Ecole Nationale Supérieure de Statistique et d'Economie Appliquée

E.N.S.E.A

Division de la Formation Continue

STATISTIQUE DESCRIPTIVE APPLIQUEE

Par KEHO Yaya

Enseignant à l'ENSEA

Mai 2006

1. Généralités statistiques–Organisation des données

1.1 Définition et objet de la statistique

La majorité des sciences, qu'il s'agisse des sciences expérimentales ou des sciences humaines, font appel à des données souvent nombreuses (issues par exemple de sondages), qu'il convient de traiter à l'aide d'une méthodologie appropriée. *La statistique est un ensemble de méthodes d'analyse numérique des ensembles comportant un grand nombre de faits*. La statistique descriptive vise à la description quantitative des ensembles d'observations ou de données. Elle procède à la description en synthétisant les données à l'aide d'indicateurs. De ce point de vue, elle permet de communiquer efficacement. La Statistique en tant que science se distingue des statistiques qui désignent des ensembles de données chiffrées sur un sujet précis. On peut par exemple étudier l'évolution des statistiques douanière en Côte d'Ivoire.

La statistique utilise des méthodes mathématiques et l'outil informatique pour sa mise en œuvre.

1.2 Quelques domaines d'application

La statistique en tant que méthode de collecte et d'analyse de données peut s'appliquer à une multitudes de domaines, notamment:

- La médecine : pour analyser les données cliniques relevées sur des patients et faire des pronostics de survie.
- La démographie: pour étudier les natalités, les mortalités, la scolarisation, les migrations, etc.
- Les assurances: pour prévoir le nombre d'accidents ou d'incendies qui se produisent dans une région pendant une période donnée. Cela en vue de déterminer la tarification optimale à appliquer.
- Les Etudes de marchés: pour savoir , par exemple pour le lancement d'un produit, quel type d'emballage et à quel prix le produit doit être vendu.
- L'économie: pour établir des relations entre faits en vue de prévoir les valeurs futures ou quantifier l'effet d'un choc sur une variable donnée (taux de chômage, taux d'inflation, taux de croissance…).

1.3 Concepts de base

- **Unité statistique et population**

Les ensembles étudiés par la statistique portent le nom général *d'univers statistiques* ou de *population*. Ainsi une population statistique peut être un ensemble de personnes humaines, d'objets ou de concepts (ensemble de matières par exemple). Par exemple, si on s'intéresse à l'infection par une maladie dans une région donnée, la population est l'ensemble des personnes vivant dans cette région. D'une façon générale, la population représente l'ensemble concerné par une étude statistique;

c'est aussi le champ de l'étude. Les éléments de la population sont appelés *unités* ou *individus statistiques*. Dans l'exemple précédent, l'entreprise est l'unité statistique.

- **Caractères/variables et modalités**

On appelle caractère ou variable ce à quoi on s'intéresse lors d'une étude statistique: c'est le thème commun à tous les individus. Les individus de la population peuvent être caractérisés par plusieurs caractères. Par exemple, en considérant la population d'une région, on peut étudier la morbidité de la maladie, les pratiques d'hyène, la fréquentation des centres de santé, l'accès aux médicaments, etc. Les caractères étudiés peuvent être constitués de plusieurs modalités qui représentent les diverses situations dans lesquelles un individu peut se trouver à l'égard du caractère étudié. Chaque individu doit présenter une seule modalité du caractère considéré.

- **Caractères qualitatifs et quantitatifs**

On distingue deux types de caractères selon la nature des modalités: les variables qualitatives et les variables quantitatives.

Les caractères quantitatifs sont des caractères dont les différentes modalités sont mesurables (quantifiables) et correspondent à des nombres. Par exemple, le poids, la taille et l'age d'une personne sont des variables quantitatives. Les caractères quantitatifs sont constitués de caractères quantitatifs discrets et continus. Les valeurs possibles d'une variable discrète sont des nombres isolés. Nous pouvons citer comme variables discrètes le nombre d'enfants d'une femme, le nombre, l'âge en années révolues.

Les valeurs possibles prises par une variable continue sont *a priori* en nombre infini et quelconques dans un intervalle de valeurs. Nous pouvons citer l'âge exact et la taille d'un individu.

La distinction entre variable continue et variable discrète est évidente en théorie, mais pose des problèmes en pratique du fait de la précision des instruments de mesure et aussi pour des raisons de commodité. Ainsi, bien que la taille soit continue, on mesure des tailles en nombres isolés : 178 cm, 189 cm, etc. On définit généralement les variables continues en classes ou tranches de valeurs possibles (pouvant avoir une amplitude constante ou variable) constituant les modalités de la variable.

Les caractères qualitatifs sont des caractères dont les modalités ne sont pas mesurables. Nous pouvons citer comme caractères qualitatifs le sexe, la nationalité, la profession, la catégorie socioprofessionnelle, etc. des patients. Pour le traitement informatique des données recueillies, on affecte généralement des codes (préétablis) aux différentes modalités des caractères qualitatifs. Par exemple pour le sexe, on affectera le code 1 à la modalité masculin et le code 2 à la modalité féminin. *Ces codes ne rendent pas le caractère quantitatif, ils servent uniquement à repérer des catégories d'individus. Ils sont arbitraires.*

1.4 Organisation des données en tableaux statistiques

En pratique, les données, ainsi que les résultats numériques relatifs à leur traitement statistiques, sont présentés sous forme de tableaux. Un tableau statistique est un tableau à deux colonnes qui précise pour, dans la première colonne, la liste des modalités, et dans la seconde colonne, le nombre d'individus correspondants. En pratique c'est à partir de ce tableau (appelé aussi tableau de dénombrement) que sont présentés tous les calculs relatifs au traitement de la variable.

