STATISTIQUES DESCRIPTIVES

Roger NOEL

2006

Table des matières

Chapitre 1

Objet et langage de la Statistique

La statistique est un ensemble de méthodes permettant de décrire et d'analyser des phénomènes repérés par des éléments de même nature, susceptibles d'être dénombrés et :/ou classés.

Le rôle d'explication et de prévision appartient à l'utilisateur et non à la statistique, qui n'est qu'un outil d'investigation.

On prendra garde à ne pas confondre **la statistique** - ensemble de méthodes scientifiques - et **les statistiques**, terme désignant les résultats numériques d'une enquête ou d'une série de mesures (après traitement éventuel par la statistique).

1.1 Population - Caractère

Toute méthode statistique considère au départ un ensemble P, appelé **population**. Cette terminologie est issue de la démographie, première science à avoir développé des méthodes statistiques ; il est cependant clair que la population que l'on envisage en statistique dépend du domaine que l'on traite, et peut donc aussi bien être constituée d'êtres humains que d'animaux, d'objets, voire d'événements.

On appelle **individu** tout élément de la population, quelle que soit sa nature.

Il faut souligner que si l'on veut donner un caractère scientifique à la notion de population, il est indispensable qu'il s'agisse effectivement d'un ensemble au sens mathématique du terme, i.e. que cette population soit très exactement définie de sorte que l'on puisse, sans ambigüité, décider si un élément quelconque appartient ou non à cette population. Cette contrainte est souvent astreignante, mais indispensable. La description précise de la population étudiée est appelée **champ de l'enquête**.

Si l'on reprend l'exemple de la démographie, i.e. l'étude d'une population d'êtres humains, il est évident qu'on ne saurait étudier toute la complexité des individus. On va en fait ne s'intéresser qu'à certaines caractéristiques de ces individus, comme l'age, le sexe, les revenus, le nombre d'enfants, etc... Plus généralement, on appelle **caractère** toute application X de la population P dans un ensemble E, dont les éléments sont appelés **modalités** du caractère X.

Résumons la terminologie :
Soit X une application de P dans E.

$$\begin{array}{rccl} X & : & P & \longrightarrow & E \\ & & x & \mapsto & X(x) = \omega \end{array}$$

X	:	caractère
P	:	population
E	:	ensemble des modalités
x	:	individu
ω=X(x)	:	modalité

1.2 Caractères discrets et caractères continus

Un caractère est dit **qualitatif** lorsque ses modalités sont des qualités (au sens intuitif du terme, sans son aspect subjectif). Ainsi le sexe des individus (modalités : homme, femme), les intentions de vote (modalités : les candidats ou OUI-NON s'il s'agit d'un référendum), la couleur des yeux (modalités : brun, vert, bleu, ?) sont des caractères qualitatifs. Les modalités ne sont dans ce cas ni classées (si ce n'est de manière arbitraire), ni susceptibles d'opérations algébriques. Il arrive fréquemment que dans le cas de caractères qualitatifs les modalités soient codées - par une **nomenclature** - qui assure un classement, certes purement arbitraire, mais s'imposant alors à tous (pensez à la nomenclature des catégories socio-professionnelles), avec le double avantage de permettre des comparaisons aisées et de définir parfaitement les modalités.
Dans le cas des caractères qualitatifs, surtout lorsqu'une nomenclature est utilisée, on parlera souvent de **rubrique** au lieu de modalité.

Un caractère est dit **quantitatif** lorsque ses modalités sont des nombres (en pratique l'ensemble $\mathbb{N}$ des entiers naturels, l'ensemble des réels $\mathbb{R}$ ou l'un de leurs sous-ensembles), et lorsque la valeur du caractère X découle d'une mesure (i.e. ayant une véritable signification scientifique et ne découlant pas d'un simple codage). Les modalités sont alors naturellement classées et susceptibles d'opérations algébriques (somme, etc...). Ainsi le nombre d'enfants, la taille, le poids (en fait la masse !), la longueur, le volume et d'une manière générale toutes les grandeurs physiques sont des caractères quantitatifs.

Notons cependant que la statistique appliquée à la physique ne traite pas seulement des caractères quantitatifs. Ainsi, pour contrôler une machine-outil, on peut très bien envisager de n'étudier que deux modalités : la pièce est défectueuse ou non. Nous sommes alors en présence d'un caractère clairement qualitatif.

Un caractère quantitatif est souvent appelé également **variable statistique**.

On distingue les variables statistiques **discrètes**, dont les modalités sont isolées (en général nombres entiers) des variables **continues** (rien à voir avec la notion de continuité des fonctions) dont les modalités sont un sous-ensemble non dénombrable de $\mathbb{R}$ (en pratique un intervalle réel).

Ainsi le nombre d'enfants d'une famille est une variable discrète, tandis que la taille d'un individu définit une variable continue. La distinction est cependant parfois moins évidente qu'il n'y parait, et il faudra alors faire un choix clair sur la nature des modalités. En général on considérera comme continue toute variable définie par une mesure non essentiellement entière, même si l'on pourrait considérer que, compte tenu des instruments de mesures utilisés, elle est discrète (le poids déterminé au kg près ne prend que des valeurs entières !).

1.3 Echantillon, regroupement des données, série statistique

Lorsque l'on veut étudier une population, on peut évidemment recourir à un **recensement**, i.e. déterminer les valeurs d'un ou de plusieurs caractères pour chaque individu de la population (ce qui est fait lors des recensements de la population française, même si les résultats ne sont guère exploités au delà de la moitié des renseignements collectés).

Notons à ce propos la différence entre recensement et **dénombrement** : un recensement collecte les valeurs d'un certain nombre de caractères pour tous les individus, tandis qu'un dénombrement se contente de déterminer le nombre d'individus de la population.

Un recensement, s'il est sans aucun doute le meilleur moyen d'étudier une population, est cependant très lourd (et très onéreux !) à mettre en place (imaginer le contrôle de toutes les pièces fabriquées par une machine-outil débitant des clous ou des vis par exemple !), lorsqu'il n'est pas pratiquement impossible à réaliser (vérification de la qualité de l'eau dans une rivière !).

On est donc amené à effectuer une collecte d'information sur une partie (souvent une petite partie) de la population : un telle collecte est appelée **sondage** ou **enquête**.

Le résultat d'un sondage, i.e. les renseignements ainsi recueillis, est appelé **échantillon**.

Les données recueillies - appelées **données brutes** - sont soumises à un premier traitement afin d'en faciliter à la fois la présentation et l'exploitation. Pour chaque valeur de modalité constatée, on détermine le nombre d'individus ayant présenté cette valeur du caractère, nombre appelé **effectif** associé à la modalité. L'ensemble des couples (modalité, effectif) ainsi déterminé est parfois appelé **distribution statistique** ou **série statistique** ou encore **variable statistique**. On dit alors que l'on a effectué un **regroupement** des données brutes.

Un distribution statistique est représentée par un **tableau statistique** faisant apparaître sur deux lignes (ou deux colonnes) les couples modalité-effectif, la nature du caractère, de la population et les caractéristiques de sondage (date, etc...) étant clairement indiqués en titre et/ou en commentaire.

Est également indiqué dans un tableau statistique l'**effectif total** - ou **taille** de l'échantillon - qui est à la fois le nombre d'individus de l'échantillon et la somme des effectifs de toutes les modalités observées (une vérification de cette égalité théorique est toujours à effectuer afin d'éliminer - autant que faire se peut - les erreurs de décompte et/ou l'oubli d'une ou plusieurs valeurs de modalité).

La **fréquence** d'une modalité est le rapport de l'effectif associé et de l'effectif total : si n_i est l'effectif associé à la modalité x_i et si n désigne l'effectif total, alors la fréquence f_i associée à la modalité x_i est

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

Les fréquences sont également souvent indiquées dans le tableau statistique, exprimées en %. L'utilisation des fréquences en remplacement des effectifs a ses avantages et ses inconvénients : indépendantes de l'effectif total, les fréquences permettent de "comparer" des échantillons de tailles différentes, mais en même temps le fait que la taille soit ainsi masquée peut mener à des interprétations abusives (60% de 3 a-t-il la même signification que 60% de 10000 ?). Il est donc préférable de maintenir dans le tableau statistique à la fois les effectifs et les fréquences, sauf dans le cas d'un échantillon de grande taille (≥ 1000) où les fréquences peuvent suffire.

Le regroupement des données n'est en théorie réalisable que dans le cas d'un caractère qualitatif ou quantitatif discret. Dans le cas d'une variable continue, il est clair que sur un échantillon fini seul un nombre fini de modalités peut être observé, et un regroupement est donc pratiquement possible, si tant est que l'on ait observé plusieurs fois la même valeur. Bien plus, on peut se demander si, étudiant par

exemple la taille des individus, il est vraiment nécessaire de distinguer une taille de $1,76\ m$ d'une taille de $1,77\ m$. On effectue alors un regroupement par **classes**, i.e. par intervalles réels. Le choix de ces intervalles est bien entendu quelque peu arbitraire, le "bon" choix étant affaire de "métier", c'est-à-dire d'expérience ; quelques règles cependant : ni trop, ni trop peu de classes, effectif de chaque classe ni trop petit ni trop grand. Le lecteur comprendra aisément que, si 90% des individus de l'échantillon ont une taille comprise entre $1,60\ m$ et $1,80\ m$, on découpera cet intervalle en plusieurs classes, alors que si seulement 1% des individus ont une taille comprise entre $2,10\ m$ et $2,30\ m$, on pourra peut-être considérer une classe plus grande !. En d'autres termes le choix des classes est souvent une affaire de bon sens après observation des résultats.

Lorsque l'on a regroupé en classes un caractère continu, on dit que l'on a **discrétisé** la variable continue ; l'on n'oubliera pas cependant qu'il s'agit toujours d'une variable continue et non discrète.

1.4 Séries simples, doubles, multiples - Chroniques

Une série statistique simple, ou à une dimension, est obtenue lorsque nous nous intéressons à un caractère élémentaire, dont l'ensemble des modalités est un sous-ensemble de R s'il est quantitatif.

Une série double (ou multiple) est obtenue lorsque à chaque individu sont associés deux (ou plusieurs) caractères élémentaires, plus précisément un couple (ou un n-uplet) de caractères élémentaires, ou encore un caractère à valeurs dans un produit cartésien ($\mathbb{N}^2$, $\mathbb{R}^2$ ou $\mathbb{N}^n$, $\mathbb{R}^n$ ou autres).

La différence entre l'étude de deux caractères simples sur la même population et d'un caractère double sur cette population peut paraître artificielle : elle est cependant essentielle. En effet, dans le cas d'une série double, nous nous intéressons pour chaque individu au couple (x,y) de réponses et nous effectuons le regroupement des données par rapport à ces couples, alors que dans le cas de l'étude des deux séries simples associées nous effectuons le regroupement des données séparément sur chacun des deux caractères X et Y ; nous obtenons alors des résultats plus concis, mais au prix d'une perte d'information.

Prenons un exemple concret :

nous effectuons un sondage auprès de nos étudiants en leur demandant leur note de mathématique au baccalauréat et le nombre de redoublements au cours de leur scolarité primaire et secondaire. Les résultats bruts obtenus sont les suivants (résultats fictifs) :

14,0	12,1	11,0	10,2	15,0	13,1	11,2	10,3
11,0	12,1	13,1	14,0	13,0	11,1	12,0	13,1

Soit X le caractère "note au bac" et Y le caractère "nombre de redoublements".

Les tableaux statistiques regroupant les données de X et de Y sont :

X	10	11	12	13	14	15
eff	2	4	3	4	2	1

et

Y	0	1	2	3
eff	7	6	2	1

alors que le tableau statistique regroupant les données du couple (X, Y) est :

Y/X	10	11	12	13	14	15
0	0	2	1	1	2	1
1	0	1	2	3	0	0
2	1	1	0	0	0	0
3	1	0	0	0	0	0

Nous pouvons remarquer que si les tableaux statitiques des variables X et Y ne permettent pas de reconstruire le tableau du couple (X, Y), ce dernier par contre permet de retrouver les tableaux de X et de Y en effectuant la somme des effectifs par colonne et par ligne :

Y/X	10	11	12	13	14	15	
0	0	2	1	1	2	1	7
1	0	1	2	3	0	0	6
2	1	1	0	0	0	0	2
3	1	0	0	0	0	0	1
	2	4	3	4	2	1	

Lorsque l'on étudie la loi du couple (X, Y) les distributions de X et de Y sont appelées **distributions marginales**, la distribution du couple étant la **distribution conjointe**

L'intérêt de la distribution conjointe par rapport aux distributions marginales est de permettre l'étude de la corrélation entre les deux caractères X et Y, i.e. de savoir s'il existe un rapport (et éventuellement quel rapport) entre la note au bac et le nombre de redoublements antérieurs. Les distributions doubles seront étudiées plus en détail au chapitre 3.

Les **chroniques** sont des séries statistiques simples très particulières pour lesquelles les modalités sont le **temps**, les effectifs étant remplacés par les valeurs d'un caractère quantitatif, les fréquences n'ayant alors aucune signification. Ainsi l'évolution de la température d'un patient au cours du temps constitue une chronique.

L'étude des chroniques étant très particulière (plus proche des séries doubles que des séries simples), on lui consacrera un chapitre propre.

1.5 Exercices

Exercice 1

Parmi les caractères suivants, indiquer ceux qui sont qualitatifs, quantitatifs, discrets, continus, et tenter de définir les individus et la population correspondante (préciser s'il s'agit d'une chronique) :

1. nombre d'élèves d'une classe
2. catégorie socio-professionnelle
3. salaire
4. nombre de mariages par année
5. défectueux ou non
6. vitesse d'un véhicule
7. durée de vie d'un disque dur
8. volume d'un objet

9. couleur des automobiles

Exercice 2

On s'intéresse à l'age des étudiants de première année. Indiquer le caractère, préciser s'il est qualitatif, quantitatif discret ou continu. Prélever un échantillon dont on précisera la taille.

Déterminer les effectifs et les fréquences des modalités observées.

Exercice 3

La population étant constituée par les étudiants de première année, on veut étudier la répartition de la taille et du sexe. Indiquer les caractères, préciser s'ils sont qualitatifs, quantitatifs discrets ou continus. Prélever un échantillon dont on préciser l'effectif total. Déterminer les effectifs et les fréquences des modalités observées.

Exercice 4

Lancer 50 fois une pièce de monnaie et observer le nombre de "pile" et de "face" obtenus. Quel est le caractère étudié ? est-il qualitatif, quantitatif discret ou continu ? Déterminer les effectifs et les fréquences des modalités observées. Quelle est la taille de votre échantillon ?

Exercice 5

On lance une pièce de monnaie jusqu'à obtenir "pile" et on note le nombre d'essais nécessaires. On répète cette opération 20 fois.
Quel est le caractère étudié ? est-il qualitatif, quantitatif discret ou continu ? Déterminer les effectifs et les fréquences des modalités observées. Quelle est la taille de votre échantillon ?

Chapitre 2

Distributions à un caractère

2.1 Représentations graphiques

Un tableau, aussi clair soit-il, n'est jamais aussi parlant qu'une représentation graphique, même si par ailleurs il recèle davantage d'informations. Dans le cas des statistiques, on parlera souvent de **diagrammes** au lieu de représentation graphique.

Les développements récents de l'informatique et notamment la vulgarisation de logiciels tels que tableurs, grapheurs, voire logiciels statistiques, ont rendu cette opération aisée, et surtout offrent un choix très variés de diagrammes, plus au moins esthétiques, mais aussi plus ou moins appropriés. N'oublions pas que si la "beauté" d'un graphisme peut aider à la compréhension d'un phénomène, alors c'est un plus ; mais si cet aspect plaisant masque les informations sous-jacentes plus qu'il ne les met en valeur, alors il faut le proscrire. Les revues, économiques et commerciales notamment, regorgent de diagrammes certes très agréables à l'oeil, mais parfois totalement inappropriés au sujet qu'ils sont censés étudier.

