STATISTIQUE

En statistique descriptive, on possède les données relatives à **toute** la population, on parlera de **fréquences**, alors qu'en statistique inférentielle, on travaille sur un échantillon, on parlera de **probabilités**.

I. STATISTIQUE DESCRIPTIVE A UNE DIMENSION

La statistique descriptive permet de présenter les données d'observation sous différentes formes: les **tableaux**, les **diagrammes** et les **paramètres**.

I.1. Les distributions de fréquences

I.1.1. Les tableaux

Les ***tableaux*** permettent de présenter les données sous la forme de distributions de fréquences. (Voir le tableau I ci–dessous)

- i représente le numéro de l'observation 1, 2, ..., p.
- x_i représente la valeur observée de la variable x.
- n_i représente le nombre de fois que la valeur x_i a été observée. C'est la ***fréquence absolue*** [1] de x_i.
- La ***distribution de fréquences*** est formée de l'ensemble des valeurs observées $x_1, x_2, ..., x_i, ..., x_p$, rangées par ordre croissant, et des fréquences correspondantes $n_1, n_2, ..., n_i, ..., n_p$.
- Le nombre total d'observations, n, est appelé ***effectif***:

$$\boxed{n = \sum_{i=1}^{i=p} n_i}$$

- f_i est la ***fréquence relative*** [2], elle est définie par:

$$\boxed{f_i = \frac{n_i}{n} \text{ ou en } \% \qquad f_i = 100\,\frac{n_i}{n}} \quad \text{et alors} \quad \boxed{\sum_{i=1}^{i=p} f_i = 1}$$

---

[1] En anglais: *absolute frequency, frequency.*

[2] En anglais: relative frequency, proportional frequency.

- Les fréquences (absolues ou relatives) peuvent aussi être additionnées de proche en proche, définissant ainsi les ***fréquences cumulées*** [1]. (N_i ou F_i).

  La distribution des fréquences cumulées est une fonction non décroissante, variant de O à 1 par bonds successifs de n_i ou f_i.

Exemple

En agriculture, on plante des graines et on s'intéresse au nombre de plants obtenus par m^2.

Tableau I

Nombre de pieds d'Asphodelus cerasifer observés dans 512 carrés de 1 m^2: distribution de fréquences.

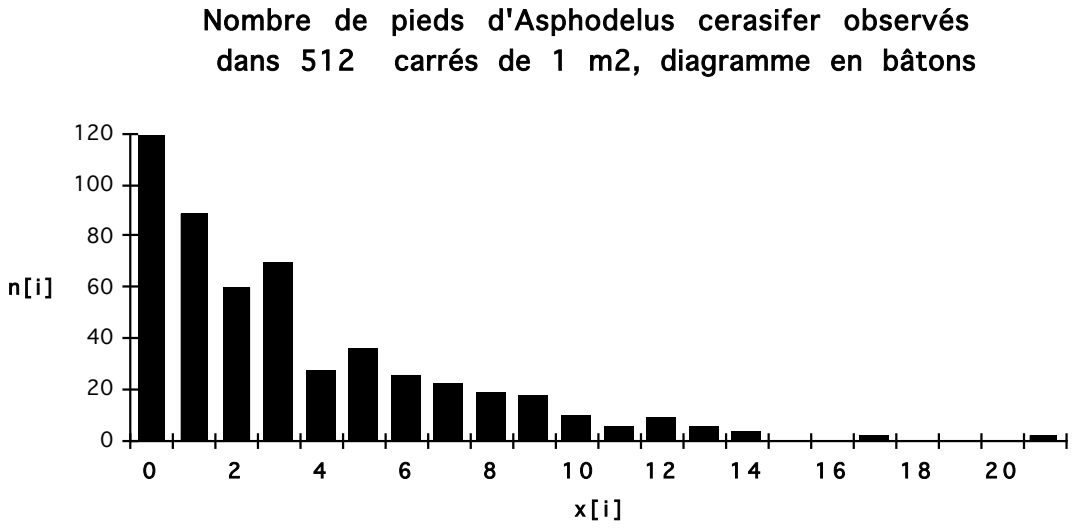
Numéro i	Nombres de pieds x_i	Fréquences absolues (Nb de carrés) n_i	Fréquences relatives f_i	Fréquences absolues cumulées N_i	Fréquences relatives cumulées F_i
1	0	119	0.232	119	0.232
2	1	88	0.172	207	0.404
3	2	59	0.115	266	0.520
4	3	69	0.135	335	0.654
5	4	27	0.053	362	0.707
6	5	36	0.070	398	0.777
7	6	25	0.049	423	0.826
8	7	22	0.043	445	0.869
9	8	18	0.035	463	0.904
10	9	17	0.033	480	0.938
11	10	9	0.018	489	0.955
12	11	5	0.010	494	0.965
13	12	8	0.016	502	0.980
14	13	5	0.010	507	0.990
15	14	3	0.006	510	0.996
16	15	0	0.000	510	0.996
17	16	0	0.000	510	0.996
18	17	1	0.002	511	0.998
19	18	0	0.000	511	0.998
20	19	0	0.000	511	0.998
21	20	0	0.000	511	0.998
22	21	1	0.002	512	1.000
Totaux		512	1		

[1] En anglais: cumulative frequency.

I.1.2. Les diagrammes

Différents types de ***diagrammes*** permettent de présenter graphiquement ces distributions.
Les distributions de fréquences non cumulées (absolues ou relatives) sont représentées par des diagrammes en bâtons, par des histogrammes ou par des polygones de fréquences. Les valeurs observées sont portées en abscisses et les fréquences en ordonnées.
Les ***diagrammes en bâtons*** [1] sont établis en traçant parallèlement à l'axe des ordonnées, et en regard de chaque valeur observée x_i, un segment de longueur égale à la fréquence de cette valeur. La somme des longueurs des bâtons est donc égale à n (pour les fréquences absolues) ou à 1 (pour les fréquences relatives).

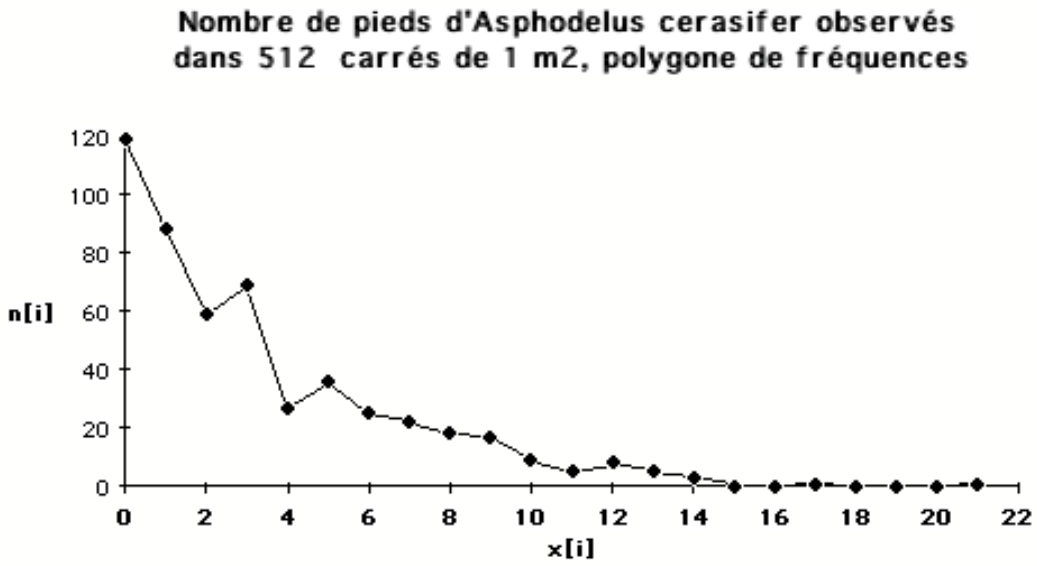
Figure I



Les ***polygones de fréquences*** sont construits en joignant par une ligne brisée les extrémités des segments voisins des diagrammes en bâtons.

