1ère partie : STATISTIQUE DESCRIPTIVE

CHAPITRE 1 : COLLECTE DE L'INFORMATION, TABLEAUX ET GRAPHIQUES.

I. Définition et vocabulaire

Définition : la statistique est une méthode scientifique qui consiste à réunir des données chiffrées sur des ensembles nombreux, puis à analyser, à commenter et à critiquer ces données.

Cette science n'a pas pour objet la connaissance des éléments des ensembles dans ce qui fait leur individualité, mais au contraire dans ce qu'ils ont en commun : il s'agit d'obtenir des résultats globaux. Ainsi, une enquête statistique portant sur des personnes n'a pas besoin de faire intervenir leurs noms, mais seulement les renseignements que l'on désire étudier : elle permet de connaître la répartition de ces personnes par âge, par sexe, groupe sanguin …

Comme toute science, la statistique fait appel à un vocabulaire spécialisé :

- Les ensembles sont appelés *populations*. Comme un ensemble, une population statistique doit être clairement définie.
- Les éléments de la population sont appelés *individus ou unités statistiques*, (que ce soient des hommes ou des automobiles).
- La population est étudiée selon un ou plusieurs *caractères*.
- Un caractère permet de déterminer une partition de la population selon diverses *modalités*. Ainsi le sexe est un caractère à deux modalités : masculin et féminin.
- Lorsque les modalités du caractère sont des nombres, le caractère est dit *quantitatif* ; on lui donne souvent le nom de variable statistique. Une variable statistique peut être *discrète* si elle ne prend que des valeurs isolées ou *continue* si elle peut prendre n'importe quelle valeur intermédiaire entre deux valeurs données.
- Lorsque les modalités du caractère ne sont pas mesurables, le caractère est dit *qualitatif*. Les modalités d'un caractère qualitatif peuvent faire l'objet d'une nomenclature ou énumération ; la nomenclature doit en principe être courte (une dizaine pour une étude statistique simple). Mais les exigences de l'étude sont parfois telles que la nomenclature occupe des volumes entiers : c'est le cas de nomenclatures codées des catégories socioprofessionnelles et des professions faites par l'I.N.S.E.E.

Exemples : Pour des chômeurs, l'âge est un caractère quantitatif continu ; le nombre d'enfants est un caractère quantitatif discret. Pour des automobiles, la couleur est un caractère qualitatif.

II. La collecte des informations

Le premier objet de la méthode statistique est de réunir les informations avant de les traiter.

Voici quelques généralités :

- Objectif de l'information. Enquête **:** Il importe, dès le départ, de bien définir le ou les objectif(s) avant de réaliser l'enquête. Si un élément est oublié dans les premières recherches, il risque d'être long et coûteux de le rechercher ensuite.

Exemple : Si l'on réalise une enquête sur l'emploi dans un secteur déterminé, il ne faut oublier aucune variable. On peut demander aux personnes interrogées leur qualification professionnelle, leur âge, etc… Mais si ensuite on s'aperçoit que le salaire est un caractère important, il est trop tard et il faut refaire l'enquête.

- Quantité d'information : Cependant il ne faut pas être trop ambitieux. Il ne doit pas y avoir de lacunes dans l'information mais il ne doit pas y avoir trop d'informations, car il devient alors impossible d'en tirer une synthèse.

- Collecte de données : Les données sont recueillies soit par observation directe, soit indirectement.
- observation *directe* : enquête menée par les statisticiens à l'aide de questionnaires qui sont ensuite dépouillés.
- Observation *indirecte* statistiques d'une entreprise tirées de sa comptabilité, statistiques de naissances et des décès tirées de l'état civil ….

- Différents modes de collecte de l'information :
    - Les résultats statistiques peuvent être obtenus à partir d'une enquête exhaustive *instantanée* (dénombrement instantané ou recensement) ou d'un relevé *continu* (état civil).
    - De même, l'enquête peut être *exhaustive* ou *partielle*. L'enquête exhaustive porte sur toutes les unités de la population ; elle est utile mais souvent coûteuse. C'est pourquoi on a recours à des enquêtes partielles faites sur un *échantillon* de la population : il s'agit alors de sondage, et il faut déterminer un échantillon représentatif, de manière que les résultats statistiques trouvés sur cet échantillon soient voisins de ceux que l'on aurait obtenus si on avait étudié la population entière.

III. Dépouillement des observations

Lorsque les observations sont obtenues, elles doivent être classées et exploitées. Auparavant une *critique* des réponses doit être faite afin d'éliminer les contradictions et les invraisemblances. Pour chaque caractère à étudier, on définit un certain nombre de classes selon les modalités, puis on fait le tri des observations, c'est à dire une répartition par classes. Ces opérations peuvent être faites à la main ou à l'aide d'un ordinateur. Le document d'enquête doit être au moins partiellement codé pour éviter la surcharge des mémoires.

IV Tableaux statistiques

On peut représenter les données brutes d'une étude dans un tableau. Mais il est possible d'en déduire un tableau plus clair, en faisant un **regroupement par classes**. On choisit les classes pas trop nombreuses, mais suffisamment pour qu'il n'y ait pas de perte d'information. Il

importe que les classes recouvrent tous les résultats et aient une intersection vide, d'où les formulations du type « de … à moins de …» ; la différence entre les deux extrémités est appelé **amplitude de la classe**.

On peut fixer le nombre de classes selon l'un des deux formules suivantes :

i) Règle de Sturge : nb. de classes = 1+ (3.3 log n)

ii) Règle de Yule :

$$\text{nb. de classes} = 2.5\sqrt[4]{n}$$

Avec n = effectif de l'échantillon.

L'amplitude de classe est alors donnée par :

$$\frac{\text{valeur max. - valeur min.}}{\text{nb. de classes}}$$

L'effectif d'une classe est le nombre d'éléments de la population observés dans cette classe.

La fréquence est le rapport de cet effectif à l'effectif total de la population. La fréquence est exprimée en pourcentage.

Exemple 1: On s'intéresse à la charge de rupture d'un fil en grammes.

711	862	851	912	922	791	825	935	895	758	8462
915	873	926	864	800	931	722	774	903	925	8633
853	700	885	857	844	907	917	786	820	930	8499
789	790	753	910	847	784	936	706	758	887	8160
941	909	784	882	859	903	925	704	792	888	8587
890	925	895	768	869	892	895	912	850	920	8816
763	805	796	759	916	853	789	942	712	764	8099
892	893	915	890	888	865	909	931	710	798	8691
914	794	931	701	772	935	887	880	933	905	8652
889	791	782	713	724	868	842	892	905	792	8198
										84797

On va regrouper ces données en classes. Nous avons un effectif de 100 ce qui nous donne en nombre de classes d'après les règles de Sturge et de Yule : 7 classes. En fait dans l'exemple ils en prennent 6.

Charge en grammes	Effectifs	Fréquences
700 à moins de 750	10	0,1
750 à moins de 800	23	0,23
800 à moins de 840	4	0,04
840 à moins de 880	15	0,15
880 à moins de 920	32	0,32
920 et plus	16	0,16
TOTAL	100	1

IV. Graphiques

4.1. Cas de distributions quantitatives

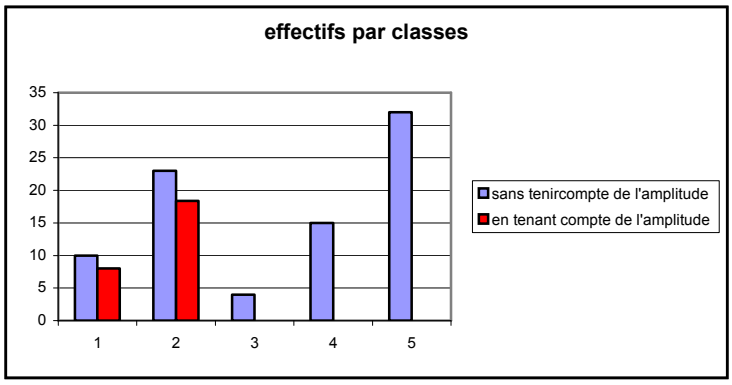
Les graphiques sont normalement réalisés en portant en abscisse la variable observée et en ordonnée l'effectif ou la fréquence.

