Initiation aux statistiques descriptives : cours

Anahita Basirat
anahita.basirat@gipsa-lab.inpg.fr

Janvier 2009

Table des matières

1 Vocabulaire de base **2**
- 1.1 Statistique descriptive . . . 2
- 1.2 Statistique inférentielle . . . 2
- 1.3 Population . . . 2
- 1.4 Echantillon . . . 2
- 1.5 Variables ou caractères statistiques . . . 2
- 1.6 Effectif et fréquence . . . 3
- 1.7 Effectifs cumulés croissants et décroissants . . . 3
- 1.8 Série statistique . . . 3

2 Représentations graphiques **4**
- 2.1 Variables qualitatives . . . 4
- 2.2 Variables quantitatives . . . 5

3 Paramètres caractéristiques d'une variable : paramètres de position **6**
- 3.1 Mode . . . 6
- 3.2 Moyenne . . . 6
  - 3.2.1 Moyenne arithmétique . . . 6
  - 3.2.2 Moyenne pondérée . . . 6
  - 3.2.3 Propriétés . . . 7
- 3.3 Médiane . . . 7
  - 3.3.1 Calcul de médiane pour une variable discrète . . . 7
  - 3.3.2 Calcul de médiane pour une variable continue . . . 7

4 Paramètres caractéristiques d'une variable : paramètres de dispersion **9**
- 4.1 Etendue . . . 9
- 4.2 Quartiles . . . 9
- 4.3 Déciles . . . 9
- 4.4 Diagramme de Tukey . . . 10
- 4.5 Variance . . . 10
- 4.6 Ecart-type . . . 11
  - 4.6.1 Propriété . . . 11

5 Série statistique à deux variables **12**
- 5.1 Représentation graphique . . . 12
- 5.2 Covariance . . . 13
- 5.3 Coefficient de corrélation linéaire . . . 13
- 5.4 La droite d'ajustement . . . 13

Ce document est conçu à partir de différents supports accessibles sur Internet. Voir la bibliographie.

Chapitre 1

Vocabulaire de base

1.1 Statistique descriptive

Il s'agit d'organiser et résumer des observations. On ne fait pas de comparaisons et on s'intéresse en général à un seul groupe, échantillon ou population. Exemple : Lorsqu'on calcule la fréquence des mots dans un texte, on fait de la statistique descriptive [J. Vérnois].

1.2 Statistique inférentielle

Partie de la statistique qui, contrairement à la statistique descriptive, ne se contente pas de décrire des observations, mais extrapole les constatations faites à un ensemble plus vaste, permet de tester des hypothèses sur cet ensemble, et de prendre des décisions le concernant. Exemple : Lorsqu'on cherche à décider si la différence entre les fréquences d'un mot dans deux textes est **significative** ou bien si elle est imputable au hasard, on fait de la statistique inférentielle [J. Vérnois].

1.3 Population

La population désigne un ensemble d'unités statistiques. Les unités statistiques, aussi appelées individus, sont les entités abstraites qui représentent des personnes, des animaux ou des objets. La statistique sert à décrire l'ensemble des unités statistiques qui composent la population.

1.4 Echantillon

Lorsque la population est trop importante, on étudie un échantillon, c'est-à-dire un sous-ensemble, beaucoup plus petit, de la population. L'échantillon doit être bien choisi pour pouvoir représenter la population.

1.5 Variables ou caractères statistiques

Un individu donné de la population peut être étudié selon certaines propriétés. Ces propriétes sont appelées caractères ou variables statistiques. Exemple : une étude sur les étudiants de l'université Stendhal peut porter sur les différentes variables : leur âge, leur sexe, leur nationalité, leur moyenne de l'année, etc.

- Une variable **qualitative** est une variable qui ne prend pas de valeur numérique. Exemple : sexe, nationalité. Chaque variable qualitative a plusieurs modalités. Exemple : pour la variable état matrimonial, les modalités sont célibataire, marié, veuf, divorcé.

  Une variable qualitative **ordinale** prend des valeurs qui sont ordonnées, hiérarchisées. On peut classer les modalités les unes par rapport aux autres mais on ne peut pas dire à partir de cet ordre de "combien" est la différence entre deux modalités. . Exemple : les réponses à un sondage, du type "pas du tout", "un peu", "assez", "beaucoup" [J. Vérnois].

- Une variable **quantative** est une variable qui est de la forme d'une variable numérique. Exemple : âge, moyenne de l'année.

Une variable quantitative **discrète** peut prendre des valeurs dénombrables. Exemple : le nombre d'enfants d'un ménage tandis qu'une varibale quantitative **continue** peut prendre toutes les valeurs à l'intérieur d'un intervalle. Exemple : la taille.