Nous allons présenter ici les types de diagrammes les plus courants, et sans aucun doute les plus parlant (au moins pour le spécialiste), en précisant pour chacun d'eux à quel type de distribution il s'applique. Mais auparavant, précisons les éléments communs à tout diagramme statistique :

- un **titre** présentant clairement la population et le caractère étudié, ainsi que les conditions de l'enquête ; au besoin l'on complètera le titre en **commentaires**
- la nature des éléments portés sur chaque axe : modalités en abscisses, effectifs ou fréquences (ou effectifs ou fréquences cumulés - voir plus loin), ainsi que les unités utilisées.
- si plusieurs échantillons sont étudiés, une **légende** présentera les symboles utilisés pour chaque échantillon.

Pour présenter les différents diagrammes utilisables, et pouvoir comparer l'intérêt de ces diagrammes, nous considérerons trois caractères particuliers d'une même population (étudiante) :

- un caractère qualitatif X, les départements de naissance
- un caractère quantitatif discret Y, le nombre de redoublement antérieurs à l'obtention du baccalauréat
- un caractère quantitatif continu, la taille (en cm)

Les résultats obtenus, après regroupement, sont représentés par les tableaux statistiques suivants :

X	57	54	38	81	99
eff	7	6	1	1	1

(la nomenclature utilisée pour le codage des départements est celle de l'INSEE, les départements ayant été ordonnés par ordre croissant d'éloignement au département de prélèvement de l'échantillon (57))

Y	0	1	2	3
eff	7	6	2	1

Z	163	170	171	175	178	180	181	182	185
eff	1	1	3	1	3	2	1	2	6

2.1.1 Diagramme à bâtons - diagramme en tuyaux d'orgues - diagramme linéaire

Ces diagrammes sont essentiellement utilisés pour des caractères qualitatifs, et éventuellement quantitatifs discrets. Les modalités sont portées en abscisse, régulièrement espacées, dans l'ordre que l'on a choisi (naturel ou non). Les effectifs sont portés en ordonnées et marqués

- par un trait vertical depuis l'axe des abscisses jusqu'au point dont l'ordonnée est l'effectif : **diagramme à bâtons**
- par un rectangle vertical dont la base est centrée sur la modalité concernée et de largeur telle que les bases de deux rectangles voisins ne se chevauchent pas, et dont la hauteur indique l'effectif : **diagramme en tuyaux d'orgue**
- par un symbole (cercle, carré, losange, etc...) centré au point de coordonnées (modalité, effectif) : **diagramme linéaire**.

Les tableurs, logiciels usuellement utilisés pour tracer des diagrammes, n'offrent en général pas de diagrammes à bâtons, les diagrammes en tuyaux d'orgue sont improprement appelés histogrammes, et les diagrammes linéaires offrent plusieurs options suivant que les points sont ou non reliés entre eux : il est clair que nous envisageons le cas des points isolés, non reliés.

Notons que ces diagrammes peuvent tout aussi bien représenter la distribution des effectifs que celle des fréquences ; en réalité le diagramme est le même dans les deux cas, seules les unités sont différentes. Il est donc possible de faire apparaître deux axes des ordonnées, l'un - à gauche du graphe - étant gradué de manière à présenter les effectifs, l'autre - à droite - étant gradué de manière à représenter les fréquences associées.

La figure 2.1 présente le diagramme en tuyau d'orgue de X et le diagramme linéaire de Y

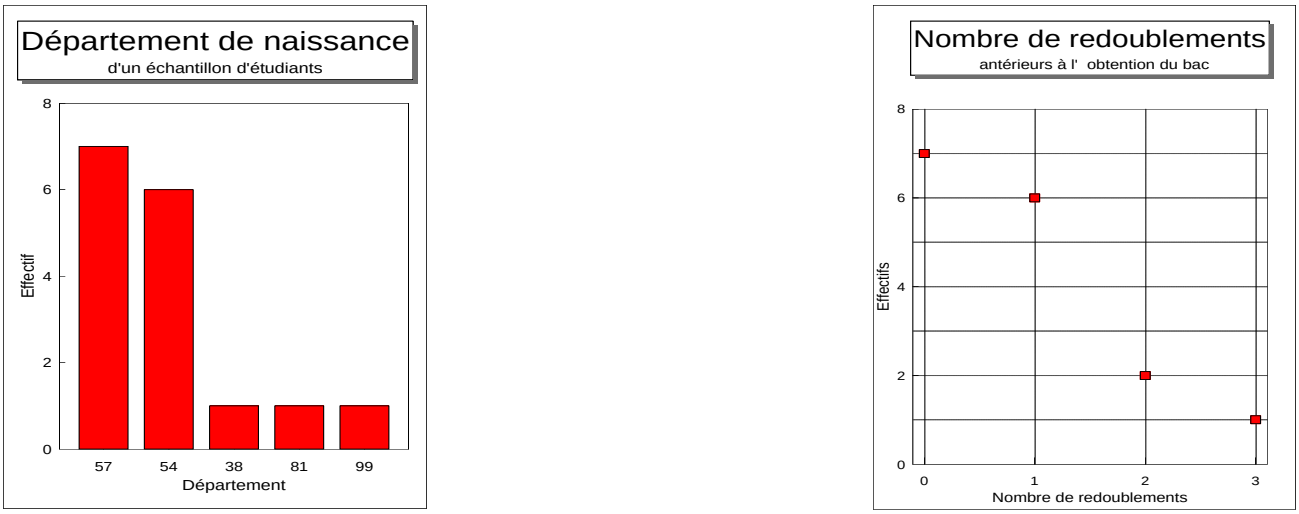


FIG. 2.1 – Diagramme en tuyaux d'orgue et diagramme linéaire

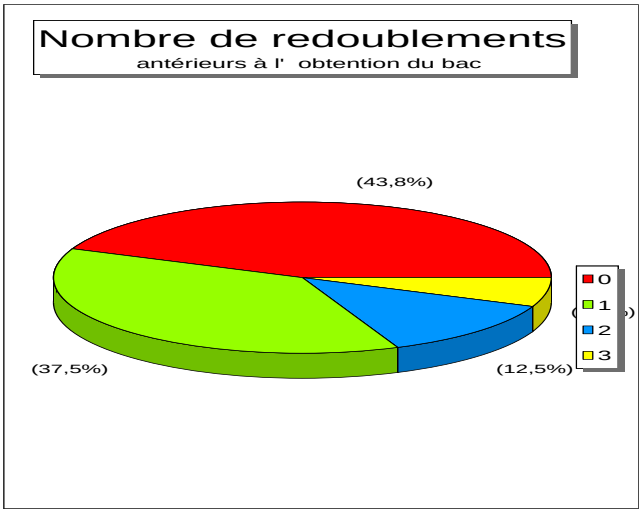


FIG. 2.2 – Diagramme circulaire

On constatera que le diagramme en tuyaux d'orgue est plus "clair" que le diagramme linéaire. Le lecteur se convaincra cependant aisément que dans le cas d'un nombre relativement important de modalités le diagramme linéaire l'emportera en lisibilité.

2.1.2 Diagrammes circulaires (camenberts) et semi-circulaires

Ces diagrammes sont très utiles dans le cas de caractères qualitatifs ou quantitatifs discrets pour représenter la distribution des fréquences. Chaque modalité y est représentée par un secteur angulaire dont la surface (et donc également l'angle au centre) est proportionnelle à sa fréquence.

On utilisera ces diagrammes lorsque le nombre de modalités représentées n'est pas trop important, et souvent dans le cas d'un caractère regroupé en classes. Le diagramme circulaire sera utilisé lorsque les modalités ne sont pas ordonnées naturellement (ou alors lorsqu'elles sont ordonnées "circulairement" comme les mois de naissance) ; dans le cas contraire on préférera le diagramme semi-circulaire.

Notons également que la lisibilité de ce type de diagramme est nettement améliorée par l'utilisation de couleurs.

La figure 2.2 donne le diagramme circulaire de la distribution Y.

2.1.3 Cartogrammes

Les cartogrammes - que l'on retrouve dans tous les journeaux les lendemains d'élection - sont très utiles pour représenter la distribution d'un caractère "géographique", comme les départements de naissance : le cartogramme serait alors constitué d'une carte géographique des départements coloriés de couleurs différentes selon les effectifs correspondants.

2.1.4 Histogrammes - Diagrammes en aires

Les histogrammes sont utilisés dans le cas de modalités continues regroupées par classes. Chaque modalité est représentée par un rectangle dont la base est l'intervalle de classe, et dont l'aire est l'effectif (ou la fréquence). Ces rectangles sont donc jointifs.

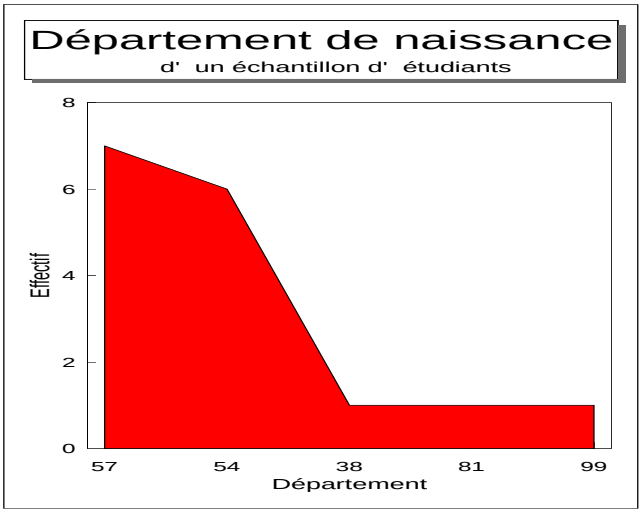


FIG. 2.3 – Diagramme en aire

Les tableurs ne permettent pas en général le tracé d'histogrammes (bien que ce nom soit utilisé, mais pour les diagrammes en tuyaux d'orgue) sauf dans le cas de classes de même amplitude (et encore les rectangles ne seront-ils pas jointifs).

On pourrait penser que les diagrammes en aires offerts par les tableurs comblent cette lacune, mais ce n'est pas le cas. Ce qu'ils appellent diagramme en aire n'est qu'un diagramme linéaire dans lequel on a relié les points, et marqué (par un grisé ou une couleur) la surface située sous la "courbe" ainsi obtenue, comme on peut le constater sur la figure 2.3 relative à la distribution X.

2.1.5 Diagrammes XY - polygone

Un diagramme XY ou polygone (des effectifs ou des fréquences) est utilisé dans le cas d'un caractère quantitatif. Il ressemble au diagramme linéaire à cette différence que

- les points correspondants aux observations sont reliés par les segments de droite (on suppose la distribution linéaire entre deux modalités observées)
- les modalités sont de véritables nombres et donc sont positionnés en fonction de leur valeur.

La figure 2.4 qui présente les polygones des effectifs des caractères X et Z permet de comprendre son intérêt dans le cas de Z et son absurdité dans le cas de X !

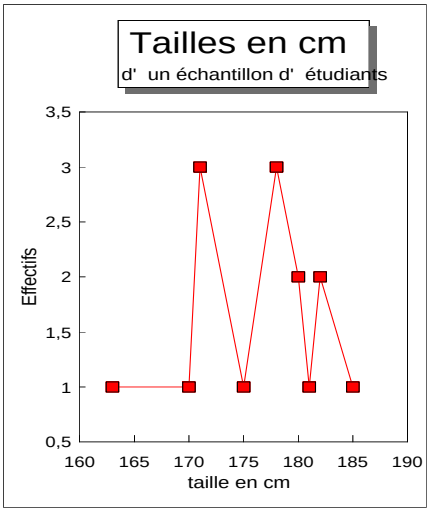
2.2 Indicateurs de position

Appelés également **caractéristiques de position** ou de **tendance centrale**, les indicateurs (soyons modestes, comme toujours en statistique) de position tentent de donner une information sur la valeur de la modalité "autour de laquelle se situent les autres modalités" (d'où le terme de tendance "centrale").

Cette notion suppose donc que les modalités constituent un ensemble ordonné, ou pour le moins que la notion de "distance" entre deux modalités ait un sens.

Il est bien évident que seuls les caractères quantitatifs permettent de donner un sens précis à ces notions intuitives.

Nous allons cependant débuter par un indicateur prenant un sens même dans le cas de caractères qualitatifs (même lorsque les modalités sont non ordonnées).



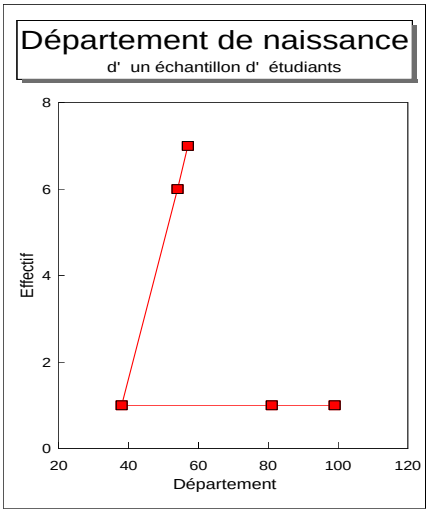


FIG. 2.4 – diagrammes XY !

2.2.1 Modes

Définition 1 *On appelle* **mode** *d'une série statistique toute valeur de la modalité dont l'effectif est maximum.*
Si $X = (x_i, n_i)_{i \in 1..p}$ *est une variable statistique, son mode est toute modalité* x_i *telle que*

$$n_i = \max_{j \in 1..p} n_j$$

Remarque

On trouve parfois la définition suivante : **le mode d'une série statistique est la valeur la plus fréqente**. Cette définition, qui a l'avantage de la concision et de la clarté intuitive, a cependant un inconvénient, à savoir l'affirmation - purement gratuite - de l'unicité d'un mode. Il est par contre clair que pour la plupart des séries statistiques il y a effectivement unicité.

Définition 2 *On appelle* **classe modale** *d'une série statistique regroupée en classes toute classe dont le rapport* $\frac{effectif}{largeur\ de\ la\ classe}$ *est maximum.*

Remarque

Si les largeurs des classes sont toutes égales, une classe modale est une classe dont l'effectif est maximum (classe la plus fréquente).

Notons également que le rapport (effectif)/(largeur de la classe) est la hauteur du rectangle associé à cette classe dans un histogramme ; on repèrera donc la ou les classes modales à partir d'un histogramme lorsque les données sont regroupées en classes.

Définition 3 *Une série statistique n'admettant qu'un seul mode est dite* **unimodale** *; elle est dite* **multimodale** *dans le cas contraire, plus précisément* **bimodale** *dans le cas de deux modes,* **trimodale** *dans le cas de trois modes, etc...*

Remarque

Le fait qu'une série soit multimodale traduit souvent l'existence de plusieurs populations ayant des caractéristiques différentes au sein de la population étudiée (caractère prenant des valeurs différentes chez les hommes et chez les femmes par exemple). Il est cependant clair que dans de telles situations, il est peu vraissemblable que les effectifs de deux modalités soit strictement égaux : nous aurons plus vraissemblablement deux modalités à effectif plus important que les autres, égaux ou non.

Dans le cas d'une variable continue, cette notion correspondrait à celle de maximum relatif par opposition à la notion stricte de mode qui est un maximum absolu. Nous allons donc tenter de préciser :

Définition 4 *On appelle* **mode secondaire** *toute valeur de la modalité dont l'effectif, sans être maximum, est cependant significativement plus important que les autres.*

Cette définition ne saurait bien entendu être considérée comme une définition mathématique, le terme "significativement plus important" étant subjectif. Elle sort donc du cadre de **la** statistique, mais ne peut être négligée dans le cadre d'une interprétation statistique des résultats.

Remarque

Nous verrons plus loin (indicateurs de dispersion) l'intérêt qu'il y a à définir un mode numérique au lieu d'une classe modale dans le cas de données regroupées. Les procédés proposés sont cependant alors plus ou moins artificiels, et l'indicateur de position restera la classe modale (contrairement à ce qu'affirment certains ouvrages - d'où la position de cette remarque dans le cours -).

Toujours dans le cas de données regroupées, il sera parfois utile de "découper" une classe modale en plusieurs sous-classes de manière à mieux préciser cette notion de mode (d'où l'intérêt de toujours conserver les données brutes !).

