bernoulli (1700-1782) qui énonça le théorème d'hydraulique qui porte son nom.

traitement et d'extrapolation des données soient apparues en Angleterre dès la seconde moitié du $17^{ème}$ siècle.

Une fois de plus la démographie est à l'honneur. Initiés par l'astronome Edmond Halley (1656-1742), les travaux sur les probabilités de la durée de vie humaine sont approfondis en Hollande et en France. Les résultats en sont utilisés dès le début du $19^{ème}$ siècle pour calculer le montant des rentes et des viagers.

Au $19^{ème}$ siècle et au début du $20^{ème}$ siècle, les statistiques et le calcul des probabilités acquièrent les moyens de leur efficacité présente.

Les travaux de Laplace* et ceux de Gauss* permettent d'établir la loi normale. Une théorie des erreurs peut alors être construite en toute généralité, grâce au théorème central limite† dont les premières versions découlent des travaux de Moivre* et de Laplace.

Bienaymé* et Tchebychev* démontrent la célèbre inégalité qui porte leurs noms, tandis que les travaux de Francis Galton (1822-1911) et Karl Pearson‡ (1857-1936), cofondateurs de la revue *Biometrika,* conduisent aux notions de corrélation et de régression ainsi qu'aux tests d'hypothèses.

A ces travaux sont encore associés les noms de Egon Pearson (le fils de Karl), Ronald Fischer* , William Gosset* et George Snedecor* qui furent, avec bien d'autres, les artisans du remarquable développement des statistiques.

La véritable révolution se produit cependant au coeur du $20^{ème}$ siècle. C'est alors que se généralisent les techniques de sondage et d'échantillonnage et que probabilités et statistiques entrent massivement, de façon opérationnelle dans des domaines les plus variés. On assiste au développement de nombreuses méthodes, qui s'adaptent de façon spécifique à des activités précises et diverses ; parmi celles-ci, les méthodes d'ajustement en physique qui généralisent la méthode des moindre carrés due initialement à Gauss et à Legendre* et deviennent "analyse factorielle" pour le marketing, la publicité ou la sociologie. Les études statistiques et les tests de toutes natures permettent d'attribuer des causes aux phénomènes incompris et d'apprécier la validité des hypothèses posées, concernant aussi bien l'anthropologie que l'industrie, la médecine que la politique. Elles fournissent une aide à la décision et deviennent recherche opérationnelle pour la gestion des files d'attente, le commerce ou les transports.

Mentionnons tout particulièrement l'extraordinaire développement de la physique statistique et des méthodes d'analyse du signal tout au long du siècle dernier.

Complétées par la puissance des ordinateurs, les statistiques apparaissent aujourd'hui comme un outil magnifique. Mais derrière les progrès enregistrés pointent des dangers qui requièrent la vigilance de chacun. Nous avons appris à nous méfier de la biométrie dans un contexte que nous souhaitons révolu. L'enfer est pavé de bonnes intentions dit-on ; maintenons donc notre méfiance des outils statistiques, des fichiers pléthoriques et des moyens de les "croiser", des cartes à puces, des passeports infalsifiables, de la numérisation de l'oeil et de la voix, car nulle loi "informatique et liberté" ne saurait protéger, à elle seule, le citoyen de cette prison sans murs qui se construit jours après jour et qui, pour enfermer ceux qui doivent l'être, enferme aussi chacun de nous.

---

*Pierre-Simon de Laplace : 1749-1827. - Carl Friedrich Gauss : 1777-1855. -Abraham de Moivre :1667-1754. - Irénée-Jules Bienaymé : 1796-1878. - Pafnuti Lvovich Tchebychev : 1821-1894. - Ronald Aylmer Fischer 1890-1962 (Sir Ronald). – William Gosset (dit *Student*) 1876-1937. - Georges Waddel Snedecor : 1881-1974. - Adrien-Marie Legendre : 1752-1833.

†Théorème dont les hypothèses furent élargies par Lyapounov et Lindeberg au début du 20ème siècle.

‡K. Pearson est en outre le fondateur d'un laboratoire de biométrie et d'un laboratoire d'eugénique.

*Les statistiques constituent un outil universellement utilisé dans le domaine scientifique, mais aussi dans des domaines aussi variés que la production industrielle, la politique ou le marketing. La démonstration de l'existence d'un phénomène ou d'une loi physique, la mesure d'une grandeur quelconque, l'évaluation des risques aussi bien que l'essentiel des informations qui nous parviennent sont de nature statistique. Quelles que soient vos activités professionnelles futures, vous serez confrontés à cette réalité de façon passive ou critique (voire active) selon votre formation. Ce constat justifie un enseignement spécifique, dans le cadre d'une formation générale scientifique approfondie.*

Les motivations pédagogiques de l'unité d'enseignement *LP*341 de l'Université Pierre et Marie Curie sont ainsi formulées, tandis que le programme est le suivant

- *Statistiques descriptives*
- *Théorie des probabilités, variables aléatoires*
- *Principales lois : binomiale, Poisson, Gaussienne, chi2, Student*
- *Loi des grands nombres et théorème de la limite centrale*
- *Statistique inférentielle : estimation statistique, intervalle de confiance*
- *Description de données, analyse chi2 de la description, analyse chi2 de l'indépendance, régression linéaire*

Ainsi exprimés, les motifs qui président à l'enseignement des statistiques sont clairs. Cependant, pour en préciser le contexte et achever cette introduction, nous attirons l'attention sur un contresens à éviter.

Avec les statistiques nous disposons d'un outil puissant ; mais ce n'est qu'un outil qui permet de qualifier les réponses à certaines questions jugées pertinentes. En aucun cas cet outil ne peut apprécier la pertinence d'une question : si ce sont les séries chronologiques des cours de la bourse que l'on présente dans les banques de préférence aux séries chronologiques des licenciements ce n'est pas dû à la nature des statistiques mais à la seule question posée du plus grand profit possible, jugé comme essentiel.

Même si les bonnes questions sont posées, il reste encore quelques écueils à éviter. L'expérience montre que l'observateur est le plus souvent partisan : "nous allons faire une étude pour montrer que..." ; combien de fois n'avons nous entendu une telle entrée en matière([†]). Pour éviter que l'observateur ne soit acteur, il faut prendre deux précautions essentielles :

1. Il faut s'assurer d'une collecte "honnête" des données. Si on étudie une population à travers un échantillon, il faut s'assurer que l'échantillon est représentatif et qu'il a été choisi suivant les règles qui permettent l'usage des méthodes scientifiques employées. Dans un sondage d'opinion, il convient de ne pas influencer la réponse par la façon de poser la question. Dans tous les cas il faut se retenir d'éliminer les observations dérangeantes ; etc.
2. Le protocole qui conduit à l'interprétation des données doit être "honnêtement" défini *a priori*, antérieurement à tout résultat.

Si ces conditions n'étaient pas toutes deux remplies, alors, avec Benjamin Disraeli[‖], nous devrions convenir qu'après le mensonge et le fieffé mensonge, ce sont effectivement les statistiques qui constituent le suprême degré du mensonge.

[†]Il faut poser le doute en postulat ce qui conduit à utiliser l'expression "pour vérifier si..." et non "pour montrer que...".

[‖]Homme d'état anglais (1804-1881).

# Chapitre 1
# LES STATISTIQUES À UNE DIMENSION

## 1.1 Introduction

Chaque discipline dispose de son propre vocabulaire. Les mots qui y sont employés recouvrent des concepts et désignent des préoccupations spécifiques. Une promenade lexicographique nous permettra donc d'introduire les statistiques.

L'objet des statistiques est la description et l'étude des ***populations*** nombreuses. Il faut toutefois comprendre que *population* ne désigne pas nécessairement une population d'êtres humains ni même d'êtres vivants mais tout simplement un ensemble d'objets que l'on appelle ***individus***. Ces mots sont là parce que "statistiques" et "démographie" se sont développées de concert. Ainsi, un ensemble d'automobiles constitue une population dont les individus sont des automobiles.

La population que l'on étudie doit être définie avec précision. Une population est définie ***en extension,*** on dit aussi "de façon extensive" ou "*in extenso*", lorsqu'on donne la liste complète des individus qui la composent : par exemple l'ensemble des abonnés au téléphone qui figurent dans le bottin de la Creuse de 1992.

Lorsque la population est définie par une propriété qui permet d'en reconnaître les individus, on dit qu'elle est définie ***en compréhension*** : par exemple l'ensemble des chiens non tatoués, errant sur la commune de Bergerac.

Il est vain d'espérer décrire un individu, quel qu'il soit, de façon exhaustive. On sélectionne en général un ***caractère*** dont on soupçonne la pertinence pour l'étude en cours.

Un gaz est considéré comme une population de molécules dont nous supposons que seule l'énergie cinétique de translation nous intéresse ; les autres formes d'énergie ainsi que la nature et le nombre d'atomes qui constituent les molécules nous indiffèrent. L'énergie cinétique de translation constitue le caractère étudié.

Pour décrire ce caractère on dispose d'un descripteur qui peut prendre plusieurs valeurs. A chaque individu, $i$, on fait donc correspondre la valeur $X_i$, de son descripteur. On dit que la correspondance $i \mapsto X_i$ définit une ***variable*** $X$ (dans un autre contexte cette correspondance est appelée "une fonction").

La variable peut être ***numérique***, c'est le cas de l'énergie cinétique de translation dans l'exemple précédent ; mais ce n'est pas toujours le cas.