Tableau 1 : Forme générale d'un tableau statistique

Modalité	Effectif
X1	n1
X2	n2
...	...
Xk	nk

Les *nj,* appelés effectifs ou fréquences absolues, représentent le nombre d'individus statistiques présentant la modalité *Xj* du caractère *X*. Un tableau statistique doit présenter un certain nombre de renseignements parmi lesquels :

- le titre qui indique l'objet du tableau ;
- l'unité utilisée ;
- les titres de lignes et de colonnes et qui en précisent le contenu et qui doivent être précis et concis ;
- les notes (de tableau) qui éclairent le lecteur et qui expliquent mieux certains contenus du tableau.
- la source est généralement indiquée au bas du tableau et permet de vérifier la fiabilité des données.

Fréquence relative

La fréquence relative de la modalité *Xj* représente la proportion $f_j = n_j / n$ des individus ayant pris la modalité *Xj* du caractère *X*. Les fréquences relatives vérifient la relation $\sum_{j=1,\ldots,k} f_j = 1$.

a) Caractère qualitatif

Lorsque le caractère est qualitatif, le tableau statistique se présente sous la forme générale ci-dessus. Considérons le caractère sexe étudié sur une population de 20 personnes. Le tableau statistique se présente sous la forme suivante :

Tableau 2 : Répartition des patients suivant le sexe

Sexe	Effectif
Masculin	16
Féminin	4
TOTAL	**20**

b) Caractère quantitatif discret

Etudions le nombre d'enfants dont dispose chaque malade interné dans un hôpital. C'est une variable discrète. Le tableau statistique correspondant aux données recueillies est le suivant :

Tableau 3 : Répartition des malades suivant le nombre d'enfants

Nombre d'enfants	Effectif
1	12
2	6
3	2
TOTAL	**20**

c) Caractère quantitatif

Les modalités du caractère sont des classes de valeurs possibles définies par les extrémités des classes. Considérons la variable « Age des patients ».

Tableau 4 : Répartition des patients suivant l'AGE

Classe d'âge	Effectif
21- 24	3
24-29	9
29-35	5

35-40	3
TOTAL	**20**

1.5 Représentations graphiques

Bien que les tableaux statistiques présentent toute l'information collectée, il est souvent utile de présenter les résultats sous forme graphique pour en réaliser une synthèse visuelle. Nous présentons les représentations graphiques usuelles suivant la nature du caractère.

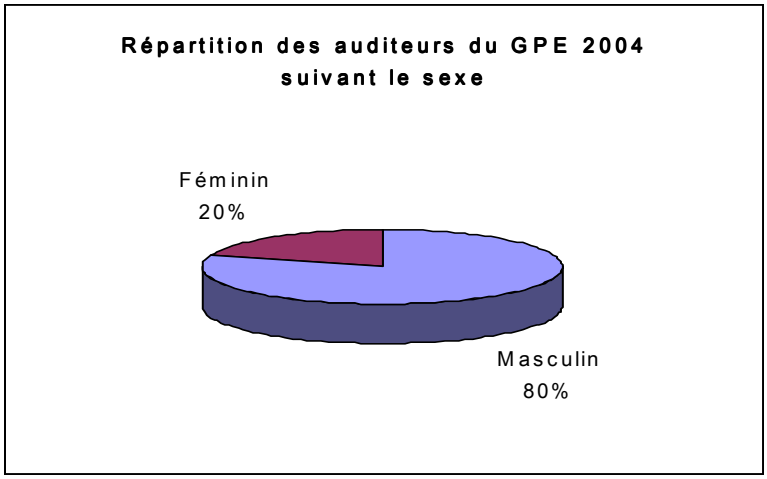
a) Caractère qualitatif

Deux types de graphiques sont généralement utilisés : les diagrammes circulaires et les diagrammes à bandes.

Tableau 5 : Répartition des patients suivant le sexe

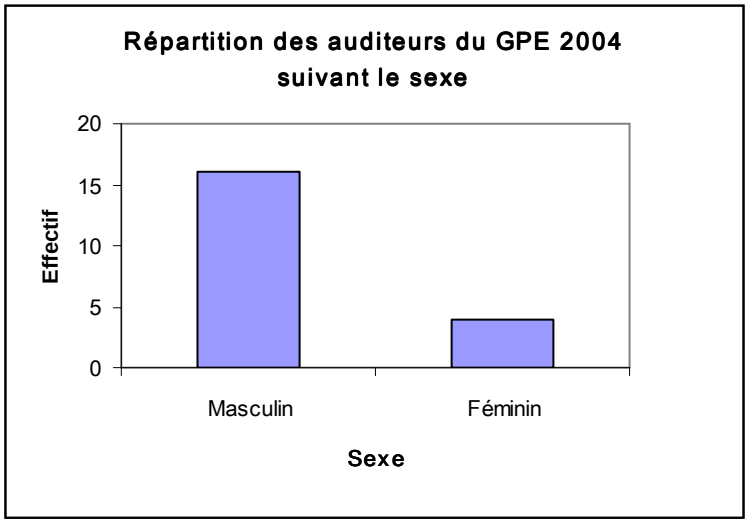
Sexe	**Effectif**	**%**
Masculin	16	80
Féminin	4	20
TOTAL	**20**	**100**

- *Diagramme circulaire*



Chaque secteur correspond à une modalité, l'angle au centre (en degré) étant proportionnelle aux fréquences relatives fj ou aux effectifs.