- dans le cas d'une variable *discrète*, le graphique est un diagramme à bâtons, ainsi apparaît la discontinuité entre deux valeurs.
- dans le cas d'une variable *continue*, le graphique est un histogramme. La surface limitée par l'histogramme doit être proportionnelle à l'effectif ou la fréquence. Il convient de prendre garde à l'amplitude des classes (on se ramène à la plus petite amplitude, amplitude élémentaire, et on divise la hauteur du rectangle par la mesure de l'amplitude de la classe par rapport à cette amplitude élémentaire).

$$\text{hauteur du rectangle} = \frac{\text{effectif ou fréquence} \times \text{amplitude élémentaire}}{\text{amplitude de la classe}}$$

Exemple : On a récolté les données suivantes :

charges en g. (classes)	effectifs	amplitude	hauteur du rectangle
700 à moins de 750	10	50	(10*40)/50=8
750 à moins de 800	23	50	(23*40)/50=18,4
800-840	4	40	4
840-880	15	40	15
880-920	32	40	32



4.2. Cas de distributions qualitatives

Diverses méthodes sont possibles, par exemple :

- on peut réaliser des diagrammes à bandes
- ou des diagrammes à secteurs

CHAPITRE 2 : ETUDE DES SERIES STATISTIQUES SIMPLES.

I. Introduction

Un tableau statistique ou un graphique sont parfois long à consulter, sans permettre d'avoir une idée suffisamment concise de la distribution statistique observée. On cherche alors à résumer celle-ci par une caractéristique de tendance centrale, c'est à dire par un seul nombre destiné à caractériser l'ensemble d'une façon objective et impersonnelle, comme par exemple la moyenne arithmétique, la médiane ou le mode.

II. La moyenne arithmétique

La moyenne arithmétique d'une série de valeurs d'une variable statistique est égale à la somme de ces valeurs divisée par leur nombre.

2.1. Cas de données énumérées

La formule générale est, pour **n** observations $x_1, x_2, x_3, \ldots x_n$:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

2.2. Cas d'une variable discrète

Si la variable est discrète on emploie la formule de la moyenne pondérée. Pour *t* classes d'effectifs n_i ou de fréquences f_i la moyenne *x* s'écrit pour les valeurs $x_1, x_2, x_3, \ldots x_t$ de la variable :

$$\overline{x} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_t x_t}{n_1 + n_2 + \ldots + n_t} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_t x_t}{f_1 + f_2 + \ldots + f_t}$$

On a l'habitude de résumer cette écriture en employant le signe Σ :

$$\overline{x} = \frac{\sum_{i=1}^{t} n_i x_i}{\sum_{i=1}^{t} n_i} = \frac{\sum_{i=1}^{t} f_i x_i}{\sum_{i=1}^{t} f_i}$$

2.3. Variable continue et données groupées.

Si la variable est continue et si les données sont groupées on ne peut que rechercher arbitrairement une moyenne à l'intérieur de chaque classe ; à défaut on choisit le « centre de

classe ». Le calcul est effectué comme si tous les individus d'une classe avaient pour caractère le centre de classe, avec toute la part d'approximation que cela comporte.

Exemple : Lors d'une étude sur la résistance d'un métal, on a réalisé 100 expériences de rupture en charge d'un fil de même épaisseur et l'on a noté les poids limites dans chaque cas. Le tableau ci-dessous représente la répartition par classes des résultats.
On calcule la moyenne de la charge de rupture d'un fil, à partir des effectifs.

Tableau 1

charge en grammes	effectifs n_i	centre de classe x_i	$n_i x_i$
700	10	725	7250
750	23	775	17825
800	4	820	3280
840	15	860	12900
880	32	900	28800
920	16	940	15040
960			
TOTAL	100		85095

$$\overline{x} = \frac{\sum n_i x_i}{\sum n_i} = \frac{85095}{100} = 850,95 \approx 851 \text{ g.}$$

Remarque : si on avait fait le calcul sur les données brutes on aurait trouvé : 848g. Cette perte de précision est due au regroupement des données en classes, et au choix des centres de classes comme moyenne de la classe, d'où perte d'information.

III. Le mode ou la classe modale

Le mode ou valeur modale est la valeur que la variable statistique prend le plus fréquemment.

- Dans le cas d'une variable *discrète*, le mode peut être trouvé immédiatement, au vu du tableau des fréquences ou des effectifs.

- Si la variable est *continue*, et si les données sont groupées en classes, on parle plutôt de classe modale : la classe ayant l'effectif le plus élevé (effectif ramené à l'unité d'amplitude). Attention il peut arriver que la classe modale ne soit pas celle où l'effectif apparaît, sur le tableau, le plus élevé. En effet, cette dernière classe peut avoir une amplitude plus grande qu'une autre dont l'effectif par unité d'amplitude, est plus élevé. Sur l'exemple précédent, si la classe 700 à moins de 800 figurait, son effectif serait 33, supérieur à celui retenu pour la classe modale. Mais ramené à l'unité d'amplitude 40, l'effectif ne serait plus que : 33*40/100=13.2. La répartition des charges de rupture d'un fil a pour classe modale la classe « 880 à moins de 920 », d'effectif 32.

IV. La médiane

La médiane d'une série statistique est une valeur de la variable telle qu'il y ait autant d'observations ayant une valeur supérieure à la médiane que d'observations ayant une valeur inférieure à la médiane.

Exemple : si nous considérons les cinq valeurs suivantes : 711 862 851 912 922.
Ces valeurs peuvent être rangées selon les grandeurs croissantes : 711 851 862 912 922.
La valeur 862 est telle que deux observations ont une valeur inférieure et deux autres une valeur supérieure : c'est la médiane.

Lorsque les observations sont toutes données, il suffit donc de les classer par ordre de grandeurs croissantes (ou décroissantes), et de prendre celle qui se trouve au milieu. Si le nombre des observations est pair, la médiane peut être théoriquement l'une quelconque des valeurs comprises entre les deux valeurs centrales observées ; le plus souvent on choisit leur demi-somme.
Si par contre les observations sont regroupées en classes , il est nécessaire de recourir aux effectifs –ou aux fréquences- cumulés.