Considérons la population de ceux qui répondirent "non" au référendum du 29 mai 2005, en France. Nous cherchons à en décrire le niveau de revenu. A ce stade, les valeurs du descripteur sont numériques, mais pour simplifier nous introduisons 4 catégories, $A$, $B$, $C$ et $D$ caractérisées par le revenu $R$ correspondant :

$$A : R \leq S,\ B : S < R \leq 3S,\ C : 3S < R \leq 6S,\ D : 6S < R$$

où $S$ est le salaire minimal interprofessionnel de croissance (SMIC). Les valeurs possibles $X = A$, $B$, $C$ ou $D$ de la variable ne sont pas des valeurs numériques mais elles peuvent

être classées par ordre croissant de revenu. On dit que $X$ est une variable ***ordinale*** (qui peut être ordonnée).

Toujours dans le même but de décrire la population de ceux qui votèrent "non", nous nous intéressons à leur "catégorie sociale" défini par leur catégorie socioprofessionnelle (CSP). La CSP constitue le caractère étudié, nous utilisons comme descripteurs la classification de l'INSEE* qui peut prendre les valeurs suivantes : $A$ = agriculteurs exploitants, $B$ = artisans, $C$ = commerçants, $D$ = chefs d'entreprise, $E$ = cadres et professions intellectuelles supérieures, $F$ = professions intermédiaires, $G$ = employés, $H$ = ouvriers, $I$ = chômeurs n'ayant jamais travaillé. Il n'est pas possible d'ordonner ces valeurs, de les classer l'une par rapport à l'autre. On dit dans ce cas que la variable est ***non numérique***. Cela n'interdit pas cependant d'utiliser un numéro de code pour distinguer les diverses classes et par conséquent de représenter une variable non numérique par un nombre.

Considérons maintenant une population constituée d'individus auxquels sont associées deux variables, $X$ et $Y$. La population est, par exemple, l'ensemble des ménages formés de couples. Les revenus respectifs de chacun des membres du couple sont $X$ et $Y$. On utilise les 4 classes, $A$, $B$, $C$ et $D$ ci-dessus pour décrire les revenus.

A chaque individu (chaque ménage), on associe les valeurs $X$ et $Y$. On observe les couples de résultats suivants, correspondant à quatre populations étudiées $P_1$, $P_2$, $P_3$ et $P_4$ :

$$\begin{array}{ll} P_1 : & (A,A),(A,A),(B,C),(B,C),(B,C),(C,C),(C,C),(D,A),(D,A),(D,A) \\ P_2 : & (A,B),(A,B),(B,A),(B,A),(B,A),(B,C),(B,C),(C,D) \\ P_3 : & (A,A),(A,D),(B,C),(C,C),(C,C),(C,D),(D,C),(D,C),(D,D),(D,D) \\ P_4 : & (A,B),(A,C),(A,C),(B,B),(B,C),(B,C),(D,B),(D,C),(D,C) \end{array}$$

Les quatre populations considérées sont formées respectivement de 10, 8, 10 et 9 individus. On dit que leur ***cardinal*** est respectivement 10, 8, 10 et 9.

Chacune de ces populations définit une relation entre $X$ et $Y$.

- $P_1$ : Dans cette population, dès que la valeur de $X$ est connue, la valeur de $Y$ s'en déduit :

$$\begin{array}{lcccc} X = & A & B & C & D \\ & \downarrow & \downarrow & \downarrow & \downarrow \\ Y = & A & C & C & A \end{array}$$

La variable $Y$ apparaît comme une fonction de $X$.

- $P_2$ : Dans cette population, dès que la valeur de $Y$ est connue, la valeur de $X$ s'en déduit :

$$\begin{array}{lcccc} X = & B & A & B & C \\ & \uparrow & \uparrow & \uparrow & \uparrow \\ Y = & A & B & C & D \end{array}$$

La variable $X$ apparaît comme une fonction de $Y$.

Dans les populations $P_1$ et $P_2$ la liaison entre $X$ et $Y$ est appelée ***liaison fonctionnelle***.

- $P_3$ : Nous mettons en évidence la relation entre $X$ et $Y$ au moyen d'un tableau à double entrée :

*Institut National de la Statistique et des Etudes Economiques.

| $X \longrightarrow$ <br> $Y \downarrow$ | $A$ | $B$ | $C$ | $D$ |
|---|---|---|---|---|
| $A$ | $\times$ | | | |
| $B$ | | | | |
| $C$ | | $\times$ | $\times\times$ | $\times\times$ |
| $D$ | $\times$ | | $\times$ | $\times\times$ |

Le tableau précédent représente le graphe de la relation entre $X$ et $Y$. A l'évidence, sauf cas particuliers, la connaissance de $X$ ne détermine pas celle de $Y$, pas plus que la connaissance de $Y$ ne détermine celle de $X$ : la relation entre $X$ et $Y$ n'est pas une relation fonctionnelle. On dit que c'est une relation ***stochastique***[†] ou une relation ***aléatoire***.

- $P_4$ : La population $P_4$ définit la relation représentée par le graphe ci-dessous

| $X \longrightarrow$ <br> $Y \downarrow$ | $A$ | $B$ | $C$ | $D$ |
|---|---|---|---|---|
| $A$ | | | | |
| $B$ | $\times$ | $\times$ | | $\times$ |
| $C$ | $\times\times$ | $\times\times$ | | $\times\times$ |
| $D$ | | | | |

Cette population est remarquable car la valeur de $X$ étant fixée, la distribution des valeurs de $Y$ est $1/3$ pour $B$ et $2/3$ pour $C$. Elle est indépendante de la valeur de $X$, que ce soit $A$, $B$ ou $D$. On démontre alors que la valeur de $Y$ étant fixée, la distribution des valeurs de $X$ ne dépend pas de la valeur de $Y$. On vérifie ici que la distribution des valeurs de $X$ est $1/3$ pour $A$, $B$, $C$ ou $D$. Dans une telle situation remarquable, on dit que ***les variables sont stochastiquement indépendantes***. De façon pratique cela signifie que la connaissance de $X$ ne nous donne aucune indication sur les valeurs possibles de $Y$.

Remarquons que ce n'est pas le cas dans la population $P_3$. Dans cette population, sans que ce ne soit une règle absolue, il apparaît que les hautes valeurs de $X$ (c'est-à-dire $X = C$ ou $D$) sont associées de préférence aux hautes valeurs de $Y$. On dit que ***les deux variables sont corrélées***.

Les populations étudiées sont souvent si nombreuses qu'il est pratiquement impossible d'en observer chaque individu. Dans ces conditions, on extrait de cette population un sous-ensemble (on en ***tire*** un ***échantillon)***. C'est l'échantillon que l'on observe afin de déduire de ces observations les propriétés de la population complète. On conçoit que la façon de choisir l'échantillon est très importante. Votre étude ne sera pas fiable si vous prétendez étudier les Français et que vous n'interrogez que des Auvergnats sous prétexte que vous habitez Clermont-Ferrand. De même, prétendre décrire toutes les étoiles alors que vos observations portent sur celles qui sont visibles à l'oeil nu, constitue une hypothèse hardie qui mérite d'être explicitée en préface de votre travail.

L'échantillon peut être établi de façon à représenter la population en respectant certains critères (critères sociologiques lorsqu'il s'agit de sondage d'opinion). Pour contrôler la qualité de la production d'automobiles, par exemple, on prélève quelques automobiles sorties de l'usine à divers moments de la semaine. Le nombre d'automobiles prélevées le lundi après-midi sera proportionnel à la production de cette demi-journée. On construit ainsi ***un modèle***. Ce modèle permet d'obtenir un échantillon de la population à étudier. Il reste à tester la représentativité de cet échantillon, c'est-à-dire la pertinence des critères retenus pour le constituer. Dans le domaine du marketing, un tel échantillon est un "*panel*".

[†]Prononcer "sto**k**astique".

Une autre manière de procéder, consiste à choisir les individus de l'échantillon en définissant une procédure d'extraction de la population. Chaque individu est numéroté et les numéros sont tirés "au hasard". Nous reviendrons sur la définition de ***"tirage au hasard"***, pour le moment admettons que nous savons ce que c'est (pour en avoir une idée "spectaculaire", il suffit de regarder le tirage du loto à la télévision).

On distingue deux sortes de tirages : ***le tirage exhaustif*** et ***le tirage avec remise.***

Dans le tirage exhaustif(*), une fois qu'un individu a été choisi, il est exclu du tirage suivant. Si le cardinal de la population est $N$, il est donc impossible d'obtenir par tirage exhaustif un échantillon de cardinal supérieur à $N$.

Dans le tirage non exhaustif (tirage avec remise), une fois qu'un individu est choisi, il est "remis" dans la population et pourra donc être choisi une seconde fois. L'échantillon peut comprendre plus de $N$ individus, rien n'en limite le cardinal.

Compte tenu de la nature des tirages au hasard effectués dans la pratique, les deux façons de procéder sont équivalentes pour les populations nombreuses lorsque le cardinal de l'échantillon est très inférieur à celui de la population. Dans ce cas, en effet, il est rare que l'on observe des répétitions dans le cas d'un tirage avec remise.

*Par la suite, sauf mention contraire, les tirages que nous évoquerons seront toujours des* ***tirages au hasard non exhaustifs****, même si nous ne le précisons pas.* Cette condition nous assure en effet que deux tirages successifs sont indépendants, c'est-à-dire que le résultat du second tirage ne dépend pas du résultat obtenu au premier tirage. Ces conditions apportent de grandes simplifications dans les calculs.

Les définitions introduites et les mots employés permettent de cerner les préoccupations du statisticien. Il faut maintenant étudier les outils qu'il s'est construit pour décrire ces populations nombreuses dont il fut question plus haut.

Nous considérons tout d'abord les variables discrètes dont les valeurs possibles sont en nombre fini ou dénombrable : variables numériques, variables ordinales ou variables non-numériques.