- *Diagramme à bandes*

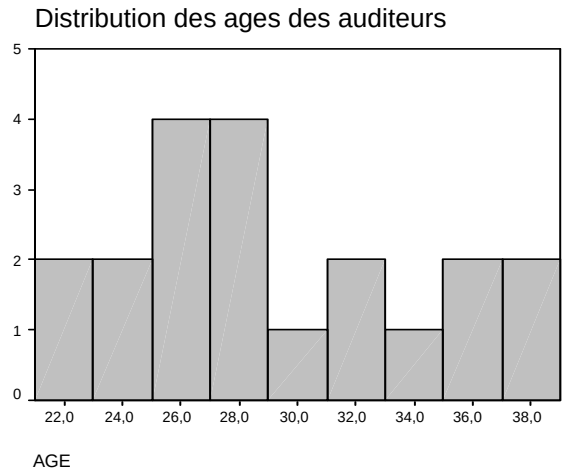


Chaque bande a une base constante et une hauteur proportionnelle à la fréquence fj ou à l'effectif nj.

b) Caractère quantitatif

- *Histogramme*

Pour représenter graphiquement un caractère quantitatif continue, il faut organiser les données sous forme de classes (surtout si le nombre d'observation est très grand). Une fois les classes déterminées, on utilise un diagramme appelé histogramme pour visualiser la répartition des effectifs au sein des classes. L'histogramme est composé d'un ensemble de rectangles contigus d'aires proportionnelles aux effectifs (ou aux fréquences) des classes et de bases déterminées par les extrémités des classes. Toutefois, en pratique le principe de sa construction diffère selon que les classes sont d'amplitudes égales ou inégales. Dans tous les cas il faut respecter le principe de proportionnalité des aires.



Lorsque les amplitudes des classes sont différentes, il est nécessaire de corriger les effectifs (ou les fréquences) pour tenir compte des différences d'amplitudes. On définit ainsi la notion de densité comme étant le nombre d'individu par unité d'amplitude. Dans ce cas, l'histogramme est construit en portant en abscisse les extrémités des classes qui servent de bases à des rectangles dont la hauteur est définie en ordonnée par les densités d'effectifs (ou de fréquences).

Le calcul des densités permet de comparer objectivement des effectifs ou des fréquences pour des classes d'amplitudes inégales. Une comparaison sans une telle précaution peut conduire à des erreurs d'interprétation. Par exemple, peut-on considérer que la ville d'Abidjan est plus peuplée que la ville de Gagnoa relativement à leur superficie et à leur nombre d'habitants?

- *Polygone de fréquences (ou d'effectifs)*

Le polygone de fréquences (ou d'effectifs) est une courbe polygonale obtenue en joignant les milieux des segments supérieurs de chaque rectangle de l'histogramme. Par principe, l'aire du polygone doit être égale à l'aire de l'histogramme (principe de conservation des aires).

Le polygone de fréquences permet de résumer l'allure de l'histogramme.

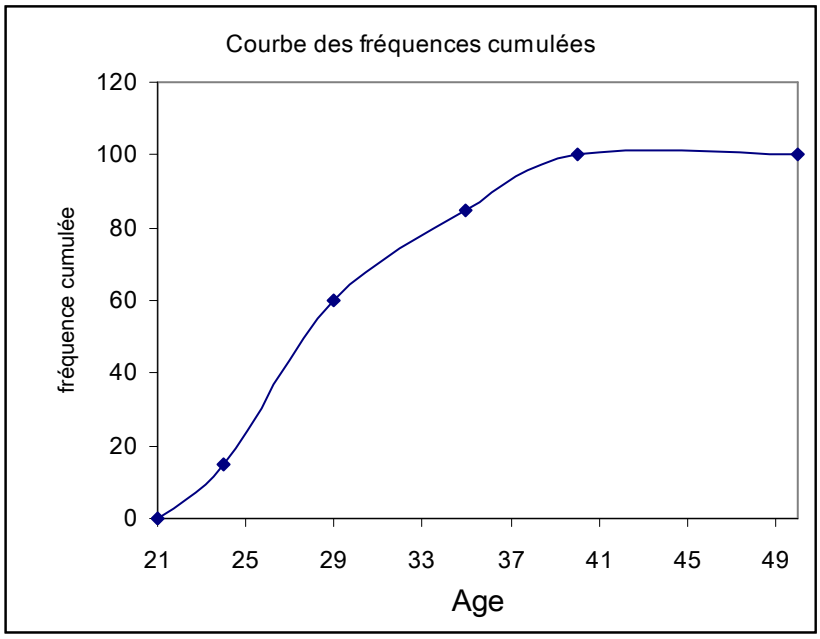
- *Courbe des fréquences cumulées*

On appelle fonction cumulative F(x) au point x du caractère X, la proportion des individus de la population dont la valeur du caractère est inférieur à x. On l'appelle aussi courbe des fréquences cumulées. Pour un caractère discret, cette fonction est constante dans l'intervalle séparant deux valeurs possibles consécutives. Pour un caractère continu, la courbe des fréquences cumulées croissante est obtenue en joignant dans un système d'axes orthogonaux, en abscisse, les extrémités des classes et, en ordonnées, les fréquences cumulées correspondantes.