2.2.2 Médiane - distribution des effectifs et des fréquences cumulées - Fonction de répartition expérimentale

Si le mode est le seul indicateur de position s'appliquant à toute série statistique, la médiane est le seul indicateur (avec le mode) s'appliquant aux séries qualitatives ordonnées. Il s'applique bien sûr également aux séries quantitatives.

Remarquons que lorsque nous parlons de modalités ordonnées, nous sous-entendons munies d'un **ordre total**, i.e. tel que deux modalités soient toujours dans un ordre défini (une hiérarchie par exemple est bien un ordre, mais non total). Nous considérons bien entendu également un ordre ayant une véritable signification par rapport au caractère étudié, et non artificiel comme dans le cas d'un codage ou d'une nomenclature : la notion de médiane perdrait alors toute signification pratique.

Cas d'un caractère discret

Définition 5 *On appelle* **médiane** *d'une série statistique ordonnée discrète* $X = (x_i, n_i)_{i\in 1..p}$ *(dire que la série est ordonnée c'est dire que* $i < j \Longrightarrow x_i < x_j$*) toute valeur* x_i *de la modalité telle que*

$$\sum_{j\ /\ x_j<x_i} n_j \leq \frac{N}{2} \text{ et } \sum_{j\ /\ x_j>x_i} n_j \leq \frac{N}{2}$$

N désignant l'effectif total, i.e. $N = \sum_{i=1}^{p} n_i$

Remarque

Les deux sommes intervenant dans la définition précédente sont liées entre elles, en ce sens que

$$n_i + \sum_{j\ /\ x_j<x_i} n_j + \sum_{j\ /\ x_j>x_i} n_j = N$$

par conséquent

Proposition 1 x_i *est une médiane d'une série statistique ordonnée discrète* $X = (x_j, n_j)_{j\in 1..p}$ *si et seulement si*

$$\frac{N}{2} - n_i \leq \sum_{j\ /\ x_j<x_i} n_j \leq \frac{N}{2}$$

N désignant l'effectif total.

démonstration

il suffit de remarquer que $\sum_{j\ /\ x_j>x_i} n_j = N - n_i - \sum_{j\ /\ x_j<x_i} n_j$ et donc que

$$\sum_{j\ /\ x_j>x_i} n_j \leq \frac{N}{2} \iff N - n_i - \sum_{j\ /\ x_j<x_i} n_j \leq \frac{N}{2}$$
$$\iff \frac{N}{2} - n_i \leq \sum_{j\ /\ x_j<x_i} n_j$$

Cette proposition met en évidence l'importance de la somme

$$\sum_{j\ /\ x_j\leq x_i} n_j$$

(en réalité $\sum_{j\ /\ x_j<x_i} n_j$)
ce qui justifie la définition

Définition 6 *Etant donné une série statistique ordonnée discrète* $X = (x_i, n_i)_{i\in 1..p}$ *on appelle* **effectif cumulé** *associé à la modalité* x_i *et on note* n_{c_i} *le nombre*

$$n_{c_i} = \sum_{j\ /\ x_j \leq x_i} n_j$$

On définit de même la notion de fréquence cumulée

Définition 7 *Etant donné une série statistique ordonnée discrète* $X = (x_i, n_i)_{i\in 1..p}$ *on appelle* **fréquence cumulée** *associée à la modalité* x_i *et on note* fc_i *le nombre*

$$f_{c_i} = \sum_{j\ /\ x_j \leq x_i} f_j = \frac{1}{N} n_{c_i}$$

La notion de médiane peut alors s'exprimer par

Proposition 2 x_i *est une médiane d'une série statistique ordonnée discrète* $X = (x_i, n_i)_{i\in 1..p}$ *si et seulement si*

$$\frac{1}{2} - f_i \leq f_{c_{i-1}} \leq \frac{1}{2}$$

ou encore si et seulement si

$$\frac{1}{2} \leq f_{c_i} \leq \frac{1}{2} + f_i$$

Pour déterminer la ou les médianes d'une série statistique, on calculera donc la série des fréquences cumulées, et on pourra alors tracer un diagramme des fréquences cumulées. Les médianes sont alors l'une, ou les deux valeurs de modalité suivant la fréquence cumulée 1/2. Plus précisément

- si la valeur 1/2 n'est pas obtenue, seule la modalité suivant immédiatement la fréquence cumulée 1/2 est médiane.
- si la valeur 1/2 est obtenue, la valeur de la modalité correspondant à 1/2 et la suivante sont toutes deux médianes.

Prenons l'exemple du caractère Y :

Y	0	1	2	3
eff	7	6	2	1

Le tableau des fréquences cumulées est

Y	0	1	2	3
$f.c.$	7/16	13/16	15/16	1

La valeur 1/2 n'est pas obtenue ; seul 1 est médiane.

Remarque

Certains auteurs préfèrent définir une médiane unique, ce qui les invite à considérer comme médiane, dans le cas d'un caractère discret, la demi-somme des valeurs des modalités dont les fréquences cumulées

encadrent 1/2 lorsque 1/2 n'est pas atteint, et la valeur de la modalité dont l'effectif cumulé est 1/2 lorsque cette valeur est atteinte.

Cette convention a deux inconvénients majeurs : elle ne correspond pas à notre définition lorsque 1/2 est atteint et dans le cas contraire la médiane n'est plus une valeur de la modalité, si tant est que la somme de deux modalités ait un sens (ce qui n'est le cas que pour un caractère quantitatif) !

Je voudrais à ce propos attirer l'attention du lecteur sur une différence de définition des fréquences ou des effectifs cumulés suivant les auteurs, essentiellement entre les conventions francophones et anglo-saxones. Lorsque nous considérons une inégalité large, les anglo-saxon considèrent souvent une inégalité stricte. Si donc la fréquence cumulée de x_i est la somme des fréquences des modalités strictement inférieure à x_i, la définition que nous venons de citer est conforme à la nôtre (à l'unicité près). Prenez donc garde aux ouvrages traduit de l'anglais ou de l'américain, surtout lorsque le traducteur a francisé certaines définitions, mais a oublié d'en tenir compte dans des résultats connexes.

Cas d'un caractère qantitatif continu - Fonction de répartition expérimentale

Définition 8 *Si X est un caractère continu sur une population finie, on appelle* **fonction de répartition** *de X la fonction F de $\mathbb{R}$ dans $\mathbb{R}$ définie par*

$$F(x) = \frac{1}{N} \sum_{x_i \leq x} n_i$$

N désignant l'effectif total, n_i l'effectif associé à la modalité x_i.
Si X est la variable statistique déduite d'un échantillon, F est appelée **fonction de répartition expérimentale**.

On déduit immédiatement de cette définition les propriétés suivantes :

Proposition 3 *La fonction de répartition (expérimentale) d'une série statistique est une fonction en escalier croissante, et*

$$\lim_{x \to -\infty} F(x) = 0 \quad et \quad \lim_{x \to +\infty} F(x) = 1$$

Remarque

La fonction de répartition expérimentale a une représentation graphique proche du diagramme à bâtons des fréquences cumulées :

- pour une valeur x de la modalité, $F(x)$ est la fréquence cumulée de x, et
- entre deux valeurs de la modalité, F a pour valeur la fréquence cumulée de la plus petite des deux modalités.

Cependant, sur une population infinie (ou finie d'effectif total très élevé), on peut raisonnablement supposer qu'entre deux valeurs observées consécutives l'évolution de la fonction de répartition est progressive (continue ou non). Il est donc raisonnable, à défaut d'autre indication, d'approcher la fonction de répartition réelle par une ligne polygonale joignant les points d'observation. Nous obtenons alors cette approximation de la fonction de répartition réelle en construisant le diagramme XY des fréquences cumulées avec lignes intermédiaires. F désignant la fonction ainsi obtenue, il est naturel de prendre pour

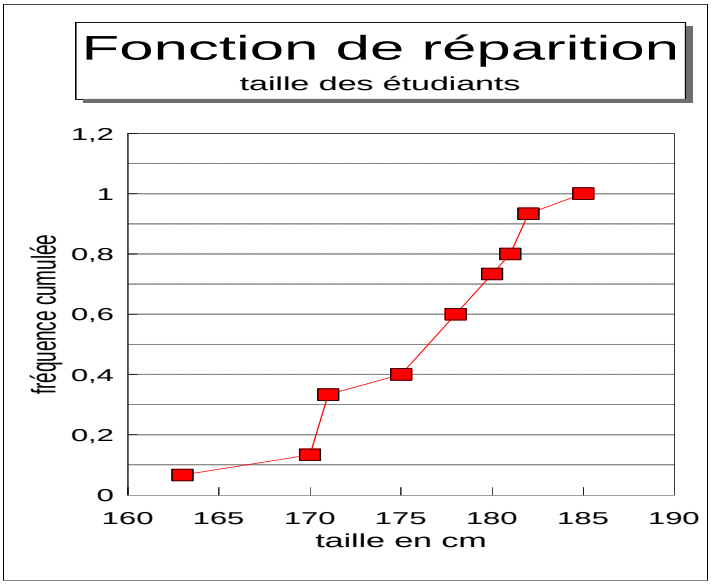


FIG. 2.5 – Fonction de répartion de la distribution Z

médiane de la série statistique la valeur de x pour laquelle $F(x) = 1/2$ (voir figure 2.5 cette fonction pour la variable Z).

D'où

Définition 9 *Si $X = (x_i, n_i)$ est une série statistique correspondant à un caractère continu, sa médiane m est donnée par*

$$m = x_i + \frac{f_{c_{i+1}} - f_{c_i}}{x_{i+1} - x_i} \times (\frac{1}{2} - f_{c_i})$$

où i vérifie

$$f_{c_i} \leq \frac{1}{2} < f_{c_{i+1}}$$

Si nous prenons l'exemple de la distribution Z, la distribution des fréquences cumulées est

Z	163	170	171	175	178	180	181	182	185
$fréq.cum.$	0.07	0.13	0.33	0.40	0.60	0.73	0.80	0.93	1

Sa médiane est comprise entre $x_i = 175$ et $x_{i+1} = 178$: elle vaut donc très exactement

$$m = 175 + \frac{178 - 175}{0.6 - 0.4} \times (0.5 - 0.4) = 175 + 1.5 = 176.5$$

2.2.3 Moyennes

La moyenne est sans doute l'indice de position le plus connu, ne serait ce que parce que la moyenne de leurs notes est d'une importance primordiale pour les élèves et les étudiants. Il reste cependant que pour eux le mot de moyenne est toujours associé à la notion de moyenne arithmétique, éventuellement

pondérée. En réalité il existe d'autres moyennes (**géométriques, harmoniques, quadratiques**, etc..) et, selon les caractères étudiés et les conditions d'expérience, il s'agira d'employer la "bonne" moyenne.

Tentons de donner une définition générale de la notion de "moyenne" :

Définition 10 *Si nous mesurons deux grandeurs a et b, leur moyenne est la grandeur constante m qui, remplaçant les grandeurs a et b dans les mêmes conditions d'expérience, produit le même résultat.*

Explicitons cette définition par quelques exemples.

Exercice 6

|| Nous montons deux résistances R_1 et R_2 en série. Quelle est la résistance moyenne ?

La résistance moyenne est la résistance R_m qui, remplaçant les deux résistances R_1 et R_2, donne le même résultat, i.e. la même tension aux bornes pour une intensité donnée, ou la même intensité pour une tension donnée. Dans tous les cas de figure les lois de l'électricité nous indiquent que

$$2R_m = R_1 + R_2$$

ou

$$R_m = \frac{R_1 + R_2}{2}$$

La résistance moyenne de deux résistances montées en série est donc leur moyenne arithmétique.

Exercice 7

|| Nous montons deux résistances R_1 et R_2 en parallèle. Quelle est la résistance moyenne ?

Dans ce cas les lois de l'électricité nous indiquent que

$$\frac{1}{R_1} + \frac{1}{R_2} = 2\frac{1}{R_m}$$

d'où

$$\frac{1}{R_m} = \frac{1}{2}\left(\frac{1}{R_1} + \frac{1}{R_2}\right)$$

ou encore

$$R_m = \frac{2R_1R_2}{R_1 + R_2}$$

ce qui exprime que la résistance moyenne de deux résistances montées en parallèle est leur **moyenne harmonique**.

Remarque

Dans les deux cas précédents, la notion de résistance moyenne se rapproche de celle de résistance équivalente, la différence étant que la résistance équivalente est la résistance unique qui remplaçant les deux autres fournit le même résultat. On remarquera que dans le cas d'un montage en série la résistance moyenne de n résistance est leur résitance équivalente divisée par n, tandis que dans le cas d'un montage en parallèle, la résistance moyenne de n résistances est leur résistance équivalente multipliée par n.

Exercice 8

> Votre patron vous tient le langage suivant :
> "l'entreprise a de grosses difficultés financières cette année, mais l'an prochain sera excellent grâce à un nouveau contrat que nous venons de conclure. C'est pourquoi je vous propose de diminuer votre salaire de 40% cette année et de l'augmenter de 50% l'an prochain, ce qui vous fera une augmentation moyenne de 5% sur les deux ans à venir."
> Signez-vous ce contrat ?

Soit S votre salaire annuel. Avec les conditions du contrat proposé, votre salaire sera, en posant $t_1 = -40\%$ et $t_2 = 50\%$, $S(1+t_1)$ au bout d'un an et $S(1+t_1)(1+t_2)$ au bout des deux ans. Avec une augmentation constante t_m sur chacune des deux années, votre salaire au bout de deux ans aurait été $S(1+t_m)(1+t_m)$. L'augmentation moyenne t_m vérifie donc $(1+t_m)^2 = (1+t_1)(1+t_2)$, d'où

$$1 + t_m = \sqrt{(1+t_1)(1+t_2)}$$

ce qui, avec les termes du contrat, nous donne une dimimution moyenne d'un peu plus de 5% !

On notera que les spécialistes de l'économie parlent plus souvent de coefficient d'augmentation que de taux d'augmentation, ce coefficient étant égal au taux plus 1. Dans ces conditions le coefficient moyen c_m est égal à $\sqrt{c_1 c_2}$, c'est-à-dire la **moyenne géométrique** des coefficients c_1 et c_2

Exercice 9

> On effectue deux mesures de l'intensité dans un même circuit pour une même tension continue aux bornes (cette intensité est théoriquement constante mais les différentes "erreurs" de mesure font que nous lirons deux intensités distinctes I_1 et I_2). Quelle est l'intensité moyenne ?

La question que nous devons nous poser est de savoir quelle est la grandeur dont l'addition a un sens physique. Il s'agit bien entendu ici de l'énergie, ou de la puissance. La puissance moyenne sera donc la moyenne arithmétique des puissances, et l'intensité moyenne celle qui correspond à la puissance moyenne. D'où

$$Z\, I_m^2 = \frac{1}{2}(Z\, I_1^2 + Z\, I_2^2)$$

où Z désigne l'impédance du circuit. Donc

$$I_m = \sqrt{\frac{I_1^2 + I_2^2}{2}}$$

L'intensité moyenne est la **moyenne quadratique** des deux intensités lues.

Exercice 10

> On prélève deux échantillons de même volume V d'un acide, et on mesure leur PH de valeur PH_1 et PH_2 respectivement. Quel est le PH moyen.

Il est clair que la concentration moyenne C_m est la moyenne arithmétique des concentrations C_1 et C_2. Comme $PH = -\log C$, nous obtenons

$$10^{-PH_m} = \frac{1}{2}\left(10^{-PH_1} + 10^{-PH_2}\right)$$

soit

$$PH_m = -\log\left(\frac{1}{2}\left(10^{-PH_1} + 10^{-PH_2}\right)\right)$$

Nous terminerons cette section par les définitions des moyennes les plus usuelles :

Définition 11 *Etant donné une série $X = (x_i, n_i)$, on appelle* **moyenne arithmétique** *de la série et on note $\overline{X}$ ou $E(X)$ le nombre*

$$E(X) = \frac{1}{N}\sum_{i=1}^{n} n_i x_i$$

où N est l'effectif total.