## 1.2 Représentations graphiques des variables discrètes

### *1.2.1 Séries statistiques*

Considérons une population d'élèves dont nous mesurons la taille, $X$. Les mesures sont regroupées en classes de 5 cm en 5 cm ; à chaque classe est attribué sa valeur centrale. Ainsi, deux élèves dont les tailles sont $1,47$ m et $1,43$ m seront regroupés dans la même classe(†) : il leur sera attribuée la même taille, $1,45$ m ; l'erreur n'excède pas $2,5$ cm.

Les observations sont les suivantes :

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $1,40$ m | $1,50$ m | $1,50$ m | $1,60$ m | $1,40$ m | $1,50$ m | $1,55$ m | $1,55$ m | $1,45$ m |
| $1,50$ m | $1,55$ m | $1,50$ m | $1,45$ m | $1,50$ m | $1,45$ m | $1,45$ m | $1,55$ m | $1,50$ m |
| $1,50$ m | $1,55$ m | | | | | | | |

Tableau 1

Cette "suite" d'observations constitue une ***série statistique***‡. Pour y voir plus clair, on classe les observations par ordre croissant et à chacune des valeurs de $X$ observées, on associe le nombre d'observations correspondantes, $n$.

*Ce procédé épuise la population qui diminue progressivement au profit de l'échantillon (en Anglais : to exhaust = épuiser)

†Remarquons que ce regroupement en classe est automatique avec les variables continues dès lors que l'on s'impose de décrire les résultats avec un nombre fini de chiffres significatifs.

‡De même, la description des diverses populations de $P_1$ à $P_4$ étaient donné page 2 sous forme de séries statistiques, qui cependant n'étaient pas des séries numériques.

| variable : $X_i =$ | | $1,40\,\text{m}$ | $1,45\,\text{m}$ | $1,50\,\text{m}$ | $1,55\,\text{m}$ | $1,60\,\text{m}$ | |
|---|---|---|---|---|---|---|---|
| fréquence absolue | $: n_i =$ | 2 | 4 | 8 | 5 | 1 | Total $= 20 = N$ |
| fréquence relative | $: \frac{n_i}{N} =$ | 0,1 | 0,2 | 0,4 | 0,25 | 0,05 | Total = 1 |

Tableau 2

Il apparaît cinq classes caractérisées par la valeur $X_i$ de la variable ($X_i = 1,40\,\text{m}$, $1,45\,\text{m}$, $1,50\,\text{m}$, $1,55\,\text{m}$ ou $1,60\,\text{m}$). Le nombre de fois que la valeur $X_i$ a été observée est $n_i$, c'est l'***effectif*** de la classe $X_i$, on dit aussi ***fréquence absolue.*** Ici, les effectifs des cinq classes représentées sont 2, 4, 8, 5, et 1. Le nombre total d'observations, $N$, est le cardinal de la population observée ; ici $N = 2 + 4 + 8 + 5 + 1 = 20$.

La proportion des observations qui entrent dans la classe $X_i$ est $n_i/N = P_i$ : c'est la ***fréquence relative***. Remarquons la relation, toujours satisfaite

$$\boxed{\sum_k P_k = 1}$$

Dès lors que les observations ont été regroupées en classes, les notions précédentes ont un sens, que la variable soit numérique ou non. Par contre, le concept d'***effectifs cumulés*** que nous définissons maintenant ne concerne que les variables numériques ou, à la rigueur, les variables ordinales.

Donnons-nous la valeur numérique $x$ et posons la question suivante : "Combien y a-t-il d'observations pour lesquelles $X \leq x$ ?" La réponse à cette question est un nombre que l'on note $F(x)$ : c'est l'effectif cumulé jusqu'à la valeur $x$ (incluse).

On peut définir également les ***fréquences relatives cumulées*** $f(x) = F(x)/N$.

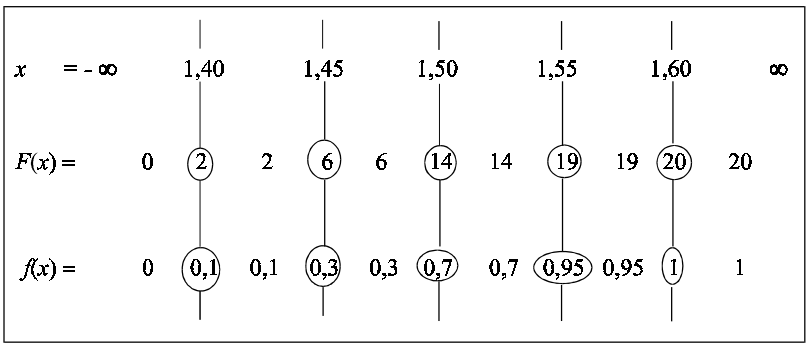



Figure 1-1.

La figure 1-1 donne les valeurs des effectifs cumulés et des fréquences relatives cumulées en fonction de $x$.

*1.2.2 Diagrammes en bâtonnets et secteurs sphériques*

A la fin du XVIII$^{ème}$ siècle et au tout début du XIX$^{ème}$ apparurent, en Angleterre, les premiers diagrammes en bâtonnets (ou diagrammes en bâtons) et les diagrammes en secteurs. Depuis, sociologues, économistes, policiers, médecins, biologistes, physiciens et plus généralement statisticiens de toutes sortes ne se lassent point de les utiliser à tous propos. Ces outils sont très commodes, aussi allons-nous les présenter maintenant.

Considérons une variable numérique discrète susceptible de prendre les valeurs $X_k$. Nous portons sur un axe les valeurs de $X_k$. Compte-tenu de la précision finie des

mesures, les valeurs que l'on observe d'une variable numérique continue sont toujours discrètes. Par conséquent, nous ne distinguons pas, ici, les variables continues et les variables discrètes.

A l'abscisse $X_k$, on dessine un bâtonnet dont la taille représente l'effectif, $n_k$, de la classe $X = X_k$

On obtient la représentation en bâtonnets de la statistique.

En modifiant l'échelle des ordonnées, le même diagramme représente les fréquences relatives. La figure 1-2 donne la représentation en bâtonnets de la statistique précédente.

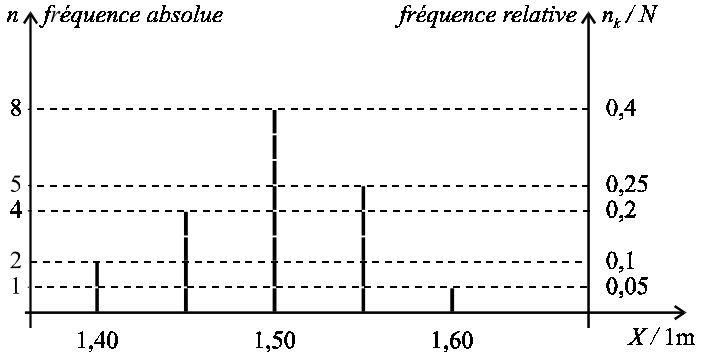



Figure 1-2.

Une telle représentation est encore acceptable lorsque la variable est ordinale, par contre lorsque la variable est non-numérique, l'usage d'un axe introduit sournoisement une relation d'ordre qui n'existe pas entre les valeurs de la variable (et là commence un "fieffé mensonge" !). Il est préférable de donner ***une représentation en secteurs*** (sous réserve bien sûr que les schémas restent lisibles).

Les résultats d'une élection sont les suivants :

| | Clovis | Clothilde | Blancs ou nuls | Total |
|---|---|---|---|---|
| nombre de voix | 200 | 500 | 100 | 800 |
| proportion | 0,250 = 25% | 0,625 = 62,5% | 0,125 = 12,5% | 1 = 100% |

On représentera la statistique de la façon suivante, donnée sur la figure 1-3.

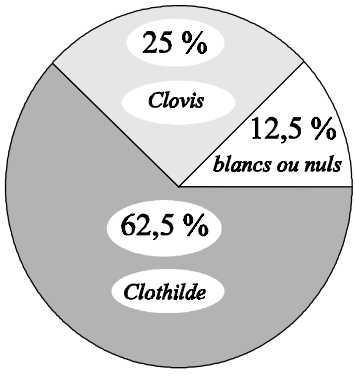



Figure 1-3

L'aire de chacun des secteurs circulaires est proportionnelle à la fréquence relative observée. Il en est de même de l'angle au centre de chaque secteur.

Remarquons que le même diagramme représente aussi les fréquences absolues, sous réserve que la proportion indiquée dans chacun des secteurs soient remplacée par l'effectif du secteur.

### *1.2.3 La fonction de répartition*

Considérons les fréquences relatives cumulées, $f(x)$, de la variable numérique discrète $X$, telles que nous les avons définies ci-dessus. Nous représentons sur la figure 1-4 le graphe de la fonction $x \mapsto f(x)$ correspondant aux données du tableau 2, précédent.

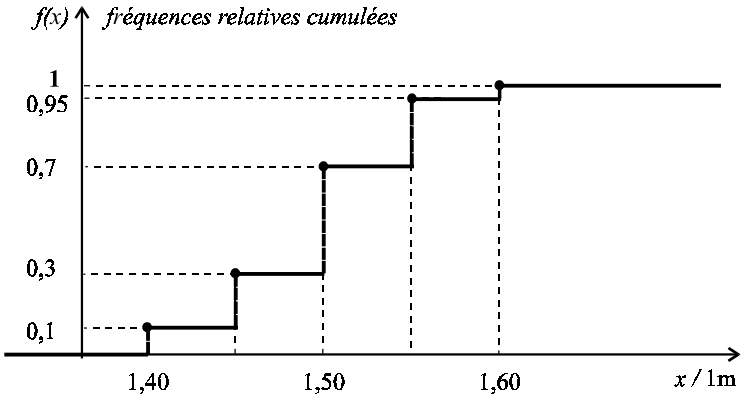



Figure 1-4 : $f(x)$ est la fonction de répartition.