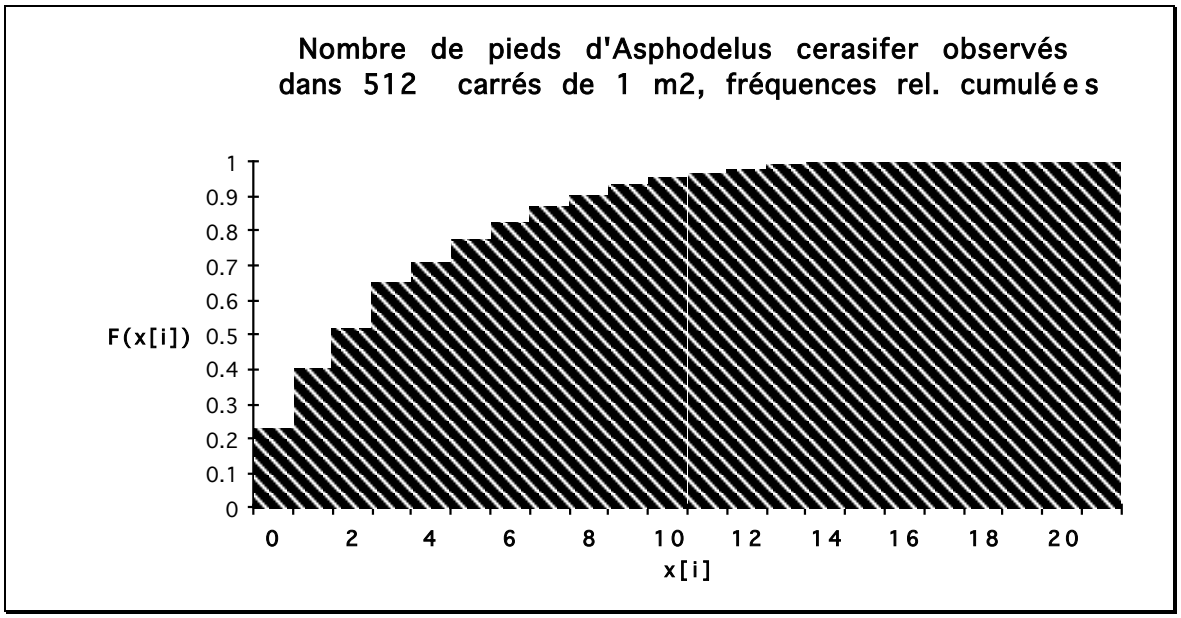
[1] En anglais: bar diagram

Figure II



Le ***polygone de fréquence cumulée*** est construit « en escalier ». On dessine des segments de droites de longueurs proportionnelles aux fréquences, mais en les décalant progressivement vers le haut, de telle sorte que l'origine de chacun d'eux soit située à hauteur de l'extrémité du précédent; on joint ensuite les différents segments verticaux par des segments horizontaux.

Figure III



A chaque point de l'axe des abscisses correspond sur le polygone de fréquences un point dont l'ordonnée indique, en valeur absolue ou relative, le nombre d'observations inférieures ou égales à l'abscisse considérée.

I.2. Les distributions groupées

Quand le nombre de valeurs observées distinctes est élevé, il est souvent nécessaire de condenser encore les tableaux statistiques, en groupant les observations en ***classes***.

I.2.1. Les tableaux

Chacune des classes est caractérisée par ses ***limites*** . La limite inférieure d'une classe, confondue avec la limite supérieure de la classe précédente, est située à mi–chemin entre la valeur la plus petite de la classe considérée et la valeur la plus élevée de la classe précédente.

L'écart entre les limites des classes est appelé ***l'intervalle de classe*** . Lorsque cet intervalle est constant, chacune des classes peut également être caractérisée par son ***point central*** situé à mi–chemin entre les limites de classes correspondantes.

<u>Exemple</u>

Tableau II

Etendue des 198'706 exploitations agricoles belges de plus de 1 ha, recensement 1959, distribution groupées en classes d'intervalles variables

Etendues des exploitations (en hectares) x_i	Limites de classes	Fréquences absolues n_i	Fréquences relatives f_i	Fréquences relatives cumulées F_i	Fréquences unitaires f'_i
	1				
1– 3 ha	3	58'122	0.2925	0.2925	0.1463
3–5 ha	5	38'221	0.1923	0.4849	0.0962
5–10 ha	10	52'684	0.2651	0.7500	0.0530
10–20 ha	20	35'188	0.1771	0.9271	0.0177
20–30 ha	30	8'344	0.0420	0.9691	0.0042
30–50 ha	50	3'965	0.0200	0.9890	0.0010
50–100 ha	100	1'873	0.0094	0.9984	0.0002
> 100 ha	>100	309	0.0016	1.0000	
Totaux		198'706	1.0000		

Lors de la répartition des observations en classes, l'intervalle de classe, souvent constant, et le nombre de classes , généralement compris entre 10 et 20, sont déterminés de manière à donner à la distribution une allure aussi régulière que possible, sans en dissimuler les caractéristiques essentielles. Lorsque les fréquences varient considérablement d'une extrémité à l'autre de la distribution, on utilise parfois un intervalle de classe variable , de manière à diminuer les écarts entre les fréquences des différentes classes.
Dans certains cas aussi, la limite inférieure de la première classe ou la limite supérieure de la dernière classe n'est pas précisée. Une telle classe est appelée ***classe ouverte*** .

Les fréquences, absolues et relatives, sont fonction des intervalles de classe. Pour éviter cet inconvénient, on peut utiliser des nombres purs, comparables d'une classe à l'autre et d'une distribution à l'autre, quels que soient les effectifs considérés et les intervalles de classe, en divisant les fréquences relatives par les intervalles correspondants. Les quotients ainsi obtenus:

$$\boxed{f'_i = \text{Erreur !}}$$

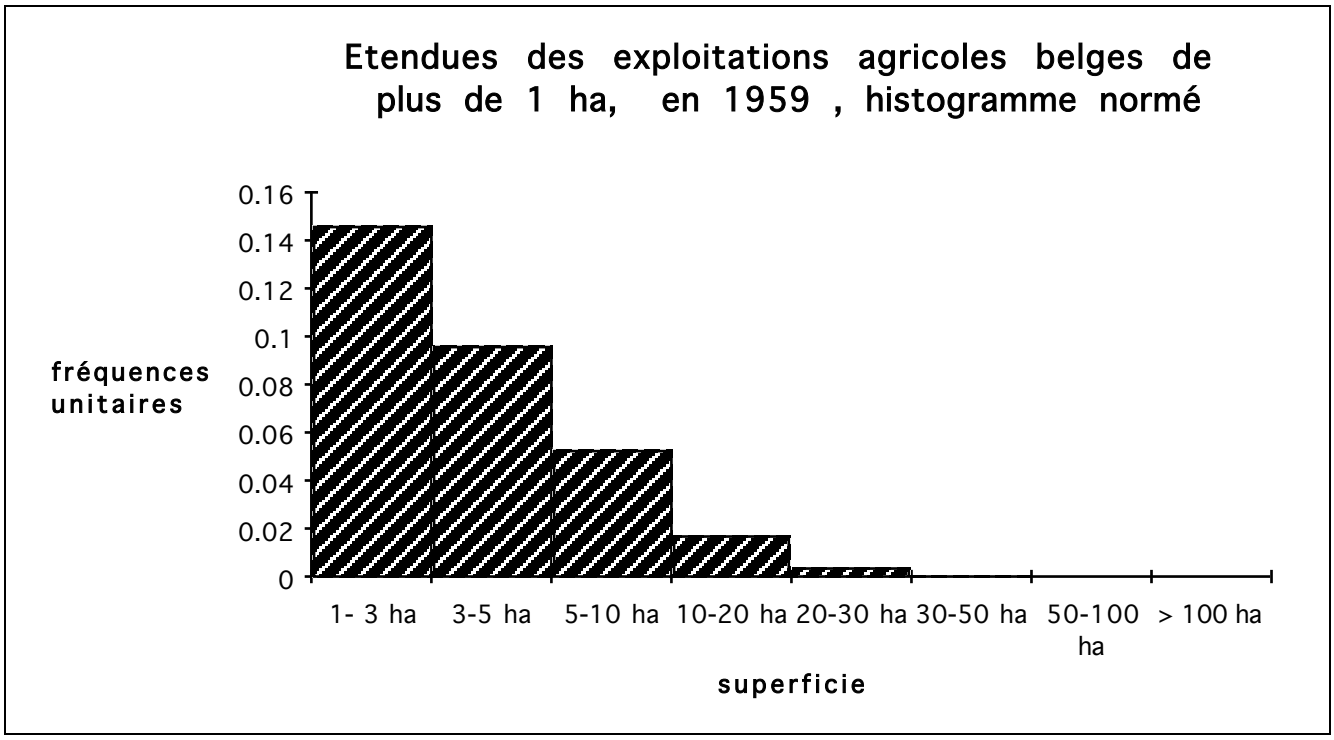
sont appelés ***fréquences unitaires***.