Tableau 6 : Répartition des auditeurs suivant l'AGE

Classe d'âge	Effectif	Fréquence (en %)	Fréquence cumulée(en %)
21- 24	3	15	15
24-29	9	45	60
29-35	5	25	85
35-40	3	15	100
TOTAL	**20**	**100**	

Ce tableau indique que 3 auditeurs sur les 20 ont moins de 24 ans, ce qui représente 15 % des auditeurs. Exactement 60% de l'effectif est âgé de moins de 29 ans.



2. Caractéristiques numériques d'un caractère quantitatif

Le traitement statistique d'une variable ne doit pas se limiter aux seules représentations graphiques dont l'interprétation est parfois subjective. Il est nécessaire de résumer les informations par des indicateurs numériques appropriés. L'objet de cette partie est de présenter quelques-uns de ces indicateurs. Les plus utilisés sont le mode, la médiane, les quartiles, la moyenne, l'écart-type et le coefficient de variation.

2.1 Caractéristiques de tendance centrale

Les caractéristiques de tendance centrale , encore appelées caractéristiques de position, servent à synthétiser l'information au moyen d'un petit nombre de valeurs numériques autour desquelles se répartissent les observations d'un caractère. Elles en fournissent un ordre de grandeur.

a) Le mode (Mo)

Le mode Mo d'une distribution statistique représente la valeur de la variable pour laquelle la fréquence relative est la plus élevée ; c'est la valeur qu'on rencontre le plus souvent dans la série.

Lorsque la variable est discrète, le mode est défini avec précision. Quand deux valeurs consécutives ont la fréquence la plus élevée, on parle d'intervalle modal. La distribution est dite bi-modale lorsque la série présente deux modes non consécutifs.

Dans le cas d'une variable continue, la classe modale est définie comme celle ayant la plus grande densité (fréquence par unité d'amplitude). On peut aussi représenter le mode par le centre de la classe modale.

b) La médiane (Me)

La médiane est la valeur du caractère qui divise la population en deux parties d'effectifs égaux égales. Cela signifie que la moitié des individus observés prennent des valeurs inférieures ou égales à la médiane. Contrairement à la moyenne, la médiane est moins sensible aux valeurs extrêmes et aberrantes. Elle ne change pas lorsqu'on modifie les valeurs extrêmes, pourvu que cette modification ne change pas le rang de l'individu médian. Cette propriété sympathique rend la médiane très utile pour l'étude de certaines distributions (distributions de revenus par exemple).
Cependant, la médiane n'existe pas toujours pour une variable discrète. On peut être amené dans ce cas à accepter pour médiane un nombre qui n'est pas une valeur du caractère.

Exemple: Considérons les notes obtenues par 7 élèves d'une classe : 4, 6, 5, 11, 8, 9, 10. pour déterminer la médiane on réordonne la série par ordre croissant (ou décroissant):4, 5, 6, 8,9,10,11. la médiane est donc égale à 8.

Exemple : Considérons les notes des 8 élèves d'une classe : 3, 4, 5, 6, 7, 8, 9, 10. L'intervalle médian est [6, 7] ou le centre médian est Me = 6,5.

Dans le cas d'un caractère continu…

Lorsqu'on étudie un caractère quantitatif continu mis sous forme de classes, la médiane peut être déterminée à partir de la courbe des fréquences cumulées. En effet Me est la valeur x du caractère solution de l'équation $F(x)=1/2$. Cette solution est unique et appartient à une classe appelée classe médiane (ou intervalle médian).

On peut aussi déterminer la médiane par interpolation linéaire suivant la formule :

$$Me=a+(b-a)\frac{0,5-F(a)}{F(b)-F(a)}$$

où [a,b] est l'intervalle médian.

Cette formule repose sur l'hypothèse d'une répartition uniforme des observations à l'intérieur des classes, et notamment de la classe médiane.

Exemple: Reprenons les données sur la distribution des âges des auditeurs. La classe médiane est [24,29[. En appliquant la formule précédente, on trouve $Me=24+(29-24)\frac{0,5-0,15}{0,60-0,15}=27,89$.

c) Les quartiles

La médiane appartient à la famille des quartiles qui peuvent être utilisés comme des caractéristiques de dispersion. On définit pour une distribution trois quartiles notés q1, q2 et q3. q1 est tels que 25 % des observations sont inférieures à q1, 50% des individus ont leur valeur inférieure à q2 et 75 % ont leur valeur inférieure à q3. On remarque ainsi que la médiane correspond au deuxième quartile.

Outre les quantiles, on définit également les déciles et les centiles d'une distribution. Les déciles sont les 9 valeurs qui partagent la population (ordonnée par valeurs croissantes ou décroissantes) en 10 parties contenant chacune 10% de l'effectif. Les 99 centiles partagent la population en 100 parties contenant chacune 1% de l'effectif. A ces 99 centiles sont associées les fréquences cumulées 1%, 2%,…, 98%, 99%.

Lorsque les données ont été groupées en classes, la détermination des quartiles, déciles et centiles se fait par interpolation linéaire ou à partir de la courbe des fréquences cumulées.

d) La moyenne

La moyenne représente la valeur qu'aurait chaque individu si la variable était distribuée équitablement dans la population. Cet indicateur renseigne sur le niveau moyen du caractère dans la population. La moyenne se calcule par la formule:

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{x_i+x_2+\ldots+x_n}{n}$$

En prenant en compte le fait que plusieurs individus peuvent présenter une même modalité du caractère, on peut définir la moyenne comme la somme des modalités pondérée par les fréquences relatives simples (part des individus de la population présentant une modalité donnée):

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} n_j x_j = \sum_{j=1}^{k} f_j x_j$$

Lorsque les valeurs du caractère (quantitatif) sont regroupées en classes, il convient de prendre comme valeurs des individus d'une classe le centre de classe. En effet, dans cas on suppose les individus uniformément répartis dans une classe.