Les variables quantitatives peuvent être regroupées en classes (intervalles). Exemple : le nombre d'enfant d'un ménage peut être regroupé en 4 classes, [0 ;1],[2 ;3],[4 ;5], plus de 5 enfants. La taille d'un échantillon d'étudiants en CP peut être classée en moins d'un mètre, [1 ;1.20] mètre, plus de 1.20 mètre. Pour une classe, on peut parler de son **amplitude** : soit [a ;b[une classe d'une variable quantitative, on dit que $b - a$ est l'amplitude de cette classe.

1.6 Effectif et fréquence

L'effectif d'une valeur donnée d'une variable est l'ensemble d'individus présentant cette valeur. L'effectif total est la somme de tous les effectifs d'une varaible.

La fréquence d'une valeur donnée est le rapport de l'effectif correspondant à l'effectif total. La fréquence totale est toujours égale à 1.

Exemple : une étude sur l'état matrimonial des salariés de la société X.
- Population : salariés de la société X.
- Unité statistique (individu) : chaque salarié de la societé X.
- Variable (caractère) étudiée : état matrimoniale avec 4 modalités : célibataire, marié, veuf, divorcé.
- Effectif : l'effectif de la modalité célibataire = n_1, marié = n_2, veuf = n_3, divorcé = n_4.
- Effectif total : N = $n_1 + n_2 + n_3 + n_4$.
- Fréquence : fréquence de la modalité célibataire = $\frac{n_1}{N}$, marié = $\frac{n_2}{N}$, veuf = $\frac{n_3}{N}$, divorcé = $\frac{n_4}{N}$.
  Fréquence totale = $\frac{N}{N} = 1$.

1.7 Effectifs cumulés croissants et décroissants

Quand les modalités ou les classes d'une variable sont rangées dans l'ordre croissant(décroissant), les effectifs cumulés croissants (ou décroissants) d'une valeur s'obtient en ajoutant à chaque effectif les effectifs des valeurs qui la précédent. Les fréquences cumulées s'obtiennent en divisant les effectifs cumulés par l'effectif total.

Note sur 20	Moins de 5	[5 ;10[	[10 ;12[	[12 ;15[	[15 ;17[	[17 ;20]
Effectif	2	3	7	5	3	1
Fréquence	0.09	0.14	0.33	0.24	0.14	0.05
Effectif cumulé croissant	2	5	12	17	20	21
Effectif cumulé décroissant	21	19	16	9	4	1

TAB. 1.1 – Exemple d'effectif cumulé : notes d'une population de 21 étudiants.

1.8 Série statistique

Une séries statistique est la suite des observations d'une (voire plusieurs) variable, relevées sur les individus d'une population. Exemple : les notes des étudiants présentées sur le tableau 1.1.

Chapitre 2

Représentations graphiques

Une bonne représentation graphique est très utile pour comprendre les observation d'une étude statistique. Ce chapitre présente quelques graphiques classiques pour représenter les effectifs observés dans une étude statistiques.

2.1 Variables qualitatives

Diagramme en barre : dans ce diagramme, les modalités de la variable sont placées sur une droite horizontale et les effectifs (ou les fréquences) sont placés sur un axe vertical. La hauteur de la barre est proportionnelle à l'effectif (figure 2.1). Les barres ont une certaine épaisseur pour qu'il n'y ait pas de confusion avec les diagrammes en bâtons réservés à des variables quantitatives discrètes(figure 2.3).

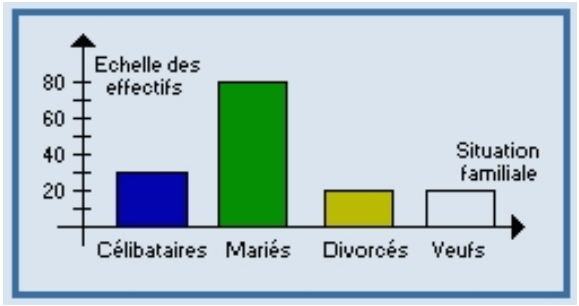

FIG. 2.1 – Exemple de diagramme en barre (figure extraite de [Technique de la statistique]).

Diagramme circulaire ou camembert : L'effectif total est représenté par un disque. Chaque modalité est représentée par un secteur circulaire dont la surface (pratiquement : l'angle au centre) est proportionnelle à l'effectif correspondant (figure 2.2). L'angle de chaque modalité se calcule par :

$$\frac{\text{effectif de chaque modalité}}{\text{effectif total}} \times 360°$$

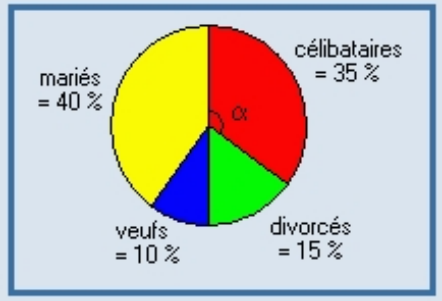

FIG. 2.2 – Exemple de diagramme circulaire (figure extraite de [Technique de la statistique]).