V. Effectifs ou fréquences cumulés.

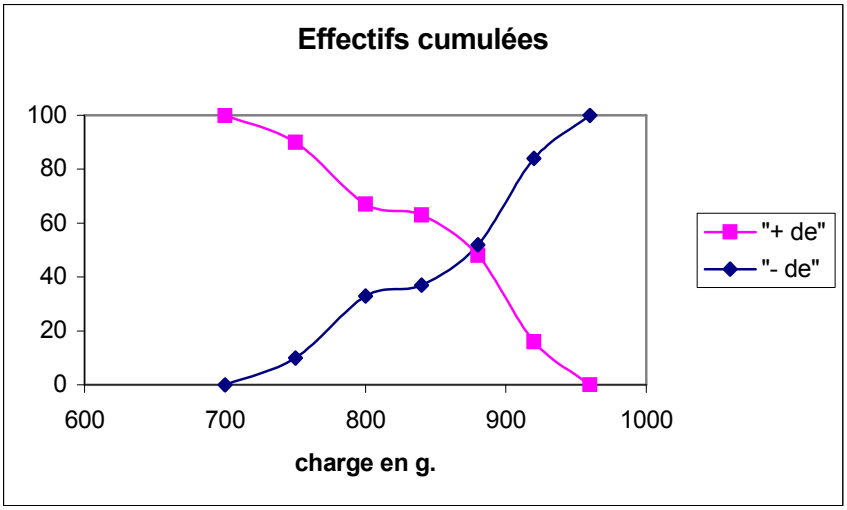
Il est souvent intéressant, devant une série statistique, de pouvoir dire « il y a tant d'observations » ou « il y a tel pourcentage d'observations » inférieures à telle valeur (ou supérieures). C'est à ce genre de préoccupation que répond le calcul des fréquences ou des effectifs cumulés.

5.1. Variable continue

On ajoute l'effectif d'une classe à l'effectif cumulé précédent, en commençant par le haut du tableau pour l'effectif cumulé « moins de » et le bas pour l'effectif « plus de », voir tableau 2.
La seule difficulté est de bien commencer ; pour cela, il suffit de se référer à la signification des résultats ; si l'on cherche combien de fil n'ont pu résister à un poids de moins de 700g, le tableau 2 permet de répondre qu'il n'y en a aucun, on écrit l'effectif cumulé 0 en face du poids 700g.
On lit par exemple que 67 fils ont supporté une charge de plus de 800g. Il est possible d'effectuer une représentation graphique des effectifs cumulés.

Tableau 2

charge en grammes	effectifs n_i	centre de classe x_i	$n_i x_i$	effectifs cumulés "- de"	"+ de"
700	10	725	7250	0	100
750	23	775	17825	10	90
800	4	820	3280	33	67
840	15	860	12900	37	63
880	32	900	28800	52	48
920	16	940	15040	84	16
960				100	0
TOTAL	100		85095		



5.2. Variable discrète

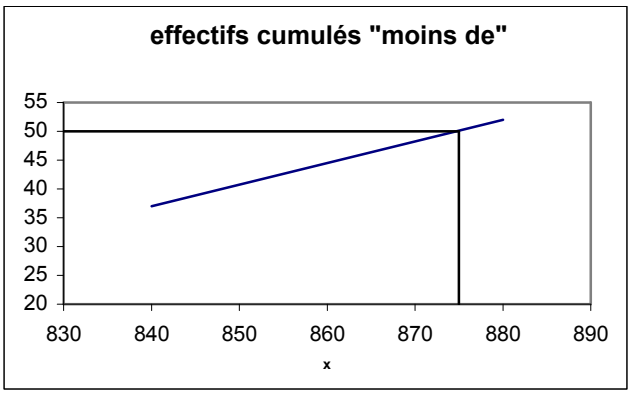
Si la variable est discrète, une petite difficulté supplémentaire apparaît, au niveau de la signification. Plusieurs définitions des effectifs ou fréquences cumulés sont possibles Il faut faire attention si on parle au sens stricte ou au sens large.

VI. Calcul de la médiane

La médiane est la valeur de la variable qui correspond à la fréquence cumulée 50% ou à l'effectif cumulé *n /2*.

On calcule la valeur de la variable correspondant à la fréquence cumulée 50%.
Considérons les données du tableau 2 : la médiane M correspond à l'effectif 100/2=50. On procède à l'interpolation linéaire sur les effectifs cumulés « moins de » (ou éventuellement « plus de ») :

840	37
880	52



Le point de coordonnées (M,50) est sur la droite passant par les points A et B.
Trouvons l'équation de cette droite : $y=ax+b$.

$$\text{A(840,37)} \quad \text{et} \quad \text{B(880,52)} \quad \rightarrow \begin{cases} 37 = 840\,a + b \\ 52 = 880\,a + b \end{cases} \Rightarrow \begin{cases} 15 = 40\,a \Rightarrow a = 3/8 \\ b = -278 \end{cases}$$

$$x = \frac{y-b}{a} \Rightarrow M = \frac{50+278}{3/8} \cong 874{,}67 \approx 875\,\text{g.}$$

Remarque : Cette valeur peut être lue sur le graphique des effectifs cumulés ; c'est la valeur de la variable correspondant à l'effectif 50 ; on lit environ 875g.

VII. Etendue.

L'étendue est la différence entre la valeur maximale et la valeur minimale d'une série.

VIII. Caractéristiques de dispersion

On considère les deux séries de données suivantes : 95 97 100 103 105
50 75 100 125 150
Elles ont même moyenne arithmétique et même médiane (100). Cependant elles diffèrent profondément. Ce qui fait leur différence, c'est ce qu'en statistique on nomme la dispersion ; la deuxième série est beaucoup plus dispersée que la première.
Il est donc important de résumer une série statistique non seulement par des caractéristiques de tendance centrale, mais aussi par des caractéristiques de dispersion. Nous en définirons de deux sortes : celle liées à la moyenne : écart absolu moyen et écart-type ; celles liées à la médiane : écart interquartile, écart interdécile, etc…

IX. Ecart absolu moyen

On calcule d'abord l'écart à la moyenne. Pour chaque valeur de la variable *x*, on calcule l'écart de cette valeur à la moyenne $\bar{x}$; on cherche ensuite à résumer ces écarts en calculant une moyenne.
Pour les deux séries du VII, les écarts sont : -5 -3 0 3 5
-50 -25 0 25 50

Il est impossible de résumer ces écarts par leur moyenne arithmétique, puisque par définition même de x :

$$\sum_{i=1}^{n}(x_i - \bar{x}) = -n\bar{x} + \sum_{i=1}^{n} x_i = -n\bar{x} + n\bar{x} = 0$$

Cependant, la simple vue des deux lignes d'écart calculées ci-dessus montre que ceux-ci caractérisent convenablement la dispersion. On a alors recours à la moyenne des valeurs absolues des écarts, c'est l'écart absolu moyen :

$$e = \frac{\sum |x_i - \bar{x}|}{n}$$

Ou, si les observations sont réparties par classes :

$$e = \frac{\sum n_i |x_i - \bar{x}|}{n}$$

Pour la première série observée on a :

$$e_1 = \frac{16}{5} = 3,2$$

Et pour la deuxième :

$$e_2 = \frac{150}{5} = 30$$

Cette caractéristique rend convenablement compte de la différence de dispersion entre les deux séries. Elle est cependant peu utilisée. En outre, la formulation des lois statistiques fait appel à une autre caractéristique : **l'écart type**.