Moyenne d'un mélange de populations

Supposons que la population est formée de m sous-populations (ces sous-populations forment une partition de la population). Soit P_l la sous-population l , n_l l'effectif et $\bar{x_l}$ la moyenne du caractère X dans la sous-population P_l. La moyenne de la population totale P est donnée par :

$$\bar{x} = \frac{1}{n}\sum_{l=1}^{m} n_l \bar{x_l} = \sum_{l=1}^{m} p_l \bar{x}_l$$

La moyenne arithmétique d'un mélange de populations est égale à la somme des moyennes arithmétiques observées dans chaque sous-population pondérée par la proportion des individus.

2.2 Caractéristiques de dispersion

Considérons la distribution des salaires des ouvriers pour deux entreprises:

Tableau 7: Distribution des salaires dans deux entreprises (x100 FCFA)

Entreprise A	528	540	550	560	572
Entreprise B	450	544	550	556	650

Peut-on considérer que les deux entreprises accordent des salaires comparables?

Les deux séries ainsi définies ont la même moyenne, la même médiane. Elles correspondent cependant à des observations qui se distribuent très différemment: dans l'entreprise B, les salaires s'écartent notoirement de la valeur 550, ce qui n'est pas le cas dans l'entreprise A. Réduire par une même caractéristique de tendance centrale ces deux séries aurait pour conséquence de masquer leur différence.

L'objet de ce paragraphe est de présenter quelques indicateurs permettant d'évaluer la dispersion des valeurs d'un caractère autour d'une valeur centrale.

a) L'étendue (range)

L’étendue représente la différence entre la plus grande et la plus petite valeur observées.

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

L'étendue donne la plage de variation des valeurs du caractère; elle est peu utilisée à cause de sa très grande sensibilité aux valeurs extrêmes et aberrantes.

b) Les écarts interquantiles

On appelle intervalle interquartile, l'intervalle [q1;q3], l'amplitude de cette intervalle est q3-q1 et mesure l' écart interquantile. Cet intervalle contient 50% des observations. On peut donc la comparer à la moitié de l'étendue pour apprécier la dispersion des observations dans la population. En effet si la distribution est uniforme, la moitié de l'entendue de la distribution contiendrait 50% des observations. Ainsi si l'intervalle interquartile est plus petit que la moitié de l'étendue, c'est qu'il y a accumulation des valeurs dans l'intervalle interquartile (un petit intervalle) et donc la dispersion est forte. On peut le même raisonnement avec l'écart interdécile et l'écart intercentile.

c) L'écart type

L'écart type permet d'apprécier la dispersion des observations d'un caractère autour de sa moyenne; c'est un indicateur d'hétérogénéité. L'écart type (noté σ) se calcule comme la racine carrée de la variance:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

Comme on peut le remarquer, l'écart type s'exprime avec la même unité de mesure que les valeurs observées. Il est d'autant plus grand que la dispersion des observations autour de la moyenne est importante. Lorsque les observations du caractère sont *normalement* distribuées, l'intervalle centré sur la moyenne et de longueur deux fois l'écart type, $\left[\bar{x}-\sigma;x+\sigma\right]$, contient 68% des observations. Si cet intervalle contient plus de 68% des observations, la dispersion peut être considérée comme faible. Elle sera forte dans le cas contraire (règle empirique).

d) Le coefficient de variation

Le coefficient de variation des observations se calcule comme le rapport de l'écart type à la moyenne:

$$cv = \frac{\sigma}{\bar{x}}$$

Les autres indicateurs de dispersion s'expriment avec la même unité de mesure que le caractère. De plus ils dépendent de l'ordre de grandeur des observations. De ce fait, ils ne sont pas adéquats pour comparer la dispersion de deux distributions qui ne s'expriment pas dans la même unité ou ayant des ordres de grandeurs différents.

Le coefficient de variation supprime ce défaut. C'est un indicateur de dispersion sans unité. Cette propriété fait du coefficient de variation un indicateur utile pour comparer les distributions. De deux distributions, la plus homogène sera celle qui a le coefficient de variation le plus faible. On peut ainsi comparer la dispersion des salaires entre les employés de deux entreprises, ou entre les hommes et les femmes d'une même entreprise.

Exemple: On considère les données du tableau 7 sur la distribution des salaires dans deux entreprises. Le traitement des données donne le tableau suivant:

Tableau 8: Caractéristiques sur la distribution des salaires dans deux entreprises

Caractéristiques	Entreprise A	Entreprise B
Moyenne	550	550
Ecart type	17,08	70,84
Coefficient de variation (en %)	3,10%	12,88%

Les deux entreprises ont le même salaire moyen. Mais l'écart type de la distribution des salaires est plus élevé dans l'entreprise B. Le coefficient de variation indique que les salaires sont plus proches de leur moyenne dans l'entreprise A que dans l'entreprise B.