2.2 Variables quantitatives

Diagramme en bâtons : Les valeurs discrètes x_i prises par les variables sont placées sur l'axe des abscisses, et les effectifs (ou les fréquences) sur l'axe des ordonnées. La hauteur du bâton est proportionnelle à l'effectif (figure 2.3).

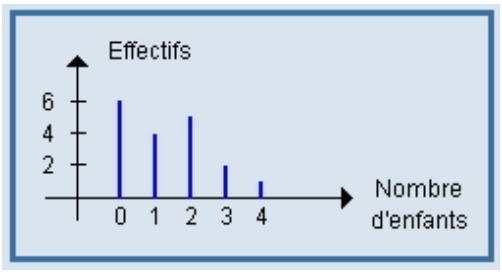

FIG. 2.3 – Exemple de diagramme en bâtons (figure extraite de [Technique de la statistique]).

Histogramme : on utilise l'histogramme pour les variables classées. C'est un ensemble de rectangles. Chaque rectangle est associé à une classe et il a une surface proportionnelle à l'effectif (ou fréquence) de cette classe.

- Amplitudes égaux : Si les classes ont la même amplitude, on reporte en ordonnée l'effectif (ou fréquence) des classes (voir figure 2.4 à gauche).
- Amplitudes diverses : si les amplitudes sont différentes, on reporte en ordonnée la **densité** d_i (effectif divisé par l'amplitude de la classe) pour que la surface de chaque rectangle soit proportionnelle à l'effectif (ou fréquence) (voir figure 2.4 à droite).

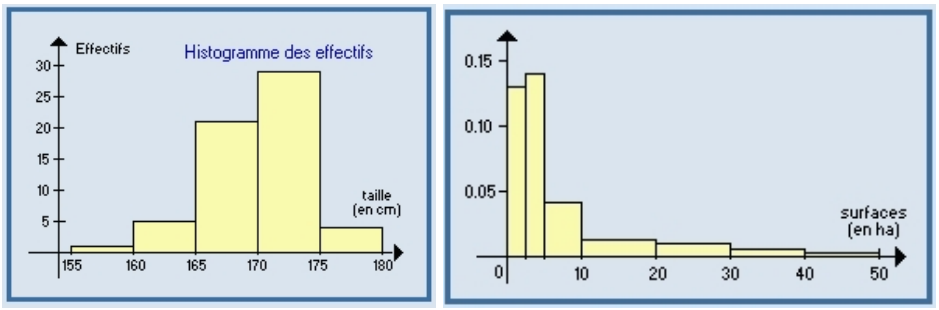

FIG. 2.4 – Exemples d'histogrammes. A gauche : des classes de même amplitude et à droite : des classes de différentes amplitudes (figures extraites de [Technique de la statistique]).

Chapitre 3

Paramètres caractéristiques d'une variable : paramètres de position

Les paramètres de position (ou de tendance centrale) permettent de savoir autour de quelles valeurs se situent les valeurs d'une variable statistique.

3.1 Mode

Pour une variable discrète, le mode est la modalité qui represente le plus grand effectif. Exemple : sur la figure 2.3, le mode est 0 enfant.

Pour une variable quantitative continue nous parlons de **classe modale** : c'est la classe dont l'effectif est maximum. Exemple : dans le tableau 1.1, la classe modale est la classe [10..12[.

3.2 Moyenne

3.2.1 Moyenne arithmétique

La moyenne arithmétique d'une série statistique est la somme des valeurs divisée par le nombre total des valeurs. Par exemple, la moyenne de l'année est la somme des notes de tous les examens divisée par le nombre d'examen. La moyen de X se calcule par $\overline{x} = \frac{x_1+x_2+\cdots+x_N}{N}$. Dans cette formule, $x_1, x_2, \cdots, x_N$ sont les notes et N est le nombre total des notes.

3.2.2 Moyenne pondérée

Lorsque les valeurs sont affectées de coeficients (ici d'effectifs), on parle de la moyenne pondérée (regarder la tableau 3.1). La moyenne pondérée de X se calcule par :

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

Dans cette formule, $n_1, n_2, \cdots, n_N$ sont les effectifs correspondants aux modalités $x_1, x_2, \cdots, x_N$, si la série est discrète ou aux centres de chaque classe, si la série est continue.

Voici l'exemple d'une série discrète, tableau 3.1 et d'une série continue, tableau 3.2.