I.2.2. Les diagrammes

Les ***histogrammes*** se composent de rectangles contigus dont les intervalles de classes sont les bases et les fréquences les hauteurs, de sorte que les ***aires*** des rectangles soient proportionnelles aux fréquences. Ce type de graphique est applicable surtout au cas des distributions groupées.

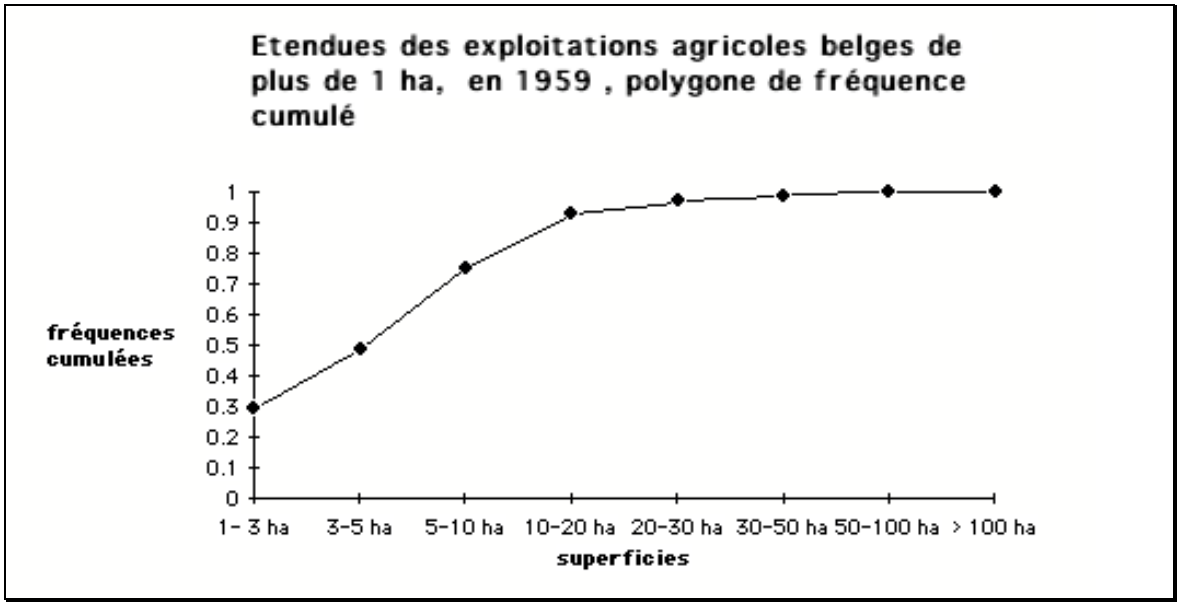
Lorsque l'intervalle de classe est variable, il est indispensable de porter en ordonnées les fréquences unitaires n'_i , de telle sorte que **les aires des rectangles des histogrammes restent proportionnelles aux fréquences des classes**. De tels histogrammes sont dits ***normés*** .

Figure IV



Le ***polygone de fréquence cumulé*** s'obtient en joignant par une ligne brisée les points obtenus en portant, en regard des limites supérieures de classes, des ordonnées égales aux fréquences cumulées, absolues ou relatives.

Figure V



A chaque point de l'axe des abscisses correspond sur le polygone de fréquences un point dont l'ordonnée indique, en valeur absolue ou relative, le nombre d'observations inférieures ou égales à l'abscisse considérée.

Les ***histogrammes de fréquences cumulées*** sont construits comme les histogrammes de fréquences, **les hauteurs des rectangles étant cette fois proportionnelles aux fréquences cumulées**.

I.3. Les paramètres

Les données peuvent être condensées sous la forme de ***paramètres***, les plus couramment utilisés sont:

1. Les ***paramètres de position***:: la **moyenne arithmétique** et la **médiane**, .

2. Les ***paramètres de dispersion*** , qui permettent de chiffrer la variabilité des valeurs observées, autour d'un paramètre de position: la **variance**, l'**écart–type** et l'**amplitude**.

I.3.1. Les paramètres de position

La ***moyenne arithmétique***, ou tout simplement ***moyenne*** [1], est désignée par le symbole $\bar{x}$ (xbarre).

Elle est égale à la somme des valeurs observées multipliées par leurs fréquences absolues et divisée par le nombre d'observations:

$$\bar{x} = \frac{n_1x_1+n_2x_2+...+n_ix_i+...+n_px_p}{n} = \frac{n_1}{n}x_1 + ...+ \frac{n_p}{n}x_p = f_1x_1 + ...+ f_px_p.$$

$$\boxed{\bar{x} = \frac{1}{n}\sum_{i=1}^{i=p} n_ix_i = \sum_{i=1}^{i=p} f_ix_i}$$

C'est la **moyenne pondérée** des valeurs x_i affectées des "poids" n_i.

Pour les distributions groupées, on remplace les valeurs observées par les points centraux des classes, (on commet ainsi une certaine erreur et ça ne marche pas s'il y a une classe ouverte).

[1] En anglais: arithmetic mean, mean, average.

La moyenne jouit des propriétés suivantes:

- La somme des écarts $x_i - \bar{x}$ entre les valeurs observées et la moyenne est nulle.
- Un changement d'origine des données $x_i' = a + x_i$ induit une translation des moyennes:

$$\bar{x}' = a + \bar{x} .$$

- Un changement d'échelle des données $x_i' = bx_i$ induit un changement d'échelle des moyennes:

$$\bar{x}' = b\,\bar{x} .$$

La ***médiane*** [1] $\tilde{\mathbf{x}}$ (xtilda) est un un nombre tel que la moitié des observations x_i lui sont inférieures ou égales et la moitié supérieures ou égales.

Lorsque le nombre d'observations n est **impair**, la médiane est l'observation numéro (n+1)/2:

$$\boxed{\tilde{\mathbf{x}} = \mathbf{x}_{(n+1)/2}}$$

Lorsque n est **pair**, on convient généralement de prendre comme valeur de la médiane la moyenne arithmétique des deux observations de numéros n/2 et 1 + n/2:

$$\boxed{\tilde{\mathbf{x}} = \frac{1}{2}(\mathbf{x}_{n/2} + \mathbf{x}_{1+n/2})}$$

<u>Exemple</u> Nombre de pieds d'Asphodelus cerasifer observés dans 512 carrés de 1 m2 (Tableau I):

$$\tilde{x} = 2, \text{ car: } x_{256} = x_{257} = 2.$$

La médiane peut également être déterminée graphiquement en recherchant sur le polygone de fréquences cumulées absolues (resp relatives) l'abscisse du point d'ordonnée n/2 (resp 1/2).