X. Ecart-type

10.1. Définition

La caractéristique de dispersion la plus usuelle est en effet l'écart-type. Puisque la moyenne arithmétique des écarts à la moyenne est nulle, on a recours à la moyenne quadratique de ces écarts. On définit :

- *la variance* d'une série : c'est une moyenne arithmétique des carrés des écarts à la moyenne :

$$V = \frac{\sum n_i (x_i - \bar{x})^2}{\sum n_i}$$

- *L'écart type* d'une série : c'est la moyenne quadratique des écarts à la moyenne, autrement dit, c'est la racine carrée de la variance.

$$\sigma = \sqrt{V} = \sqrt{\frac{\sum n_i (x_i - \bar{x})^2}{\sum n_i}}$$

En reprenant les séries du §7, on a pour la première :

$$v_1 = \frac{25+9+25+9}{5} = 13{,}6 \text{ et } \sigma_1 = 3{,}69$$

Et pour la deuxième :

$$v_2 = \frac{2500+625+625+2500}{5} = \frac{6250}{5} = 1250 \text{ et } \sigma_2 = 35{,}36$$

10.2. Méthode de calcul

Les calculs risquent de devenir laborieux si la moyenne n'est pas entière : on a à traiter des « écarts à la moyenne » non entiers avec d'inévitables arrondis, d'où des calculs lourds et forcément peu précis. Pour alléger les calculs, on se sert du théorème de Koenig.
Développons :

$$\begin{aligned} S &= \sum n_i (x_i - \bar{x})^2 \\ &= \sum n_i x_i^{\,2} - \sum 2 n_i x_i \bar{x} + \sum n_i \bar{x}^2 \\ &= \sum n_i x_i^{\,2} - 2\bar{x} \underbrace{\sum n_i x_i}_{n\bar{x}} + n\bar{x}^2 \quad \text{car} \quad \bar{x} = \left(\frac{\sum n_i x_i}{n} \right) \\ &= \sum n_i x_i^{\,2} - n\bar{x}^2 \end{aligned}$$

On exprime souvent ce théorème à partir de la formule de la variance qui s'en déduit :

$$V(X) = \frac{\sum n_i x_i^{\,2}}{n} - (\bar{x})^2$$

La variance est égale à la moyenne des carrés moins le carré de la moyenne. Ce résultat simplifie considérablement les calculs nécessaires pour obtenir la variance et l'écart-type ; c'est sous cette forme que le théorème de Koenig est utilisé dès qu'on dispose d'une machine à calculer.

Remarque : cette dernière formulation de la variance limite les erreurs d'arrondis car la moyenne n'intervient qu'une seule fois alors que dans la formulation précédente elle intervient i fois.

10.3. Exemples

Il est possible de calculer la variance et l'écart type sur l'exemple du §2.
Pour la rupture en charge des fils, sur les données groupées du tableau 1. On utilise la formule :

$$V(X) = \frac{\sum n_i x_i^2}{n} - (\overline{x})^2$$

Tableau 3

charge en grammes	effectifs n_i	centre de classe x_i	$n_i x_i$	$n_i x_i^2$
700	10	725	7250	5256250
750	23	775	17825	13814375
800	4	820	3280	2689600
840	15	860	12900	11094000
880	32	900	28800	25920000
920	16	940	15040	14137600
960				
TOTAL	100		85095	72911825

$$V(X) = \frac{\sum n_i x_i^2}{\sum n_i} - (\overline{x})^2 = \frac{72911825}{100} - (850{,}95)^2 = 5002{,}35$$

$$\Rightarrow \sigma = \sqrt{5002{,}35} = 70{,}73 \text{ g.}$$

10.4. Signification de l'écart type

Remarque : Il existe une autre quantité représentante de la dispersion d'une série, c'est l'étendue :

étendue = valeur maximale – valeur minimale.

Lorsque l'on compare deux séries de même nature, celle qui a l'écart type le plus élevé est la plus dispersée.

Cependant, par référence à une loi statistique usuelle, la loi normale, il est possible de préciser un peu la signification de l'écart type. Lorsqu'une série statistique satisfait à la loi normale, 95% des observations sont comprises entre $\bar{x} - 2\sigma$ et $\bar{x} + 2\sigma$: plus l'écart type est élevé, plus les observations sont dispersées. Si la série statistique étudiée, sans suivre une loi normale, n'est pas trop dissymétrique, la même propriété est approximativement vraie.

On déduit de la propriété énoncée ci-dessus dans le cas de la loi normale, la règle de vérification suivante : l'étendue d'une série statistique (différence entre observation la plus élevée et la plus basse) est du même ordre de grandeur que quatre écart-types.

Par exemple : pour la rupture en charge de fils, l'étendue est certainement inférieure à 960-700=260g (en effet 960 et 700 sont des extrémités de classes dont on sait par les données brutes qu'elles ne sont pas toutes deux atteintes) et 4σ=283g : les deux nombres ne sont pas égaux mais ils sont du même ordre de grandeur.

10.5. Coefficient de variation

L'étendue, la variance et l'écart type sont des paramètres de dispersion absolue qui mesurent la variation absolue des données. Cependant, un écart type de 6 mm n'a pas la même signification s'il se rapporte à des mesures de l'ordre de 160 mm ou à des mesures de l'ordre de 80 mm. Aussi dispose-t-on d'un indice de dispersion relative appelé coefficient de variation, noté CV. Par définition, le coefficient de variation est égal à

$$CV = \frac{100\sigma}{\bar{x}}$$

Remarque : ce coefficient cesse d'être efficace pour $\bar{x}$ petit.

Ce coefficient de variation a l'avantage d'être comparable pour toutes les séries statistiques.

Exemple : (rupture en charges) le coefficient de variation de la série étudiée ci-dessus est :

$$\frac{70{,}73}{850{,}95} = 0.08$$

La série de poids apparaît peu dispersée, parce que toutes les observations sont « relativement » voisines de la moyenne.

XI. Caractéristiques de position : quartiles, déciles, centiles

Les quartiles, déciles et centiles sont des caractéristiques qui correspondent au même genre de préoccupation que la médiane.
Il s'agit des valeurs de la variable qui correspondent aux effectifs cumulés :

n/4, 2n/4, 3n/4	pour des quartiles, le 2ème quartile est la médiane,
n/10, 2n/10,…9n/10	pour les déciles ; le 5ème décile est la médiane,
n/100, 2n/100,…99n/100	pour les centiles ; le 50ème centile est la médiane.

On les appelle caractéristiques de position, puisqu'elle permettent de placer les valeurs de la variable.

- *Cas d'une variable continue*

Les calculs s'effectuent comme ceux concernant la médiane.
Reprenons l'exemple de rupture des fils. Les quartiles peuvent être déterminés soit graphiquement, soit par un calcul d'interpolation linéaire. La premier quartile Q1 correspond à l'effectif cumulé 25% soit :

A	750	10
?	Q_1	25
B	800	33

$$\begin{cases} 10 = 750a + b \\ 33 = 800a + b \end{cases} \Rightarrow 23 = 50a \Rightarrow a = \frac{23}{50} \text{ et } b = 10 - 750 \times \frac{23}{50} = 335$$

Ce qui donne :

$$25 = Q_1 \times \frac{23}{50} - 335 \Rightarrow Q_1 = (25 + 335) \times \frac{50}{23} = 782{,}61 \approx 783\,\text{g}.$$

De même on peut trouver pour le 3ème quartile : (Q_3,75)

$$Q_3 = 908{,}75\text{g} \approx 909\text{g}.$$

On calculerait de la même manière les déciles. Pour le 1er et le 9ème décile, on obtient :

D_1=*750* g et D_9=*935* g.

- *Cas d'une variable discrète*

Le principe est le même.

Pour des populations nombreuses, on calculerait de même certains centiles, particulièrement les centiles extrêmes, C_1 et C_{99}.

XII. Caractéristiques de dispersion : étendue, écarts interdéciles, écarts interquartiles

Les caractéristiques de position définies au §X suggèrent une manière de caractériser la dispersion sensiblement différente de celle qui aboutit à l'usage de l'écart-type. En effet, un intervalle dans lequel on trouve toute la population étudiée, ou un intervalle à l'intérieur duquel se situe 80% de cette population, les 10% extrêmes (les plus aberrants) étant éliminés des deux côtés, peut donner une idée de la façon dont se réparti une série.

Le premier intervalle ainsi défini est l'étendue, différence entre l'observation la plus élevée et l'observation la plus faible. Le second est l'écart interdécile : $D_9 - D_1$. On définit de la même manière l'écart interquartile : $Q_3 - Q_1$.