Considérons la valeur $x = 1,52\,\text{m}$ par exemple. Combien vaut $f(x)$ ? Combien y a-t-il d'individus pour lesquels $X \leq 1,52\,\text{m}$? La réponse est "$8+4+2=14$" (*cf.* tableau 2). La valeur de $f(x)$ est donc 14/20=0,7.

$f(x)$ est ***la fonction de répartition*** : c'est la fréquence relative du sous ensemble dont chaque individu satisfait la relation $X \leq x$. On rencontre aussi dans la littérature une autre définition où $X \leq x$ est remplacé par $X < x$. Cette distinction est sans importance pratique, sauf aux points de discontinuité.

La fonction $f$ est une fonction en escalier, non négative ; en construisant son graphe point par point, il nous vient les remarques suivantes.

1. La figure 1-4 représente aussi le graphe des effectifs absolus cumulés, sous réserve de modifier l'échelle des ordonnées.
2. Pour chaque valeur observée, $X_k$, la fonction $f$ est discontinue. La discontinuité en $x = X_i$ est égale à la fréquence relative de la classe $X = X_i$ :

$$\Delta_i f = f(X_i+) - f(X_i-) = \frac{n_i}{N}$$

3. $f(X_i-) \leq f(X_i) = f(X_i+)$. La fonction $f$ est donc monotone, jamais décroissante.
4. Remarquons que tous les individus satisfont la relation $X < +\infty$. On en déduit la relation $\lim_{x\to\infty} [f(x)] = 1$. De même, $X > -\infty$ implique $\lim_{x\to-\infty} [f(x)] = 0$.

Dans de nombreux cas, les valeurs de $X_k$ les plus voisines sont "très proches" et les discontinuités "très petites". La fonction de répartition peut alors être remplacée par une fonction continue qui décrit convenablement la statistique (figure 1-5).

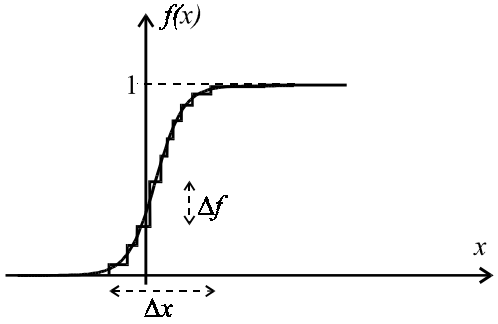



Figure 1-5

L'étendue de variation de $X$ est notée $\Delta x$. Considérons deux valeurs voisines de $X$, les valeurs $X_k$ et $X_{k+1}$; on considère que ces valeurs sont "très proches" si $|X_{k+1} - X_k| << \Delta x$. C'est le cas par exemple des revenus mensuels des Français estimés au centime près. Pour fixer les idées, supposons que l'étendue est $\Delta x = 2000$€. La relation 0,01€<<2000€ est satisfaite ; on peut donc considérer les valeurs voisines de $X_k$ comme très proches.

Si nous définissons l'étendue comme la différence entre le revenu le plus élevé et le revenu le plus faible, il est clair que celle-ci est très supérieure à 2000€ la relation $|X_{k+1} - X_k| = 0,01$€ $<< \Delta x$ est donc satisfaite *a fortiori*. Si nous avons estimé ici, l'étendue à 2000€ c'est parce qu'en 2005, moins de 20% de la population a un revenu supérieur à cette somme. On peut donc "oublier" 20% de la population pour s'intéresser au plus grand nombre. Pour être plus précis, nous introduirons ultérieurement un paramètre bien défini et bien adapté à la mesure de $\Delta x$ : l'écart quadratique moyen.

Considérons maintenant la discontinuité maximale de $f(x)$, notée $\Delta f$. Supposer que $\Delta f$ est "très petit", c'est admettre la relation $\Delta f << 1$. Dans l'exemple que nous considérons, cela signifie que la proportion d'individus dont le revenu est dans la classe $X_k$ est très petite devant l'unité. Les classes de revenus étant définies au centime près, cette propriété est bien vérifiée.

Dans ces conditions, on peut considérer que la variable est continue et que la fonction de répartition est elle aussi continue (voir page 104).

## 1.3 Représentation graphique des variables continues

### *1.3.1 Fonction de répartition et densité de répartition*

Considérons la variable continue $X \in (-\infty, \infty)$ [‡]. Nous notons $f(x)$ la fonction de répartition et, pour fixer les idées, nous supposons qu'elle est continue et dérivable.

Rappelons que $f(x)$ est la proportion d'individus pour lesquels $X \leq x$. Cette définition a certaines conséquences.

1. $f(x)$ est positive ; elle n'est jamais décroissante.
2. Lorsque $x$ tend vers $\pm\infty$, les limites de $f(x)$ sont les suivantes :

$$\boxed{\lim_{x\to-\infty} f(x) = 0} \text{ et } \boxed{\lim_{x\to\infty} f(x) = 1} \tag{1.1}$$

Posons-nous la question de savoir quelle est la proportion d'individus, $P_{]a,b]}$ pour lesquels $X \in ]a, b]$, c'est-à-dire pour lesquels $a < X \leq b$. La définition de $f(x)$ nous donne la réponse :

$$\boxed{P_{]a,b]} = f(b) - f(a)} \tag{1.2}$$

[‡]Rappelons que $\in$ se lit "appartient à".

Ces propriétés sont également satisfaites si la fonction de répartition n'est pas continue ; elles sont très générales.

Lorsque la fonction de répartition est dérivable, on définit la densité de répartition $p(x)$ :

$$\boxed{p(x) = \frac{d}{dx} f(x)} \tag{1.3}$$

Remarquons que $f(x)$ est une proportion, c'est donc une grandeur sans dimension ; par contre $p(x)$ a les mêmes dimensions physiques que $1/x$.

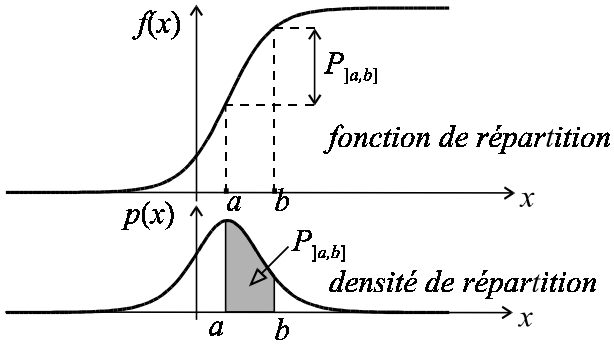



Figure 1-6

Les propriétés de $p(x)$ se déduisent de celles de $f(x)$ :

1. la fonction $p(x)$ n'est pas négative ; elle tend vers zéro lorsque $x$ tend vers $\pm\infty$ :

$$\boxed{\lim_{x \to \pm\infty} p(x) = 0}$$

2. $\boxed{P_{]a,b]} = f(b) - f(a) = \int_a^b p(x)\ dx}$ $P_{]a,b]}$ peut être noté $P_{(a,b)}$ en l'absence de discontinuité car dans ce cas $P_{]a,b]} = P_{[a,b]} = P_{]a,b[} = P_{[a,b[}$ ; c'est l'aire représentée sur la figure 1-6.

3. Sous forme différentielle, il vient :

$$\boxed{P_{(x,x+dx)} = dP = p(x)\ dx} \tag{1.4}$$

Remarquons que les relations 1.1 impliquent :

$$\boxed{\int_{-\infty}^{\infty} p(x)dx = 1}$$

### *1.3.2 Diagramme en bâtonnets, histogrammes et classes*

Etant donnée une fonction de répartition continue, on peut la remplacer par une fonction en escalier et transformer ainsi la statistique de la variable continue, $X$, en statistique d'une variable discrète, susceptible de prendre les seules valeurs $X_k$.

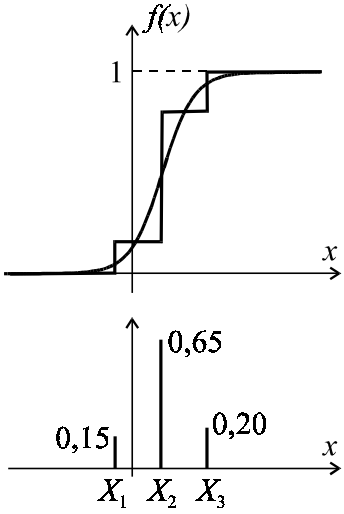



Figure 1-7

Cette façon de procéder permet de simplifier considérablement les calculs, tout en leur conservant une approximation satisfaisante.

On peut remplacer le graphe de la fonction de répartition par une courbe formée de segments de droites. La densité de répartition est alors une fonction constante par morceaux.

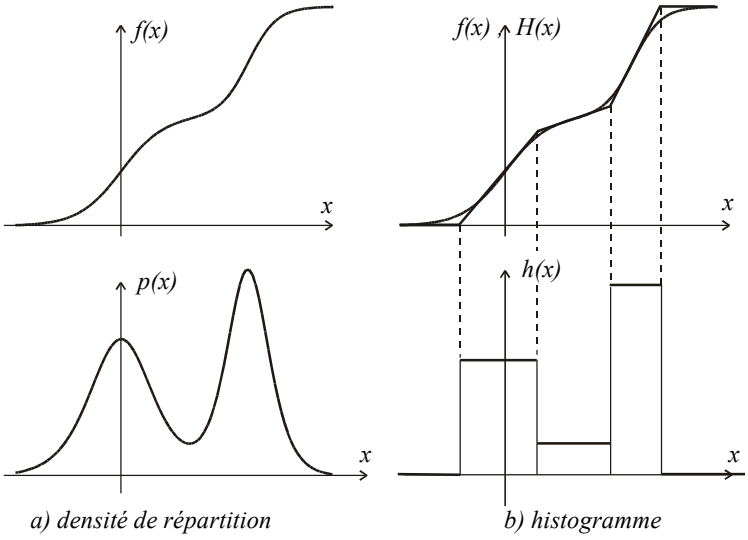



Figure 1-8

Sur la figure 1-8 *a*) nous représentons une fonction de répartition $f(x)$ et la densité de répartition correspondante $p(x)$. Sur la figure 1-8 *b*) nous avons représenté la même fonction de répartition. Nous avons donné de $f(x)$ une approximation polygonale que nous notons $H(x)$. La densité de répartition correspondant à la fonction de répartition $H(x)$ est ***un histogramme***, $h(x)$; c'est une approximation de la densité $p(x)$. Nous avons représenté cet histogramme sur la figure 1-8 *b*).