La notation $\overline{X}$ est davantage statistique tandis que la notation $E(X)$ est davantage probabiliste. Nous utiliserons indifféremment l'une ou l'autre de ces deux notations.

Définition 12 *Etant donné une série $X = (x_i, n_i)$ telle que les x_i soient tous non nuls, on appelle* **moyenne harmonique** *de la série le nombre m tel que $1/m$ soit la moyenne arithmétique de la série $(1/x_i, n_i)$.*

$$\frac{1}{m} = \frac{1}{N}\sum_{i=1}^{n} n_i \frac{1}{x_i}$$

où N est l'effectif total.

Définition 13 *Etant donné une série $X = (x_i, n_i)$ telle que les x_i soient tous positifs, on appelle* **moyenne géométrique** *de la série le nombre m tel que $\ln m$ soit la moyenne arithmétique de la série $(\ln x_i, n_i)$.*

$$\ln m = \frac{1}{N}\sum_{i=1}^{n} n_i \ln x_i$$

où N est l'effectif total.

Définition 14 *Etant donné une série $X = (x_i, n_i)$ telle que les x_i soient tous positifs, on appelle* **moyenne quadratique** *de la série le nombre positif m tel que m^2 soit la moyenne arithmétique de la série (x_i^2, n_i).*

$$m^2 = \frac{1}{N}\sum_{i=1}^{n} n_i x_i^2$$

où N est l'effectif total.

On remarquera que toutes ces moyennes peuvent se ramener au calcul d'une moyenne arithmétique, quitte à modifier la grandeur étudiée. C'est pourquoi en statistique on étudie exclusivement la moyenne arithmétique ; c'est à l'utilisateur qu'il appartient de choisir correctement la grandeur étudiée.

2.2.4 Moyenne arithmétique : propriétés

Proposition 4 *Soit $X = (x_i, n_i)$ une série statistique. Nous noterons $X + a$ la série $X = (x_i + a, n_i)$ et λX la série $X = (\lambda x_i, n_i)$. Alors*

$$E(X + a) = E(X) + a$$

et

$$E(\lambda X) = \lambda E(X)$$

Remarque

Les propriétés précédentes expriment, dans le cas de mesures de grandeurs, que la moyenne arithmétique ne dépend pas ni choix de l'origine ni de l'unité. Ainsi une moyenne de températures exprimées en degrés Celsius sera obtenue en degrés Celsius et une moyenne de températures exprimées en degrés Kelvin sera obtenue en degré Kelvin. Ceci est évidemment la moindre des choses que l'on peut attendre d'une notion ayant un sens "physique".

Ces propriétés sont également utilisées pour faciliter les calculs et l'entrée des données dans un programme de calcul. Si nous considérons par exemple la série Z donnant la taille d'un échantillon d'étudiants, on peut considérer la variable $Z - 100$ (si l'on ne veut pas faire des calculs "de tête" trop compliqués, ou mieux $Z - 180$ en choisissant une valeur "centrale" pour le décalage. Nous obtenons alors le tableau suivant :

Z	163	170	171	175	178	180	181	182	185
$Z - 180$	-17	-10	-9	-5	-2	0	1	2	5
eff	1	1	3	1	3	2	1	2	6

d'où

$$E(Z - 180) = (-17 - 10 - 3 \times 9 - 5 - 3 \times 2 + 1 + 2 \times 2 + 6 \times 5)/20 = -30/20 = -1.5$$

et donc

$$E(Z) = 180 - 1.5 = 178.5$$

Nous aurions pu également regrouper en classes les résultats de la série Z (notamment dans le cas d'un échantillon de taille plus importante), et considérer des classes de longueur 0.5 dont les centres sont donnés dans le tableau suivant ; on peut alors se ramener à la série $(Z - 180)/5$:

Z_1	165	170	175	180	185
$\frac{Z-180}{5}$	-3	-2	-1	0	1
eff	1	4	1	8	6

D'où

$$E\left(\frac{Z_1 - 180}{5}\right) = (-3 - 4 \times 2 - 1 + 6)/20 = -6/20 = -0.3$$

et donc

$$E(Z_1) = 5E\left(\frac{Z_1 - 180}{5}\right) + 180 = 180 - 5 \times 0.3 = 178.5$$

On notera que nous obtenons $E(Z) = E(Z_1)$, ce qui est bien sûr une pure coïncidence.

2.2.5 Quantiles : quartiles, déciles, centiles

La notion de quantile généralise celle de médiane. Si nous introduisons cette notion à la fin des indices de position et juste avant les indices de dispersion, c'est parce que, bien qu'indices de position, les quantiles introduisent surtout des indices de dispersion et ne sont pas (la médiane exceptée) des indices de tendance centrale.

Définition 15 *Soit $X = (x_i, n_i)_{i\in 1..n}$ une série statistique quantitative.*
Soit k un entier compris entre et 99.
Le k-ième **centile** *est toute valeur x_i telle que*

$$\sum_{j=1}^{i-1} n_j \leq k/N$$

et

$$\sum_{j=i+1}^{n} n_j \leq (N-k)/N$$

N désignant l'effectif total, i.e.

$$N = \sum_{i=1}^{n} n_i$$

Il s'ensuit immédiatement que

Proposition 5 *La médiane est le 50-ème centile.*

Définition 16 *Le premier* **quartile** *est le 25-ème centile ;*
le deuxième quartile est le 50-ème centile, donc la médiane ;
le troisième quartile est le 75-ème centile.
k étant un entier compris entre 0 et 9, le k-ème **décile** *est le centile d'ordre $10k$.*

Quantile est un terme général pour désigner aussi bien un centile, qu'un décile ou la médiane.

2.3 Indices de dispersion

Les indices de dispersion sont censés renseigner sur la manière dont les données se répartissent autour d'une valeur centrale, en terme d'éloignement. Ce sont donc des nombres, et par conséquent cette notion ne peut avoir de sens que pour les caractères quantitatifs.

2.3.1 Etendue

Définition 17 *On appelle* **étendue** *d'une distribution statistique quantitative la différence entre la plus grande et la plus petite valeur observée.*

L'étendue de la distribution Z qui nous sert d'exemple est donc $185 - 163 = 22$. Cette étendue est bien sûr exprimée dans l'unité utilisée pour les modalités x_i, ici le cm.

Bien que très "primitif", cet indice de dispersion n'est pas à négliger.

2.3.2 Intervalle et écart interquartile - boite de dispersion (boxplot)

Définition 18 *On appelle* **écart interquartile** *d'une distribution statistique la différence entre le troisième et le premier quartile.*
On appelle **intervalle interquartile** *d'une distribution statistique l'intervalle $[q_1, q_3]$, où q_1 et q_3 désignent respectivement le premier et le troisième quartile.*

Remarque

On prendra garde au fait que certains auteurs utilisent intervalle interquartile pour écart interquartile.

Proposition 6 *Soit q_1 et q_3 le premier et le troisième quartile respectivement d'une distribution X. Si le caractère est quantitatif continu, l'intervalle interquartile $[q_1, q_3]$ contient alors 50% de la population expérimentale, et contient la médiane $q_2 = m$.*

Définition 19 *Le* **diagramme en boite interquartile** *ou* **boite de dispersion** *(traduction - laborieuse - de l'anglais* **IQboxplot***) d'une distribution X est construit de la manière suivante :*

- *sur un axe gradué on porte la médiane, le premier et le troisième quartile (q_1 et q_3) ;*
- *on construit ensuite autour de cet axe - et centré sur l'axe - un rectangle de hauteur arbitraire et de longueur correspondant aux deux quartiles ; la médiane est repérée par un trait plein dont la hauteur est celle du rectangle ;*
- *0n porte de chaque coté des quartiles une longueur égale à 1.5 fois l'intervalle interquartile et on marque (par un trait orthogonal à l'axe qui est prolongé jusqu'à ces points) les modalités observées d_1 et d_3 les plus proches des points ainsi obtenus (vers les quartiles) ; on obtient ainsi des "moustaches" (whiskers en anglais)*
- *on place - et on marque par un disque - toutes les valeurs observées de la modalité extérieures à l'intervalle $[d_1\ ,\ d_3]$; ces valeurs sont dites* **abérantes**

La figure 2.6 montre la boite de dispersion de notre distribution Z. On remarquera la forte dissymétrie de ce diagramme, avec l'absence complète de moustache vers la droite, ce qui se produit lorsque au moins le quart de l'effectif total se retrouve sur la plus grande valeur observée.

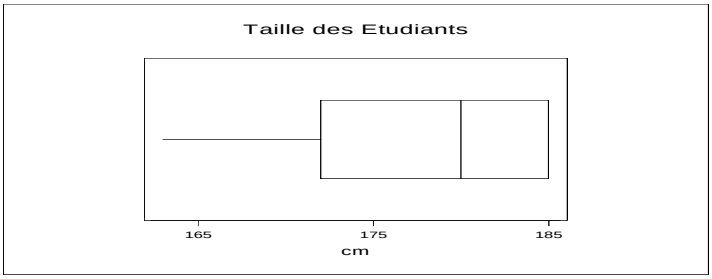


FIG. 2.6 – boite de dispersion de la distribution Z

Remarque

Outre le fait de marquer les points abérants, ces diagrammes sont très utiles pour comparer deux distributions, comme par exemple les durées de vie des hommes et des femmes d'une même population, ou des caractéristiques de pièces fabriquées par des machines théoriquement identiques.

2.3.3 Ecart absolu moyen

Définition 20
Etant donné une série statistique quantitative $X = (x_i, n_i)_{i \in ..p}$*, on appelle* **écart absolu moyen** *par rapport à une valeur centrale* α *le nombre*

$$e_\alpha = \frac{1}{N} \sum_{i=1}^{p} n_i |x_i - \alpha|$$

où N *désigne l'effectif total.*

Il s'agit donc d'une moyenne arithmétique des écarts absolus à la valeur α. Si nous prenons l'exemple d'un village - supposé linéaire - tel que à la position x_i (abscisse sur un axe) logent n_i personnes, et si α désigne l'abscisse de l'arrêt d'autobus, e_α est la distance moyenne que parcourent les habitants du village pour prendre le bus.
Le problème du choix de α assurant la minimisation de cette distance est donc en pratique important.

Proposition 7 *L'écart absolu moyen par rapport à* α *est minimum lorsque* α *est la médiane de la série.*

Remarque

Le lecteur vérifiera sans difficulté que l'écart absolu moyen, comme la médiane, est indépendant du choix de l'origine et du choix de l'unité.

2.3.4 Variance et écart type

Définition 21
Etant donné une série statistique quantitative $X = (x_i, n_i)_{i\in 1..p}$, on appelle **écart quadratique moyen** *par rapport à une valeur centrale α le nombre*

$$\sigma_\alpha = \sqrt{\frac{1}{N}\sum_{i=1}^{p} n_i(x_i-\alpha)^2}$$

où N désigne l'effectif total.

Remarque

L'écart quadratique moyen est donc la moyenne quadratique des écarts.

Proposition 8 Théorème de KÖNIG
Etant donné une série $X = (x_i, n_i)_{i\in..p}$,

$$\frac{1}{N}\sum_{i=1}^{p} n_i(x_i-\alpha)^2 = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i-\overline{X})^2 + (\overline{X}-\alpha)^2$$

N désignant l'effectif total.

démonstration

Soit

$$S = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i-\alpha)^2$$

Nous obtenons

$$\begin{aligned}
S &= \frac{1}{N}\sum_{i=1}^{p} n_i(x_i-\overline{X}+\overline{X}-\alpha)^2 \\
&= \frac{1}{N}\sum_{i=1}^{p} n_i(x_i-\overline{X})^2 + \frac{1}{N}\sum_{i=1}^{p} n_i(\overline{X}-\alpha)^2 + 2\frac{1}{N}\sum_{i=1}^{p} n_i(x_i-\alpha)(\overline{X}-\alpha) \\
&= \frac{1}{N}\sum_{i=1}^{p} n_i(x_i-\overline{X})^2 + (\overline{X}-\alpha)^2\frac{1}{N}\sum_{i=1}^{p} n_i + (\overline{X}-\alpha)\frac{N}{\sum}{}_{i=1}^{p}\, n_i(\overline{X}-x_i) \\
&= \frac{1}{N}\sum_{i=1}^{p} n_i(x_i-\overline{X})^2 + (\overline{X}-\alpha)^2
\end{aligned}$$

Il s'ensuit immédiatement

Proposition 9 *L'écart quadratique moyen par rapport à α est minimum lorsque α est la moyenne arithmétique de la série.*

démonstration

Cela découle directement du théorème précédent.

Définition 22 *On appelle* **écart type** *de la série $X = (x_i, n_i)_{i\in..p}$ l'écart quadratique moyen par rapport à la moyenne, i.e. le nombre noté σ_X défini par*

$$\sigma_X = \sqrt{\frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \overline{X})^2}$$

Définition 23 *Le carré de l'écart type est appelé* **variance** *de X et noté $Var(X)$.*

Proposition 10

$$\begin{aligned} Var(X) &= \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - \overline{X}^2 \\ Var(X+a) &= Var(X) \quad \forall a \in R \\ Var(\lambda X) &= \lambda^2 Var(X) \quad \forall \lambda \in R \end{aligned}$$

démonstration

La première propriété découle du théorème de König en choisissant $\alpha = 0$

Les deux propriétés suivantes découlent directement de la définition de la variance et des propriétés de la moyenne $E(X + a) = E(X) + a$ et $E(\lambda X) = \lambda E(X)$.

Proposition 11 *Soit $(X_k = (x_{i,k}, n_{i,k})_{i\in1..p_k})_{k\in1..p}$ p échantillons d'une même population et X l'échantillon constitué par la réunion de ces p échantillons.*
Soit $N_k = \sum_{i=1}^{p_k} n_{i,k}$ l'effectif de l'échantillon
X_k, $M = (M_k, N_k)_{k\in1..p}$ la distribution des moyennes des X_k ($M_k = E(X_k)$) et
$V = (V_k, N_k)_{k\in1..p}$ la distribution des variances des X_k ($V_k = Var(X_k)$). Alors

$$E(M) = E(X)$$

et

$$Var(X) = E(V) + Var(M)$$

$Var(M)$ est appelé **variance interpopulation**
$E(V)$ est appelé **variance intrapopulation**

démonstration

Soit N l'effectif total de X. Alors $N = \sum_{k=1}^{p} N_k$.
Par ailleurs

$$M_k = E(X_k) = \frac{1}{N_k} \sum_{i=1}^{p_k} n_{i,k} x_{i,k}$$

Donc

$$\begin{aligned} E(M) &= \frac{1}{N} \sum_{k=1}^{p} N_k M_k \\ &= \frac{1}{N} \sum_{k=1}^{p} N_k \frac{1}{N_k} \sum_{i=1}^{p_k} n_{i,k} x_{i,k} \\ &= \frac{1}{N} \sum_{k=1}^{p} \sum_{i=1}^{p_k} n_{i,k} x_{i,k} \\ &= E(X) \end{aligned}$$

et

$$Var(M) = \frac{1}{N} \sum_{k=1}^{p} N_k M_k^2 - E(M)^2 = \frac{1}{N} \sum_{k=1}^{p} N_k M_k^2 - E(X)^2$$

Par ailleurs

$$V_k = Var(X_k) = \frac{1}{N_k} \sum_{i=1}^{p_k} n_{i,k} x_{i,k}^2 - M_k^2$$

et donc

$$\begin{aligned} E(V) &= \frac{1}{N} \sum_{k=1}^{p} V_k \\ &= \frac{1}{N} \sum_{k=1}^{p} \left(\sum_{i=1}^{p_k} n_{i,k} x_{i,k}^2 - N_k M_k^2 \right) \\ &= \frac{1}{N} \sum_{k=1}^{p} \sum_{i=1}^{p_k} n_{i,k} x_{i,k}^2 - \frac{1}{N} \sum_{k=1}^{p} N_k M_k^2 \end{aligned}$$

En définitive

$$\begin{aligned} Var(M) + E(V) &= \frac{1}{N} \sum_{k=1}^{p} N_k M_k^2 - E(X)^2 + \frac{1}{N} \sum_{k=1}^{p} \sum_{i=1}^{p_k} n_{i,k} x_{i,k}^2 - \frac{1}{N} \sum_{k=1}^{p} N_k M_k^2 \\ &= \frac{1}{N} \sum_{k=1}^{p} \sum_{i=1}^{p_k} n_{i,k} x_{i,k}^2 - E(X)^2 \\ &= Var(X) \end{aligned}$$

Remarque

Si un même type de pièces est fabriqué par p machines,
• la variance interpopulation du lot de pièces fabriquées est la moyenne des dispersions (variances) des machines, et dépend donc de la **justesse** ou de la **précision** des machines (erreur aléatoire) ;
• la variance intrapopulation du lot de pièces fabriquées est la dispersion (variance) des moyennes des machines, et dépend donc de la **fidélité** des machines (adéquation du réglage de la machine).

