CHAPITRE V

LES DISTRIBUTIONS STATISTIQUES A DEUX CARACTERES

V. A – Les tableaux statistiques

Considérons une population de N individus décrits simultanément suivant deux caractères A et B.

Supposons, dans un premier temps, que les caractères A et B sont des variables statistiques discrètes qu'on appellera X et Y (nous verrons que le traitement est le même pour les variables continues ou pour les caractères qualitatifs).

Désignons par $X_1, X_2, ..., X_i, ..., X_k$ les k modalités du caractère X et par $Y_1, Y_2, ..., Y_j, ..., Y_p$ les p modalités du caractère Y.

Soit n_{ij} le nombre des individus de la population qui présentent à la fois la modalité X_i du caractère X et la modalité Y_j du caractère Y.

Le tableau statistique décrivant les N individus est un tableau à double entrée où figurent en lignes les modalités de X et en colonnes les modalités de Y.
(tableau de k lignes et de p colonnes).

X / Y	Y_1	...	Y_j	...	Y_p	Total
X_1	n_{11}	...	n_{1j}	...	n_{1j}	$n_{1.}$
...	...	...	...	...	...	...
X_i	n_{i1}	...	n_{ij}	...	n_{ip}	$n_{i.}$
...	...	...	...	...	...	...
X_k	n_{k1}	...	n_{kj}	...	n_{kp}	$n_{k.}$
Total	$n_{.1}$	...	$n_{.j}$	...	$n_{.p}$	$N = n_{..}$

on désigne par un point une totalisation suivant l'indice i ou l'indice j

$n_{i.}$ est le total des effectifs n_{ij} suivant j

$n_{.j}$ est le total des effectifs n_{ij} suivant i

EXEMPLE :

Afin d'étudier la relation existante entre le nombre de feuilles et le nombre de fruits d'une certaine variété de fraises, 150 arbrisseaux ont été sélectionnés dans un champ. On a dénombré les feuilles et les fruits de chaque arbrisseau et le tableau suivant a été obtenu.

X/Y	6	11	14	16	18	Total
8	8	2	1	1	0	12
10	4	16	2	0	1	23
16	3	10	15	8	5	41
20	2	4	14	3	1	24
24	4	5	6	5	2	22
32	3	1	2	8	2	16
42	0	0	6	4	2	12
Total	24	38	46	29	13	150

X représente le nombre de fruits
Y représente le nombre de feuilles

Nous avons 7 valeurs différentes pour la variable X, $k = 7$
Nous avons 5 valeurs différentes pour la variable Y, $p = 5$

$n_{1.} = 12$ est le total des effectifs n_{1j} des arbrisseaux ayant 8 fruits (1ère ligne)
$n_{3.} = 41$ est le total des effectifs n_{3j} des arbrisseaux ayant 16 fruits (3ème ligne)

de même :

$n_{.1} = 24$ est le total des effectifs n_{i1} des arbrisseaux ayant 6 feuilles (1ère colonne)
$n_{.5} = 13$ est le total des effectifs n_{i5} des arbrisseaux ayant 18 feuilles (5ème colonne)

Tableaux particuliers :

La forme générale d'un tableau statistique à double entrée est celle que nous avons présenté plus haut ; mais il peut arriver que l'on soit dans des situations particulières (mais assez fréquentes) où chaque modalité X_i n'est observée qu'une seule fois. En plus de cela elle n'est observée que conjointement à une et une seule modalité Y_j. Autrement dit, les valeurs des variables X et Y sont appariées deux à deux.

Par exemple :

Nous disposons d'un groupe de 10 personnes sur chacune desquelles nous avons mesuré la taille X et le poids Y. Les mesures ont été les suivantes
(1.55 ; 58.2) (1.60 ; 58.1) (1.62 ; 61.3) (1.64 ; 65.3) (1.65 ; 69.5) (1.70 ; 69.7)
(1.72 ; 70.3) (1.73 ; 75.4) (1.76 ; 74.2) (1.78 ; 82.0)

Si nous disposons ces données dans un tableau statistique à double entrée nous obtenons