Ainsi pour la série des charges de rupture du fil, l'étendue est : *e= 960-700 =260* g.

L'écart interdécile est : *D9-D1 = 935-750=185* g
Interprétation de ce résultat : En éliminant les 10% les plus résistants et les 10% les moins résistants, les charges de rupture des fils sont réparties à l'intérieur d'une plage de 185 g.

L'écart interquartile est : *Q3-Q1 = 909-783=126* g.
Interprétation : 50% de la population des fils a une charge de rupture répartie sur 126g. Cet écart est élevé par rapport au précédent : mais la répartition des charges de rupture fait apparaître en quelque sorte deux populations distinctes ; l'élimination des 20% ou des 50% de l'ensemble qui se trouvent aux extrémités ne fait pas disparaître la classe centrale « 800 à moins de 840 » dont l'effectif est très faible.

On peut faire les mêmes calculs pour une variable discrète. Les résultats sont sensiblement moins intéressants. En effet, il est fréquent que des quartiles ou des déciles soient égaux à la médiane.

XIII. Quelques conseils pour l'étude de séries statistiques simples

Il est nécessaire de séparer clairement deux types de calculs :

- moyenne, écart-type .. à réaliser à partir des centres de classes et des effectifs de classes.
- médiane, quartiles, intervalles interquartiles .. à réaliser à partir des extrémités de classes et des effectifs cumulés.

CHAPITRE 3 : ETUDE DES SERIES STATISTIQUES DOUBLES

I. Position du problème

Dans les chapitres précédents on étudiait une population selon un seul caractère. Cependant il est souvent utile de considérer à la fois plusieurs caractères de la même population : taille, âge, poids d'un groupe d'enfants ; qualification et salaire de salariés ; température et pression d'un milieu à différentes heures …
Nous nous limiterons ici à l'étude simultanée de deux caractères ; l'analyse des données permet d'en étudier un grand nombre.

II. Notations et représentation des séries statistiques doubles

Une série statistique double peut être donnée comme l'énumération d'un certain nombre de résultats. La tableau ci-dessous donne la consommation en milliers de calories de douze familles en moyenne par jour. Chaque homme adulte est compté pour une « unité de consommation » ; un enfant est compté pour une part d'unité, dépendant de son âge et de son sexe.

Tableau 1

n° de famille	unité de consommation x_i	calories par jour y_i
1	5,3	13
2	7,2	18
3	5,6	9,4
4	7,1	15,4
5	5	7,8
6	3,3	9,3
7	5,2	10,1
8	4,5	7,1
9	4	8,9
10	2	4,4
11	5,7	12,1
12	4,7	11,5
TOTAL	**59,6**	**127**

On peut avoir des données groupées : on parle alors de tableaux carrés ou de tableaux à double entrée. Il est alors nécessaire d'employer des notations précises.
Soient x et y deux caractères (quantitatifs ou non). Les classes du caractère x sont désignées par les indices 1,…,j,….,p, celles du caractère y par 1,…,i,…,q.
n_{ij} est le nombre d'unités représentant la modalité y_i de y et la modalité x_j de x.
Les sommes des effectifs de la ligne i, de la colonne j et de l'ensemble sont notés respectivement :

$$\sum_{j=1}^{p} n_{ij} = n_{i.} \;,\; \sum_{i=1}^{q} n_{.j} = n_{.j} \;,\; \sum_{j=1}^{p}\sum_{i=1}^{q} n_{ij} = n_{..}$$

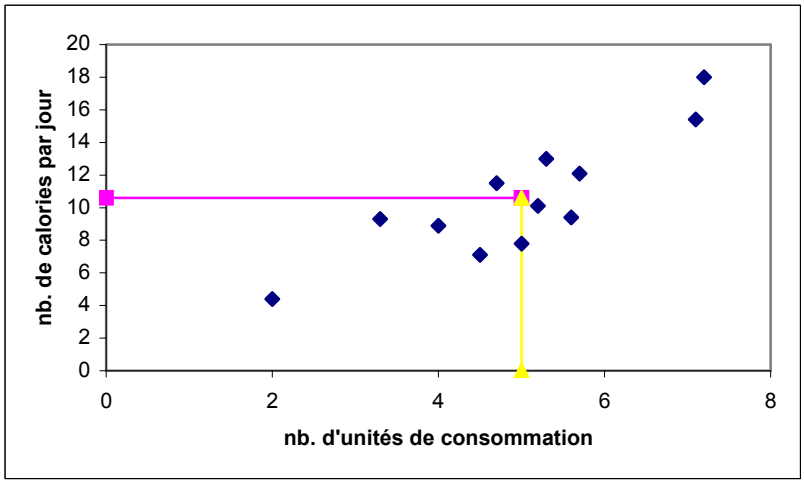
Tableau 2 : Notation des tableaux carrés

		caractère x						
		x_1	x_2		x_j		x_p	total
caractère y	y_1	n_{11}	n_{12}		n_{1j}		n_{1p}	$n_{1.}$
	y_2	n_{21}	n_{22}		n_{2j}			$n_{2.}$
	:				:			:
	:	:	:		:			:
	y_i	:	:		n_{ij}			$n_{i.}$
	:	:	:					:
	:	:						:
	y_q	n_{q1}					n_{qp}	$n_{q.}$
	total	$n_{.1}$	$n_{.2}$	…….	$n_{.j}$	……..	$n_{.p}$	$n_{..}$

La dernière ligne et la dernière colonne du tableau représentent *les distributions marginales*, c'est à dire la distribution de x sans tenir compte du caractère y ou celle de y sans tenir compte de x.
La distribution marginale des y_i , comme celle des x_j, peut être traitée comme une série simple. On définit en particulier la moyenne y, la variance V(y) et l'écart type $\sigma(y)$.
De même, l'une quelconque des lignes ou des colonnes du tableau peut être interprétée comme *une distribution conditionnelle*.
Il est possible d'étudier les distributions conditionnelles comme des séries statistiques simples.

III. Ajustement linéaire. Principe de la méthode des moindres carrés

Les données du tableau 1 peuvent être représentées graphiquement :



On constate sur la figure ci-dessus que les points représentatifs de la série des consommations ne sont pas rigoureusement alignés, mais qu'ils forment un nuage de points allongé.
Il n'est pas alors dépourvu de sens de chercher si l'on peut trouver une droite qui résume approximativement l'ensemble des points. La recherche d'une telle droite est un ajustement linéaire.

3.1. Ajustement graphique

Théoriquement, diverses sortes d'ajustement linéaires sont possibles. La plus simple est l'ajustement graphique, réalisé par le dessinateur. L'inconvénient majeur de l'ajustement graphique est qu'il est subjectif.

3.2. Autres ajustements

D'autres ajustements peuvent être réalisés de façon plus objective, par exemple en utilisant les points extrêmes ou les moyennes de certains groupes de résultats. Lorsqu'il s'agit de séries chronologiques, il est usuel de réaliser un ajustement linéaire par de telles méthodes.

3.3. Méthode des moindres carrés

La méthode des moindres carrés présente un caractère plus rigoureux que les précédentes. Elle consiste à rechercher une droite telle que la somme de ses distances aux différents points représentant les données soit minimale. Le mot distance est pris au sens large. La distance choisie est le carré de la différence des ordonnées entre chaque point et le point de la droite ayant même abscisse.