2.3.5 Coefficient de variation

Définition 24 *On appelle* **coefficient de variation** *d'une série statistique X le rapport de son écart-type sur sa moyenne :*

$$C_X = \frac{\sigma_X}{E(X)}$$

Ce coefficient est sans unité, souvent exprimé en pourcentage (même s'il n'a pas de raison particulière d'être inférieur à 1). Il permet la comparaison entre des séries différentes et correspond à la notion d'erreur relative, σ_X correspondant à l'erreur absolue.

2.3.6 Distribution réduite centrée - Moments

Définition 25 *Soit X une distribution statistique de moyenne $E(X)$ et d'écart-type σ. On appelle* **distribution centrée** *associée la distribution $X - E(X)$ et* **distribution réduite centrée** *associée la distribution*

$$\frac{X - E(X)}{\sigma}$$

Les distributions réduites centrées constituent une sorte de normalisation des distributions : l'origine des modalités est ramenée à la moyenne et l'unité de mesure est prise égale à l'écart-type. Cette forme facilite donc l'étude des caractéristiques d'une distribution autres que celles de tendance centrale et de dispersion, et permet la comparaison entre distributions.

Si l'on dit que deux distributions sont équivalentes lorsqu'elles ont même distribution réduite centrée associée, on définit une relation d'équivalence sur l'ensemble des distributions statistiques et une grande part de la statistique consiste en l'étude des classes d'équivalence ainsi obtenues.

Définition 26 *Soit $X = (x_i, n_i)$ une distribution statistique. Pour un entier naturel r on pose $X^r = (x_i^r, n_i)$ et on appelle*

- **moment (simple) d'ordre r** *le nombre $m_r = E(X^r)$*
- **moment centré d'ordre r** *le nombre $\mu_r = E((X - E(X))^r)$, i.e. le moment simple d'ordre r de la variable centrée associée*
- **moment centré réduit d'ordre r** *le nombre*

$$\nu_r = \frac{E((X - E(X))^r}{\sigma_X}$$

i.e. le moment simple d'ordre r de la variable réduite centrée associée

Nous en déduisons immédiatement les résultats suivants

Proposition 12

$$m_0 = E(X),\ \mu_0 = \nu_0 = 0$$
$$m_1 = Var(X) + m_0^2,\ \mu_1 = Var(X),\ \nu_1 = 1$$
$$\nu_r = \frac{\mu_r}{\sigma^r}$$

2.3.7 Autres coefficients de dispersion

L'on peut bien évidemment introduire bien d'autres indices de dispersion en considérant d'autres indicateurs de tendance centrale (moyennes diverses, médiane) et d'autres moyennes (géométriques, harmoniques, etc...) des écarts à une position centrale.

Signalons en particulier l'**écart géométrique** par rapport à α, que nous noterons $s(\alpha)$, défini pour une série $X = (x_i, n_i)$ par

$$\log s(\alpha) = \sqrt{\frac{1}{N}\sum_{i=1}^{n} n_i(\log x_i - \log \alpha)^2}$$

N désignant comme toujours l'effectif total ($N = \sum n_i$).

$\log s(\alpha)$ est donc l'écart quadratique moyen de la série $\log X = (logx_i, n_i)$ par rapport à $\log \alpha$; son minimum est donc atteint lorsque $\log \alpha$ est la moyenne arithmétique de la série $\log X$.

L'**écart médian** par rapport à α est la médiane de la série des écarts absolus par rapport à α. Lorsque α est la médiane de la série, l'écart médian est appelé **écart probable**. D'autres auteurs, et notamment certains tableurs, appellent écart médian la moyenne de la série des écarts absolus à la moyenne. Ces différences de terminologie sont encore trop fréquentes en statistique, qui est une science encore jeune. Nous ne pouvons donc qu'inciter le lecteur à toujours vérifier la définition exacte d'une notion utilisée par un auteur ou par un logiciel.

2.4 Caractéristiques de forme

2.4.1 Coefficients d'asymétrie

Définition 27 *On dit qu'une distribution est* **symétrique** *- autour de sa moyenne - si chaque fois que la distribution centrée associée prend la valeur x avec un effectif n elle prend également la valeur $-x$ avec le même effectif n.*
Elle est dite **oblique** *dans le cas contraire.*

Il s'ensuit immédiatement

Proposition 13 *Si une distribution est symétrique, tous ses moments centrés d'ordre impair sont nuls*

Il est donc naturel de "mesurer" - au sens d'évaluer numériquement - l'asymétrie d'une distribution statistique par un moment d'ordre impair.

Définition 28 *On appelle* **coefficient d'asymétrie de Fisher** *le troisième moment réduit centré ν_2 (noté γ_1 par Fisher).*
On appelle **deuxième coefficient d'asymétrie de Pearson** *le carré de ν_2 (noté β_1 par Pearson).*

Proposition 14 *Si γ_1 est nul, la distribution est symétrique. Elle est oblique à gauche (ou étalée à droite) si $\gamma_1 > 0$ et oblique à droite (ou étalée à gauche) si $\gamma_1 < 0$*

La figure 2.7 montre la forme générale d'une distribution oblique à gauche et d'une distribution oblique à droite.

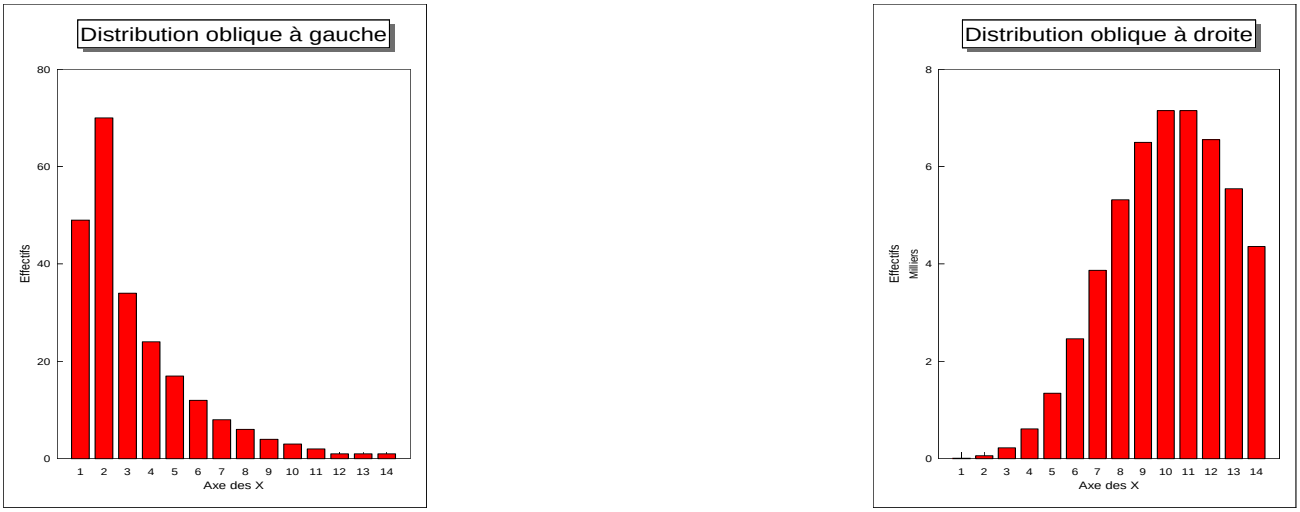


FIG. 2.7 – Distributions obliques

Le coefficient d'asymétrie de Fisher est de loin le plus utilisé, mais d'autres coefficients peuvent être utilisés :

Définition 29 *On appelle* **premier coefficient d'asymétrie de Pearson** *le nombre*

$$s = \frac{\overline{X} - m}{\sigma}$$

où m désigne la médiane de la distribution.

Proposition 15 *Le premier coefficient de Pearson n'est significatif que pour les distributions unimodales faiblement asymétriques. Si s est nul la distribution est symétrique, elle est oblique à droite si $s < 0$ et oblique à gauche si $s > 0$*

Définition 30 *On appelle* **coefficient de Yule** *le rapport*

$$\frac{(q_3 - q_2) - (q_2 - q_1)}{q_3 - q_1}$$

q_1 et q_3 désignant respectivement le premier et le troisième quartile, q_2 le deuxième quartile, i.e. la médiane.

Proposition 16 *Si s désigne le coefficient de Yule, alors si s est nul la distribution est symétrique, elle est oblique à droite si $s < 0$ et oblique à gauche si $s > 0$*

2.4.2 Coefficients d'aplatissement

On mesure l'aplatissement d'une distribution par référence à la loi mormale (loi de Gauss-Laplace) que l'on étudiera plus loin.

Définition 31 *Une distribution est dite* **platicurtique** *si elle est plus aplatie que la distribution normale et* **leptocurtique** *si elle est moins aplatie que la distribution normale. La distribution normale est dite* **mesocurtique**

Remarque

La terminologie adoptée vient du grec "kurtosis" qui signifie *bosse*, "platus" qui signifie *large*, "leptos" qui signifie *mince* et "mesos' qui signifie *qui est au milieu*.

Définition 32 *On appelle* **coefficient d'aplatissement de Pearson** *ou* **kurtosis** *le quatrième moment réduit centré $\nu_4 = \mu_4/\sigma^4$ noté β_2.*
On appelle **coefficient d'aplatissement de Fisher** $\gamma_2 = \beta_2 - 3$.

On montrera que le coefficient d'aplatisseme,t de Pearson de la loi normale est égal à 3. De là résulte la définition de Fisher qui ramène la comparaison à 3 à l'étude du signe.

Proposition 17 *La distribution est leptocurtique si $\beta_2 > 3$ ou $\gamma_2 > 0$ et platocurtique si $\beta_2 < 3$ ou $\gamma_2 < 0$*

2.4.3 Exemples

Exercice 11

Etude de la distribution G définie par

G	1	2	3	4	5	6	7	8	9	10	11	12	13	14
eff	49	70	34	24	17	12	8	6	4	3	2	1	1	1

Le diagramme en tuyaux d'orgue de cette distribution est donné à la figure 2.7 (diagramme de gauche).

Le calcul des différents indices de cette distribution nous donne

INDICES DE POSITION

mode	:	2
médiane	:	2
quartiles	:	2 2 4
moyenne arithmétique	:	3.34
moyenne harmonique	:	2.11
moyenne géométrique	:	2.63
moyenne quadratique	:	4.17

INDICES DE DISPERSION

étendue	:	13
écart interquartile	:	2
écart absolu moyen	:	1.90
écart type	:	2.49
coefficient de variation	:	0.77

COEFFICIENTS D'ASYMETRIE

coefficient de Fisher	:	1.62
coefficient de Yule	:	1.00
1*er coefficient de Pearson*	:	.539

COEFFICIENT D'APLATISSEMENT

kurtosis	:	5.70

Nous considérons cette série comme une série discrète, et n'effectuons donc pas d'interpolation linéaire pour le calcul des quartiles.
Les autres résultats sont bien entendu approchés

Nous noterons les différences notables entre les différentes moyennes, ce qui confirme l'intérêt qu'il y a à choisir la "bonne" moyenne. Dans le cas présent, la nature de la distribution et les conditions d'expérience n'étant pas précisées nous ne pouvons bien entendu faire de choix entre ces moyennes.

Les indices de dispersion sont eux relativement proches et indiquent une dispersion assez importante (coefficient de variation de 77%). Notons à ce sujet que l'écart-type et l'écart moyen absolu ne sont significatifs que si la "bonne" moyenne est arithmétique.

Les coefficients d'asymétrie de Fisher et de Yule sont positifs, ce qui confirme l'obliquité à gauche, i.e. l'étalement à droite, "évidente" sur la représentation graphique.

Le coefficient d'aplatissement (Kurtosis), égal à 5.7, indique ($5.7 > 3$) que la distribution est leptocurtique (leptokurtic en anglais)

Exercice 12

Etude de la distribution P définie par

P	1	2	3	4	5	6	7	8	9	10	11	12	13	14
eff	11	61	222	610	1342	2461	3867	5316	6498	7148	7148	6552	5544	4356

Le diagramme en tuyaux d'orgue de cette distribution est donné à la figure 2.7 (diagramme de droite).

Le calcul des différents indices de cette distribution nous donne

INDICES DE POSITION

mode	:	10 11
médiane	:	10
quartiles	:	8 10 12
moyenne arithmétique	:	10.1
moyenne harmonique	:	9.24
moyenne géométrique	:	9.69
moyenne quadratique	:	10.4

INDICES DE DISPERSION

étendue	:	13
écart interquartile	:	4
écart absolu moyen	:	2.07
écart type	:	2.53
coefficient de variation	:	0.25

COEFFICIENTS D'ASYMETRIE

coefficient de Fisher	:	-0.36
coefficient de Yule	:	0
1*er coefficient de Pearson*	:	.025

COEFFICIENT D'APLATISSEMENT

kurtosis	:	2.53

L'aspect le plus intéressant de l'étude de cette série est l'observation des coefficients d'asymétrie, pour lesquels règne la confusion la plus totale. Yule indique une distribution symétrique, Pearson une distribution oblique à gauche et Fisher une distribution oblique à droite. La seule chose qui soit claire est que, si asymétrie il y a, celle-ci est très faible.

Une étude plus détaillée montre que la distribution n'est pas symétrique malgré le coefficient de Yule égal à zéro. Ceci n'est pas étonnant car le coefficient de Yule est nul dès que les premier et troisième

quartiles sont symétriques par rapport à la médiane, ce qui ne préjude bien entendu pas du comportement de la distribution complète.

L'aspect visuel du diagramme de la figure 2.7 semble indiquer une obliquité à droite, donc donner raison à Fisher (dont le coefficient est effectivement le plus utilisé en pratique). La vrai conclusion sera cependant qu'il y a une très faible obliquité à droite.

En ce qui concerne le coefficient d'aplatissement ($2.53 < 3$), il indique que la courbe est platicurtique.

Les diagrammes de la figure 2.7 ne mettent pas ce phénomène en évidence, essentiellement du fait que les unités sur l'axe (Oy) sont différentes. Nous allons donc considérer un polygone des fréquences, en positionnant également la loi normale de même moyenne et de même écart-type. (figure 2.8)

FIG. 2.8 – Distributions G et P comparées à la loi normale

Notons que la comparaison est ici significative car les valeurs de la modalité sont espacées de 1 en 1 ; dans le cas contraire, nous aurions utilisé un histogramme (figure 2.9)

FIG. 2.9 – Histogrammes des distributions G et P comparées à la loi normale

2.5 Exercices

Exercice 13

Les données suivantes représentent l'intensité solaire (en watts/m^2) dans différentes villes du sud de l'Espagne :
(562, 869, 708, 775, 704, 809, 856, 655, 806, 878, 909, 918, 558, 768, 870, 918, 940, 946, 661, 820, 898, 935, 952, 957, 693, 835, 905, 939, 955, 960, 498, 653, 730, 753.