X/Y	58.1	58.2	61.3	65.3	69.5	69.7	70.3	74.2	75.4	82.0
1.55	0	1	0	0	0	0	0	0	0	0
1.60	1	0	0	0	0	0	0	0	0	0
1.62	0	0	1	0	0	0	0	0	0	0
1.64	0	0	0	1	0	0	0	0	0	0
1.65	0	0	0	0	1	0	0	0	0	0
1.70	0	0	0	0	0	1	0	0	0	0
1.72	0	0	0	0	0	0	1	0	0	0
1.73	0	0	0	0	0	0	0	0	1	0
1.76	0	0	0	0	0	0	0	1	0	0
1.78	0	0	0	0	0	0	0	0	0	1

Ici les effectifs ne sont égaux qu'à zéro ou à un (une seule fois par ligne ou par colonne).
Dans ce genre de situations il est préférable de dresser un tableau à deux colonnes seulement, dans le style de celui qui suit

X_i	Y_i
X_1	Y_i
X_2	Y_2
...	...
X_i	Y_i
...	...
X_N	Y_N

Remarquons que de cette façon nous n'avons plus qu'un seul indice *i*.

Le nombre de valeurs différentes est égal à l'effectif total (N) de la population.

Exemple :

Taille (en mètres) X_i	Poids (en Kg) Y_i
1.55	58.2
1.60	58.1
1.62	61.3
1.64	65.3
1.65	69.5
1.70	69.7
1.72	70.3
1.73	75.4
1.76	74.2
1.78	82.0

Ce tableau s'appelle, lui aussi, tableau à double entrée.

V. B – Distributions marginales

Considérons la dernière colonne du tableau statistique.

Les effectifs $n_{i.}$ définissent ce qu'on appelle la **distribution marginale** selon le caractère X seul. On les appelle les effectifs marginaux de X.

La distribution marginale est donc une distribution à un seul caractère.

La **fréquence relative marginale** de la modalité X_i est notée $f_{i.}$, c'est le rapport de l'effectif marginal $n_{i.}$ à l'effectif total N.

$$f_{i.} = \frac{n_{i.}}{N}$$

Nous avons ainsi le tableau de la distribution marginale de X :

X	Effectifs	Fréquences relatives
X_1	$n_{1.}$	$f_{1.}$
...	...	...
X_i	$n_{i.}$	$f_{i.}$
...	...	...
X_k	$n_{k.}$	$f_{k.}$
Total	$N = n_{..}$	1

Exemple :

Distribution marginale de X
(nombre de fruits)

X	Effectifs	Fréquences relatives
8	12	0.08
10	23	0.153
16	41	0.273
20	24	0.16
24	22	0.147
32	16	0.107
42	12	0.08
Total	**150**	**1**

Considérons, maintenant, la dernière ligne du tableau statistique.

Les effectifs $n_{.j}$ définissent la distribution marginale selon le caractère Y seul.

C'est une distribution à un seul caractère.

La fréquence relative marginale de la modalité Y_j est notée $f_{.j}$ et c'est le rapport de l'effectif marginal $n_{.j}$ à l'effectif total N.

$$f_{.j} = \frac{n_{.j}}{N}$$

Le tableau de la distribution marginale de Y est :

Y	Effectifs	Fréquences relatives
Y_1	$n_{.1}$	$f_{.1}$
…	…	…
Y_j	$n_{.j}$	$f_{.j}$
…	…	…
Y_p	$n_{.p}$	$f_{.p}$
Total	**N = $n_{..}$**	**1**

Exemple :
Distribution marginale de Y
(nombre de feuilles)

Y	Effectifs	Fréquences relatives
6	24	0.16
11	38	0.253
14	46	0.307
16	29	0.193
18	13	0.087
Total	**150**	**1**

Remarque :

Remarquez que la somme des effectifs marginaux de X est égale à l'effectifs total de la population. Il en est de même des effectifs marginaux de Y.
Autrement dit,

$$\sum_{i=1}^{k} n_{i.} = \sum_{j=1}^{p} n_{.j} = N$$

V. C – Distributions conditionnelles

i) Le tableau statistique à double entrée contient p colonnes intérieures.