Qualité de service	Effectif	Produit n_ix_i
1	1	1
2	3	6
3	5	15
4	2	8
5	1	5
total	12	35

TAB. 3.1 – Moyenne d'une variable discrète, $\overline{x} = \frac{35}{12} = 2.9$

Notes	Effectifs	Centre	Produit $n_i x_i$
[0..5[	10	2.5	25
[5..8[	8	6.5	52
[8..12[	12	10	120
[12..15[	11	13.5	148.5
[15..20]	9	17.55	157.5
Total	50		503

TAB. 3.2 – Moyenne d'une variable continue, $\overline{x} = \frac{503}{50} = 10.06$

3.2.3 Propriétés

1. Considérons une série statistique S_1 de modalités $x_1, x_2, \cdots, x_N$ avec des effectifs $n_1, n_2, \cdots, n_N$ de moyenne $\overline{x}$ et la série statistique S_2 de modalités $y_1, y_2, \cdots, y_N$ avec des effectifs $n_1, n_2, \cdots, n_N$ telle que pour tout i appartenant à $\{1, 2, \cdots, N\}$: $y_i = ax_i + b$. Alors la moyenne de la série statistique S_2 est : $\overline{y} = a\overline{x} + b$. Exemple : la moyenne de notes d'une classe de 22 étudiants est 12.5. En ajoutant 0.5 point à toutes les notes, on obtient une moyenne de 13.

2. Soient S_1 et S_2 deux séries statistiques d'effectifs totaux respectifs N_1 et N_2 et de moyennes respectives $\overline{x_1}$ et $\overline{x_2}$. Alors la moyenne de la série S regroupant les deux séries S1 et S2 est : $\overline{x} = \frac{N_1\overline{x_1}+N_2\overline{x_2}}{N_1+N_2}$. Exemple : la moyenne de notes d'une classe de 22 étudiants est 12.5 et celle d'une classe de 18 étudiants est 13.2. La note moyenne de ces deux classes est

$$\overline{x} = \frac{22\times 12.5+18\times 13.2}{22+18} = 12.81.$$

3.3 Médiane

C'est le nombre qui permet de couper la population étudiée (ou une série statistique) en deux groupes contenant le même nombre d'individus (ou en deux sous groupes égaux). Exemple : Soit la série statistiques suivante : 15, 7, 22, 4, 12, 30, 9, 18, 6. Pour déterminer la médiane, il faut ordonner la série : 4, 6, 7, 9, 12, 15, 18, 22, 30. La médiane est le 12 car dans cette série, il y a 4 nombres inférieure et 4 supérieure de 12.

On peut utiliser les effectifs cumulés pour déterminer la médiane. Exemple : dans le tableau 1.1, la médiane se trouve dans la classe [10 ;12[car l'individu 11 qui distingue deux groupes egaux de la population (deux groupes de 10 individus), se trouve dans cette classe.

3.3.1 Calcul de médiane pour une variable discrète

Si l'effectif total est impair (2n+1), la médiane est parfaitement déterminée : la modalité correspondant à n+1. Exemple : dans le tableau 3.3, une étude sur le nombre d'enfant d'une échantillon de 51 individus $(2 \times 25 + 1)$ est présentée. La médiane est la modalité "1 enfant" qui correspond au foyer 26.

Nombre d'enfant	0	1	2	3	4
Effectif	20	16	10	5	0
Effectif cumulé croissant	20	36	46	51	51

TAB. 3.3 – Calcul de médiane en utilisant les effectifs cumulés croissants : cas d'une variable discrète

Si l'effectif total est pair (2n), on ne peut pas définir précisément la médiane : On définit un intervalle médian. Exemple : une série représantant les notes d'une classe : 15, 7, 20, 4, 12, 20, 9, 18, 6, 4 (séire ordonée : 4, 4, 6, 7, 9, 12, 15, 18, 20, 20), l'interval madéiane est 9 et 12. Dans ce cas là, la médianne est $\frac{9+12}{2} = 10.5$.

3.3.2 Calcul de médiane pour une variable continue

Pour une variable continue, on détermine la classe médiane de même façon que pour une variable discrète en utilisant les effectifs cumulés. Exemple : dans le tableau 1.1, la **classe médiane** est la classe [10 ;12[. On détermine la médiane au sein d'une classe par l'interpolation linéaire.

Soit une étude sur la note d'une population de 50 étudiants (tableau 3.4) [Math Web]. D'après la colonne "effectif cumulé", 18 personnes ont moins de 8 et 30 personnes ont moins de 12. La médiane se trouve donc

dans l'intervalle [8 ;12[.

Notes	Effectifs	Effectif cumulés
[0 ;5[	10	10
[5 ;8[	8	18
[8 ;12[	12	30
[12 ;15[	11	41
[15 ;20]	9	50

TAB. 3.4 – Calcul de médiane en utilisant les effectifs cumulés croissants : cas d'une variable continue

Sur la figure 3.1, les points A, X, B sont alignés et les droites AX, BX et AB ont le même coefficient directeur (la pente est la même). Le coefficient directeur d'une droite est détérminé par deux de ces points. Le coefficient directeur de la droite AB se calcule par :

$$m = \frac{y_B - y_A}{x_B - x_A}$$

Pour trouver la valeur Me, on peut calculer m_{AX} et m_{AB} et résoudre la régle de trois suivante :

$$m_{AX} = m_{AB} \text{ donc } \frac{Me-8}{25-18} = \frac{12-8}{30-18}$$

La médiane Me est donc 10.33. Cela signifie que environ %50 des personnes ont eu moins de 10.33 et %50 plus de 10.33.