Pour les distributions groupées, la classe médiane est celle qui contient la médiane.

La médiane est égale à la moyenne dans le cas d'une distribution symétrique.

---

[1] En anglsid: median.

En raison des propriétés précédentes, la moyenne très utilisée.
Mais le calcul de la médiane reste possible, même en présence de classes ouvertes, et il ne dépend pas des observations extrêmes, qui peuvent être parfois très anormales ou douteuses.

I.3.2. Les paramètres de dispersion

La ***variance*** [1] d'une distribution de fréquences est la moyenne des carrés des écarts à la moyenne:

$$\boxed{s^2 = \frac{1}{n}\sum_{i=1}^{i=p} n_i(x_i-\bar{x})^2 = \sum_{i=1}^{i=p} f_i(x_i-\bar{x})^2}$$

L'***écart–type*** [2] **s** (ou écart quadratique moyen ou déviation standard) est la racine carrée de la variance.

La variance et l'écart–type jouissent des propriétés suivantes:

- Ils sont nuls si et seulement si tous les écarts $x_i - \bar{x}$ sont nuls:

$$x_1 = x_2 = ... = x_n = \bar{x}\ .$$

- Un changement d'origine des données $x_i' = a + x_i$ ne les modifie pas:

$$s_{x'}^2 = s_x^2$$

- Un changement d'échelle des données $x_i' = bx_i$ induit des changements d'échelles:

$$s_{x'}^2 = b^2 s_x^2 \quad \text{et} \quad s_{x'} = |b|\, s_x$$

- L'intérêt de l'écart–type est qu'il possède la même unité que celle des données.

<u>Exemple:</u> Tableau I

La moyenne vaut ≈ 3,3 et l'écart–type vaut ≈ 3,43

---

[1] En anglais: variance.
Certaines calculettes donnent la variance corrigée s' qui vérifie la relation $(n-1)s'^2 = ns^2$. Dès que n est assez grand, les deux variances ne diffèrent que de très peu.

[2] En anglais: standard deviation.

Exemple Hauteur totale (en mètres) de 12 arbres::

20,4	25,4	25,6	25,6	26,6	28,6	28,7	29,0	29,8	30,5	30,9	31,1

$\overline{x} = 27,68$ $s^2 = 8,85$ $s = 2,97$

II. STATISTIQUE DESCRIPTIVE A 2 DIMENSIONS

La statistique descriptive à deux dimensions a essentiellement pour but de mettre en évidence les relations qui existent entre deux séries d'observations considérées **simultanément**.

II.1. Les distributions de fréquence

Les observations relatives à deux variables et à plusieurs individus (n par exemple), se présentent le plus simplement sous la forme d'une ***série statistique double*** , c'est–à–dire de la suite des n couples de valeurs observées (x_i,y_i),

$$x_1, x_2,...,x_i,..., x_n \qquad \text{et} \qquad y_1, y_2,...,y_i,...,y_n.$$

<u>Exemple</u>

TABLEAU III

Longueur du corps (x) et profondeur de poitrine (y) de 22 vaches laitières, en centimètres.

x	y	x	y
168	71	176	74
169	68	159	70
150	65	159	73
148	67	151	69
154	67	155	71
145	66	169	74
165	69	158	70
163	69	157	71
148	68	161	73
161	69	146	71
151	70	150	65

Lorsque *n* est grand, il est nécessaire de condenser les données en une distribution de fréquences. Celle–ci se présente sous la forme d'un tableau statistique à double entrée [1], dont une ligne est réservée par exemple à chaque valeur observée x_j de x et une colonne à chaque valeur observée y_i de y

[1] En anglais: bivariate frequency table, two–way table.

(tableau IV). Les **fréquences absolues n_{ij}** représentent, pour chaque cellule du tableau, le nombre de couples d'observations (x_j, y_i). On parle d'une distribution de fréquences à deux dimensions [1].

TABLEAU IV

Représentation schématique d'une distribution de fréquences à deux dimensions.

x_j / y_i	x_1	x_2	...	x_j	...	x_q	TOTAUX
y_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	$n_{1.}$
y_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	$n_{2.}$
...	...	...	...	...	...	...	...
y_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	$n_{i.}$
...	...	...	...	...	...	...	...
y_p	n_{p1}	n_{p2}	...	n_{pj}	...	n_{pq}	$n_{p.}$
TOTAUX	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.q}$	n

Lorsque le nombre de lignes et ou de colonnes est encore trop grand, il est possible de condenser les observations en une distribution groupée, en réunissant en classes les diverses valeurs observées de x et de y. Les symboles x_j et y_i représentent alors les points centraux des classes. En général, les intervalles de classe sont choisis de telle sorte que le nombre de classes ne dépasse pas 10 ou 15 pour chacune des deux variables.

En calculant les sommes relatives aux différentes lignes ou aux différentes colonnes, on obtient les ***fréquences marginales*** $n_{i.}$ et $n_{.j}$ définies comme suit[2]:

$n_{i.}$ = **Erreur !** $n_{.j}$ = **Erreur !**

Un point remplace dans chaque cas l'indice suivant lequel on a sommé.

[1] En anglais: bivariate distribution.

[2] En anglais: marginal frequency

Ces fréquences sont liées par les relations:

$$n = \sum_{i=1}^{i=p} n_{i.} = \sum_{j=1}^{j=q} n_{.j} = \textbf{Erreur !Erreur !}$$

Exemple

Tableau V

Poids des feuilles (x) et des racines (y) de 1000 individus de Cichorium intybus, en grammes: distribution de fréquences, moyennes et variances conditionnelles.

y_i \ x_j	59.5	99.5	139.5	179.5	219.5	259.5	299.5	339.5	Totaux $n_{i.}$	$\bar{x}\ \mid y_i$	$s^2_{x\ \mid y_i}$
39.5	2								**2**	60	3677
119.5	49	46	5	2					**102**	84	2046
199.5	86	137	46	11					**280**	97	1538
279.5	27	153	89	25	7				**301**	117	1179
359.5	5	45	91	40	6				**187**	139	1454
439.5		10	33	21	16	1	1		**82**	164	3661
519.5		1	4	11	10	3			**29**	193	6816
599.5			2	1	2	4		1	**10**	228	14982
679.5				1		3	2		**6**	260	21021
759.5	0	0	0	0	1	0	0	0	**1**	220	9872
Totaux $n_{.j}$	**169**	**392**	**270**	**112**	**42**	**11**	**3**	**1**	**1000**		
$\bar{y}\ \mid x_j$	192	247	315	357	436	585	600	600	281		
$s^2_{y\ \mid x_j}$	11890	6905	8328	15876	34147	97740	113975	101175	11696		

Dans ce tableau $\Delta x = 40$, $\Delta y = 80$, ainsi, par exemple, 59.5 représente le milieu de la classe $40 \leq x \leq 79$ et 679.5 celui de la classe $640 \leq y \leq 719$.

En négligeant un des caractères, on obtient des tableaux à une variable analogues à ceux du chapitre 1. Ce sont les **tableaux marginaux**.

Tableau Va

Poids des racines de 1000 individus de Cichorium intybus, en grammes: distribution de fréquence à une dimension.

Racines	59.5	99.5	139.5	179.5	219.5	259.5	299.5	339.5	**Total**
Totaux	**169**	**392**	**270**	**112**	**42**	**11**	**3**	**1**	**1000**

Tableau Vb

Poids des feuilles de 1000 individus de Cichorium intybus, en grammes: distribution de fréquence à une dimension.