1. Construire un histogramme de ces données
2. Calculer la moyenne, la médiane, le mode ainsi que la variance et l'écart type de ces données

Exercice 14

Dans une entreprise, on a contrôlé le poids d'un produit fabriqué, et on a obtenu les résultats suivants (poids donnés en grammes)

47.7	43.9	45.1	43.9	44.5	44.6	45.2	45.0	45.1
45.0	46.1	45.5	45.0	45.1	44.3	45.4	45.0	43.9
45.1	46.3	45.8	45.1	44.9	45.2	44.3	45.6	43.9
45.5	44.2	44.9	43.5	46.0	44.6	44.9	45.0	45.9
45.5	45.3	45.4	44.4	45.0	44.4	44.6	44.4	45.6
44.9	44.5	44.7	45.3	44.4	45.1	44.2	44.8	43.9
44.5	44.6	45.0	45.5	44.9	45.5	45.1	44.7	44.8
45.0	45.3	44.4	45.6	45.1	44.7	44.5	45.7	44.6
45.2	44.2	44.5	45.2	44.8	45.6	44.2	45.1	44.9
43.3	45.5	45.3	44.4	45.5	43.3	44.6	44.6	45.2
45.5	44.7	45.3	44.0	43.2	42.8	45.1	44.5	45.9
44.2	45.1	44.4	44.8	45.0	46.3	46.3	45.0	43.3
45.0	45.4	43.5	44.5	45.1	44.6	45.2	45.5	44.7
43.5	46.1	45.5	45.1	43.6	44.5	45.6	46.6	45.4
44.9	45.3	45.7	44.9	44.1	44.6	46.5	44.3	46.1
43.0	43.6	46.0	45.0	44.1	45.3	43.3	43.0	46.2
44.3	44.7	44.5	44.0	44.2	43.7	45.4	45.2	44.9
45.1	44.5	45.7	43.3	42.9	44.4	43.6	44.8	43.9
44.6	45.0	45.4	43.4	42.9	43.5	42.1	41.8	42.8
43.2	43.8	43.7	44.2	44.4	44.3	45.1	44.9	43.9
45.2	44.9	45.3	44.8	45.7	45.8	45.0	43.3	45.8
46.0	45.6	44.9	4.50	44.7	44.8	44.4	43.8	43.3

Etudier ces résultats.

Exercice 15

Nous avons jeté 20 fois une paire de dés, et noté la somme des points obtenus. Nous avons ensuite répété trois fois cette opération (lancer 20 fois deux dés) et obtenu les échantillons $X_1,\ X_2,\ X_3$ *et* X_4 suivants :

X_1	2	3	4	5	6	7	8	9	10	11	12
$effectifs$	0	0	1	0	3	3	5	2	4	2	0

X_2	2	3	4	5	6	7	8	9	10	11	12
$effectifs$	1	1	1	5	0	3	3	3	0	0	3

X_3	2	3	4	5	6	7	8	9	10	11	12
$effectifs$	0	0	1	3	3	5	3	3	0	1	1

X_4	2	3	4	5	6	7	8	9	10	11	12
$effectifs$	1	1	2	2	1	2	4	1	3	3	0

Tracer des diagrammes de ces séries statistiques, déterminer leurs éléments caractéristiques (moyenne, médiane, mode, variance, écart-type, etc..)

Considérer la distribution X constituée par la réunion des observations de ces quatre séries ; construire le tableau des effectifs et des fréquences de X, une représentation graphique et déterminer ses indices de position, de dispersion, d'asymétrie et d'aplatissement.

Déterminer les variances inter et intrapopulation et vérifier que la variance de X en est la somme.

Exercice 16

Etudiants des hautes écoles suisses selon le lieu de domicile avant le début de leurs études, en 1982-1983 :

Valeurs absolues :

Université	Lausanne	St Gall	Zürich	EPFL	EPFZ
Canton de la Haute Ecole	3104	354	6985	725	2150
Autres cantons	1391	1254	7155	872	5086
Etranger	1324	480	1661	881	747
Total	5819	2088	15801	2478	7983

Compléter le tableau des valeurs relatives :

Université	Lausanne	St Gall	Zürich	EPFL	EPFZ
Canton de la Haute Ecole					
Autres cantons					
Etranger					
Total					

En utilisant EXCEL :

1. En utilisant les valeurs absolues, faire un graphique en barre simple (sans tronçons) montrant la répartition totale des étudiants entre les différentes hautes écoles.
2. Compléter le tableau en valeurs relatives et construire le graphique en barre à tronçon avec ces valeurs.
3. Construire les graphiques circulaires présentant la distribution des étudiants des universités de Lausanne et de l'EPFL selon les trois critères suivants :
   (a) Etudiants provenant du canton où la haute école est située
   (b) Etudiants provenant d'autres cantons
   (c) Etudiants ayant leur domicile à l'étranger.
4. Refaire la même chose en utilisant Scilab.

Exercice 17

Une entreprise a effectué une étude statistique sur la durée de vie des compresseurs qu'elle fabrique. Le tableau ci-dessous rend compte de cette étude.

Durée de vie en heures	Nombre de compresseurs
[50 000 ; 60 000 [	1 800
[60 000 ; 70 000 [	3 200
[70 000 ; 80 000 [	5 300
[80 000 ; 90 000 [	2 700
[90 000 ; 100 000 [	2 000

- . Calculer la durée de vie moyenne des compresseurs et l'écart-type de cette série statistique (les résultats seront arrondis à l'unité).
- Quel est le nombre de compresseurs qui appartient à l'intervalle $[\overline{x} - \sigma\,;\,\overline{x} + \sigma]$. L'exprimer en pourcentage du nombre total de compresseurs.

Exercice 18

Les diamètres exprimés en mm de 40 pièces usinées sont réparties de la manière suivante :

Diamètres en mm	Effectif
[194,5 ; 195,5 [	2
[195,5 ; 196,5 [	7
[196,5 ; 197,5 [	6
[197,5 ; 198,5 [	8
[198,5 ; 199,5 [	7
[199,5 ; 200,5 [	4
[200,5 ; 201,5 [	6

- Pour la série statistique associée au tableau précédent :
  1. Calculer la moyenne et l'écart-type (donner les valeurs arrondies au dixième de mm).
  2. Construire le polygone des fréquences cumulées croissantes sur le repère ci-après.
- Par une lecture graphique, déterminer le nombre de pièces dont le diamètre appartient à l'intervalle [196,3 ; 200]. Exprimez le en pourcentage de l'effectif total.

Exercice 19

Vérification de réglage d'une remplisseuse

Sur une ligne de conditionnement en flacons, on veut vérifier qu' 'une remplisseuse est bien réglée. Le contrôle du remplissage d'un échantillon de 100 flacons a donné la série statistique S suivante :

Mesure, en mL de la quantité de liquide contenue dans le flacon	ni : nombre de flacons
[995 ; 1 000 [	1
[1 000 ; 1 005 [	58
[1 005 ; 1 010 [	25
[1 010 ; 1 015 [	10
[1 015 ; 1 020 [	5
[1 020 ; 1 025 [	1

- On tolère que la quantité de liquide contenue dans un flacon ait une valeur comprise entre 1 000 mL et 1 020 mL (1 000 accepté ; 1 020 refusé). Quel est le pourcentage, par rapport au nombre total de flacons examinés, qui contiennent une quantité de liquide conforme à cette tolérance ?

- On estime que chaque flacon de l'échantillon contient une quantité de liquide égale au centre de la classe dans laquelle il est répertorié. Pour la série statistique ainsi obtenue, calculer sa moyenne et son écart-type . (A cet effet, on pourra utiliser le tableau suivant. Les résultats demandés seront exprimés en millilitres et arrondis au dixième de millilitre.)
  Mesure, en mL de la quantité de liquide contenue dans le flacon : (x_i : centre des classes, n_i : nombre de flacons)

Classes	x_i	n_i	$n_i x_i$	$n_i x_i^2$
[995 ; 1 000 [		1		
[1 000 ; 1 005 [		58		
[1 005 ; 1 010 [		25		
[1 010 ; 1 015 [		10		
[1 015 ; 1 020 [		5		
[1 020 ; 1 025 [		1		

- On appelle **Coefficient d'Aptitude Machine** (C.A.M) le rapport $\frac{IT}{6\sigma}$ où IT est l'amplitude de l'intervalle de tolérance et où σ est l'écart-type de l'échantillon (calculé à la question précédente).
  - Sachant que IT=(1020-1000) mL, calculer la valeur du rapport (prendre pour la valeur de 4,7 mL).
  - Si le Coefficient d'Aptitude Machine est strictement supérieur à 1 (CAM>1), la machine est considérée comme performante et les flacons sont correctement remplis. Est-ce le cas dans l'exemple retenu

Exercice 20

On a relevé le nombre de mots frappés en un quart d'heure par des secrétaires en formation :

520 514 471 419 460 380 430 507 450 397
496 433 415 439 450 427 454 432 462 516
470 424 398 437 485 450 431 435 429 459
467 387 450 458 460 423 499 409 456 409

Traduire cette série statistique par un tableau après avoir rangé les valeurs en classes d'amplitude 25, puis calculer la moyenne et l'écart type de cette série.

Exercice 21

On fait passer à 100 sujets une épreuve notée sur une échelle continue de 0 à 20. On obtient la distribution suivante :

0 0,5 1 1 1,5 2 2 2 3 3 4 4 4 4 4 4 4 4

4 4 4 4 4 4 4,5 4,5 4,5 4,5 4,5 4,5 4,5 5 5 5 5 5

5 5 5 5 5 5 5 5 5 5,5 5,5 6 6 6 6 6 6 6

6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6,5 6,5 6,5

6,5 6,5 6,5 6,5 6,5 6,5 6,5 6,5 6,5 6,5 6,5 6,5 6,5 6,5 6,5 7 7 7 7 7

7 7 7 7 7 8 8 8 8 8 8,5 8,5 8,5 9 10

- Donner le(s) mode(s), les valeurs extrêmes et l'étendue de cette série statistique.
- Calculer la médiane ainsi que les quartiles

Chapitre 3

Distributions à deux caractères

Toute "loi" physique (ou économique, ou...) est la plupart du temps déterminée de manière empirique, au moins dans un premier temps (avant l'élaboration d'une "théorie"). Nous nous proposons ici de déterminer, à partir de données expérimentales, s'il existe une relation liant entre elles les différentes variables mesurées, et, le cas échéant, de déterminer cette relation.

3.1 Les tableaux de contingence

Nous nous limiterons au cas de deux variables que nous noterons X et Y, X étant appelée **variable explicative** et Y étant appelée **variable expliquée**, ces rôles pouvant d'ailleurs être échangés comme nous le verrons par la suite. Nous considérerons une série de n mesures $(x_1, y_1) \dots (x_n, y_n)$. Notons que chaque mesure doit effectivement correspondre à un couple de valeurs si nous voulons qu'une liaison éventuelle entre les variables ait véritablement une signification ; ainsi, si nous voulons étudier une relation (éventuelle) entre le poids et la taille des hommes, il est bien clair qu'une mesure (x, y) doit correspondre au poids et à la taille d'un même individu !

Les modalités de X aussi bien que de Y peuvent être regroupées en classes, comme dans le cas des distributions à un caractère.

Nous regroupons toutes les données dans un tableau à double entrée appelé **tableau de contingence** et comportant en ligne et en colonne les modalités (x_i) et (y_j) des deux variables X et Y, l'effectif $n_{i,j}$ du couple (x_i, y_j) étant indiqué à l'intersection de la ligne i et de la colonne j.
De plus, une colonne et une ligne supplémentaires indiquent les distributions statistisques de X et de Y (les **distributions marginales**).

L'effectif associé à la modalité x_i (resp. y_j) de la variable X est noté $n_{i,\bullet}$ (resp. $n_{\bullet,j}$) et donc

$$n_{i,\bullet} = \sum_j n_{i,j} \quad et \quad n_{\bullet,j} = \sum_i n_{i,j}$$

ce qui signifie que les distributions marginales sont obtenues en effectuant les sommes des effectifs par ligne et par colonne respectivement.

L'**effectif total** d'un échantillon double (X,Y) est noté $n_{\bullet,\bullet}$ et donc

$$n_{\bullet,\bullet} = \sum_{i,j} n_{i,j}$$

Nous pouvons également considérer les **fréquences conjointes** $f_{i,j}$ définies par

$$f_{i,j} = \frac{n_{i,j}}{n_{\bullet,\bullet}}$$

et les **fréquences marginales** définies naturellement par

$$f_{\bullet,j} = \frac{n_{\bullet,j}}{n_{\bullet,\bullet}}$$

et

$$f_{i,\bullet} = \frac{n_{i,\bullet}}{n_{\bullet,\bullet}}$$

Nous avons alors

Proposition 18

$$f_{\bullet,j} = \frac{n_{\bullet,j}}{n_{\bullet,\bullet}} = \sum_i \frac{n_{i,j}}{n_{\bullet,\bullet}} = \sum_i f_{i,j}$$

et

$$f_{i,\bullet} = \frac{n_{i,\bullet}}{n_{\bullet,\bullet}} = \sum_j \frac{n_{i,j}}{n_{\bullet,\bullet}} = \sum_j f_{i,j}$$

ce qui signifie que l'on obtient les fréquences marginales à partir du tableau de contingence des fréquences, en effectuant les sommes par lignes et par colonnes .

Lors de l'étude des distributions à deux caractères, nous nous intéresserons également aux **distributions conditionnelles**. Plus précisément

Définition 33 *Etant donné une série double (X, Y), on appelle*
distribution de X conditionnée par $Y = y_j$ *la distribution* $(x_i, n_{i,j})$*notée* $X_{/Y=y_j}$*, et*
distribution de Y conditionnée par $X = x_i$ *la distribution* $(y_j, n_{i,j})$ *notée* $Y_{/X=x_i}$
Plus généralement, si A est un sous-ensemble de l'ensemble des modalités de Y, nous appelons distribution de X conditionnée par $Y \in A$ la distribution $(x_i, n_{i,A})$ où

$$n_{i,A} = \sum_{y_j \in A} n_{i,j}$$

Définition 34 *La* **fréquence de** x_i **conditionnée par** $Y = y_j$ *est notée $f_{i/j}$ et définie par*

$$f_{i/j} = \frac{n_{i,j}}{n_{\bullet,j}}$$

De même la **fréquence de** y_j **conditionnée par** $X = x_i$ *est notée $f_{j/i}$ et définie par*

$$f_{j/i} = \frac{n_{i,j}}{n_{i,\bullet}}$$

Proposition 19 *Les fréquences conditionnées peuvent être obtenues à partir des fréquences conjointes et marginales par*

$$f_{i/j} = \frac{f_{i,j}}{f_{\bullet,j}} \text{ et } f_{j/i} = \frac{f_{i,j}}{f_{i,\bullet}}$$

démonstration

$$\begin{aligned} f_{i/j} &= \frac{n_{i,j}}{n_{i,\bullet}} \\ &= \frac{f_{i,j} * n_{\bullet,\bullet}}{n_{i,\bullet}} \\ &= \frac{f_{i,j}}{\frac{n_{i,\bullet}}{n_{\bullet,\bullet}}} \\ &= \frac{f_{i,j}}{f_{\bullet,j}} \end{aligned}$$

Le deuxième résultat se démontre de manière analogue.

□

Reprenons l'exemple (que nous allons développer tout au long de ce chapitre) du sondage auprès de nos étudiants leur demandant la note de mathématique au baccalauréat et le nombre de redoublements au cours de leur scolarité primaire et secondaire. Nous avions obtenu le tableau de contingence

Y/X	10	11	12	13	14	15	
0	0	2	1	1	2	1	7
1	0	1	2	3	0	0	6
2	1	1	0	0	0	0	2
3	1	0	0	0	0	0	1
	2	4	3	4	2	1	

où X désigne le caractère "note au bac" et Y le caractère "nombre de redoublements".

On s'attend à priori à ce que le nombre de redoublements (Y) "explique" la note de mathématique au baccalauréat (X), auquel cas Y sera la variable explicative et X la variable expliquée. Mais est-on vraiment sûr que ce n'est pas la note (et donc le niveau) de mathématique qui explique le nombre de redoublements ? Nous serions alors amenés à considérer X comme variable explicative et Y comme variable expliquée. La meilleure des solutions est pratiquement toujours de considérer les deux cas possibles !