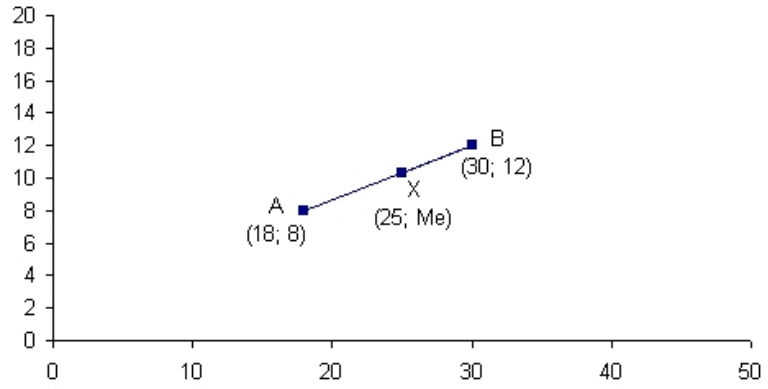

FIG. 3.1 – Calcul de médiane pour une variable continue. En abscisse : effectifs cumulés et en ordonnée : notes.

Chapitre 4

Paramètres caractéristiques d'une variable : paramètres de dispersion

Les paramètres de dispersion nous rensiengent sur la dispersion des valeurs autour de la valeur centrale de référence.

4.1 Etendue

L'étendue d'une série statistique quantitative est la différence entre la plus grande valeur de la variable (discrète ou continue) et la plus petite valeur. Exemple, dans le tableau 1.1, l'étendue est $20 - 0 = 20$.

4.2 Quartiles

On appelle **premier quartile** d'une série la plus petite valeur Q1 des termes de la série pour laquelle au moins un quart (25%) des données sont inférieures ou égales à Q1.

On appelle **troisième quartile** d'une série la plus petite valeur Q3 des termes de la série pour laquelle au moins trois quarts (75%) des données sont inférieures ou égales à Q3.

On appelle **intervalle interquartile** l'intervalle [Q1 ;Q3].

On appelle **écart interquartile** le nombre Q3-Q1.

Exemple [Math Web] : la **série ordonnée par ordre croissant** S a 12 termes :

$$S = \{11, 12, 13, 15, 16, 16, 17, 17, 18, 19, 20, 22\}$$

Rang	1	2	3	4	5	6	7	8	9	10	11	12
Série	11	12	13	15	16	16	17	17	18	19	20	22

TAB. 4.1 – Calcule des quartiles

Un quart (25%) des données correspond à : $12 \times 0.25 = 3$. Le premier quartile est alors, par définition, la plus petite valeur Q1 pour laquelle les valeurs de 3 termes de la série sont inférieurs ou égales à Q1. Le premier quartile est donc la valeur du 3ème terme de la série c'est-à-dire 13[1].

Trois quarts (75%) des données correspondent à : $12 \times 0.75 = 9$. Le troisième quartile est alors, par définition, la plus petite valeur Q3 pour laquelle les valeurs de 9 termes de la série sont inférieurs ou égales à Q3. Le troisième quartile est donc la valeur du 9ème terme de la série c'est-à-dire 18.

L'intervalle interquartile est [3 ;9]. L'écart interquartile est 18-13 = 5.

4.3 Déciles

On appelle **premier décile** d'une série la plus petite valeur D1 des termes de la série pour laquelle au moins un dixième (10%) des données sont inférieures ou égales à D1

[1] Notez qu'il existe d'autres méthodes de calcules des quartiles et déciles. Vous pouvez, par exemple, consulter ce document http ://www.math.unicaen.fr/irem/stat/quartile.pdf

On appelle **neuvième décile** d'une série la plus petite valeur D9 des termes de la série pour laquelle au moins neuf dixièmes (90%) des données sont inférieures ou égales à D9.
On appelle intervalle **interdécile l'intervalle** [D1 ;D9].
On appelle **écart interdécile** le nombre D9-D1.