3.4. Notion de corrélation linéaire

La méthode des moindres carrés peut être utilisée pour n'importe quelle série double. Quelle que soit cette série, il existe une droite d'estimation par la méthode des moindres carrés. Pour s'assurer de façon objective que l'ajustement est valable, on calcule le coefficient de corrélation linéaire :

$$r = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

Ce coefficient est compris entre –1 et +1. S'il est voisin en valeur absolu de 1, l'ajustement est valide ($0.70 < |r| < 1$). La covariance joue un rôle analogue à la variance dans les séries statistiques simples, elle est définie par :

$$\mathbf{cov}(x,y) = \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum x_i y_i - \bar{x} \times \bar{y}$$

Remarque : si on fait $x=y$, on retrouve la formule de la variance.

Sur l'exemple du tableau 1, calculons le coefficient de corrélation linéaire.

A l'aide de ce tableau, on peut effectuer les calculs suivants :

Tableau 3

n° de famille	unité de consommation x_i	calories par jour y_i	x_i^2	y_i^2	$x_i y_i$
1	5,3	13	28,09	169	68,9
2	7,2	18	51,84	324	129,6
3	5,6	9,4	31,36	88,36	52,64
4	7,1	15,4	50,41	237,16	109,34
5	5	7,8	25	60,84	39
6	3,3	9,3	10,89	86,49	30,69
7	5,2	10,1	27,04	102,01	52,52
8	4,5	7,1	20,25	50,41	31,95
9	4	8,9	16	79,21	35,6
10	2	4,4	4	19,36	8,8
11	5,7	12,1	32,49	146,41	68,97
12	4,7	11,5	22,09	132,25	54,05
TOTAL	**59,6**	**127**	**319,46**	**1495,5**	**682,06**

$$\bar{x} = \frac{59,6}{12} = 4,97 \approx 5 \text{ unités de consommation}$$

$$\bar{y} = \frac{127}{12} = 10.58 \approx 10,6.10^3 \text{ calories.}$$

$$V(x) = \frac{1}{n}\sum x_i^2 - \bar{x}^2 = \frac{319,46}{12} - (4,97)^2 = 1,95$$

$$\Rightarrow \sigma_x = 1.4 \text{ unités de consommation}$$

$$V(y) = \frac{1}{n}\sum y_i^2 - \bar{y}^2 = \frac{1495\ .5}{12} - (10\ .58\ )^2 = 12\ .62$$

$$\Rightarrow \sigma_y = 3,55\ .10^3 \text{ calories}$$

$$\mathbf{cov}(x, y) = \frac{1}{n}\sum x_i y_i - \bar{x} \times \bar{y} = \frac{682.06}{12} - 4.97 \times 10.58 = 4.26$$

Le coefficient de corrélation est alors :

$$r = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y} = \frac{4.26}{1.4 \times 355} = 0.86$$

Sans indiquer une bonne corrélation (il faudrait qu'il soit supérieur à 0.95), ce coefficient autorise l'ajustement linéaire.

IV. Application de la méthode des moindres carrés à des données individuelles

4.1. Principe de la méthode

La droite définie au paragraphe (3.3) a pour équation :

$$\hat{y} = ax + b$$

On recherche les paramètres *a* et *b*. La différence des ordonnées entre un point(x_i,y_i) et le point de la droite ayant même abscisse est :

$$y_i - \hat{y}_i = y_i - ax_i - b$$

La somme des carrés de ces différences doit être minimum :

$$S = \sum_{i=1}^{n} (y_i - ax_i - b)^2 \quad \text{minimum}$$

Pour définir les coefficients *a* et *b*, on développe S et on le considère successivement comme un trinôme en *b*, puis *b* étant déterminé, comme un trinôme en *a*. On trouve :

$$b = \bar{y} - a\bar{x}$$

$$a = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\bar{x}^2}$$

On reconnaît au numérateur la covariance de *x* et de *y* et au dénominateur la variance de *x*, au coefficient n près.

$$a = \frac{\mathrm{cov}(x, y)}{V(x)}$$

La forme du coefficient *b* permet de constater que la droite d'ajustement passe par le « point moyen » (x,y). Son équation est :

$$\hat{y} = \bar{y} + a(x - \bar{x})$$

4.2. Application à l'exemple du tableau 3 :

En utilisant les formules ci-dessus, on trouve :

$$a = \frac{4.26}{1.95} = 2.18$$

$$\hat{y} = ax + b = ax + \bar{y} - a\bar{x} = \bar{y} + a(x - \bar{x})$$

$$\Rightarrow \hat{y}_1 = 10.6 + 2.2(x - 5) \text{ droite d'estimation de y en x.}$$

4.3. Droite d'estimation de x en y

Le calcul précédent fait jouer un rôle dissymétrique à *x* et à *y*. Or rien au plan statistique ne permet de dire si une variable dépend de l'autre. Il est alors aussi logique de recommencer les calculs précédents, mais en inversant les rôles des deux variables.

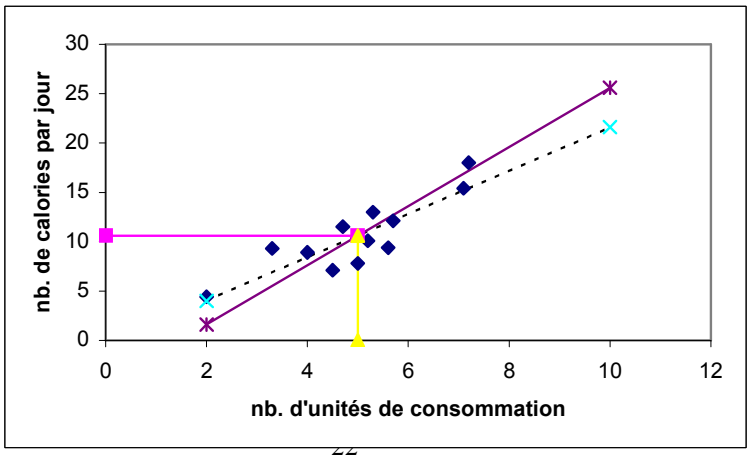
On définit une droite d'estimation de *x* en *y*, d'équation :

$$\hat{x} = \bar{x} + a'(y - \bar{y}) \text{ avec } a' = \frac{\mathrm{cov}(x,y)}{V(y)} = \frac{4.26}{12.62} = 0.34$$

$$\hat{x} = 5 + 0.34(y - 10.6) \quad \text{qui peut s'écrire} \quad y = f(x)$$

$$\Rightarrow \hat{y}_2 = \frac{x - 5}{0.34} + 10.6 \Leftrightarrow \hat{y}_2 = 10.6 + 3(x - 5)$$

Elle diffère de la précédente par sa pente.



4.4. Retour sur le coefficient de corrélation linéaire

Les deux droites d'estimation trouvées sont différentes. Le carré du coefficient de corrélation linéaire est précisément égal au produit des pentes.

$$aa' = \frac{\mathrm{cov}(x,y)^2}{V(x)V(y)} = \left(\frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y}\right)^2 = r^2$$

$$r^2 = aa' = 0.86^2 = 0.75$$

Si les deux droites étaient identiques, r serait en valeur absolue égal à 1. Si les droites sont proches, | r | est voisin de 1. Par contre si | r | est voisin de 0, les deux pentes sont loin d'être inverse l'une de l'autre, et par conséquent les droites d'ajustement sont sensiblement différentes : les points représentatifs sont loin d'être réellement alignés.