La $j^{ème}$ colonne de ce tableau s'appelle la **distribution conditionnelle** de la variable X lorsque la variable Y est égale à Y_j

X / Y	Y_1	…	Y_j	…	Y_p	Total
X_1	n_{11}	…	n_{1j}	…	n_{1p}	$n_{1.}$
…	…	…	…	…	…	…
X_i	n_{i1}	…	n_{ij}	…	n_{ip}	$n_{i.}$
…	…	…	…	…	…	…
X_k	n_{k1}	…	n_{kj}	…	n_{kp}	$n_{k.}$
Total	$n_{.1}$	…	$n_{.j}$	…	$n_{.p}$	$N = n_{..}$

Il y a p distributions conditionnelles selon le caractère X.

Toute distribution conditionnelle n'est autre qu'une distribution à un caractère.

Donc, avec ces p distributions conditionnelles nous pouvons générer avec la variable statistique X (en lui affectant les effectifs conditionnels correspondants) p variables statistiques à une dimension.

La fréquence conditionnelle de la modalité X_i liée par Y_j s'écrit :

$$f_i^j = \frac{n_{ij}}{n_{.j}}$$

Nous avons ainsi le tableau de la distribution conditionnelle de X liée à la valeur Y_j :

X	Effectifs	Fréquences
X_1	n_{1j}	f_1^j
…	…	…
X_i	n_{ij}	f_i^j
…	…	…
X_k	n_{kj}	f_k^j
Total	$n_{.j}$	1

Lorsque, par exemple, Y = 6 la distribution conditionnelle est

X	Effectifs	Fréquences
8	8	0.333
10	4	0.167
16	3	0.125
20	2	0.083
24	4	0.167
32	3	0.125
42	0	0
Total	24	1

Pour l'exemple que nous avons considéré, il y a 5 distributions conditionnelles de X liées aux différentes valeurs de Y.

ii) Le tableau statistique à double entrée contient k lignes intérieures.

La $i^{ème}$ ligne de ce tableau s'appelle la distribution conditionnelle de la variable Y lorsque la variable X est égale à X_i

X / Y	Y_1	…	Y_j	…	Y_p	Total
X_1	n_{11}	…	n_{1j}	…	n_{1p}	$n_{1.}$
…	…	…	…	…	…	…
X_i	n_{i1}	…	n_{ij}	…	n_{ip}	$n_{i.}$
…	…	…	…	…	…	…
X_k	n_{k1}	…	n_{kj}	…	n_{kp}	$n_{k.}$
Total	$n_{.1}$	…	$n_{.j}$	…	$n_{.p}$	$N = n_{..}$

La fréquence conditionnelle de la modalité Y_j liée par X_i s'écrit :

$$f_j^i = \frac{n_{ij}}{n_{i.}}$$

Nous avons ainsi le tableau de la distribution conditionnelle de Y liée à la valeur X_i :

Y	Effectifs	Fréquences
Y_1	n_{i1}	f_1^i
...	...	...
Y_{ji}	n_{ij}	f_j^i
...	...	...
Y_p	n_{ip}	f_p^i
Total	$n_{i.}$	1

Il y a k distributions conditionnelles selon le caractère Y.

Pour l'exemple que nous avons considéré, Il y a 7 distributions conditionnelles de Y liées aux différentes valeurs de X. Lorsque X = 16 la distribution conditionnelle est

Y	Effectifs	Fréquences
6	3	0.073
11	10	0.244
14	15	0.366
16	8	0.195
18	5	0.122
Total	41	1

V. D – Caractéristiques de position et de dispersion

Il n'y a, en fait, aucune nouveauté dans les définitions qui vont suivre. Nous avons déjà vu ce qu'est une moyenne arithmétique ou une variance dans le cas de variables statistiques à une dimension.

Or, dans les paragraphes précédents, nous avons constaté que l'étude de la variable à deux dimensions se ramène à considérer des variables à une dimension. En effet, le tableau à double entrée nous donne p+2 colonne, la première colonne est celle des valeurs de X. Si on accompagne cette colonne par celle qui est à la marge du tableau (les effectifs marginaux) on obtient ce qu'on a appelé la distribution marginale de X qui n'est autre qu'une variable statistique à une dimension. Nous saurons, en oubliant le reste du tableau à double entrée, calculer les caractéristiques de cette variable avec les connaissances acquises dans les chapitres précédents.