Considérons la statistique représentée par l'histogramme de la figure 1-9 où $p(x)$ est la densité de répartition.

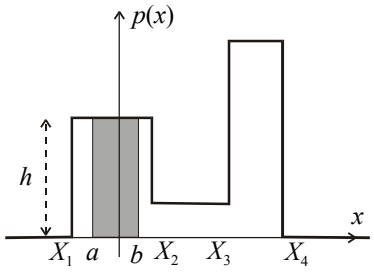


Figure 1-9

La proportion d'individus, $P_{(a,b)}$ pour lesquels $X \in (a,b)$ est l'aire grise de la figure 1-9. Soit $P_{(a,b)} = h \times (b-a)$. Dans ce cas, $P_{(a,b)}$ est proportionnel à l'intervalle de valeurs considérés $(b-a)$. On dit que la variable $X$ est **répartie uniformément** entre $X_1$ et $X_2$.

Une statistique représentée par un diagramme en bâtonnets est la statistique d'une variable discrète. La proportions d'individus pour lesquels $X = X_k$ est donnée par la hauteur du bâtonnet.

Une statistique représentée par un histogramme est la statistique d'une variable continue. Etant donné un intervalle $(a,b)$, la proportion d'individus pour lesquels $a < X \leq b$ est donnée par l'aire sous l'histogramme, limitée par les abscisses $a$ et $b$.

Un autre type de représentation est fréquemment utilisé, en physique par exemple.

Répétons de multiples fois la même mesure par exemple la position de l'impact d'un photon sur un écran. L'écran est considéré comme la réunion de petites surfaces jointives (les pixels). Sur chacune de ces surfaces élémentaires, on mesure le nombre d'impacts pendant le temps $\tau$. Ce procédé permet d'obtenir l'éclairement moyen de l'écran pendant le temps $\tau$. La mesure effectuée ici fournit deux valeurs, l'abscisse et l'ordonnée de l'impact.

Considérons seulement l'abscisse $X$. C'est une variable numérique continue cependant la précision des mesures étant finie, on ne cherche pas à connaître la valeur précise de $X$ mais seulement à classer les valeurs de $X$ dans des "boîtes" (*bins* en Anglais). Le résultat d'une mesure est donc le numéro d'ordre de la boîte dans laquelle se situe la valeur de $X$ observée.

Le plus souvent, les boîtes présentent toutes la même largeur $2\delta$. Par exemple, la boîte n°$i$ est l'intervalle $]x_i - \delta, x_i + \delta]$, centré en $x_i$, avec $x_{i+1} - x_i = 2\delta$; l'incertitude sur la valeur de $X$ est alors $\delta$.

Le résultat des mesures peut être représenté par un diagramme en bâtonnet. Les valeurs $x_i$ sont les abscisses des bâtonnets, tandis que l'effectif de la classe n°$i$ est représenté par la hauteur du bâtonnet.

Comme on ignore tout de la répartition des valeurs de $x$ dans chaque boîte on peut aussi supposer que les valeurs observées sont réparties uniformément et représenter le résultat des mesures sous la forme d'un histogramme. Bien sûr, cette représentation n'a de sens que dans la mesure où l'effectif, $n_i$, de chaque classe est élevé ; dans le cas contraire ($n_i = 1$ par exemple) le concept de répartition uniforme à l'intérieur de la boîte (et par conséquent le concept d'histogramme) est dénué de sens.

De façon générale, on prend pour convention de représenter l'effectif de la classe n°$i$ par un rectangle dont la surface est proportionnelle à $n_i$, sans préjuger de la nature, uniforme ou non, de la répartition des valeurs de $x$ dans chaque boîte.

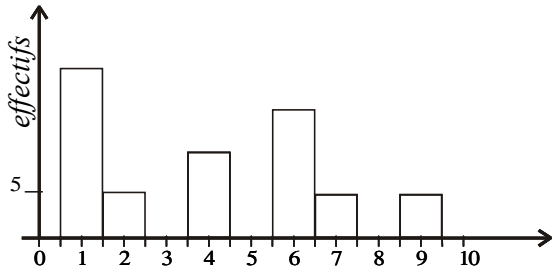


Figure 1-10

La largeur des boîtes étant la même pour toutes les boîtes, la hauteur de chaque rectangle est proportionnelle à l'effectif de la boîte représentée. Une telle représentation est appelée "*histogramme*" par abus de langage.

### 1.4 Pot-pourri de remarques et critiques diverses

Il n'est jamais recommandé d'écrire ce que l'on ne doit pas écrire, de fixer sur le papier ce qu'il ne faut pas faire. Mais pour une fois....

Vous trouverez ci-dessous quelques mauvais exemples que nous avons rencontré assez souvent pour qu'ils soient désignés à l'opprobre publique.

- Parfois, dans des rapports "sérieux", apparaissent des tableaux de la forme ci-après où $N$ est le nombre de ménages qui ont dépensé la somme $X$.

| Variable $X$ (en millier d'euros) | entre 1 et 2 | entre 3 et 4 | entre 5 et 6 | etc |
|---|---|---|---|---|
| Observations $N$ : | 7 | 7 | 12 | 2 |

Présentation **à prohiber** : " entre 2 et 3 :$N = ????$"

Il y a plusieurs façons de comprendre ce tableau.

1. Dans la présentation des effectifs de classes correspondant à des valeurs discrètes de la variable, on compte en général les bords de l'effectif dans l'effectif lui-même. Par exemple on dira "prenez 9 jours de vacances, du 8 au 16". Pour faire 9 jours il faut compter le 8 et le 16. Crucifié le vendredi, Jésus ressuscita trois jours après, le dimanche, nous dit l'Eglise catholique qui compte le vendredi et le dimanche pour faire 3 jours.

   On peut donc comprendre ce tableau en supposant que 7 est l'effectif d'une classe correspondant à la variable $X \in [1, 2]$, intervalle dans lequel $X$ est réparti uniformément. Avec cette interprétation, aucun individu ne correspond à l'intervalle $[2, 3]$.

2. En réalité, l'interprétation était beaucoup plus compliquée. Le tableau précédent était un tableau de dépenses (représentées par une variable continue $X$) et $N$ était un nombre de ménages. L'auteur avait regroupé les valeurs de $N$ correspondant à l'intervalle $[0, 5k€ - 1, 5k€]$ [§] pour construire un bâtonnet à la valeur $1k€$. Il avait opéré de même pour l'intervalle $[1, 5k€ - 2, 5k€]$ afin de créer un bâtonnet

[§] $1k€$ représente un kiloeuro, soit 1000 euros

pour $2k€$. Puis il avait regroupé ces deux intervalles pour créer une classe, réunion des deux classes précédentes. En réalité $N = 7$ représentait l'effectif de la classe $X \in [0,5€ - 2,5k€]$, c'est-à-dire le nombre de ménages qui avaient consenti une dépense comprise entre $0,5k€$ et $2,5k€$.

Présenté sous cette forme, le tableau aurait été compréhensible alors que sous la forme ci-dessus on ne peut pas le lire sans connaître la manière dont il a été obtenu.

Dans le cas cité ici, l'auteur expliquait ce qu'il avait fait : il était sérieux (mais pas très malin !), dans de nombreux cas les auteurs ne citent même pas la façon dont ils procèdent. Les statistiques sont alors inutilisables : c'est le suprême degré du bluff sinon du mensonge.

Chaque fois que c'est possible il faut présenter les tableaux de telle sorte qu'on puisse les comprendre sans se référer au texte (la légende sert à ça !)[¶] et donc se référer à des méthodes de représentation connues. Il faut être aussi précis et complet que possible : s'il n'y a pas d'observations entre 2 et 3 il faut le préciser et en particulier distinguer l'absence d'observation due à une impossibilité (c'est un constat d'ignorance) et l'observation de $N = 0$ (c'est une information).

- Une autre ânerie[‖] que l'on a rencontrée est représentée ci-dessous.

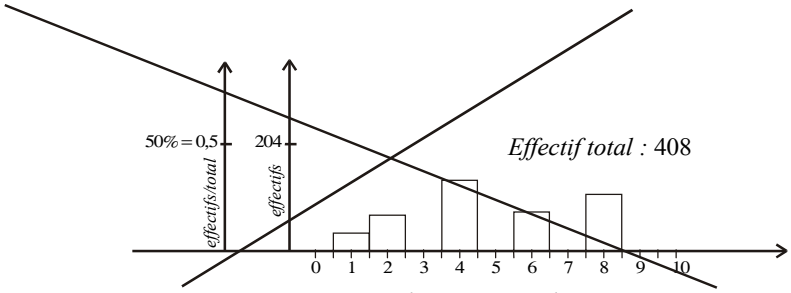



*Diagramme interprété tantôt comme un diagramme en bâtonnets représentant les effectifs, tantôt comme un histogramme représentant les fréquences relatives.*

> Pour éviter les confusions, représenter un bâtonnet par un bâtonnet et un histogramme par des rectangles.