3. Description bidimensionnelle et mesures de liaison entre deux caractères

3.1 L'intérêt de l'analyse bidimensionnelle

La description unidimensionnelle faite jusque là, si elle opère un premier traitement des données, reste toutefois insuffisante car elle laisse de côté les relations qui peuvent exister entre les variables, et qui constituent un aspect aussi important de la description statistique.
L'analyse bidimensionnelle (ou bivariée) étude donc les liaisons entre les variables observées: c'est ce que l'on appelle communément l'étude des corrélations. Les méthodes et les indices de liaison varient selon la nature des variables étudiées.

3.2 Liaison entre deux caractères qualitatifs: le test d'indépendance du Khi-deux

a) Tableau croisé

On considère un tableau croisé (encore appelé tableau de contingence ou à double entrée) ventilant un échantillon *aléatoire* de taille n selon deux critères qualitatifs x et y ayant respectivement l et c modalités.

Exemple: On considère la répartition d'un échantillon de 120 travailleurs suivant le sexe et la catégorie socioprofessionnelle.

Tableau 9: Répartition des individus suivant le sexe et la CSP

	SEXE		
CSP	Masculin	Féminin	**TOTAL**
Cadre	14	6	**20**
Employé	25	45	**70**
Ouvrier	52	8	**60**
TOTAL	**91**	**59**	**150**

Un tableau croisé exprime les effectifs conjoints des modalités des deux caractères. Ici, on observe 14 cadres de sexe masculin. Par la suite ces effectifs sont notés n_{ij} .

Existe-t-il une relation entre le sexe et la catégorie socioprofessionnelle des travailleurs?Si oui, quelle quelles sont les affinités de modalités qu'on observe?

b) Mesures de liaison

Les deux caractères sont indépendants signifie que la connaissance de l'un n'apporte rien à la connaissance de l'autre (i.e. ne change pas les distributions conditionnelles de l'autre) . Dans ce cas, la proportion de la population totale qui possèdent la modalité *j* est identique à celle des individus possédant la modalité *j* dans la sous-population des individus possédant la modalité *i*. En d'autres termes, lorsqu'il y a indépendance, les *l* sous-populations caractérisées par les modalités de la variable *x* se répartissent selon les *c* modalités de la variable *y* avec les mêmes pourcentages. Il s'ensuit que toutes les lignes ou colonnes sont proportionnelles. La réciproque est vraie : lorsque toutes les lignes (resp. colonnes) sont proportionnelles, les deux caractères sont indépendants.

L'indépendance (empirique) se traduit donc par $\frac{n_{ij}}{n_{i.}}=\frac{n_{.j}}{n}$ ou encore $n_{ij}=\frac{n_{i.}n_{.j}}{n}$.

On définit ainsi les quantités $T_{ij}=\frac{n_{i.}n_{.j}}{n}$, qui représentent les effectifs théoriques des cases (*i,j*), c'est-à-dire les effectifs qu'on aurait observé s'il y avait indépendance entre les deux caractères. L'indépendance se traduit par l'égalité entre les effectifs observés et les effectifs théoriques.

En pratique, on observe le plus souvent un écart entre ces effectifs, dû aux fluctuations d'échantillonnage; cet écart exprime l'écart à l'indépendance .

La statistique du χ^2 permet d'évaluer la distance entre ces deux tableaux d'effectifs. Elle est définie par :

$$d^2=\sum_{i,j}\frac{\left(n_{ij}-\frac{n_{i.}n_{j}}{n}\right)^2}{\frac{n_{i.}n_{j}}{n}}=\sum_{i,j}\frac{\left(O_{ij}-T_{ij}\right)^2}{T_{ij}}.$$

En cas d'indépendance, cette quantité devrait être voisine de zéro.

Diverses mesures de liaison liées au d^2 ont été proposées pour obtenir une mesure comprise entre 0 (indépendance) et 1 (liaison fonctionnelle). Citons:

- le coefficient de contingences dit Φ^2 Pearson $\Phi^2=d^2/n$
- le coefficient de contingences dit *T* de Tschuprow $T=\sqrt{\frac{\Phi^2}{(l-1)(c-1)}}$
- le coefficient de contingences dit V de Cramer $V=\sqrt{\frac{\Phi^2}{\inf\{(l-1);(c-1)\}}}$
- le coefficient de contingences dit C de Cramer $C=\sqrt{\frac{d^2}{n+d^2}}$

b) Test d'indépendance du Khi-deux

A partir de quelle valeur "critique" peut-on considérer que la liaison est significative? Le test d'indépendance consiste en fait à apprécier le caractère significatif de l'écart entre le tableaux des effectifs théoriques (tableau *T*) et le tableau des effectifs observés (tableau *O*), donc du d^2. L'idée du test est la suivante: si l'échantillon était prélevé dans une population où *x* et *y* sont indépendantes quelles seraient les valeurs probables de d^2? On montre qu'alors d^2 est une réalisation d'une variable aléatoire suivant approximativement une loi du χ^2 à *v=(l -1)(c-1)* degrés de liberté. Il suffit alors de se fixer un risque d'erreur α, c'est-à-dire une valeur qui, s'il y avait indépendance, n'aurait qu'une probabilité faible d'être dépassée. $1-\alpha$ représente le niveau de confiance du test. On prend usuellement $\alpha=5\%$ ou *1%*.

La valeur critique χ^2_v est alors la valeur qu'un χ^2 à *v=(l -1)(c-1)* degrés de liberté a une probabilité α de dépasser. On rejettera donc l'hypothèse d'indépendance si la valeur calculée de d^2 est supérieure à χ^2_v. Dans ce cas, il existe une différence significative entre les distributions d'effectifs et les écarts constatés entre les effectifs empiriques et effectifs théoriques sont trop grands pour être attribués au hasard (fluctuations d'échantillonnage).