V. Application de la méthode des moindres carrés à des données groupées

5.1. Méthode

Le principe de la méthode est le même, calculer si nécessaire les centres de classes.
Le tableau ci-dessous représente la répartition des distances parcourues par des véhicules après un coup de frein selon la vitesse.

y distance (m)		x vitesse (km/h) 70	80	90	100	110	120	$n_{i.}$	y_i	$n_{i.}y_i$	$n_i y_i^2$	$n_{ii}x_i$	x_v	y_i
		1	2	3	4	5								
20	1	3						3	25	75	1875	225	75	5625
30	2	25	5					30	35	1050	36750	2300	76,67	80500
40	3	13	29	24	12			78	50	3900	195000	6980	89,49	349000
60	4	6	12	21	12	5		56	70	3920	274400	5300	94,64	371000
80	5		2	10	16	5		33	90	2970	267300	3375	102,3	303750
100	$n_{.j}$	47	48	55	40	10		200		11915	775325	18180		1109875
	x_j	75	85	95	105	115								
	$n_{.j}x_j$	3525	4080	5225	4200	1150		18180						
	$n_{.j}x_j^2$	264375	346800	496375	441000	132250		1680800						
	$n_{ij}y_i$	2020	2645	3570	2880	800		11915						
	y_x	42,97872	55,10417	64,90909	72	80								
	x_j	151500	224825	339150	302400	92000		1109875						

Les colonnes $n_{i.}y_i$, $n_{i.}y_i^2$ et les lignes $n_{.j}x_j$ et $n_{.j}x_j^2$ permettent de calculer les caractéristiques des deux distributions marginales.

$$\bar{x} = \frac{18180}{200} = 90.9 \text{ km/h} \qquad \bar{y} = \frac{11915}{200} = 59.58 \text{ m}$$

$$V(x) = \frac{1680800}{200} - (90.9)^2 = 141.19 \qquad V(y) = \frac{775325}{200} - (59.58)^2 = 327.44$$

$$\Rightarrow \sigma_x = 11.9 \text{ km/h} \qquad \Rightarrow \sigma_y = 18.1 \text{ m}$$

5.2. Distributions conditionnelles : courbes de régression

Sur ce tableau il est possible d'analyser les distributions conditionnelles.
En calculant sur une ligne

$$\sum_j n_{ij} x_j$$

Il est possible de calculer x_y, c'est à dire la moyenne (conditionnelle) de *x* pour un *y* donné (plus exactement, pour y compris entre les limites de classes).
Ainsi pour y compris entre 30 et 40 m. (ligne 2) :

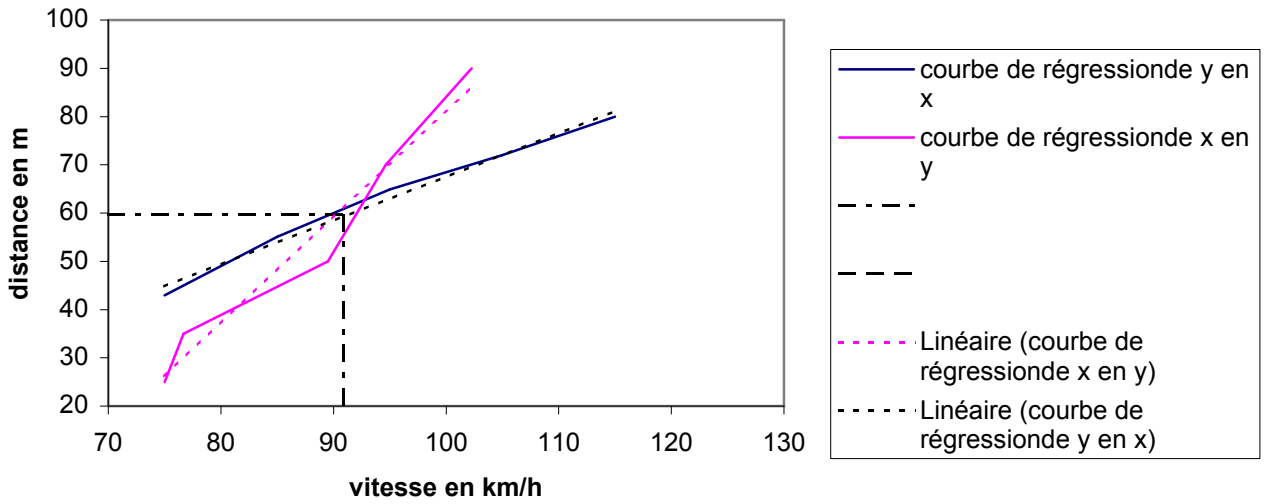
$$\bar{x}_y = \frac{\sum_j n_{2j} x_j}{\sum_j n_{2j}} = \frac{2300}{30} = 76.67 \text{ km/h.}$$

De même, en colonne, on calcule les moyennes conditionnelles de *y* pour *x* donné.
Pour x compris entre 100 et 110 km/h :

$$\bar{y}_x = \frac{\sum_i n_{i4} y_i}{\sum_i n_{i4}} = 72 \text{ m.}$$

Il serait possible de même de calculer les écart-types conditionnels.
L'ensemble des points de coordonnées ($\bar{y}_x$, x_i) constitue la courbe de régression de *y* en *x*, et l'ensemble des points de coordonnées ($\bar{x}_y$,y_i) constitue la courbe de régression de *x* en *y*.
Ces deux courbes représentent valablement la distribution.
On peut en déduire que si on roule à telle vitesse, on s'arrête en moyenne en tant de mètres.
Ou si l'on a besoin de tant de mètres pour s'arrêter, c'est qu'on roulait en moyenne à telle vitesse.
Il n'est guère possible de décrire efficacement l'ensemble des 200 observations.

Figure 2



5.3. Coefficient de corrélation linéaire

On peut constater que les deux courbes de régression sont sensiblement différentes de droites. Il est possible en effet de vérifier que le coefficient de corrélation linéaire entre *x* et *y* est assez mauvais. Pour obtenir la covariance on calcule :

$$\sum_i y_i \sum_j n_{ij} x_j = \sum_j x_j \sum_i n_{ij} y_i = \sum_i \sum_j n_{ij} x_j y_i$$

$$\mathrm{cov}(x, y) = \frac{1109875}{200} - (90.9)(59.58) = 133.55$$

D'où le coefficient de corrélation linéaire :

$$r = \frac{133.55}{18.1 \times 11.9} = 0.62$$

Ce coefficient est presque égal à 0.60, ce qui est faible. L'ajustement linéaire ne représente pas un bon résumé des observations. Puisqu'il est toujours possible de calculer les droites des moindres carrés, nous allons vérifier que l'ajustement linéaire est mauvais en calculant les équations de ces droites.

5.4. Droites d'estimation par la méthode des moindres carrés

Pour la droite d'estimation de *y* en *x* :

$$a = \frac{\mathrm{cov}(x, y)}{V(x)} = \frac{133.55}{141.19} = 0.95$$

D'où l'équation :

$$\hat{y}_1 = \overline{y} + a(x - \overline{x}) = 59.58 + 0.95(x - 90.9)$$

Pour la droite d'estimation de *x* en *y* :

$$a' = \frac{\text{cov}(x, y)}{V(y)} = \frac{133.55}{327.44} = 0.41$$

L'équation est, en rétablissant les axes :

$$\hat{x} = \overline{x} + a'(y - \overline{y}) \Leftrightarrow \hat{y} = \frac{\hat{x} - \overline{x}}{a'} + \overline{y} \Rightarrow \hat{y}_2 = 59.58 + 2.44(x - 90.9)$$

Ces deux droites, tracées sur la figure 2, sont différentes ; par contre, elles ne sont pas très éloignées chacune de la courbe de régression correspondante. Ce résultat explique la fréquente confusion entre « courbe de régression » et « droite d'ajustement par la méthode des moindres carrés ».

VI. Ajustement non linéaire

Il peut arriver que les points représentant une série double ne soient pas alignés, mais soient voisins d'une courbe connue. On se sert alors en général de la méthode des moindres carrés, mais en transformant au préalable l'une des variables. Ainsi, un ajustement entre *y* et x^n donne un ajustement de la forme $y = a\,x^n + b$; un ajustement entre *y* et *ln x* donne :
$y = b\ e^{ax}$.