Dans un diagramme en bâtonnets, les bâtonnets ne doivent pas être représentés par des rectangles afin d'éviter la méprise avec les histogrammes. Ici, la confusion est patente. On se pose même la question de savoir si ce qui est représenté ne serait pas les effectifs des boîtes élémentaires d'une mesure quelconque.

Remarquons que la question se pose des observations de la région $(2,5 - 3,5)$ ou $(4,5 - 5,5)$ par exemple. Le silence est inquiétant. Ignore-t-on tout des événements dont la variable tombe dans ces intervalles ou n'a-t-on observé aucun événement ? Là encore nous sommes dépendant d'un texte explicatif qui bien souvent n'apporte pas, lui non plus, les informations manquantes.

- Enfin, dans les discussions qui s'appuient sur des histogrammes, il arrive malheureusement que les arguments mettent en avant les ordonnées des diverses classes,

[¶] Et dans le présent polycopié où sont les légendes ? ! Dans un polycopié de cours, les figures illustrent le texte dont l'importance est primordiale. Dans un article ou un rapport, les figures et le texte, contribuent à parts égales à l'exposé. Là est la différence.

[‖] Les mots ont un sens auquel il faut se référer parfois. L'âne est sans doute l'un des animaux domestiques les plus sympathiques. Il est sobre, patient et courageux. Sans âne pas de civilisation méditerranéenne ! La connotation péjorative du mot "ânerie" est sans doute moins désobligeante pour l'âne lui-même que pour ceux qui l'emploient : les auteurs plaident coupables.

alors que ce sont seulement les aires des rectangles qui portent une signification. Là encore, la confusion règne trop souvent.

## 1.5 Position et dispersion

Dans les années 70, Valéry Giscard d'Estaing étant président de la République française, le parti communiste français avait pour secrétaire général Georges Marchais. Le premier prétendait que les Français allaient de mieux en mieux, le revenu des ménages ne cessant d'augmenter, tandis que le second affirmait que tout se dégradait, les riches devenant encore plus riches et les pauvres encore plus pauvres.

Pour comprendre ce "dialogue" de sourds, considérons la statistique des revenus, $X$, des ménages en 1975 et 1979. La variable $X$ peut être considérée comme continue ; nous en représentons la densité de répartition sur la figure 1-11.

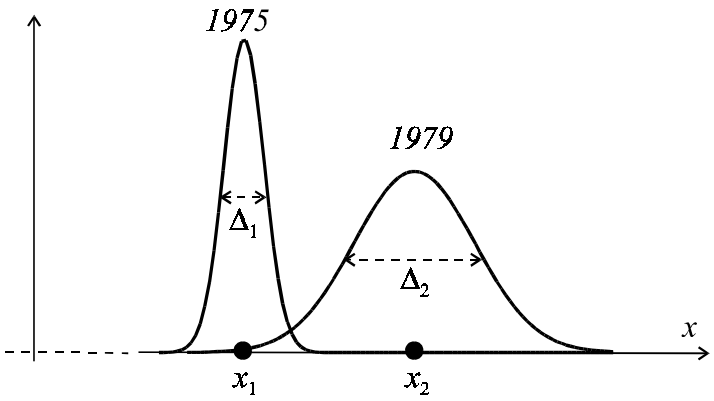



Figure 1-11

L'observation des courbes de la figure 1-11 conduisent aux observations suivantes.

1. En 1975, le revenu se situe autour de la valeur centrale $x_1$ alors qu'en 1979 il se situe autour de la valeur $x_2$. Compte tenu de la relation $x_2 > x_1$, nous devons considérer que Valéry Giscard d'Estaing n'a pas tort.
2. Pour estimer les inégalités, on introduit un indicateur (c'est plutôt un indicateur d'égalité au sens que nous donnerons à ce mot page 91). Ce pourrait être l'étendue comme dans l'exemple précédent page 8. Nous préférons choisir ici la largeur à mi-hauteur qui vaut $\Delta_1$ en 1975 et $\Delta_2$ en 1979. La relation $\Delta_2 > \Delta_1$ justifie l'affirmation de Georges Marchais.

Ainsi tout le monde à raison ! VGE considérait la position de la courbe par rapport à l'origine, sur l'axe des $x$, tandis que GM considérait la dispersion des revenus autour d'une valeur centrale*.

Pour décrire la position d'une statistique et la dispersion de la variable, nous disposons de plusieurs paramètres que nous allons introduire.

### *1.5.1 Paramètres de position*

Les statistiques de 1975 ou 1979 présentent un maximum de la fonction $p(x)$. Ce maximum est ***le mode***, ces statistiques sont unimodales. Dans un tel cas, on peut utiliser le mode comme paramètre de position ; mais les statistiques ne sont pas toujours unimodales (voir par exemple la statistique de la figure 1-8 qui est bimodale). En outre, la

*Ce ne sont pas les statistiques qui pourront décider du paramètre important.

comparaison de deux statistiques peut être compliquée par le fait que l'une serait bimodale tandis que l'autre ne présenterait qu'un seul mode.

Pour toutes ces raisons, on utilise ***la moyenne*** comme indicateur de position.

Considérons une variable numérique discrète, $X$, susceptible de prendre les valeurs $X_k$ toutes différentes. Soit $P_k$ la proportion d'individus pour lesquels $X = X_k$. On définit la moyenne $\overline{X}$ :

$$\boxed{\overline{X} = P_1X_1 + P_2X_2 + .... = \sum_k P_k \times X_k} \tag{1.5}$$

Remarquons que cette définition de la moyenne s'applique à toute fonction de $X$ ; par exemple la moyenne de $X^2$ se définit par la relation $\overline{X^2} = \sum_k P_k \times (X_k)^2$ .

Si $X$ est une constante $\lambda$, cela signifie que $X_k = \lambda$ quel que soit $k$. On trouve alors $\overline{\lambda} = \sum_k P_k \times \lambda = \lambda \times \sum_k P_k = \lambda$ car $\sum_k P_k = 1$. **La moyenne d'une constante est donc cette constante elle-même**.

Supposons que les proportions $P_1$, $P_2$,... soient des masses situées aux points d'abscisse $X_1$, $X_2$, $X_3$, ... Alors le point d'abscisse $\overline{X}$ apparaît comme le point $G$, centre de masse du système. La représentation adoptée ici est semblable à la représentation que l'on pratique en physique lorsqu'on assimile un solide à un point matériel confondu avec son centre de masse[†].

Si la variable est continue, on remplace la fonction de répartition par la fonction en escalier de la figure 1-12 dont les points de discontinuité sont pris en $X_k = k \times \delta x$ (l'entier relatif $k$, variant de $-\infty$ à $+\infty$). Lorsque $\delta x$ tend vers zéro, la fonction en escalier est confondue avec la fonction de répartition et l'erreur commise disparaît.

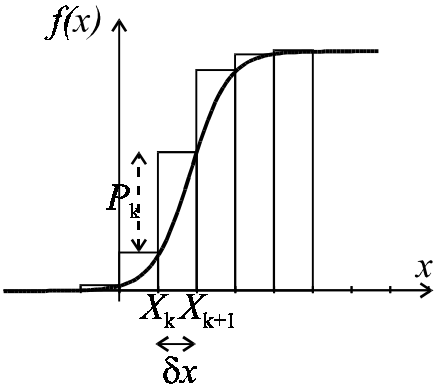



Figure 1-12

Lorsque $\delta x$ est non nul, il vient $P_k = f\left(X_{k+1}\right) - f(X_k) \simeq \left(\dfrac{df}{dx}\right)_{X_k} \times \delta x = p(X_k)\ \delta x$. La moyenne est alors $\overline{X} = \sum_k p(X_k)\ \delta x \times X_k$. A la limite continue $\delta x \to 0$, on reconnaît l'intégrale de $xp(x)dx$ étendue à l'intervalle $(-\infty, +\infty)$ :

$$\boxed{\overline{X} = \int_{-\infty}^{\infty} x\ p(x)\ dx} \tag{1.6}$$

[†]Le centre de masse est encore appelé "centre d'inertie" ou, suivant une ancienne terminologie, "centre de gravité".

**Remarque :** Pour étudier la statistique, à partir de la série statistique des observations, nous avons regroupé les individus associés à la même valeur, $X_k$ de $X$. Nous avons alors introduit la fréquence relative $P_k$ de la valeur $X_k$ et nous avons utilisé ces quantités pour définir la moyenne. Cependant, la moyenne peut être calculée directement à partir de la série statistique. Dans ce but nous additionnons toutes les valeurs observées (avec répétition). Soit $\Sigma$ la quantité obtenue. Nous divisons cette somme par $N$, cardinal de la population ; le résultat, $\Sigma/N$ est précisément la moyenne $\overline{X}$. En effet, dans la liste précédente des termes que nous additionnons, nous pouvons regrouper les termes qui se répètent. Il vient : $\Sigma = n_1X_1 + n_2X_2 + ...$ où les $X_k$ sont tous différents, tandis que $n_k$ est le nombre de fois que nous rencontrons $X_k$ ($n_k$ est la fréquence absolue de $X_k$). En divisant par $N$ il vient $\dfrac{\Sigma}{N} = \dfrac{n_1}{N}X_1 + \dfrac{n_2}{N}X_2 + ...$ On reconnaît l'expression $P_1X_1 + P_2X_2 + ...$ où $P_k = \dfrac{n_k}{N}$ est la fréquence relative. Ainsi est démontrée la relation $\dfrac{\Sigma}{N} = \overline{X}$.