Remarques:

1) Le test du Khi-deux est un test asymptotique qui n'est valable que si les effectifs théoriques sont « grands ». En pratique, il faut que ces effectifs dépassent 5. Si ce n'est pas le cas, il faut procéder à des regroupements.
2) Certains logiciels (comme Excel, SPSS, EPI-Info) donnent le niveau de probabilité critique, appelé *p-value*. C'est la probabilité de se tromper en rejetant l'hypothèse d'indépendance.

C'est également le niveau maximal pour lequel le test ne peut pas rejeter l'hypothèse d'indépendance. Les *p-values* sont directement interprétables et permettent d'effectuer plus facilement les tests sans recourir aux valeurs critiques théoriques: on rejette l'hypothèse d'indépendance lorsque la *p-value* est inférieure au niveau du risque d'erreur choisi.

Exemple: On considère les données du tableau 9. Sur cet exemple, le degré de liberté du χ_2 est (3-1)(2-1)=2. La valeur de d^2 est 35,992; la valeur critique à 5% d'un χ_2^2 est 5,99. On est donc amené à rejeter l'hypothèse d'indépendance entre le sexe et la catégorie socioprofessionnelle (au risque de 5%). Autrement dit, on peut conclure avec 5% de chances de se tromper qu'il existe une liaison significative entre le sexe et la catégorie socioprofessionnelle.

Sur cet exemple, *p-value*=1,53 10^{-8} ; ce qui signifie qu'on a moins de 0,01 % de chances de faire erreur en rejetant l'hypothèse d'indépendance. Si on fait le test à 5% ou à 1%, on conclut à l'absence d'indépendance entre le sexe et la catégorie socioprofessionnelle.

Calcul des contributions

Si l'on conclut à une liaison entre deux caractères, il faut expliquer ce qui fait la liaison, c'est-à-dire quelles modalités de *x* et *y* sont plus souvent/rarement associées. A cet égard, la construction du tableau des effectifs théoriques T_{ij} (tableau *T*) et sa comparaison avec le tableau des effectifs observés O_{ij} (tableau *O*) est en général instructive: en particulier le calcul des termes $\frac{(O_{ij}-T_{ij})^2}{T_{ij}}.\frac{1}{d^2}$ permet d'évaluer les contributions de chaque couple de modalités (*i,j*) au χ^2 et donc de mettre en évidence les associations significatives entre catégories des deux caractères. Le signe de la différence $(O_{ij}-T_{ij})$ indiquera alors s'il y a association positive ou négative entre les catégories *i* de *x* et *j* de *y*. Un tel calcul devrait être systématiquement associé à chaque test d'indépendance du χ^2 .

On remarque ainsi une contribution positive de 30,78% d'une seule case: employé et féminin, ce qui indique une sur-représentation de ce sexe dans cette catégorie socioprofessionnelle. L'association négative Ouvrier-Féminin contribue à 28,65 % au χ^2 : on rencontre rarement les femmes Ouvrières. Enfin, l'association Ouvrier-Cadre contribue également à la liaison (environ 20%) entre le sexe et la catégorie socioprofessionnelle.

3.3 Liaison entre deux caractères quantitatifs

a) Le coefficient de corrélation linéaire

La liaison (linéaire) entre deux caractères quantitatifs *x* et *y* est mesurée par le coefficient de corrélation de "Bravais-Pearson". Ce coefficient est définit par :

$r = \text{cov}(x,y)/\sigma_x\sigma_y$ où $\text{cov}(x,y) = \frac{1}{n}\sum_{i=1\ldots n}(x_i - \bar{x})(y_i - \bar{y})$.

Notez bien que r mesure le caractère plus ou moins linéaire de la liaison entre deux caractères quantitatifs.

r varie entre –1 et 1. Une valeur proche de 1 signifie que les deux variables sont corrélées positivement entre elles. Cela signifie qu'elles évoluent dans le même sens. Au contraire, un coefficient proche de –1 traduit une corrélation négative et les deux variables évoluent en sens opposés. Une corrélation nulle traduit une indépendance linéaire entre les deux variables et non une indépendance fonctionnelle.

Afin d'examiner la nature de la liaison, on représente les observations (x_i, y_i) dans un système d'axes orthogonaux. On obtient ainsi un nuage de points dont la forme est suggestive de la nature de la relation entre les deux caractères.

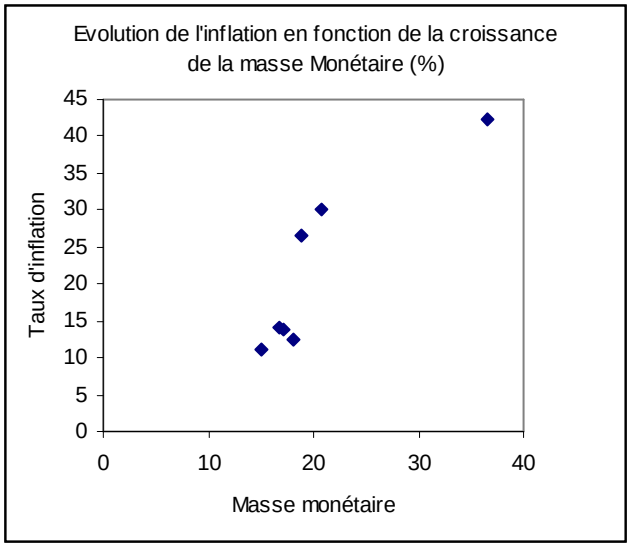
Exemple: Le tableau suivant donne le taux d'inflation et le taux de croissance de la masse monétaire d'un ensemble de pays africains.