De même, si on décide de faire accompagner la première colonne par l'une quelconque des colonnes intérieures du tableau à double entrée nous obtenons également une variable statistique discrète à une dimension (que nous avions appelée distribution conditionnelle de X). Comme il y a p colonnes intérieures nous pouvons générer p variable à une dimension qui soient de la même espèce (distributions conditionnelles). Pour toutes ces variables nous savons parfaitement calculer les caractéristiques (malgré le nouveau nom que nous leurs avons donné).

Ce même commentaire peut être repris quand on considère les lignes du tableau à double entrée.

En somme, nous obtenons (p+1) + (k+1) = p+k+2 variables statistiques discrètes à une dimension. Deux d'entre elles ont été obtenues en utilisant la colonne de la marge (pour X) et la ligne de la marge (pour Y). Le reste a été obtenu en utilisant les colonnes ou les lignes intérieures.

En résumé, les notions de moyenne arithmétique ou de variance sont les mêmes qu'on a vu précédemment, il s'agit seulement de préciser pour quelle variable on les calcule. Si c'est pour les distributions marginales, nous les appellerons **caractéristiques marginales**. Si c'est pour les distributions conditionnelles nous les appellerons **caractéristiques conditionnelles**.

a) Caractéristiques marginales

La moyenne marginale de la variable X est

$$\overline{X}_M = \frac{1}{N}\sum_{i=1}^{k} n_{i.} X_i = \sum_{i=1}^{k} f_{i.} X_i$$

La variance marginale de la variable X est

$$Var_M(X) = \left[\frac{1}{N}\sum_{i=1}^{k} n_{i.} X_i^2\right] - \overline{X}_M^2 = \left[\sum_{i=1}^{k} f_{i.} X_i^2\right] - \overline{X}_M^2$$

Son écart-type marginal est donc

$$\sigma_M(X) = \sqrt{Var_M(X)}$$

La moyenne marginale de la variable Y est

$$\overline{Y}_M = \frac{1}{N}\sum_{j=1}^{p} n_{.j} Y_j = \sum_{j=1}^{p} f_{.j} X_j$$

La variance marginale de la variable Y est

$$Var_M(Y) = \left[\frac{1}{N}\sum_{j=1}^{p} n_{.j} Y_j^2\right] - \overline{Y}_M^2 = \left[\sum_{j=1}^{p} f_{.j} Y_j^2\right] - \overline{Y}_M^2$$

Son écart-type marginal est donc

$$\sigma_M(Y) = \sqrt{Var_M(Y)}$$

Reprenons notre exemple et calculons les caractéristiques marginales pour les deux variables X et Y.

pour la distribution marginale de X, nous avons :

X	$n_{i.}$	$n_{i.} X_i$	$n_{i.} X_i^2$
8	12	96	768
10	23	230	2300
16	41	656	10496
20	24	480	9600
24	22	528	12672
32	16	512	16384
42	12	504	21168
Total	150	3006	73388

et pour la distribution marginale de Y, nous avons :

Y	$n_{.j}$	$n_{.j} Y_j$	$n_{.j} Y_j^2$
6	24	144	864
11	38	418	4598
14	46	644	9016
16	29	464	7424
18	13	234	4212
Total	150	1904	26114

Ce qui nous donne pour la variable X

$$\overline{X}_M = \frac{1}{N}\sum_{i=1}^{k} n_{i.} X_i = \frac{3006}{150} = 20{,}04$$

et

$$Var_M(X) = \left[\frac{1}{N}\sum_{i=1}^{k} n_{i.} X_i^{\,2}\right] - \overline{X}_M^2 = \frac{7\,3398}{150} - (20{,}04)^2 = 87{,}651$$

et

$$\sigma_M(X) = \sqrt{Var_M(X)} = \sqrt{87{,}651} \approx 9{,}36$$

Remarque :

J'espère que vous avez reconnu l'exemple que nous avons utilisé pour développer le chapitre sur la variable statistique discrète et que ces valeurs ne vous sont pas étrangères.

On ne prenait en considération, dans cet exemple que la variable « nombre de fruits » seulement.