Exemple [Math Web] : La **série ordonnée par ordre croissant** S a 11 termes :

$$S = \{1500, 1650, 1700, 1800, 1850, 2000, 2100, 2300, 2500, 2650, 2700\}$$

Rang	1	2	3	4	5	6	7	8	9	10	11
Série	1500	1650	1700	1800	1850	2000	2100	2300	2500	2650	2700

TAB. 4.2 – Calcule des déciles

Un dixième (10%) des données correspond à : $11 \times 0.10 = 1.1$. Le premier décile est alors, par définition, la plus petite valeur D1 pour laquelle les valeurs de 2 termes ($2 \geqslant 1.1$) de la série sont inférieurs ou égales à D1 . Le premier décile est donc la valeur du 2ème terme de la série c'est-à-dire 1650.

Neuf dixièmes (90%) des données correspondent à : $11 \times 0.9 = 9.9$. Le neuvième décile est alors, par définition, la plus petite valeur D9 pour laquelle les valeurs de 10 termes ($10 \geqslant 9.9$) de la série sont inférieurs ou égales à D9. Le neuvième décile est donc la valeur du 10ème terme c'est-à-dire 2650.

L'intervalle interdécile est [2 ;10]. L'écart interdécile est 2650 - 1650 = 1000.

4.4 Diagramme de Tukey

Ce type de diagramme est appelé diagramme de Tukey ou boîte à moustaches ou boîte à pattes. Il représente le 1er et le 3ème quartile, le 1er et le 9ème décile, les valeurs extrêmes et éventuellement la médiane d'une série (figure 4.1).

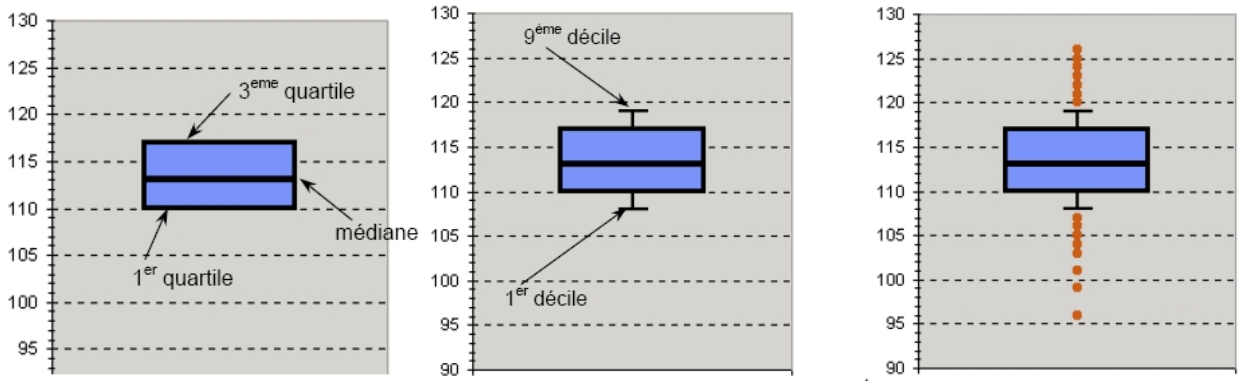

FIG. 4.1 – Diagramme de Tukey (figures extraites de [Math Web])

4.5 Variance

La variance est un indicateur de la dispersion d'une série par rapport à sa moyenne. La définition de la variance d'une série statistiques est donnée par la formule :

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

$V(x)$ désigne la variance des n valeurs associées aux n unités statistiques de la population et $\overline{x}$ est la moyenne de ces unités statistiques.

4.6 Ecart-type

La définition de l'écart-type d'une série est donnée par la formule : $\sigma(x) = \sqrt{V(x)}$. Si l'écart-type est faible, cela signifie que les valeurs sont assez concentrées autour de la moyenne et si l'écart-type est élevé, cela veut dire au contraire que les valeurs sont plus dispersées autour de la moyenne.

Exemple [F.Mazerolle] : Dans une usine, le fait d'avoir un écart-type aussi bas que possible peut constituer un objectif de contrôle de qualité. Soit une entreprise qui fabrique un certain composant et qu'un des éléments du contrôle de la qualité consiste à mesurer le diamètre du composant. Chaque composant aura donc son diamètre mesuré. On calculera ensuite le diamètre moyen, puis l'écart-type. Si l'écart-type est faible, cela signifie que les pièces ont dans l'ensemble un diamètre proche de la moyenne, donc que leur diamètre se ressemble. À la limite, un écart-type nul signifie que toute les pièces ont le même diamètre. Inversement, plus l'écart-type est élevé, plus il y a de pièces dont le diamètre s'écarte de la moyenne et qui risque de ne pas cadrer avec le système auxquelles elles sont destinées.

4.6.1 Propriété

Considérons une série statistique S_1 de modalités $x_1, x_2, \cdots, x_N$ affectées des effectifs $n_1, n_2, \cdots, n_N$ d'écart-type $\sigma(x)$, et la série statistique S_2 de modalités $y_1, y_2, \cdots, y_N$ affectées des même effectifs $n_1, n_2, \cdots, n_N$ telle que pour tout i appartenant à {1, 2,..., N} : $yi = axi + b$. Alors : l'écart-type de la série statistique S_2 est : $\sigma(y) = |a|\,\sigma(x)$.