La distribution de $Y_{/X=0}$ (distribution de X conditionnée par $Y = 0$, i.e. distribution de la note de mathématique des élèves n'ayant jamais redoublés) est donnée par la première ligne du tableau précédent :

$Y_{/X=0}$	10	11	12	13	14	15
eff	0	2	1	1	2	1

De même la distribution de $X_{/Y\neq 0}$ est

$Y_{/X\neq 0}$	10	11	12	13	14	15
eff	2	2	2	3	0	0

3.2 Caractéristiques numériques

Nous nous intéressons, comme dans le cas des distributions simples, aux indices de position et de dispersion ainsi qu'à divers moments faisant intervenir les deux lois conjointement.

3.2.1 Distributions marginales

Les différents indicateurs des lois marginales nous donnent bien évidemment un certain résumé (incomplet certes) de la distribution.

Définition 35 *Les moyennes (ou espérances) marginales sont notées $E(X)$ (resp $E(Y)$) ou $\overline{\overline{x}}$ (resp $\overline{\overline{y}}$) et définies par*

$$E(X) = \sum_i f_{i,\bullet} x_i = \frac{1}{n_{\bullet,\bullet}} \sum_i n_{i,\bullet} x_i$$

et

$$E(Y) = \sum_j f_{\bullet,j} y_j = \frac{1}{n_{\bullet,\bullet}} \sum_j n_{\bullet,j} y_j$$

De même

Définition 36 *Les variances marginales sont notées $Var(X)$ (resp $Var(Y)$) et définies par*

$$Var(X) = \sum_i f_{i,\bullet} (x_i - \overline{\overline{x}})^2 = \frac{1}{n_{\bullet,\bullet}} \sum_i n_{i,\bullet} (x_i - \overline{\overline{x}})^2$$

et

$$Var(Y) = \sum_j f_{\bullet,j} (y_j - \overline{\overline{y}})^2 = \frac{1}{n_{\bullet,\bullet}} \sum_j n_{\bullet,j} (y_j - \overline{\overline{y}})^2$$

Définition 37 *Les écarts types des lois marginales sont notés σ_X et σ_Y, et sont les racines carrées des variances marginales correspondantes.*

3.2.2 Distributions conditionnelles

Définition 38 *Etant donné une série double $(x_i, y_j, n_{i,j})$, on associe à chaque valeur y_j la série $(X_i, n_{i,j})$ notée X_j et appelée* **distribution de X conditionnée par** $Y = y_j$. *De même on associe à chaque valeur x_i la série $(Y_j, n_{i,j})$ notée Y_i et appelée* **distribution de Y conditionnée par** $X = x_j$.

Pour chacune de ces distributions conditionnelles on peut considérer leur moyenne et leur variance. Nous obtenons alors

Proposition 20 *La moyenne marginale $E(X)$ est égale à la moyenne des moyennes conditionnelles $E(X_j)$ pondérées par les effectifs marginaux et de même la moyenne marginale $E(Y)$ est égale à la moyenne des moyennes conditionnelles $E(Y_i)$ pondérées par les effectifs marginaux.*

$$\frac{1}{\sum_j n_{\bullet,j}} \sum_j n_{\bullet,j} \overline{x_j} = E(X)$$

et

$$\frac{1}{\sum_i n_{i,\bullet}} \sum_i n_{i,\bullet} \overline{y_i} = E(Y)$$

Proposition 21 *La variance marginale est la somme de la variance des moyennes conditionnelles et de la moyenne des variances conditionnelles, les moyennes et les variances conditionnelles étant pondérées par les effectifs marginaux (qui sont les tailles des distributions conditionnelles). Plus précisément*

$$Var(X) = Var(E(X_j)) + E(Var(X_j))$$

et

$$Var(Y) = Var(E(Y_i)) + E(Var(Y_i))$$

3.2.3 Moments et covariance

Définition 39 *Etant donné une série double $(x_i, y_j, n_{i,j})$, on appelle* **moment simple d'ordre** *(r, s) de la série le nombre*

$$m_{r,s} = \frac{1}{N} \sum_{i,j} n_{i,j} x_i^r y_j^s$$

On appelle **moment centré d'ordre** *(r, s) de la série le nombre*

$$\mu_{r,s} = \frac{1}{N} \sum_{i,j} n_{i,j} (x_i - \overline{\overline{x}})^r (y_j - \overline{\overline{y}})^s$$

On appelle **covariance** *de X et de Y et on note $Cov(X, Y)$ le moment centré d'odre $(1, 1)$, i.e.*

$$Cov(X, Y) = \frac{1}{N} \sum_{i,j} (x_i - E(X))(y_j - E(Y))$$

Proposition 22

$$Cov(X, X) = Var(X)$$

$$Cov(X, Y) = \frac{1}{N} \sum_{i,j} n_{i,j} x_i y_j - E(X)E(Y) = E(XY) - E(X)E(Y)$$

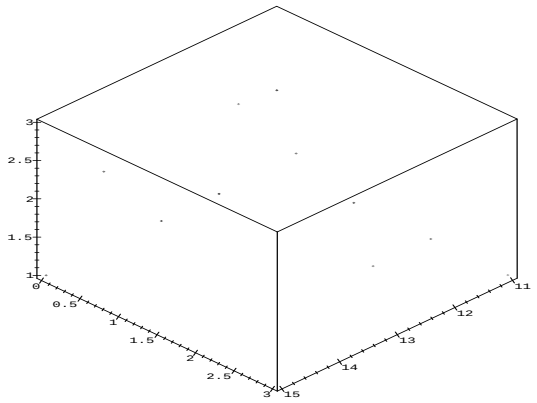


FIG. 3.1 – Nuage de points de 3 dimensions

3.3 Représentation graphique - Courbes de régression

La représentation graphique d'une série double est malaisée. Il nous faudrait en effet effectuer une représentation en 3 dimensions. La figure 3.1 nous montre les points de coordonnées $(x_i, y_j, n_{i,j})$ associés à notre exemple.

Une représentation en 2 dimensions est possible (points de coordonnées (x_i, y_j)), mais nous perdons alors l'indication de l'effectif associé à ces points. Nous parlons alors d'une représentation du **nuage** de points (ou d'un **diagramme de dispersion**).

Si nous voulons une représentation en 2 dimensions de nos données et en même temps tenir compte des effectifs, il nous faut accepter de réduire les données : pour une valeur x_i de X (resp. une valeur y_j de Y) on remplacera les différentes valeurs de y_j (resp. x_i) associées par leur moyenne. Plus précisément

Définition 40 *Etant donné une série double $(x_i, y_j, n_{i,j})$, on appelle* **courbe de régression de Y en X** *l'ensemble des points de coordonnées $(x_i, E(Y_i))$ (ou la ligne polygonale joignant ces points). on appelle de même* **courbe de régression de X en Y** *l'ensemble des points de coordonnées $(E(X_j), y_j)$ (ou la ligne polygonale joignant ces points).*

Proposition 23 *Les courbes de régression sont les courbes qui réalisent le minimum du carré des distances au nuage de points (effectifs pris en compte). Plus précisément*

$$\sum_{i,j} n_{i,j}(x_i - a_i)^2$$

est minimum pour $a_i = E(Y_i)$ et

$$\sum_{i,j} n_{i,j}(y_j - b_j)^2$$

est minimum pour $b_j = E(X_j)$ et

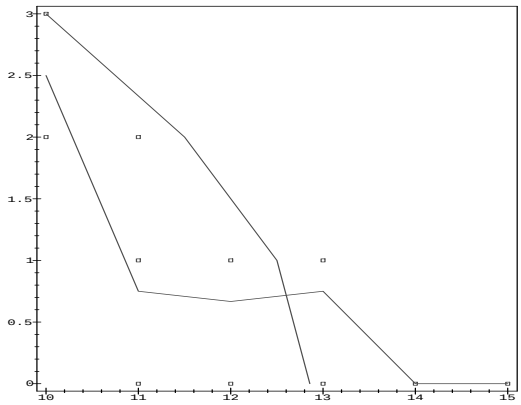


FIG. 3.2 – Nuage de points de 2 dimensions et droites de régression

La figure 3.2 montre le nuage de points (en 2D) et les droites de régression de notre distribution exemple.

3.4 Régression, Ajustement et Corrélation

Lorsque nous étudions une série double (X, Y), nous sommes inévitablement amenés à nous poser la question de la dépendance entre les caractères X et Y, et, le cas échéant, à établir la liaison entre ces variables (les mots "dépendance" et "liaison" étant prsi au sens intuitif du terme).

Comme toujours en statistique, nous nous proposons de développer un certain nombre d'outils pour "mesurer" ces notions, les conclusions découlant de l'usage de ces outils restant sous la responsabilité de l'utilisateur (qui doit donc avant tout connaître et maîtriser correctement ces outils s'il veut en tirer un réel profit).

3.4.1 Notion de dépendance et d'indépendance

Indépendance totale ou liaison nulle

Définition 41 *Deux variables X et Y sont dites* **totalement indépendantes** *si les fréquences conditionnelles $f_{i/j}$ sont indépendantes de j.*

Proposition 24 *Les variables X et Y sont totalement indépendantes si les fréquences conditionnelles sont égales aux fréquences marginales, i.e.*

$$f_{i/j} = f_{i,\bullet}$$

et

$$f_{j/i} = f_{\bullet,j}$$

Proposition 25 *Les variables X et Y sont totalement indépendantes si les fréquences conjointes sont égales au produit des fréquences marginales, i.e.*

$$f_{i,j} = f_{i,\bullet} \times f_{\bullet,j}$$

Proposition 26 *Si les variables X et Y sont totalement indépendantes, les moyennes conditionnelles sont égales aux moyennes marginales.*

Proposition 27 *Si les variables X et Y sont totalement indépendantes, les courbes de régression sont deux droites orthogonales (et parallèles aux axes).*

Dépendance totale ou liaison fonctionnelle

Définition 42 *Deux variables X et Y sont dites* **totalement dépendantes** *si à chaque valeur de X correspond une valeur de Y et une seule (Y totalement dépendant de X) ou si à chaque valeur de Y correspond une valeur de X et une seule (X totalement dépendant de Y), i.e. si*

$$\forall i \ \exists j \ / \ k \neq j \Longrightarrow n_{i,k} = 0$$

ou

$$\forall j \ \exists i \ / \ k \neq i \Longrightarrow n_{k,j} = 0$$

Si les deux conditions sont réalisées, X et Y sont dites **mutuellement dépendantes**.

Proposition 28 *Si les deux variables X et Y sont mutuellement dépendantes alors leurs courbes de régression sont confondues.*

Liaison relative

Si les variables X et Y ne sont ni totalement dépendantes ni totalement indépendantes, on dit qu'il existe une **liaison relative** entre X et Y. C'est évidemment le cas le plus fréquent !

Définition 43 *S'il existe une liaison relative entre les variables X et Y, les courbes de régression sont toutes deux croissantes ou toutes deux décroissantes : si elles sont croissantes on dit qu'il a* **corrélation positive** *entre X et Y, et* **corrélation négative** *dans le cas contraire.*

Remarque

Appelée **corrélation** dans le cas de caractères quantitatifs, la liaison éventuelle entre deux caractères est appelée **contingence** dans le cas de caractères qualitatifs.

3.4.2 Ajustement

Nous nous proposons ici, après avoir vérifié la non indépendance totale de deux variables X et Y, de déterminer s'il existe une liaison fonctionnelle entre ces variables, ou du moins si l'on peut raisonnablement approcher le nuage de points par une courbe d'équation $y = f(x)$.

Les différentes courbes usuelles d'ajustement

La relation la plus facile à observer à partir d'un nuage de points est une relation linéaire - en réalité affine ! - car seul l'alignement des points du nuage est caractéristique d'une "loi" simple.

Cependant lorsque l'alignement des points n'est pas constaté - même approximativement - nous disposons de **papier semi-log** et de **papier log-log** afin de ramener au cas de l'alignement des lois autres, fréquentes en pratique.

Utiliser une **échelle logarithmique** sur un axe de coordonnées consiste à placer $\ln x$ au lieu de x (s'il s'agit de l'axe des abscisses) ; ainsi un papier logarithmique est gradué de telle sorte que 2 (par exemple) soit marqué à la distance $\ln 2$ de l'origine (ce dans une unité quelconque) ; ainsi l'origine est-elle marquée 1. Un papier est dit semi-log lorsque l'un des axes uniquement comporte une échelle logarithmique, il est dit log-log lorsque les deux axes comportent une échelle logarithmique. Notons que tous les tableurs et les grapheurs dignes de ce nom offrent la possibilité d'utiliser une échelle (scale en anglais) logarithmique sur chaque axe.

Voyons maintenant quelles sont les lois "repérée" par l'alignement du nuage de points en échelle semi-log ou log-log.

- Si l'axe des abscisses est gradué en échelle logarithmique et l'axe des ordonnées normalement, nous obtenons une loi de la forme
$$y = a\,\ln(x) + b$$
(cas assez rare en pratique).
- Si l'axe des ordonnées est gradué en échelle logarithmique et l'axe des abscisses normalement, nous obtenons une loi de la forme
$$\ln y = a\,x + b$$
et donc
$$y = \mu\, e^{kx}$$
cas fréquent dans la pratique physique.
- Si les deux axes sont gradués en échelle logarithmique, nous obtenons une loi de la forme
$$\ln y = a\,\ln x + b$$
et donc
$$y = \mu\, x^k$$

Il existe bien entendu dans la pratique d'autres lois que les lois affines, exponentielles ou puissances. Nous donnons ici une liste non exhaustive de différentes autres lois et courbes que l'on peut rencontrer.

- courbe parabolique : $y = ax^2 + bx + c$
- cubique : $y = ax^3 + bx^2 + cx + d$
- hyperbole : $y = 1/(ax + b)$ ou $1/y = ax + b$; dans ce cas, en traçant le nuage des points de coordonnées $(x, 1/y)$, on se ramène au cas de l'ajustement linéaire.
- fonction de Gompertz : $y = pq^{b^x}$ ou $\ln y = ab^x + c$
- fonctions logistiques : $1/y = ab^x + c$ et $y = a(\ln x)^2 + b\ln x + c$

Méthode des moindres carrés

La méthode graphique d'ajustement est bien évidemment empirique et subjective. Même dans le cas d'un phénomène linéaire, une détermination graphique de la droite d'ajustement - comme droite approximant "au mieux" le nuage de points - mènera à des résultats différents selon les opérateurs. Il est donc indispensable de définir rigoureusement la notion de courbe d'ajustement, de telle sorte qu'il n'y ait qu'une seule réponse possible.

Prenons d'abord le cas d'un ajustement linéaire. Il parait naturel de considérer que la meilleure droite d'ajustement est celle qui minimise la somme des distances des points du nuage à la droite. Cette somme est, si la droite recherchée est d'équation $y = ax + b$,

$$f(a,b) = \sum_{i,j} \frac{n_{i,j}|y_j - ax_i - b|}{\sqrt{a^2+1}}$$

et la recherche "mathématique" du minimum pose de gros problèmes liés à la présence de valeurs absolues. Ce critère a donc été abandonné, et nous ne l'utiliserons pas, même si les méthodes actuelles de calcul assisté par ordinateur permettent sans difficulté majeure de résoudre ce problème à l'aide d'un algorithme. Par contre la distance d'un point à une courbe quelconque a une expression encore plus "compliquée".

La première simplification apportée à cette notion de courbe d'ajustement est de remplacer les distances par les carrés des distances, éliminant ainsi le problème de la valeur absolue. Nous laissons au lecteur le soin de déterminer la droite d'ajustement qui minimise la somme des carrés des distances des points du nuage à cette droite .

Une seconde simplification intervient ensuite pour tenir compte de la difficulté du problème dans le cas des courbes autres que les droites. Elle consiste à remplacer la distance à la courbe par la différence des ordonnées entre les points du nuage et les points de même abscisse de la courbe : cette grandeur (qui peut être positive ou négative) a une expression simple quelle que soit la courbe et est appelée **écart**, **erreur** ou **résidu**.

Notons que cette dernière simplification rompt la symétrie entre les grandeurs X et Y, et que par conséquent nous obtiendrons à priori une courbe différente selon que nous cherchons une loi donnant Y en fonction de X ou X en fonction de Y : nous reviendrons sur ce problème.