Calculons maintenant les caractéristiques marginales de Y, nous avons

$$\overline{Y}_M = \frac{1}{N}\sum_{j=1}^{p} n_{.j} Y_j = \frac{1904}{150} = 12{,}693$$

et

$$Var_M(Y) = \left[\frac{1}{N}\sum_{j=1}^{p} n_{.j} Y_j^{\,2}\right] - \overline{Y}_M^2 = \frac{26114}{150} - (12{,}693)^2 = 12{,}981$$

et

$$\sigma_M(Y) = \sqrt{12{,}981} \approx 3{,}60$$

b) Caractéristiques conditionnelles

Quand il s'agit de conditionnelle il faut préciser par rapport à quelle valeur. Quelle valeur Y_j qui conditionne la variable X ou quelle valeur X_i qui conditionne la variable Y.

La moyenne conditionnelle de la variable X liée à la valeur Y_j est

$$\overline{X}_j = \frac{1}{n_{.j}}\sum_{i=1}^{k} n_{ij} X_i = \sum_{i=1}^{k} f_i^{\,j} X_i$$

La variance conditionnelle de la variable X liée à la valeur Y_j est

$$Var_j(X) = \left[\frac{1}{n_{.j}}\sum_{i=1}^{k} n_{ij} X_i^{\,2}\right] - \overline{X}_j^{\,2} = \sum_{i=1}^{k} f_i^{\,j} X_i^{\,2} - \overline{X}_j^{\,2}$$

Son écart-type conditionnel est donc

$$\sigma_j(X) = \sqrt{Var_j(X)}$$

La moyenne conditionnelle de la variable Y liée à la valeur X_i est

$$\overline{Y}_i = \frac{1}{n_{i.}}\sum_{j=1}^{p} n_{ij} Y_j = \sum_{j=1}^{p} f_j^{\,i} Y_j$$

La variance conditionnelle de la variable Y liée à la valeur X_i est

$$Var_i(Y) = \left[\frac{1}{n_{i.}}\sum_{j=1}^{p} n_{ij} Y_j^{\,2}\right] - \overline{Y}_i^{\,2} = \sum_{j=1}^{p} f_j^{\,i} Y_j^{\,2} - \overline{Y}_i^{\,2}$$

Son écart-type conditionnel est donc

$$\sigma_i(Y) = \sqrt{Var_i(Y)}$$

Considérons, par exemple, l'une des distributions conditionnelles de X : Lorsque Y = 6
Nous avons :

X	n_{ij}	$n_{ij} X_i$	$n_{ij} X_i^2$
8	8	64	512
10	4	40	400
16	3	48	768
20	2	40	800
24	4	96	2304
32	3	96	3072
42	0	0	0
Total	24	384	7856

Nous avons

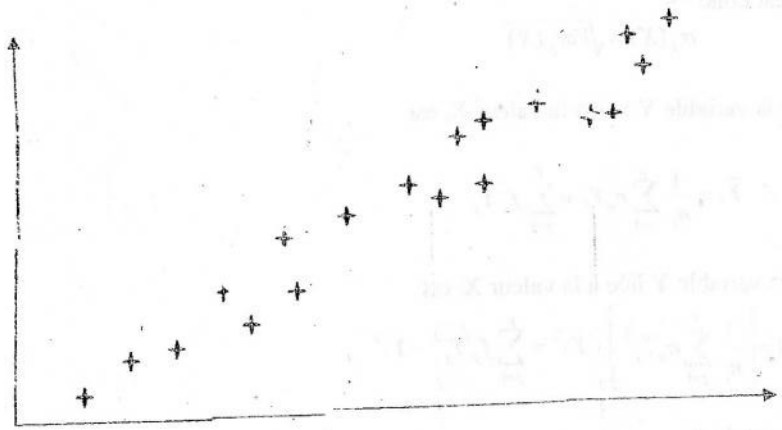
$$\overline{X}_j = \frac{1}{n_{.j}}\sum_{i=1}^{k} n_{ij} X_i = \frac{384}{24} = 16$$