Considérons maintenant le cas d'une population constituée d'individus auxquels sont attachés deux ou plusieurs variables. Une population de ménages formés de couples auxquels sont attachés les revenus de chacun des membres du couple ou, encore, une population de molécules auxquelles sont associées l'énergie cinétique de translation, l'énergie cinétique de rotation et l'énergie de vibration.

Considérons le cas de deux variables : $X$ et $Y$. A chaque individu, on peut encore associer une troisième variable $Z = X + Y$. Si, par exemple, $X$ et $Y$ sont les revenus de chacun des membres du couple, $Z$ est le revenu total du ménage.

Pour calculer la moyenne $\overline{Z}$ de $Z$, nous pouvons additionner toutes les valeurs de $Z$ et diviser le résultat par le cardinal $N$ de la population. Additionner toutes les valeurs de $Z$, c'est additionner toutes les valeurs de $X$ et toutes les valeurs de $Y$ pour obtenir deux sommes partielles que l'on additionne ensuite : $\sum Z = \sum (X + Y) = \sum X + \sum Y$. On en déduit $\overline{Z} = \dfrac{\sum Z}{N} = \dfrac{\sum X}{N} + \dfrac{\sum Y}{N} = \overline{X} + \overline{Y}$ par conséquent la moyenne d'une somme de deux variables est égale à la somme des moyennes de chacune des variables

$$\boxed{\overline{X + Y} = \overline{X} + \overline{Y}} \tag{1.7}$$

Remarquons en outre que la substitution $X \to \lambda X$ où $\lambda$ est une constante, conduit à la modification $\overline{X} \to \lambda\overline{X}$. Cette propriété se déduit immédiatement de la relation 1.5 :

$$\boxed{\overline{\lambda X} = \lambda \times \overline{X}} \tag{1.8}$$

Outre le mode et la moyenne, la ***médiane*** est parfois utilisée pour positionner les statistiques de variables numériques continues. Soit $f(x)$ la fonction de répartition (c'est la proportion d'individus pour lesquels on a $X \leq x$). La médiane $X_m$ est définie par la relation $f(X_m) = 1/2$. La médiane est la valeur, $X_m$, de la variable qui divise la population en deux sous-ensembles de même cardinal.

Sur la figure 1-13, nous avons représenté la fonction de répartition et la densité de répartition d'une variable positive, continue.

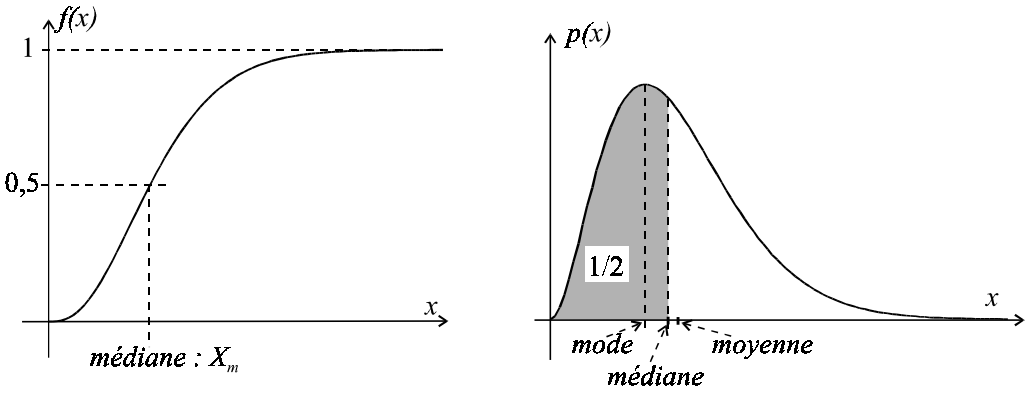



Figure 1-13

Mode, médiane et moyenne prennent des valeurs différentes mais chacune de ces grandeurs fournit une indication sur la position de la statistique.

### *1.5.2 Paramètres de dispersion*

Nous avons déjà rencontré divers paramètres de dispersion : l'étendue (page8) et la largeur à mi-hauteur pour une courbe unimodale (page 14). Le paramètre que l'on utilise le plus souvent est ***l'écart quadratique moyen***.

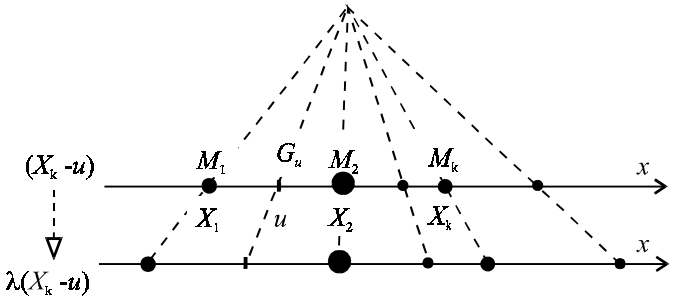
Considérons la statistique d'une variable discrète. Choisissons un point, $G_u$, d'abscisse arbitraire $u$. Pour apprécier la dispersion autour de $G_u$, nous introduisons les distances $d_k = |X_k - u|$ où $X_k$ est l'une quelconque des valeurs de $X$ observables. Il n'est jamais agréable de manipuler les valeurs absolues, aussi préférons-nous utiliser $d_k^2 = (X_k - u)^2 = (u - X_k)^2$. Cette quantité est un indicateur de la distance de $G_u$ à $M_k$, point d'abscisse $X_k$. Nous affectons $d_k^2$ du coefficient $P_k$, fréquence relative de $X_k$. Nous additionnons alors toutes les quantités ainsi obtenues pour chacune des valeurs des observations. Le résultat est une quantité positive car c'est la somme de quantités positives. C'est donc le carré d'une grandeur que nous appelons $\Sigma$ :

$$\Sigma^2 = \sum_k (X_k - u)^2 P_k$$

Comparons cette expression à l'expression 1.5, il apparaît que $\Sigma^2$ est la moyenne de $d_k^2 = (X_k - u)^2$, ce que l'on écrit $\boxed{\Sigma^2 = \overline{(X-u)^2}}$

Remarquons que dans l'expression de $\Sigma^2$ les valeurs $X_k$ les plus souvent observées présentent un poids plus important que celles rarement rencontrées. Le paramètre $\Sigma^2$ ainsi construit est un indicateur de la dispersion par rapport au point $G_u$. En effet, si $\Sigma^2$ est nul, cela signifie que $X_k = u$ quel que soit $k$. La nullité de $\Sigma^2$ se produit donc lorsque toutes les valeurs observées, $X_k$, sont égales (et égales à $u$), c'est-à-dire lorsque tous les points $M_k$ sont regroupés en $G_u$. Dans un tel cas, il n'y a pas de dispersion par rapport à $G_u$.

D'autre part, l'indicateur $\Sigma^2$ est positif et il croît lorsque les points $M_k$ s'éloignent de $G_u$. Pour s'en convaincre, il suffit de considérer la statistique obtenue en effectuant la substitution $X_k - u \to \lambda (X_k - u)$ où $\lambda$ est un nombre réel positif. Le paramètre $\Sigma^2$ est alors multiplié par $\lambda^2$. Pour $\lambda > 1$ les points s'éloignent les uns des autres, la dispersion augmente, il en va de même de $\Sigma^2$ (figure 1-14).




*$\lambda > 1$ : les points s'éloignent de G et la dispersion augmente*

$\Sigma^2 \text{ - - - - } \triangleright \lambda^2\Sigma^2$

Figure 1-14

Pour $\lambda < 1$ les points se rapprochent. La dispersion diminue de même que la valeur numérique de $\Sigma^2$.

Toutes ces propriétés confèrent à $\Sigma^2$ le statut d'***indicateur de dispersion.*** Remarquons que $\Sigma$ est aussi un indicateur de dispersion et que l'on aurait pu en construire beaucoup d'autres.

Reprenons l'image d'une répartition de points matériels de masse $P_k$ situés en $M_k$. De ce point de vue $\Sigma^2$ représente le moment d'inertie de la distribution de masse par rapport à $G_u$ (voir l'annexe page 36)

Pour donner la position de la répartition de points matériels $M_k$, nous cherchons le point géométrique $G$ autour duquel la dispersion est minimale. Pour déterminer ce point, nous considérons $\Sigma^2$ comme une fonction de $u$ et nous posons $\dfrac{d\Sigma^2}{du} = 0$. Il vient $\dfrac{d}{du}\sum_k (X_k - u)^2 P_k = -2\sum_k (X_k - u) P_k = 0$. Des relations $\sum_k P_k = 1$ et $\overline{X} = \sum_k X_k P_k$, on déduit $u = \overline{X}$. Le point $G_u = G$ autour duquel se regroupent les points matériels $M_k$ est le centre de masse du système ; son abscisse est $\overline{X}$.

Choisir $\overline{X}$ comme indicateur de position et $\Sigma^2$ comme indicateur de dispersion conduit à une forte similitude entre statistique et mécanique.

En général, ce que l'on appelle ***"la dispersion"*** est la dispersion minimale, c'est-à-dire la dispersion par rapport au point moyen, $G$. La dispersion de la variable $X$ est notée $\sigma_X^2$ où $\sigma_X$ est l'***écart quadratique moyen*** :

$$\boxed{\sigma_X^2 = \sum_k \left(X_k - \overline{X}\right)^2 P_k = \overline{\left(X - \overline{X}\right)^2} = \overline{X^2} - \overline{X}^2}$$

Dans l'expression précédente, la première égalité est une définition, la seconde est une réécriture de la définition en utilisant la définition de la moyenne de $\left(X - \overline{X}\right)^2$ quant à la dernière égalité elle se démontre de la façon suivante :

$\overline{\left(X - \overline{X}\right)^2} = \overline{\left(X^2 + \overline{X}^2 - 2X\overline{X}\right)} = \overline{X^2} + \overline{\overline{X}^2} - \overline{2X\overline{X}}$ (car la moyenne d'une somme est la somme des moyennes cf 1.7) ; en outre $\overline{X}^2$ et $2\overline{X}$ sont des constantes. On en déduit $\overline{\overline{X}^2} = \overline{X}^2$ et $\overline{2X\overline{X}} = \overline{\left(2\overline{X}\right)X} = \left(2\overline{X}\right)\overline{X} = 2\overline{X}^2$. Par conséquent il vient $\overline{\left(X - \overline{X}\right)^2} = \overline{X^2} - \overline{X}^2$.