Définition 44 *On dit qu'une courbe d'équation $y = f(x)$ ajuste les données $(x_i, y_j, n_{i,j})$* **au sens des moindres carrés** *lorsqu'elle réalise - parmi les courbes de même nature - le minimum de*

$$\sum_{i,j} n_{i,j} \left(f(x_i) - y_j\right)^2$$

Droite des moindres carrés

Nous recherchons l'ajustement d'un nuage de points par une droite d'équation $y = ax + b$. Nous sommes donc amenés à déterminer les coefficients a et b qui minimisent la fonction

$$f(a,b) = \sum_{i,j} n_{i,j}(y_j - ax_i - b)^2$$

Rappelons le résultat suivant :

Proposition 29 *Une fonction f de plusieurs variables $(x_1 \dots x_n)$, si elle est différentiable, atteint ses extrema en des points annulant* $\overrightarrow{\text{grad}}\, f$

Nous ne vérifierons pas ici les conditions de validité (triviales ?). Déterminons donc les dérivées partielles de f.

$$\frac{\partial f}{\partial a} = 2\sum_{i,j} n_{i,j}(y_j - ax_i - b) \times (-x_i)$$

et

$$\frac{\partial f}{\partial b} = 2\sum_{i,j} n_{i,j}(y_j - ax_i - b) \times (-1)$$

Par conséquent $\overrightarrow{\text{grad}}\, f$ est nul si et seulement si $\frac{\partial f}{\partial a} = 0$ et $\frac{\partial f}{\partial b} = 0$, i.e.

$$\left\{ \begin{array}{lcl} \sum_{i,j} n_{i,j}(y_j - ax_i - b) \times (x_i) & = & 0 \\ \sum_{i,j} n_{i,j}(y_j - ax_i - b) & = & 0 \end{array} \right.$$

soit

$$\sum_{i,j} n_{i,j} y_j = a\sum_{i,j} n_{i,j} x_i + \sum_{i,j} n_{i,j} b \tag{3.1}$$

et

$$\sum_{i,j} n_{i,j} x_i y_j - a\sum_{i,j} n_{i,j} x_i^2 - b\sum_{i,j} n_{i,j} x_i = 0 \tag{3.2}$$

ou encore

$$\sum_{j} n_{\bullet,j} y_j = a\sum_{i} n_{i,\bullet} x_i + n_{\bullet,\bullet} b \tag{3.3}$$

et

$$\sum_{i,j} n_{i,j} x_i y_j - a\sum_{i} n_{i,\bullet} x_i^2 - b\sum_{i} n_{i,\bullet} x_i = 0 \tag{3.4}$$

Divisons ces deux équations par $n_{\bullet,\bullet}$.

L'équation 3.3 devient

$$E(Y) = aE(X) + b \tag{3.5}$$

et l'équation 3.4

$$E(XY) = aE(X^2) + bE(X) \tag{3.6}$$

d'où, en remplaçant, dans 3.6, b par sa valeur tirée de 3.5,

$$\begin{array}{rcl} E(XY) & = & aE(X^2) + (E(Y) - aE(X))\, E(X) \\ & = & e\left(E(X^2) - E(X)^2\right) + E(X)E(Y) \end{array}$$

et donc finalement

$$a = \frac{E(XY) - E(X)E(Y)}{E(X^2) - E(X)^2} = \frac{Cov(X,Y)}{Var(X)} \tag{3.7}$$

Proposition 30 *La droite d'ajustement de Y en X est la droite d'équation $y = ax + b$ avec*

$$a = \frac{Cov(X,Y)}{Var(X)}$$

et

$$E(Y) = aE(X) + b$$

cette dernière relation exprimant le fait que la droite d'ajustement passe par le point de coordonnées $(E(X), E(Y))$, i.e. par le barycentre des points du nuage.

3.4.3 Corrélation

Corrélation linéaire

Nous avons vu précédemment que la droite d'ajustement de Y en X est d'équation $y = ax + b$ avec $\overline{y} = a\overline{x} + b$ et

$$a = \frac{Cov(X,Y)}{\sigma_X^2}$$

Cette droite peut être utilisée pour estimer la valeur de Y pour une valeur donnée de X : on parle alors de **droite de régression** de Y en X. La pente a de cette droite est appelée **coefficient de régression de Y en X**.

Si maintenant nous voulons estimer X à partir d'une valeur de Y, nous sommes amenés à considérer la droite de régression de X en Y, qui - en inversant les rôles de X et de Y dans les résultats précédants - sera d'équation $x = \alpha y + \beta$ avec $\overline{x} = \alpha\overline{y} + \beta$ et

$$\alpha = \frac{Cov(Y,X)}{\sigma_Y^2}$$

Ce coefficient α de cette droite est appelée **coefficient de régression de X en Y**.

Nous constatons que ces deux droites de régression passent toutes deux par le barycentre du nuage de points, mais ont des coefficients directeurs (a et $1/\alpha$ respectivement) à priori distincts. Si la relation entre X et Y est parfaite, nous devons avoir

$$a\alpha = 1$$

Or

$$\begin{aligned} a\alpha &= \frac{Cov(X,y)Cov(Y,X)}{\sigma_X^2\sigma_Y^2} \\ &= \left(\frac{Cov(X,Y)}{\sigma_X\sigma_Y}\right)^2 \end{aligned}$$

car il est clair que $Cov(X,Y) = Cov(Y,X)$.

Définition 45 *On appelle* **coefficient de corrélation linéaire** *entre X et Y le nombre*

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

(certains auteurs préfèrent la valeur absolue de ce nombre).
Ce coefficient vérifie toujours la relation $|\rho| \leq 1$.
Si $|\rho|$ *est proche de 1 (*≥ 0.99 *en physique), on dit que X et Y sont fortement corrélés. Si* $|\rho|$ *est proche de 0, on dit que X et Y sont faiblement corrélés (voire non corrélés).*

On remarquera que ρ est une grandeur sans dimension, indépendante des unités choisies pour X et Y.

3.4.4 Corrélation dans le cas d'un ajustement non linéaire

L'introduction précédente du coefficient de corrélation n'est pas transposable dans le cas général où nous ajustons le nuage de points par une courbe d'équation $y = f(x)$. Nous allons donc donner une autre caractérisation du coefficient de corrélation linéaire, caractérisation qui pourra se transposer au cas général et que nous prendrons alros comme définition.

Définition 46 *Si nous ajustons Y en fonction de X par une équation* $y = f(x)$ *par la méthode des moindres carrés, on appelle* **valeur estimée** *de Y pour une valeur* x_i *de X le nombre*

$$y_{i,est} = f(x_i)$$

On appelle **variance totale** *de Y le nombre*

$$Var(Y) = \frac{1}{n_{\bullet,\bullet}} \sum_j n_{i,\bullet}(y_j - \overline{y})^2$$

(i.e. la variance marginale de Y)
et **variance expliquée** *de Y en X la quantité*

$$Var_{expl}(Y) = \frac{1}{n_{\bullet,\bullet}} \sum_{i=1}^{n} n_{i,\bullet}(y_{i,est} - \overline{y})^2$$

On appelle **variance résiduelle** *de Y en X la différence entre la variance totale et la variance expliquée.*

Nous avons alors

Proposition 31 *Dans les conditions de l'ajustement linéaire par la méthode des moindres carrés,*

$$\rho = \sqrt{\frac{Var_{expl}(Y)}{Var(Y)}}$$

démonstration

Nous avons

$$\begin{aligned}
\sum_{i,j} n_{i,j}(y_{i,est} - \overline{y})^2 &= \sum_{i,j} n_{i,j}(a\,x_i + b - \overline{y})^2 \\
&= \sum_{i,j} n_{i,j}(a\,x_i + \overline{y} - a\,x_i - \overline{y})^2 \\
&= \sum_{i,j} n_{i,j}\left((a(x_i - \overline{x})\right)^2 \\
&= a^2 n_{\bullet,\bullet} Var(X) \\
&= \left[\frac{Cov(X,Y)}{Var(X)}\right]^2 n_{\bullet,\bullet} Var(X) \\
&= n_{\bullet,\bullet}\frac{Cov(X,Y)^2}{Var(X)}
\end{aligned}$$

et donc

$$r = \sqrt{\frac{Cov(X,Y)^2}{Var(X)Var(Y)}} = \rho$$

□

Proposition 32 *Dans les conditions de l'ajustement linéaire par la méthode des moindres carrés la variance résiduelle est égale à*

$$\frac{1}{n_{\bullet,\bullet}} \sum_{i,j} n_{i,j}(y_j - y_{i,est})^2 = E((Y - Y_{est})^2)$$

démonstration

La variance totale est égale à

$$\begin{aligned}
E[(Y - \overline{Y})^2] &= E[\left((Y - Y_{est}) + (Y_{est} - \overline{Y})\right)^2] \\
&= E((Y - Y_{est})^2) + E((Y_{est} - \overline{Y})) + E\left((Y - Y_{est})(Y_{est} - \overline{Y})\right) \\
&= Var_{expl}(Y) + E((Y - Y_{est})^2) + E\left((Y - Y_{est})(Y_{est} - \overline{Y})\right)
\end{aligned}$$

Or

$$\begin{aligned}
E\left((Y - Y_{est})(Y_{est} - \overline{Y})\right) &= E\left((Y - aX + b)(aX + b - \overline{Y})\right) \\
&= E\left((Y - aX + \overline{Y} - a\overline{X})(aX + \overline{Y} - a\overline{X} - \overline{Y})\right) \\
&= E\left((Y + \overline{Y} - a(X + \overline{X}))(a(X - \overline{X})\right) \\
&= a\left(E(XY) - \overline{XY} + \overline{XY} - \overline{XY} - aE(X^2 - \overline{X}^2\right) \\
&= a\left(Cov(X,Y) - aVar(X)\right) \\
&= 0
\end{aligned}$$

compte-tenu de la valeur de a.

□

Définition 47 *Dans le cas d'un ajustement de la série $(x_i, y_j, n_{i,j})$ par une courbe d'équation $y = f(x)$, on appelle coefficient de corrélation de Y en X de l'ajustement la racine carrée du rapport de la variance expliquée par la variance totale de Y, i.e.*

$$r_{Y/X} = \sqrt{\frac{Var_{expl}(Y)}{Var(Y)}}$$

Nous définissons bien sur de la même manière les variances totales et expliquées de X en Y ainsi que le coefficient de corrélation de X en Y.

Proposition 33 *Dans le cas d'un ajustement linéaire les coefficients de corrélation de Y en X et de X en Y sont égaux (et égaux au coefficient de corrélation linéaire).*

Dans le cas général cependant ces deux coefficients de corrélation n'ont aucune raison d'être égaux.

3.4.5 Rapport de corrélation

Dans le cas général où aucune "loi" classique ne semble apparente, il est intéressant de disposer d'un instrumement permettant de mesurer le degré de liaison entre les variables. Nous allons donc revenir aux courbes de régression, i.e. considérer l'ajustement par ces courbes (il s'agit bien d'un ajustement par les moindres carrés !).

Définition 48 *On appelle* **rapport de corrélation** *de Y en X le coefficient de corrélation de Y en X par rapport à son ajustement par la courbe de régression de Y en X.*

Proposition 34 *Dans le cas d'un ajustement par la courbe de regression, la variance expliquée est égale à la variance des moyennes conditionnelles et la variance résiduelle est égale à la moyenne des variances conditionnelles.*

3.5 Exercices

Traiter les exercices suivant à l'aide d'un TABLEUR.

- Prévoir deux colonnes pour les données A et B.
- Construire les graphes dans les différentes échelles et conclure quant au type de l'ajustement (lorsque la loi n'est pas donnée et que la réponse est évidente). Si la réponse n'est pas évidente, envisager les différents cas par la suite.
- Prévoir deux colonnes X et Y contenant les données transformées afin de se ramener à un ajustement linéaire (par exemple $X = A$ et $Y = \ln B$).
- Faire calculer la moyenne, la variance et l'écart-type de X et de Y, ainsi que la covariance de X et Y. En déduire le coefficient de corrélation.

- Bien que les tableurs actuels comportent presque tous des fonction adaptées, vous n'utiliserez - à des fins pédagogiques - que les fonctions :
  - @SUM (ou SOMME si la tableur est francisé) qui calcule la somme des éléments d'un domaine
  - @COUNT qui détermine le nombre de cellules non vides dans un domaine
  - @SQRT qui détermine la racine carrée d'un nombre

  Il faudra donc prévoir des colonnes pour X^2, Y^2, XY.

Rédiger - à l'aide d'un traitement de texte - un rapport présentant le problème et les données, les graphes, un rapide rappel de cours, les résultats et les conclusions que l'on peut tirer de cette étude.

Exercice 22

On se propose de déterminer la loi de variation de la résistance d'une sonde de platine en fonction de la température. Une série de mesures nous donne les résultats suivants :

température (*degrés Celsius*)	20	30	40	50	60	70	80	90	100
Résistance (Ω)	107.75	111.67	115.54	119.40	123.94	127.07	130.89	134.70	138.50

Exercice 23

On se propose de déterminer la loi de variation de la viscosité de l'eau en fonction de la température. Une série de mesures nous donne les résultats suivants :

température (*degrés Celsius*)	10	15	20	15	30	35	40	45
Viscosité (*Centipoise* : *CPo*)	1.305	1.148	1.018	0.910	0.820	0.740	0.675	0.615

Exercice 24

Pour le dosage d'un acide faible (acide acétique) par une base forte (soude), on étudie le PH en fonction du volume de soude v. Si l'on pose $x = v/v_c$ (v_c étant le volume à l'équivalence) la "théorie" nous dit que

$$PH = a + \log \frac{x}{1-x}$$

a étant le PK de l'acide.
Une série de mesures nous donne les résultats suivants :

volume de soude (cm^3)	2	4	8	12	16	18	19
PH	3.9	4.2	4.6	4.9	5.4	5.9	7

Sachant que le volume à l'équivalence est égal à 19.3 cm^3, on se propose de déterminer le PK de l'acide.

Exercice 25

Une réaction cinétique est dite d'ordre 0 si $[a] = [a_0] - kt$ et d'ordre 1 si $[a] = [a_0]\, e^{-kt}$, où $[a]$ désigne une concentration et t le temps.
La réaction

$$C_2H_5I + OH^- \longrightarrow C_2H_5OH + I^-$$

donne les résultats suivants

t *en mn*	0	906	2715	6335
$[OH^-]$ *en mol/l*	0.05	0.025	0.0125	0.00625

La réaction

$$(H_2N)_2CO \longrightarrow NH_4^+ + (CNO)^-$$

donne

t *en mn*	0	9600	18220	28600
$[urée]$ *en mol/l*	0.10	0.0854	0.742	0.0625

(urée : $(H_2N)_2CO$))
Déterminer l'ordre de chacune de ces deux réactions.

Exercice 26

Dans le cas d'un gaz diatomique (l'air par exemple), on étudie la variation de pression en fonction du volume (transformation adiabatique). Des mesures nous donnent :

volume en l	1	3	5	7	9	11	13
Pression en Pascal	101325	14523	8291	5538	4032	3112	2546

Déterminer la loi de variation de la pression en fonction du volume.

Exercice 27

On mesure la charge de rupture et la teneur en carbone de différents aciers :

teneur en carbone en %	70	60	68	64	66	64	62	70	74	62
charge de rupture en kg	87	71	79	74	79	80	75	86	95	70

Etudier l'ajustement de ces deux caractères et donner une estimation de la teneur en carbone d'un acier dont la charge de rupture est de 90kg.

On représentera le nuage de points et la courbe d'ajustement.

Exercice 28

On mesure la résistance à l'avancement d'un poids lourd et sa vitesse :

Vitesse V en km/h	10	20	30	40	50	60	70	80	90
Résistance Pr en kW	2.6	5.8	9.9	15.4	23.6	34.5	49	67.2	89.1

Etudier l'ajustement de ces deux caractères (on testera Pr en fonction de V^3) et donner une estimation de la résistance pour une vitesse de 100 km/h.

On représentera le nuage de points et la courbe d'ajustement.