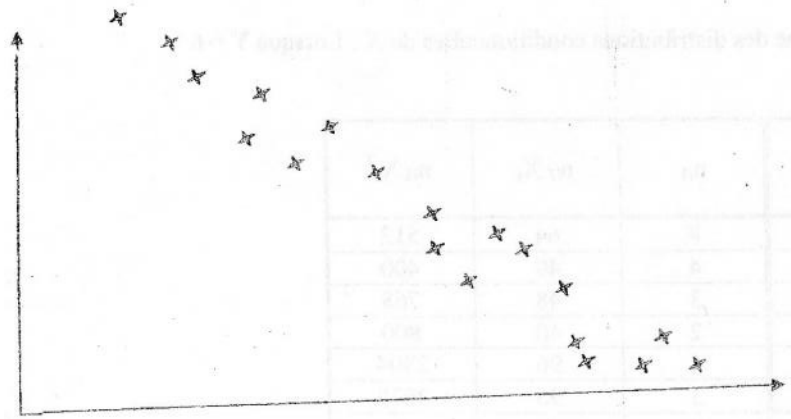
et

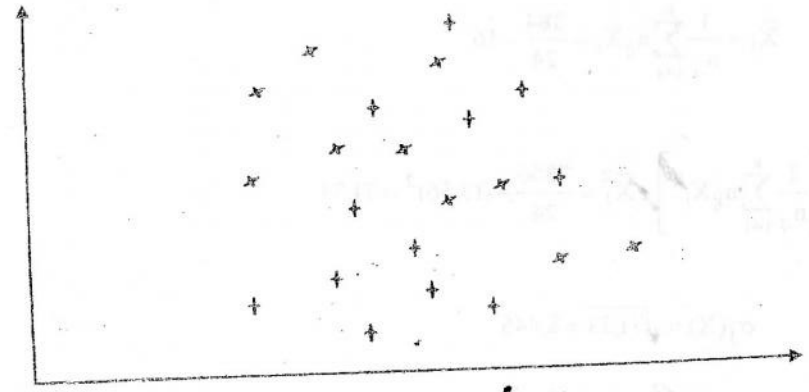
$$\mathrm{Var}_j(X) = \left[\frac{1}{n_{.j}}\sum_{i=1}^{k} n_{ij} X_i^{\,2}\right] - \overline{X}_j^2 = \frac{7856}{24} - (13{,}56)^2 = 71{,}33$$

et

$$\sigma_j(X) = \sqrt{71{,}33} \approx 8{,}445$$







Considérons, aussi, l'une des distributions conditionnelles de Y : Lorsque X = 16
Nous avons :

Y	n_{3j}	$n_{3j} Y_j$	$n_{3j} Y_j^2$
6	3	18	108
11	10	110	1210
14	15	210	2940
16	8	128	2048
18	5	90	1620
Total	41	556	7926

Nous avons

$$\overline{Y}_i = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij} Y_j = \frac{556}{41} = 13{,}561$$

et

$$Var_i(Y) = \left[\frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij} Y_j^2 \right] - \overline{Y}_i^2 = \frac{7926}{41} - (13{,}561)^2 = 9{,}417 ;$$

et

$$\sigma_i(Y) = \sqrt{9{,}417} \approx 3{,}07$$

V – E – Covariance

C'est, en réalité, la véritable innovation de ce chapitre. Jusqu'ici on ne s'est intéressé aux variables que prises isolément. Nous avons vu leurs caractéristiques de position. Nous avons vu comment elles se dispersent. Nous avons vu comment les représenter graphiquement.

Mais qu'en est-il de leur relation l'une par rapport à l'autre ? Quand peut-on dire qu'elles varient dans le même sens ou qu'elles varient dans le sens contraire ? Comment peut-on mesurer la force de leur liaison ?
Nous allons essayer de construire une formule qui pourrait répondre à ces questions.