Feuilles	Totaux
39.5	2
119.5	102
199.5	280
279.5	301
359.5	187
439.5	82
519.5	29
599.5	10
679.5	6
759.5	1
Total	**1000**

II.2. Les représentations graphiques

Dans la figure VI ci–dessous, la série statistique double est représentée sous la forme d'un ***nuage de points*** [1].

Ceux–ci sont obtenus en représentant chaque couple d'observations (x_j, y_i) par un point dans le plan Oxy. Si certains couples apparaissent plusieurs fois, la représentation n'est plus fiable.

[1] En anglais: scatter diagram.

Les échelles des diagrammes de dispersion sont généralement choisies de manière à donner à ces diagrammes une forme approximativement carrée. Dans certains cas, il peut être utile dans ce but d'employer des échelles non linéaires (logarithmiques par exemple).

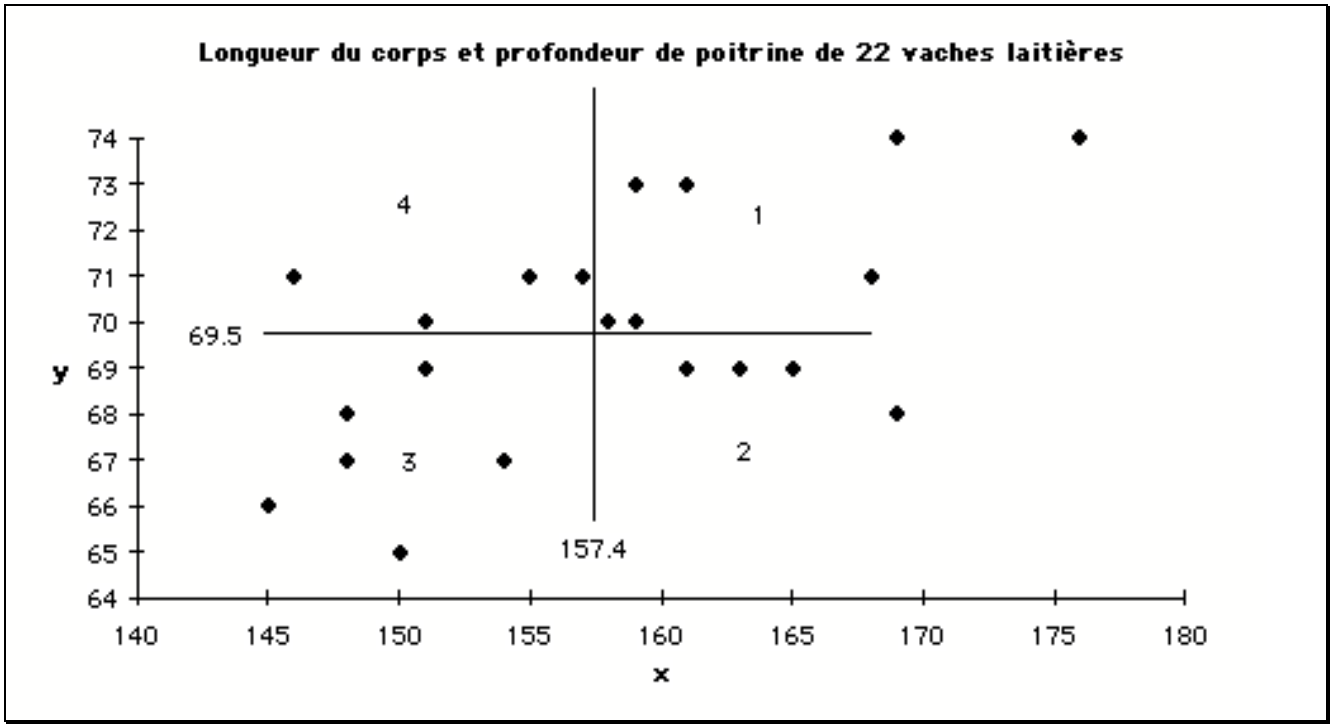
Les distributions de fréquences à deux dimensions peuvent être représentées dans un espace à trois dimensions sous la forme de diagrammes en bâtons (surtout pour les distributions non groupées) ou de stéréogrammes (pour les distributions groupées).

Les ***diagrammes en bâtons*** sont établis en traçant perpendiculairement au plan (x,y) et en chaque point (x_j, y_i) un segment de longueur égale à n_{ij} ou f_{ij}.
La somme des longueurs est alors égale à n ou à 1.

Les ***stéréogrammes*** sont composés de parallélépipèdes rectangles (ou de prismes carrés) juxtaposés, dont les bases correspondent chacune à une cellule du tableau statistique à double entrée et dont les hauteurs sont égales ou proportionnelles aux fréquences absolues ou relatives. Si on prend comme hauteur les fréquences unitaires, c'est–à–dire les fréquences relatives divisées par les hauteurs des bases $(f_{ij}/\Delta x \Delta y)$, le volume total des parallélépipèdes rectangles vaut 1 et le stéréogramme est dit ***normé***.

Exemple

Figure VI



II.3. Les paramètres

Les paramètres utilisés pour caractériser les séries statistiques doubles et les distributions de fréquences à deux dimensions sont de deux types. Les uns ne concernent qu'une variable à la fois (**paramètres marginaux**, pour "Cichorium intybus", ils sont calculés à partir des tableaux Va et Vb:

- Les ***moyennes marginales*** [1]: $\bar{y} = \frac{1}{n}\sum_{i=1}^{i=p} n_{i.}y_i$ et $\bar{x} = \frac{1}{n}\sum_{j=1}^{j=q} n_{.j}x_j$.

- Les ***variances marginales*** [2]: $s_y^2 = \frac{1}{n}\sum_{i=1}^{i=p} n_{i.}(y_i-\bar{y})^2$ et $s_x^2 = \frac{1}{n}\sum_{j=1}^{j=q} n_{.j}(x_j-\bar{x})^2$.

[1] En anglais: marginal mean.

[2] En anglais: marginal variance.

Les autres paramètres servent à décrire les relations existant entre les deux séries d'observations:

- Les ***moyennes conditionnelles*** [1]: $\bar{y}\,|x_j = \frac{1}{n_{.j}} \sum_{i=1}^{i=p} n_{ij} y_i =$ moyenne de la colonne j.

$$\bar{x}\,|y_i = \frac{1}{n_{i.}} \sum_{j=1}^{j=q} n_{ij} x_j = \text{moyenne de la ligne i.}$$

- Les ***variances conditionnelles*** [2]: $s^2_{y\,|x_j} = \frac{1}{n_{.j}} \sum_{i=1}^{p} n_{ij}(y_i - \bar{y})^2 =$ variance de la colonne j.

$$s^2_{x\,|y_i} = \frac{1}{n_{i.}} \sum_{j=1}^{q} n_{ij}(x_j - \bar{x})^2 = \text{variance de la ligne i.}$$

L'étude simultanée des deux séries d'observations fait intervenir diverses notions nouvelles, dont les principales sont: la ***covariance***, les ***droites de régression*** et le ***coefficient de corrélation***..

La ***covariance*** $\mathbf{s_{xy}}$ **caractérise simultanément les deux séries d'observations**:

$$\boxed{\mathbf{s_{xy} = \frac{1}{n}\ \text{Erreur !Erreur !}}}$$

<u>Exemple</u>

Pour cichorium intybus, on trouve $\bar{x} = 281$, $\bar{y} = 120$, $s^2_x = 11696$, $s^2_y = 2057$ et $s_{xy} = 3231$.

La covariance est **positive** ou **négative** selon que la relation entre les deux séries de données est croissante ou décroissante, c'est–à–dire selon que les valeurs élevées d'une série correspondent, dans l'ensemble, aux valeurs élevées ou aux valeurs peu élevées de l'autre.
En effet les droites $x = \bar{x}$ et $y = \bar{y}$ divisent le plan Oxy en quatre régions (Voir figure VI). Les valeurs observées supérieures aux deux moyennes ou inférieures aux deux moyennes apportent une

[1] En anglais: conditional mean.