Chapitre 5

Série statistique à deux variables

Une série statistique à deux variables (ou une série double) représente une étude statistique sur deux variables de la même population.

Exemple : Le tableau 5.1 montre deux séries mensuelles. La première indique le temps passé par une personne sur Internet chaque mois (en heures) et la seconde série indique le total de la somme dépensée sur différents sites marchands [F.Mazerolle].

Mois		Temps en heure	Dépense en euros
1	Janv 07	4.20	142.25
2	Févr 07	3.50	79.59
3	Mars 07	1	12.50
4	Avr 07	5.60	56.42
5	Mai 07	5.00	98.74
6	Juin 07	7.30	319.12

TAB. 5.1 – Exemple d'une série à deux variables : Temps passé sur Internet (heures/mois) et somme dépensée sur différents sites marchands (euros) [F.Mazerolle]

On représente une série statistique à deux variables sous la forme des couples $(x_i; y_i)$, où x et y représentent les deux variables étudiées et i représente l'individu i de la population. Le point $(\overline{x}; \overline{y})$, où $\overline{x}$ et $\overline{y}$ sont la moyenne de x et y, est appelé **point moyen** de la série statistique double.

5.1 Représentation graphique

Un graphique de dispersion ou nuage de points est un graphique qui met en relation les valeurs de deux variables sur un repère de coordonnées cartésiennes (figure 5.1).

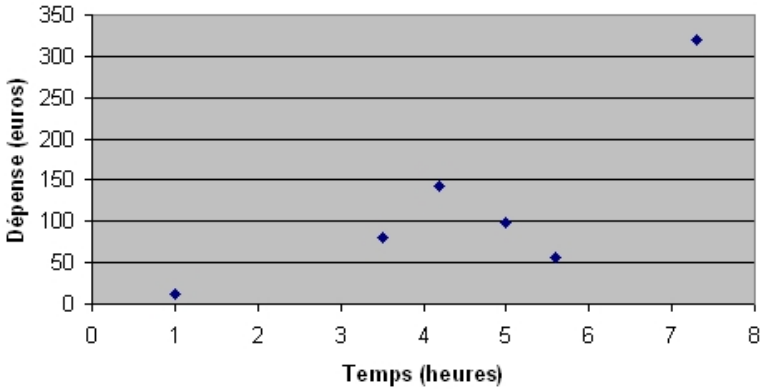

FIG. 5.1 – Nuage de de points correspondant au tableau 5.1.

5.2 Covariance

Pour mesurer la dispersion des points d'un nuage par rapport au point moyen, on utilise la covariance :

$$Cov(x;y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$$

Théorème de Huyghens-König : $Cov(x,y) = \overline{x.y} - \overline{x}.\overline{y}$

La covariance est positive si X et Y ont tendance à varier dans le même sens, et négative si elles ont tendance à varier en sens contraire. La grandeur de la covariance entre deux variable X et Y ne nous apporte pas beaucoup d'information sur la liaison entre X et Y car elle dépend de la variance des deux variables X et Y. Si les variables sont indépendantes, alors Cov(x ;y)= 0.

5.3 Coefficient de corrélation linéaire

Le coefficient de corrélation entre deux variables statistiques X et Y sur les mêmes individus est :

$$r(x;y) = \frac{Cov(x;y)}{\sigma_x \sigma_y}$$

Ce coefficient est toujours compris entre -1 et +1. S'il est proche de +1 ou -1 , X et Y sont bien corrélées, c'est-à-dire qu'elles sont liées entre elles par une relation presque affine ; le nuage de points est presque aligné le long d'une droite (croissante si r = +1, décroissante si r = -1). S'il n'y a aucun lien linéaire entre X et Y, ce coefficient est nul, ou presque nul.

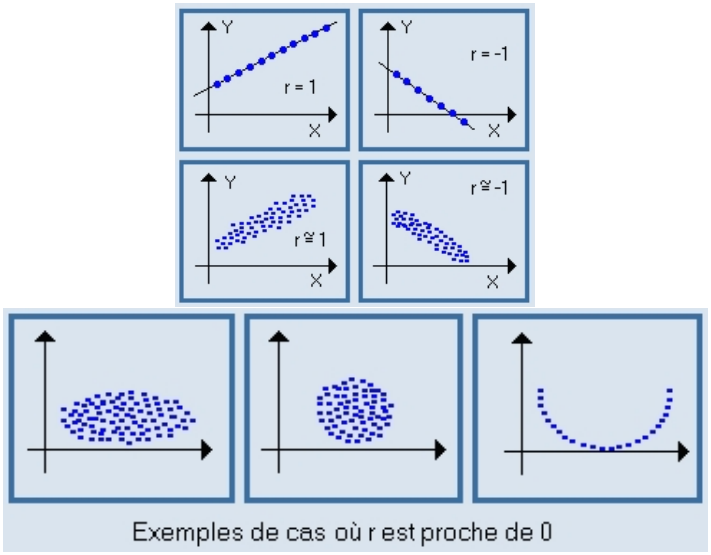

FIG. 5.2 – Coefficient de corrélation linéaire (figures extraites de [Technique de la statistique])