Observons les nuages* de points suivants

12

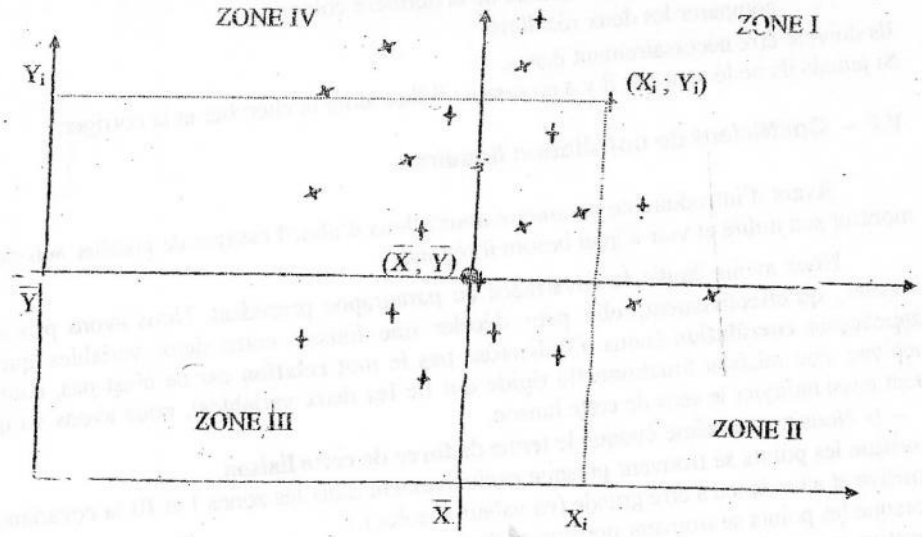
/* lorsqu'on représente deux variables graphiquement nous obtenons un repère où l'un des axes représente la première variable et l'autre axe représente la deuxième variable. De cette manière, une observation (X_i ; Y_i) est un point dans le plan généré par ces deux axes ; X_i représente son abscisse et Y_i représente son ordonnée.
Un nuage de points est donc une représentation graphique d'un tableau statistique à double entrée.

La disposition des points nous donne déjà une idée de l'évolution des deux variables l'une par rapport à l'autre. Nous constatons, en effet, que les trois nuages n'ont pas le même aspect.
La forme des deux premiers nuages est assez allongée tandis que celle du troisième est plutôt arrondie.
En observant le premier nuage nous constatons que, dans l'ensemble, les points qui ont de petites abscisses ont de petites ordonnées et lorsqu'on a des grandes valeurs pour X on a aussi des grandes valeurs pour Y. Ainsi, les variables X et Y varient dans le même sens : lorsque X croit, Y croit également.

Le deuxième nuage suggère qu'il y a des variations opposées : quand X a tendance à prendre de grandes valeurs Y a tendance à prendre de petites valeurs. Ainsi, les variables X et Y varient dans le sens contraire : lorsque X croit Y décroît.

Le troisième nuage, quant à lui, ne permet d'identifier aucune relation entre les deux variables. Nous avons tous les cas de figures, les petites valeurs de Y correspondent aussi bien aux petites comme aux grandes valeurs de X. De même les grandes valeurs de Y s'associent aux petites valeurs de X et s'associent aussi à ses grandes valeurs. On peut comprendre que ces deux variables sont "indépendantes".

Essayons de quantifier les résultats de cette analyse, c'est-à-dire de mesurer par la quantité l'existence d'une liaison entre X et Y, d'identifier son sens (variation dans le même sens ou dans le sens contraire) et voir la force de cette liaison (forte ou faible).
Pour cela, procédons de la manière suivante :
considérons le point moyen du nuage $(\overline{X};\overline{Y})$, $\overline{X}$ et $\overline{Y}$ étant les moyennes arithmétiques des variables X et Y. Ce point moyen se situe au centre du nuage.



Quand nous déplaçons le repère sur ce point nous obtenons quatre zones.
Soit, maintenant la quantité

$$\alpha_i = (X_i - \overline{X}).(Y_i - \overline{Y})$$

- Si le point $(X_i;Y_i)$ se trouve dans la zone I alors les quantités $(X_i - \overline{X})$ et $(Y_i - \overline{Y})$ son toutes les deux positives et donc leur produit α_i est positif.
- Si le point $(X_i;Y_i)$ se trouve dans la zone III alors les quantités $(X_i - \overline{X})$ et $(Y_i - \overline{Y})$ son toutes les deux négatives et donc α_i est positif.

Par contre :
- Si le point $(X_i;Y_i)$ se trouve dans la zone II alors la quantité $(X_i - \overline{X})$ est positive tandi que la quantité $(Y_i - \overline{Y})$ est négative et, par conséquent, α_i est négative.
- Si le point $(X_i;Y_i)$ se trouve dans la zone IV alors la quantité $(X_i - \overline{X})$ est négative tandi que la quantité $(Y_i - \overline{Y})$ est positive et, par conséquent, α_i est négative.