Le passage d'une variable discrète à une variable continue modifie l'expression de la moyenne : la relation 1.5 devient la relation 1.6 ci-dessus. Nous avons vu la relation $\sigma_X^2 = \overline{\left(X - \overline{X}\right)^2}$. En effectuant la substitution $x \to \left(x - \overline{X}\right)^2$, il vient

$$\begin{aligned} \sigma_X^2 &= \int_{-\infty}^{+\infty} (x - \overline{X})^2 p(x) dx = \overline{X^2} - \overline{X}^2 \\ &= \int_{-\infty}^{+\infty} x^2 p(x) dx - \left( \int_{-\infty}^{+\infty} x p(x) dx \right)^2 \end{aligned}$$

Modifions la variable en effectuant la substitution $X \to \lambda X$ où $\lambda$ est une constante. Dans ces conditions, il vient $\overline{X} \to \overline{\lambda X} = \lambda \overline{X}$ et $\overline{X^2} \to (\lambda X)^2 = \lambda^2 \overline{X^2}$. On en déduit

$$\boxed{\sigma_{\lambda X} = \lambda \times \sigma_X} \tag{1.9}$$

Considérons maintenant la somme, $Z$, des variables $X$ et $Y$. Posons $\sigma_Z^2 = \overline{Z^2} - \overline{Z}^2$. En utilisant les résultats précédents, il vient $\sigma_Z^2 = \overline{Z^2} - \overline{Z}^2 = \overline{(X+Y)^2} - \left(\overline{X} + \overline{Y}\right)^2$, soit $\sigma_Z^2 = \overline{X^2} + \overline{Y^2} + 2\overline{XY} - \overline{X}^2 - \overline{Y}^2 - 2\overline{X}\ \overline{Y} = \sigma_X^2 + \sigma_Y^2 + 2\left(\overline{XY} - \overline{X}\ \overline{Y}\right)$. On définit la ***covariance*** de $X$ et $Y$ : $cov\left[X,Y\right] \underset{déf}{=} \overline{XY} - \overline{X}\ \overline{Y}$. Il vient

$$\boxed{\left(\sigma_{X+Y}\right)^2 = \left(\sigma_X\right)^2 + \left(\sigma_Y\right)^2 + 2cov\left[X,Y\right]}$$

On vérifie aisément les relations

$$\boxed{cov\left[X,Y\right] = \overline{XY} - \overline{X}\ \overline{Y} = \overline{\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}} \quad \text{et} \quad \left|cov\left[X,Y\right]\right| \leq \left(\sigma_X\right)\left(\sigma_Y\right) \tag{1.10}$$

[[Pour démontrer la première relation, on développe $\overline{\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}$, il vient $\overline{\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)} = \overline{XY} - \overline{X\overline{Y}} - \overline{\overline{X}Y} + \overline{\overline{X}\,\overline{Y}}$ (car la moyenne d'une somme est égale à la somme des moyennes).

$\overline{\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)} = \overline{XY} - 2\overline{X}\ \overline{Y} + \overline{X}\ \overline{Y}$ (car $\overline{\lambda X} = \lambda \overline{X}$ quand $\lambda$ est une constante et que $\overline{\lambda} = \lambda$).

On en déduit donc $\overline{\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)} = \overline{XY} - \overline{X}\ \overline{Y} = cov\left[X,Y\right]$.

Pour démontrer la seconde relation, on pose $U = X - \overline{X}$ et $V = Y - \overline{Y}$. La quantité $F = \overline{(U + \lambda V)^2}$ est positive quelle que soit la constante $\lambda$ car c'est la moyenne d'un carré. Un développement de $F$ donne $F = \lambda^2 \overline{V^2} + 2\lambda \overline{UV} + \overline{U^2} \geq 0$. Le coefficient de $\lambda^2$ étant positif, le polynôme du second degré, $F$, est positif ou nul quel que soit $\lambda$, si et seulement si $F$ n'a pas de racines, c'est-à-dire pour $\overline{UV}^2 - \overline{V^2}\ \overline{U^2} \leq 0$. En utilisant les expressions précédentes de la covariance et de l'écart quadratique moyen, cette condition s'écrit $\left|cov\left[X,Y\right]\right|^2 - \left(\sigma_X\right)^2 \left(\sigma_Y\right)^2 \leq 0$, ce qui démontre la seconde relation 1.10.]]

En général, $cov\left[X,Y\right]$ n'est pas nul et $\left(\sigma_{X+Y}\right)^2 \neq \left(\sigma_X\right)^2 + \left(\sigma_Y\right)^2$. Cependant, **lorsque $X$ et $Y$ sont indépendantes,** il vient $\boxed{cov\left[X,Y\right] = 0}$ et

$\boxed{\left(\sigma_{X+Y}\right)^2 = \left(\sigma_X\right)^2 + \left(\sigma_Y\right)^2}$ (voir le paragraphe 2.5.3 page 33 ci-dessous)

Pour décrire la statistique, on introduit parfois les ***quartiles***. Ce sont les valeurs $X_q$ et $X_{q'}$ telles que $f(X_q) = 1/4$ et $f(X_{q'}) = 3/4$.

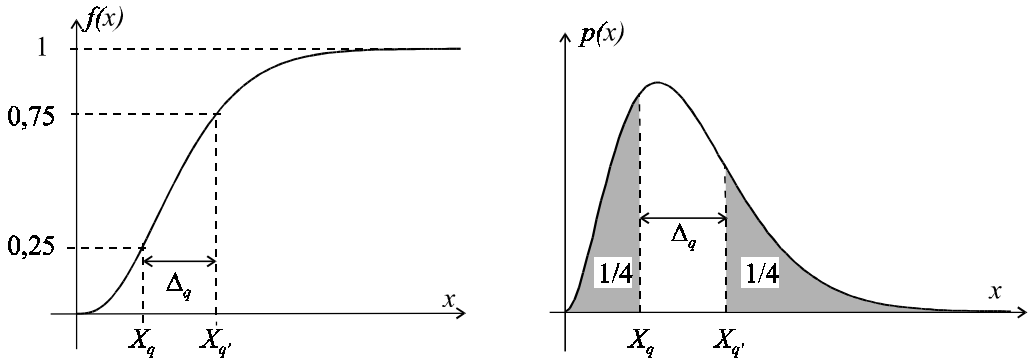



Figure 1-15 : Les quartiles.

On peut utiliser le paramètre $\Delta_q = X_{q'} - X_q$ comme paramètre de dispersion. Ce paramètre présente, par rapport à l'étendue, l'avantage d'éliminer les cas exceptionnels, très rares, correspondants aux grandes valeurs ou aux petites valeurs de la variable. Bien entendu, le choix de la proportion 1/4, utilisée pour définir les (trop) grandes et les (trop) petites valeurs de la variable est arbitraire. Nous pouvons tout aussi bien remplacer 0,25 par 20% = 0,2 comme nous l'avons déjà fait page 8.

### 1.6 Conclusion

Remarquons que les paramètres de position introduits, le mode, la moyenne $\overline{X}$ ou la médiane $X_m$, ainsi que les paramètres de dispersion, l'étendue, $\sigma$ ou $\Delta_q$, sont des grandeurs de mêmes dimensions physiques que la variable $X$. Ces grandeurs sont comparables et nous permettent de définir le concept de "grand" et le concept de "petit" que nous avons utilisé précédemment au paragraphe 1.2.3 page 8.

**"grand" signifie "rare"** : si Pierre est qualifié de "grand", c'est parce qu'il est rare de rencontrer quelqu'un d'aussi grand. Si c'était banal, si tout le monde était aussi grand que Pierre quel sens cela aurait-il de le qualifier de "grand" ? Le qualificatif "grand" et les qualificatifs de même nature ne concernent pas seulement la taille des objets. On évoque un *haut* revenu, une production *importante* de céréales, une *forte* densité de population, etc...

L'écart quadratique moyen donne un ordre de grandeur de la dispersion des valeurs de la variable $X$ autour de la moyenne. De façon grossière, on peut dire que la quasi-totalité de la population présente une valeur de $X$ comprise entre $\overline{X} - \sigma$ et $\overline{X} + \sigma$. Les valeurs $X > \overline{X} + \sigma$ sont donc "rares" ; elles pourront être qualifiées de "grandes" et les valeurs $X < \overline{X} - \sigma$ de "petites". Bien sur, les mots "grand" et "petit" se rapportent à des estimations qualitatives, on pourra donc tout aussi bien désigner comme "grandes" les valeurs de $X$ telles que $X > 2 \times \sigma$. Ce que nous devons retenir, c'est le rôle de $\sigma$ comme grandeur de référence pour estimer les positions relatives des diverses valeurs de $X$ au sein d'une même statistique.

Il existe une autre définition de "grand" et "petit" lorsqu'existe une référence objective, extérieure à la statistique elle-même. On pourra, par exemple qualifier de "faible" tout revenu ne permettant pas de survivre et de "petit" tout sujet britannique dont la taille ne lui permet pas de devenir *horse guard*. Encore que dans ce dernier cas "petit" signifie "pas assez grand" où "grand" est défini en référence à la population britannique afin que "horse guard" reste synonyme de "beau gars".