"Il faut prendre garde à la confusion fréquente entre corrélation et causalité. Que deux phénomènes soient corrélés n'implique en aucune façon que l'un soit cause de l'autre. Très souvent, une forte corrélation indique que les deux caractères dépendent d'un troisième, qui n'a pas été mesuré. Ce troisième caractère est appelé "facteur de confusion". Qu'il existe une corrélation forte entre le rendement des impôts en Angleterre et la criminalité au Japon, indique que les deux sont liés à l'augmentation globale de la population. Il arrive qu'une forte corrélation traduise bien une vraie causalité, comme entre le nombre de cigarettes fumées par jour et l'apparition d'un cancer du poumon. Mais ce n'est pas la statistique qui démontre la causalité, elle permet seulement de la détecter" [B. Ycart].

5.4 La droite d'ajustement

Le coefficient de corrélation mesure la dépendance linéaire des variables. Si cette dépendance est bonne, on peut exprimer la variable Y comme fonction linéaire de X. C'est à dire que les valeurs y_i peuvent être remplacées par des valeurs calculées qui sont fonctions des x_i. Plus précisément : $y_1 = ax_1 + b$, $y_2 = ax_2 + b$, $\cdots$. Il reste donc à déterminer les valeurs des paramètres a et b, qui désignent respectivement la pente et l'ordonnée à l'origine

de la droite d'ajustement (regarder la figure 5.2 où r=1 et r=-1).

Pour déterminer les paramètres a et b, on peut utiliser la **méthode des moindres carrés**. Cette méthode consiste à minimiser la somme des carrés des distances entre les points y_i et les valeurs calculées correspondantes y_{ic} (regarder la figure 5.3). Il est alors possible d'en déduire des formules pour a et b :

$$a = \frac{Cov(x;y)}{{\sigma_x}^2} = \frac{r_{xy}\sigma_y}{\sigma_x}$$
$$b = \overline{y} - a.\overline{x}$$

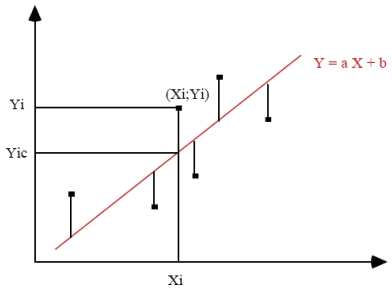

FIG. 5.3 – La droite d'ajustement corrélant Y à X (figure extraite de [O.Maggioni]).

si l'on veut exprimer X comme fonction de Y on obtiendra une autre droite, qui correspond à la minimisation des carrés de distances horizontales (regarder la figure 5.4).

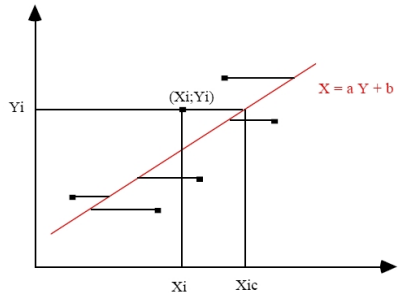

FIG. 5.4 – La droite d'ajustement corrélant X à Y (figure extraite de [O.Maggioni]).

En général on ajuste l'effet (Y) contre la cause (X). Cette relation de causalité ne provient pas de l'analyse statistique, mais bien de la connaissance que l'on a du phénomène considéré.

Bibliographie

[D.Grau] Cours de statistique descriptive de Daniel Grau, Département Gestion des Entreprises et des Administrations, IUT de Bayonne.

[O.Maggioni] Premier Chapitre de cours de statistique d'Olivier Maggioni, Université de Neuchâtel, Institut de Mathématiques, Suisse.

[Math Web] Site Internet Math Web sur http ://jellevy.yellis.net/

[F.Mazerolle] Cours de statistique descriptive de Fabrice Mazerolle, Faculté d'Economie Appliquée, Université Aix-Marseille III.

[Technique de la statistique] Collection de "Technique de la statistique" sur http ://www.agro-montpellier.fr/cnam-lr/statnet/

[J. Vérnois] Cours d'informatique et statistique I, Centre Informatique pour les Lettres et Sciences Humaines, Université Provence, Aix-en-Provence.

[B. Ycart] Cours de statistique descriptive, Projet SMEL sur http ://www.math-info.univ-paris5.fr/smel/cours/sd/sd.html.