[2] En anglais: conditional variance.

contribution positive à la covariance; alors que les valeurs observées supérieures à une moyenne et inférieures à l'autre lui apportent une contribution négative.

La covariance est donc positive ou négative selon que le nuage de points a une allure générale croissante ou décroissante.

La covariance est nulle ou presque nulle lorsqu'il y a compensation entre les deux catégories de points, il n'y a alors aucun lien entre les deux caractères observés .

Exemple Longueur du corps et profondeur de poitrine de 22 vaches laitières (Tableau III):

Les moyennes observées sont, en centimètres:

$$\bar{x} = 157{,}4 \qquad \text{et} \qquad \bar{y} = 69{,}5$$

Les couples (x_i,y_i) tels que:

$$x_i > 157{,}4 \text{ et } y_i > 69{,}5$$

ou

$$x_i < 157{,}4 \text{ et } y_i < 69{,}5$$

apportent donc une contribution positive à la covariance (régions 1 et 3 de la figure VI). Les autres valeurs observées, qui sont telles que:

$$x_i > 157{,}4 \text{ et } y_i < 69{,}5$$

ou

$$x_i < 157{,}4 \text{ et } y_i > 69{,}5$$

lui apportent une contribution négative (régions 2 et 4).

La figure VI montre qu'il y a prédominance des premières valeurs par rapport aux secondes. Il n'est pas surprenant, dans ces conditions, d'obtenir pour la covariance une valeur positive:

$$s_{xy} = 12{,}73.$$

Cette quantité serait négative si la prédominance se marquait dans l'autre sens. Elle serait nulle ou presque nulle si les points se répartissaient uniformément ou presque uniformément dans les quatre régions.

Comme les variances, la covariance est influencée par les **changements d'échelles**, mais non par les **changements d'origines**.

Si $x' = a + bx$ et si $y' = c + dy$, on a en effet: $s_{x'y'} = b\,d\,s_{xy}$.

II.4. Les courbes de régression

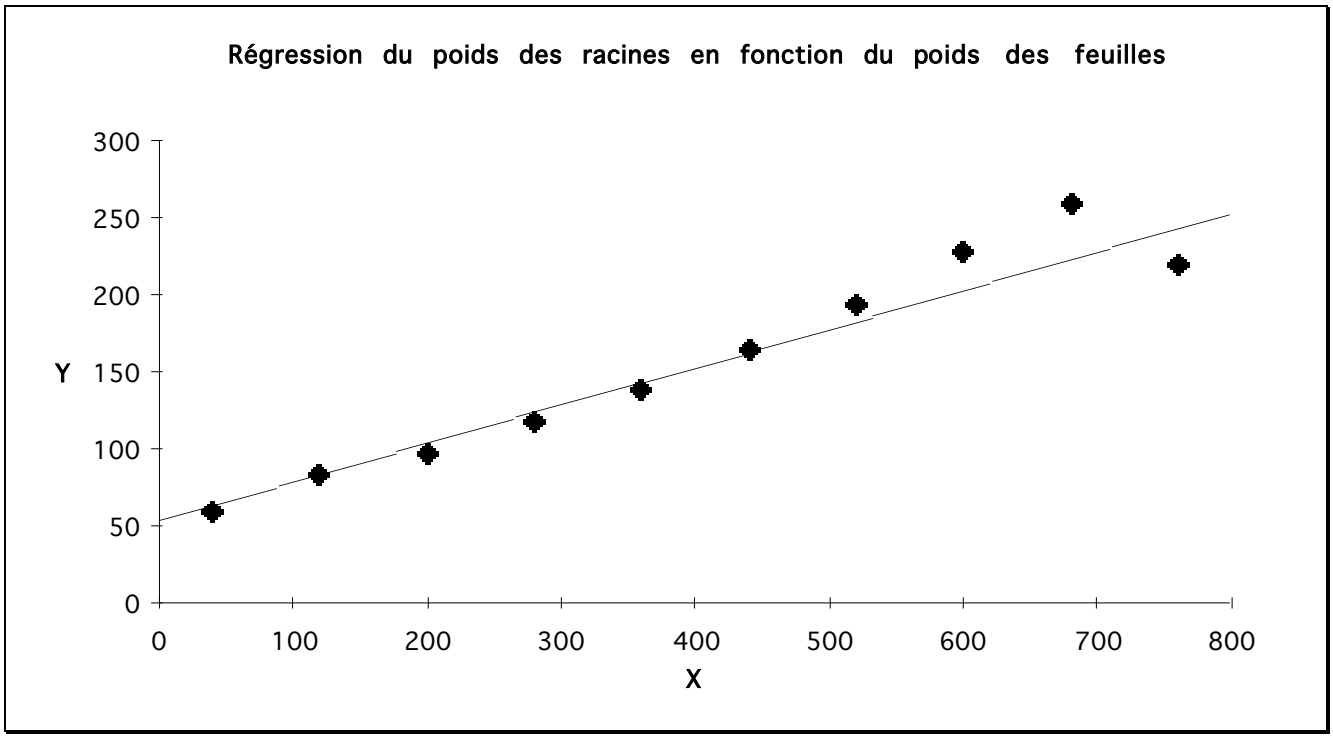
II.4.1 La droite de régression de y en x

Exemple

Poids des feuilles (x) et des racines (y) de 1000 individus de Cichorium intybus, en grammes:

Figure VII

On a représenté en abscisse les valeurs x_j et en ordonnée les moyennes conditionnelles $\bar{y}\,|x_j$.



Lorsque qu'on pense qu'il y a une relation linéaire (du premier degré) entre les variables observées

$$y = a + bx$$

Le problème est alors le suivant:

Les points mesurés $(x_j\ ;\ y_j)$ n'étant pas exactement alignés, par exemple, à cause des erreurs de mesure sur les y_i, comment trouver la meilleure droite?

- Une première tentative est de la dessiner "à l'oeil".
- Une approche calculatoire est la méthode des moindres carrés.(Voir ci–dessous)

Remarque

Si plusieurs points ont les mêmes coordonnées, on prendra comme nouveaux points les points moyens conditionnels $(x_j\ ;\ \bar{y}\ |x_j)$.

Méthode des moindres carrés:

La droite (D): y = a + bx obtenue par cette méthode est la ***droite de régression de y en x*** [1].

Elle est déterminée de manière à rendre minimale la somme des carrés des écarts entre les points observés et les points correspondants de la droite.

Pour simplifier les notations, supposons que le nuage de points soit donné par le tableau $(x_i;y_i)$.
y_i est l'ordonnée mesurée, tandis que $a + bx_i$ sera l'ordonnée calculée.

La somme des carrés des écarts à minimiser est alors:

$$S = \sum_{i=1}^{i=n} (y_i - a - bx_i))^2.$$

Calcul de a et b:

$$S = \sum_{i=1}^{n} [y_i - \bar{y} - b(x_i - \bar{x}) - (a - \bar{y} + b^{-};x)]^2 = \sum_{i=1}^{n} [y_i -^{-};y - b(x_i - \bar{x})]^2 + n(a - \bar{y} + b\bar{x})^2 - 2(a - \bar{y} + b\bar{x})$$

$$\sum_{i=1}^{n} [y_i -^{-};y - b(x_i - \bar{x})] \ .$$

Ce dernier terme est nul en raison des propriétés de la moyenne. Ainsi:

$$S = \sum_{i=1}^{n} (y_i - \bar{y})^2 + b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2b \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) + n(a - \bar{y} + b\bar{x})^2 =$$

$$n[s_y^2 + b^2 s_x^2 - 2bs_{xy} + (a - \bar{y} + b\bar{x})^2] = n[(s_y^2\, s_x^2 - s_{xy}^2)/s_x^2 + s_x^2(b - s_{xy}/s_x^2)^2 + (a - \bar{y} + b\bar{x})^2].$$

La valeur de S sera minimale lorsque

[1] En anglais: regression line.

$$\boxed{b = s_{xy}/s_x^2} \qquad \boxed{a = \bar{y} - b\bar{x}}$$

Ce minimum vaut alors S_{min} = **Erreur !** .

a est l'**ordonnée à l'origine** et b le **coefficient de régression de y en x**.