Relations usuelles :

$$y = a\exp(bx)$$
$$y = ax^b$$
$$y = a + b\log x$$
$$y = cx^a + dx^b$$
$$y = a_0 + a_1x + a_2x^2 \quad \text{(parabole)}$$
$$y = ab^x \quad \text{(géométrique)}$$
$$y = ca^{bx} \quad \text{(Gompertz)}$$
$$y = abx^{b-1}\exp(-ax^b) \quad \text{(Weibull)}$$

Il est également possible de réaliser des ajustements linéaires ou non, à plusieurs variables, toujours sur le principe de la méthode des moindres carrés.

VII. Quelques conseils pour l'ajustement linéaire

- Faire d'abord une étude graphique. On distinguera ainsi si un ajustement, linéaire ou non peut se justifier.

- Un coefficient de corrélation est compris entre –1 et +1.

- Calculer le coef. de corrélation avant d'effectuer l'ajustement, si ce coef est trop faible en valeur absolue, ne pas continuer les calculs (chercher un ajustement non linéaire)

- La covariance est du même signe que la pente de la droite ajustée.

CHAPITRE 4 : LES PRINCIPALES LOIS DE PROBABILITE

I. Loi binomiale

Lorsque les éventualités se réduisent à une alternative (« succès » ou « échec »), la variable aléatoire « nombre de succès » suit une loi de probabilité appelée loi binomiale définie par :
- chaque épreuve donne lieu à deux éventualités exclusives de probabilité constante *p* (succès) et donc *q=1-p* (échecs).
- Les épreuves répétées sont indépendantes.

La loi binomiale est notée *B(n,p)* et a pour caractéristique :

E(X)=np

$$V(X) = n(p - p^2) = np(1 - p) = npq \quad \text{et} \quad \sigma = \sqrt{npq}$$

Remarques : la loi binomiale est symétrique pour p=1/2, et dissymétrique sinon, la dissymétrie est d'autant plus forte :
- Pour n fixe, que p est différent de q
- Pour p fixe, que n est petit.

II. Loi hypergéométrique

Dans le cas de la loi binomiale, la proportion p d'éléments possédant le caractère recherché est fixe, ce qui peut changer si par exemple le tirage se fait sans remise.
Pour une population d'effectif N dont on tire un échantillon d'effectif n sans remise :

$$E(X) = np \quad \text{et} \quad V(X) = npq \frac{N - n}{N - 1}$$

Si N est grand par rapport à n et si p n'est pas trop voisin de 0 ou de 1, il est possible de faire une approximation de la loi hypergéométrique par la loi binomiale.

III. Loi de Poisson

On appelle processus de Poisson, la réalisation d'événements aléatoires dans le temps et dans l'espace, obéissant aux conditions suivantes :
- la probabilité de réalisation de l'événement au cours d'une petite période ou sur une petite portion d'espace t est proportionnelle à t, soit p t,
- la probabilité de deux apparitions sur le même t est négligeable. Ainsi des événements qui se réalisent de façon aléatoire dans le temps : appels téléphoniques sur un central, pannes de machines, arrivées à un péage d'autoroute ou à un guichet de vente, ou dans l'espace : répartition de points au hasard sur une droite … peuvent être considérés comme réalisés par un processus de Poisson.

$$E(X) = \sum_{x=0}^{\infty} x \frac{e^{-m} m^x}{x!}$$

En utilisant

$$e^m = \sum \frac{m^x}{x!} \Rightarrow E(X) = m$$

$$V(X) = m$$

Remarque : On substitue en général une loi de Poisson à une loi binomiale si l'on a à la fois : *n>50* et *np<5.*

IV. Loi normale

On parle de loi normale ou loi de Laplace – Gauss ou loi de Gauss ou encore deuxième loi de Laplace, lorsqu'on a affaire à une variable aléatoire continue dépendant d'un grand nombre de causes indépendantes, dont les effets s'additionnent et dont aucune n'est prépondérante (conditions de Borel).

Exemple : les dimensions de pièces fabriquées dépendent du réglage de l'appareil de fabrication, des vibrations auxquelles il est soumis, de l'homogénéité de la matière première, de la température, de l'humidité … Lorsque tous ces facteurs sont indépendants et qu'aucun n'est prépondérant, on peut supposer que les dimensions suivent une loi normale .

Une variable aléatoire continue X est distribuée selon une loi normale si sa densité de probabilité est

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2}, \quad x \in \Re$$

la loi de probabilité dépend de deux paramètres : *m* et σ ; on la note *N(m , σ).*
On a l'habitude d'effectuer le changement de variable :

$$T = \frac{X - m}{\sigma}$$

La loi de distribution de T est alors :

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

notée *N(0,1)*. Cette loi est dénommée loi normale centrée réduite.

Remarque : f est une fonction paire.

Voici les caractéristiques d'une variable aléatoire X distribuée selon N(0,1) :

En utilisant l'intégrale de Gauss :

$$\int_{-\infty}^{+\infty} e^{\frac{-t^2}{2}} dt = \sqrt{2\pi}$$

On démontre que : $E(X)=0$, $\sigma(X)=1$.

Sa forme est :

Remarque : La valeur m détermine l'axe de symétrie de la courbe.

Pour une loi $N(m,\sigma)$, $E(X)= m$ = médiane = mode
Ecart-type = σ

Remarque : Du fait de la multiplicité des facteurs qui interviennent dans de nombreux phénomènes physiologiques, génétiques, éco ou éthologiques, le loi normale est fondamentale en biologie.

v. Loi du χ^2 de Pearson

Définition : étant données ν variables aléatoires normales centrées, réduites et indépendantes X_i, la somme :

$$\chi^2 = \sum_{i=1}^{\nu} X_i^2$$

suit une loi du χ^2 , dite à ν degrés de liberté.
On calcule sa moyenne et sa variance :

$$E(\chi^2) = \nu \quad \text{et} \quad V(\chi^2) = 2\nu$$

La distribution du χ^2 tend à devenir symétrique quand n augmente, on peut l'assimiler à la distribution normale pour $\nu>30$.

Importante en biologie : comparaisons (moyenne, variance), calcul des intervalles de confiance, tests de conformité, d'indépendance de deux caractères.

VI. Loi de Student

Soient X et Z deux variables aléatoires indépendantes. Z suit une loi du χ^2 à ν degrés de liberté et X une loi *N(0,1).* La variable aléatoire

$$T = \frac{X}{\sqrt{\frac{Z}{\nu}}}$$

suit une loi de Student (ou loi de Student-Fisher) à ν degrés de liberté.
On calcule sa moyenne et sa variance pour $\nu > 2$:

$$E(T) = 0 \quad \text{et} \quad V(T) = \frac{\nu}{\nu - 2}$$

Il existe des tables de la loi de Student qui donnent t tel que :

Utilisée pour les comparaisons de paramètres (moyenne), estimation des paramètres d'une population à partir d'un échantillon.

VII. Loi de Fisher-Snedecor

Si X_1^2 et X_2^2 sont un couple de variables aléatoires indépendantes suivantes deux lois du χ^2 à ν_1 et ν_2 d.d.l. , alors :

$$F = \frac{X_1^2 / \nu_1}{X_2^2 / \nu_2}$$

suit une loi $F(\nu_1 , \nu_2)$.

La loi de Fisher-Snedecor s'applique lors de la comparaison de variances expérimentales et pour l'analyse de variance et covariance.
Sa fonction de densité est toujours positive :