Tableau 10: Inflation et masse monétaire

Année	Inflat	Masse Monétaire (%)
1994	42,2	36,5
1995	30	20,7
1996	26,5	18,9
1997	14,2	16,7
1998	11	15
1999	12,4	18,1
2000	13,7	17,2

Source:Statistiques de la Bad (2002), volume 4.

Le nuage de points associé à ce tableau est:



Ce graphique indique qu'il existe bien une corrélation positive entre l'inflation et la masse monétaire. La valeur du coefficient de corrélation est ici de *r=0,8871* .

b) Ajustement linéaire et prévision

Quel est l'impact de la masse monétaire sur l'inflation? De combien augmentent l'inflation si la masse monétaire augmente de 15%? Ce sont là des questions intéressantes que l'on se pose très souvent.

L'analyse de la corrélation ne permet pas de répondre à ce genre de questions. En effet, le coefficient de corrélation ne renseigne que sur **l'intensité de la liaison (**linéaire) entre deux variables, il ne permet pas de mesurer quantitativement l'effet d'une variable sur l'autre.

En fait, pour répondre aux questions précédentes, il faut expliciter puis estimer la relation mathématique liant l'inflation à la masse monétaire: c'est l'objet de la régression. La forme de la relation peut être linéaire, exponentielle ou quadratique. En général, elle est suggérée par le nuage de points obtenu en croisant la masse monétaire et l'inflation dans un système d'axes orthogonaux.

Le graphique précédent suggère un ajustement linéaire entre la masse monétaire et le taux d'inflation. Cette relation s'écrit $y=a+bx$ où x représente le taux de croissance de la masse monétaire et y le taux d'inflation. Dans cette relation, y (le taux d'inflation) est la variable expliquée et x (le taux de croissance de la masse monétaire) est la variable qui permet d'expliquer le niveau de l'inflation. Dans le contexte de la régression on l'appelle variable explicative.

L'objectif de la régression est de fournir une estimation des coefficients a et b à partir d'un échantillon d'observations. La méthode de régression généralement utilisée à cet effet est celle des *Moindres Carrés Ordinaires (MCO)*. Le principe de cette méthode est d'ajuster le nuage de points par une droite qui passe le plus près possible de tous les points du nuage. Il s'agit de rendre minimale la distance de la droite d'ajustement aux points.

Les estimations des coefficients sont données par :

$$\hat{b}=\frac{\sum_{t=1}^{T}(x_t-\bar{x})(y_t-\bar{y})}{\sum_{t}(x_t-\bar{x})^2} \text{ et } \hat{a}=\bar{y}-\hat{b}\bar{x}\,.$$

Exemple: Reprenons l'exemple de l'inflation et la masse monétaire. L'équation de la droite d'ajustement est *y = 1,4314x - 7,8341,* avec un coefficient de détermination de *R2 = 0,787,* soit *78,7%.*

Le coefficient de détermination R^2 mesure le **pouvoir explicatif** de l'ajustement. Ici, ce coefficient est proche de 1, ceci traduit un très bon ajustement. En effet, *78,7%* des variations du taux d'inflation est expliquée par la masse monétaire. On peut utiliser cette équation pour faire des prévisions à court terme. En effet, si la masse monétaire croit à un taux de 13,9%, l'inflation devrait se maintenir en baisse en 2001, et le taux d'inflation devrait se situer à 12%.

3.3 Liaison entre une variable quantitative et une variable qualitative

On cherche par exemple à étudier la liaison entre le salaire (*y*) et la catégorie socioprofessionnelle (*x*) d'un échantillon d'individus. Soit *k* le nombre de modalités du caractère qualitatif *x*. On note $n_1,\ n_2,...,n_k$ les effectifs observés correspondants à ces modalités et $\bar{y}_1,\ \bar{y}_2,...,\bar{y}_k$ les moyennes de *y* pour chaque catégorie (il est indispensable qu'au moins un des n_i soit supérieur à 1) et $\bar{y}$ la moyenne totale. Le caractère *y* sera lié à *x* si pour chaque modalité *j* de *x*, les n_j individus ont tous la même valeur y_j de *y* . Inversement, l'absence de dépendance en moyenne implique l'égalité des moyennes $\bar{y}_1,\ \bar{y}_2,...,\bar{y}_k$ de chaque catégorie.

L'intensité de la liaison entre *x* et *y* est mesurée par le rapport de corrélation η défini par :

$$\eta^2=\frac{\sum_{j=1}^{k} n_j\left(\bar{y}_j-\bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}\,.$$

η varie de 0 (absence de liaison en moyenne) à 1 (dépendance en moyenne). Un test de Fisher peut être effectué pour tester la significativité de η.

LECTURES CONSEILLEES

P. CHAREILLE ET Y. PINAULT: *Statistique descriptive*, Montchrestien, 2e édition, 1998.

G. CALOT *: Cours de Statistique Descriptive, éditions,* Dunod, Paris, 1973.

M. JAMBU: *Méthodes de base de l'analyse des données,* Dunod, Paris, 1989.

G. SAPORTA: *Probabilité, analyse des données et statistique*, Technip, Paris, 1989

M TENENHAUS: *Méthodes statistiques en gestion,* Dunod, Paris, 1994.