La droite de régression de y en x passe par le **point moyen ($\bar{x}$,$\bar{y}$)** , son équation peut s'écrire:

$$\boxed{(D_{yx}): \quad y - \bar{y} = b(x - \bar{x})}$$

<u>Exemple</u>

Reprenons l'exemple précédent et calculons l'équation de la droite de régression de y en x:

On trouve $\bar{x}$ = 281,4 et $\bar{y}$ = 120,1, s_x^2 = 11,690 et s_{xy} = 3,232.

Nous en déduisons que (D_{yx}): y = 0,2765x + 42,3.

<u>Exemple</u>

Pour cichorium intybus, on trouve y = 0.276 x + 42.4.

II.4.2 La régression par l'origine

Dans certains cas, on impose à la droite de régression de passer par l'origine.
L'équation est alors:

$$y = b\,x$$

Le coefficient b devant assurer le minimum de la somme des carrés des écarts est donné par:

b = **Erreur !**

Remarquons que ce n'est pas la même formule que celle donnat la pente de la droite de régression de y en x. Cette droite **ne passe généralement pas par le point moyen ($\bar{x}$,$\bar{y}$).**

II.4.3. La droite de régression de x en y

La ***droite de régression de x en y*** s'obtient en minimisant la somme des carrés des écarts parallèlement à l'axe des abscisses Ox. Cette droite admet l'équation:

$$(D_{xy}): \qquad x - \bar{x} = b_{xy}(y - \bar{y}), \text{ où } b_{xy} = s_{xy}/s_y^2$$

b_{xy} est le ***coefficient de régression de x en y***:

En général, cette droite est différente de la droite de régression de y en x.
On choisira celle-ci si les erreurs de mesure sont plus importantes sur la variable x que sur la variable y.

Finalement, il existe aussi la droite de régression orthogonale qui minimise la distance des points à la droite.
Un autre type de droite minimise la somme des valeurs absolues des écarts.
Son avantage est d'être moins sensible à un point très mal mesuré.

<u>Exemple</u>

Pour cichorium intybus, on trouve x = 1.57 y + 92.74.

II.4.4. Le coefficient de corrélation r

On prouve facilement que la covariance est toujours, en valeur absolue, inférieure ou égale au produit des écarts–types:

$$|s_{xy}| \leq s_x s_y.$$

Et que l'égalité a lieu si et seulement si tous les points observés sont sur la droite de régression de y en x:

Le ***coefficient de corrélation*** **r** est défini par:

r =
Erreur !

et on doit évidemment supposer pour le définir que s_x et s_y sont tous deux différents de zéro.

Cette définition montre que **le coefficient de corrélation possède le même signe que la covariance et qu'il est toujours compris entre —1 et + 1**. Il ne peut d'ailleurs être égal à 1, en valeur absolue, que si tous les points observés se trouvent sur une même droite non parallèle aux axes de coordonnées.

Tenant compte également du signe de la covariance, on en conclut que:

r = 1	tous les points observés se trouvent sur une même droite de coefficient angulaire positif
r ≈ 1	tous les points observés sont situés a proximité d'une telle droite
0 < r < 1	le nuage de points est allongé parallèlement à une telle droite
r ≈ 0	le nuage de points est allongé parallèlement à l'un des axes de coordonnées ou s'il a une forme arrondie
–1 < r < 0	le nuage de points est allongé parallèlement à une droite de coefficient angulaire négatif
r ≈ –1	tous les points observés se trouvent a proximité d'une telle droite
r = –1	tous les points observés sont situés exactement sur une telle droite

Le coefficient de corrélation mesure donc la **netteté de la liaison** existant entre les deux séries d'observations, pour autant que cette liaison soit linéaire ou approximativement linéaire.

D'autre part, on peut facilement démontrer que le coefficient de corrélation est indépendant de toute transformation linéaire positive, c'est–à–dire que:

$$r_{x'y'} = r_{xy}$$

lorsque:

$$x' = a + bx \quad \text{et} \quad y' = c + dy \ (b \text{ et } d > 0).$$

Contrairement à la variance et à la covariance, le coefficient de corrélation est un nombre pur, indépendant des unités de mesure.

Les équations des droites de régression peuvent s'écrire:

$$(D_{yx}):\ \frac{y-\bar{y}}{s_y} = r\,\frac{x-\bar{x}}{s_x} \qquad \text{et} \qquad (D_{xy}):\ \frac{x-\bar{x}}{s_x} = r\,\frac{y-\bar{y}}{s_y}$$

Exemple

Poids des feuilles et des racines de 1.000 individus de *Cichorium intybus:* .

$$r = 0,66.$$

Cette valeur est l'indice d'une relation positive entre les deux variables considérées.

Malgré l'existence d'une relation assez étroite entre les deux variables, la prévision qui pourrait être faite pour l'une d'entre elles en fonction de l'autre serait peu précise.

II.4.5. Corrélation et causalité

Afin d'éviter toute interprétation abusive, il ne faut pas perdre de vue que l'existence d'une corrélation même élevée entre deux séries d'observations n'implique pas nécessairement l'existence d'une relation de cause à effet entre les deux variables considérées. Souvent en effet, les corrélations observées sont dues au fait que les variables étudiées sont toutes deux soumises a des influences communes, qui peuvent en modifier simultanément les valeurs, soit dans le même sens (corrélation positive), soit en sens opposés (corrélation négative).

Pour un ensemble d'exploitations agricoles, par exemple, l'existence d'une corrélation positive entre le revenu par unité de main–d'oeuvre et l'importance de la main–d'oeuvre (nombre d'ouvriers) n'implique pas qu'il suffit d'augmenter l'importance de la main–d'oeuvre pour améliorer la rentabilité de l'exploitation! En réalité, ces deux variables sont elles–mêmes fonction d'une troisième caractéristique importante: l'étendue des exploitations. La corrélation observée provient du fait que les grandes exploitations nécessitent normalement une main–d'oeuvre plus nombreuse et qu'elles sont aussi économiquement les plus rentables.

2.10. La régression curvilinéaire

2.10.1. Généralites

Dans de nombreux problèmes, une relation nette apparaît entre les variables étudiées, sans que cette relation soit linéaire. Il peut alors être utile de procéder à l'ajustement[1] d'une courbe de régression au nuage des points observés. Dans certaines conditions, des tests de linéarité permettent de contrôler si l'ajustement est satisfaisant ou non.

[1] En anglais: fitting, smoothing.

Lorsqu'une courbe doit être ajustée, deux problèmes distincts se posent: d'une part, le choix de l'équation de cette courbe, c'est–à–dire le choix d'un certain type de fonction, et d'autre part, la détermination des paramètres intervenant dans cette équation.

2.10.2. Le choix d'une équation

Le choix d'une équation de régression peut être basé soit sur des considérations théoriques, soit sur des considérations empiriques. D'une facon générale, l'utilisation de bases théoriques est de loin préférable. On obtient en effet, de cette facon, des équations qui s'adaptent mieux à la réalité des problèmes étudiés, et des paramètres qui possèdent le plus souvent une signification précise.

Les considérations théoriques qui interviennent dans ce domaine conduisent fréquemment à la résolution d'équations différentielles.

Une relation du type

$$y = a\, e^{x} + b\, x^{2} + \textbf{Erreur !}$$

est dite **linéaire en les paramètres** a,b,c, car la fonction $f(a,b,c) = a\, e^{x} + b\, x^{2} +$ **Erreur !** vérifie les conditions de linéarité:

$f(a_1+a_2,b,c) = f(a_1,b,c) + f(a_2,b,c)$	$f(\lambda a,b,c) = \lambda\, f(a,b,c)$
$f(a,b_1+b_2,c) = f(a,b_1,c) + f(a,b_2,c)$	$f(a,\mu b,b) = \mu\, f(a,b,b)$
$f(a,b,c_1+c_2) = f(a,b,c_1) + f(a,b,c_2)$	$f(a,b,\nu\, c) = \nu\, f(a,b,c)$

Les logiciels utilisent en général le même algorithme pour déterminer les courbes de régression linéaire en les paramètres.
Dans Mathematica, par exemple, la commande Fit[data , {Exp[x] , x^2 , 1/x } , x] , où data est le tableau de points de mesures, résout le problème posé.

Une régression du type $y = e^{ax} + b\, x + c$, elle, n'est pas linéaire en les paramètres, les méthodes de résolution sont alors beaucoup plus délicates. (NonLinearFit dans Mathematica).

Certaines régressions peuvent être linéarisées en prenant le logarithme des données. (Voir ci-dessous).

Les commandes Regress et NonLinearRegress de Mathematica fournissent en plus une étude statistique de la qualité du modèle.

2.10.3. L'ajustement d'une fonction puissance

L'ajustement d'une fonction du type: $y = ax^b$ peut être réalisé en effectuant une **transformation logarithmique** $\log y = a' + b \cdot \log x$. On efectue ensuite une régression linéaire entre logx et logy.

Ce type de régression est couramment utilisé, en matière biologique comme en matière économique.

2.10.4. L'ajustement à une fonction exponentielle $y = c \cdot 10^{bx}$

L'ajustement d'une courbe exponentielle peut se faire grâce à un simple **changement de variable**, en remplacant les valeurs observées y_i par leurs logarithmes.

L'équation à rechercher est: $y = c \cdot 10^{bx}$, c'est–à–dire $\log y = a + bx$, où $a = \log(c)$.

En posant $\log y = y'$, on retrouve l'équation d'une droite: $y' = a + bx$.

<u>Exemple</u>

Tableau VII

Décroissance des résidus fongicides appliqués sur la laitue: Ajustement d'une fonction exponentielle log(y) = a+bx.

Dates x	Résidus y	log(y)
1.0	8.96	0.952
1	11.5	1.061
1	13.12	1.118
3	6.72	0.827
3	7.68	0.885
3	8	0.903
7	3.84	0.584
7	3.84	0.584
7	6.08	0.784
14	1.92	0.283
14	2.24	0.350
14	2.56	0.408

Figure VIII

Décroissance des résidus de fongicides appliqués sur laitue:
diagramme de dispersion et courbe exponentielle de régression

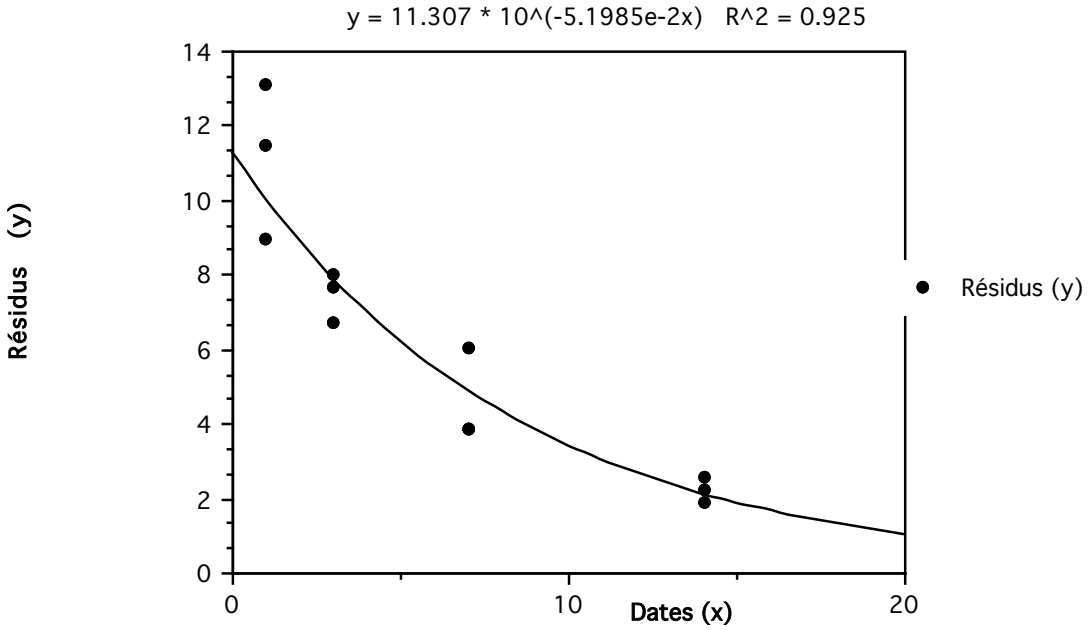
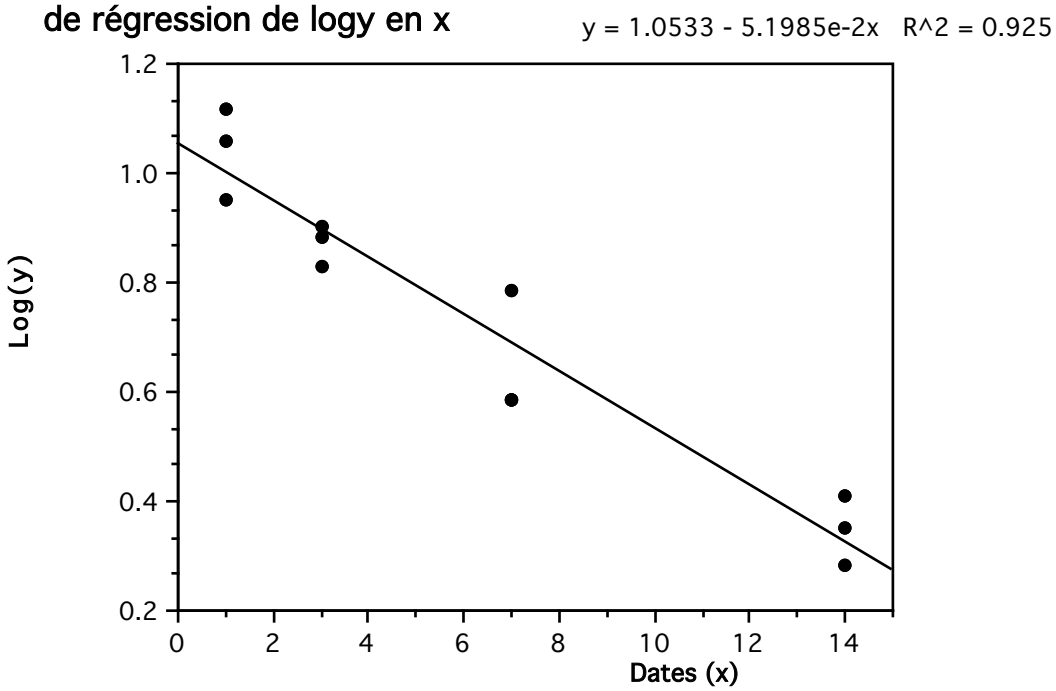


Figure IX

Décroissance des résidus de fongicides appliqués sur laitue:
diagramme semi-log de dispersion et droite
de régression de logy en x



L'équation de la droite est donc y = 1,053 – 0,052x, donc b = –0,052, a = 1,053.

L'équation de la courbe est alors $y = 10^{1,053} \cdot 10^{-0,052x}$.