Regardons le nuage dans son ensemble et considérons la quantité

$$\beta = \frac{1}{N}\sum_{i=1}^{N} \alpha_i$$

β tient compte de tous les points du nuage. C'est la moyenne des valeurs α_i

- α_i est positive pour tous les points des zones I et III, et donc chaque point de ces deu zones apporte une contribution positive à β .
- α_i est négative pour tous les points des zones II et IV, et donc chaque point de ces deux zones apporte une contribution négative à β .

Revenons, maintenant à nos trois nuages du début.
Les points du premier nuage appartiennent presque tous aux zones I et III, et donc apporten presque tous une contribution positive à β qui sera ainsi positive et grande en valeur absolue.
Les points du deuxième nuage appartiennent presque tous aux zones II et IV, et donc apportent presque tous une contribution négative à β qui sera ainsi négative et grande en valeur absolue.
Mais les points du troisième nuage se dispersent sur les quatre zones et les contributions positives ramenés par certains points seront compensées par les contributions négatives ramenées par les autres points. Nous devons donc nous attendre, dans ce cas, que β soi proche de zéro.
En conclusion, la quantité β nous renseigne bien sur la liaison entre les variables X et Y, son sens et sa force.
la quantité β est appelé la **covariance** entre les deux variables X et Y. On la note **Cov(X ; Y)**
Ainsi

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X}).(Y_i - \overline{Y})$$

$$= \frac{1}{N}\sum_{i=1}^{N} X_i Y_i - \overline{X}.\overline{Y}$$

Remarques :

1. La formule de la covariance établie ci-dessus correspond au cas où tous les points se démarquent les uns des autres. C'est-à-dire, où tous les effectifs ne sont égaux qu'a un ou à zéro. C'est le cas particulier des tableaux statistiques à double entrée vu au paragraphe 1.Dans le cas général, lorsque les points peuvent être regroupés sur les mêmes positions. C'est-à-dire lorsque les effectifs n_{ij} peuvent être quelconques. Le tableau statistique est le tableau à double entrée défini dans le cas général et la formule de la covariance est

$$\mathrm{Cov}(X;Y)=\left[\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p}n_{ij}(X_i-\overline{X}_M).(Y_j-\overline{Y}_M)\right]$$

Remarquez que la formule précédente n'est qu'un cas particulier de celle-ci, lorsque $i=j$

2. En développant le produit sous le signe somme on obtient la formule suivante qui est celle que l'on utilise dans la pratique

$$\mathrm{Cov}(X;Y)=\left[\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p}n_{ij}.X_i.Y_j\right]-\left[\overline{X}_M.\overline{Y}_M\right]$$

3. Lorsque la variable Y s'identifie à la variable X (la même variable), C'est-à-dire lorsque $X_i=Y_i$ pour tout i, alors la formule précédente devient

$$\mathrm{Cov}(X;Y)=\left[\frac{1}{N}\sum_{i=1}^{k}\sum_{i=1}^{k}n_{ii}.X_i.X_i\right]-\left[\overline{X}_M.\overline{X}_M\right]$$

c'est-à-dire

$$\mathrm{Cov}(X;Y)=\left[\frac{1}{N}\sum_{i=1}^{k}n_i X_i^2\right]-\left[\overline{X}_M^2\right]$$

Nous reconnaissons là la formule de la variance de la X.
Ainsi la variance d'une variable statistique n'est autre que la covariance de cette variable avec elle même.

Exemple de calcul de la covariance:

Reprenons notre exemple du début du chapitre. Nous avons deux variables : X (le nombre de fruits) et Y (le nombre de feuilles). Nous allons utiliser cet exemple pour voir comment on calcul la covariance entre X et Y.
Il s'agit de dresser le tableau qui répond à cette demande. Soit donc le tableau suivant :

X / Y	6	11	14	16	18	Total
8	8 384	2 176	1 112	1 128	0 0	800
10	4 240	16 1760	2 280	0 0	1 180	2460
16	3 288	10 1760	15 3360	8 2048	5 1440	8896
20	2 240	4 880	14 3920	3 960	1 360	6360
24	4 576	5 1320	6 2016	5 1920	2 864	6696
32	3 576	1 352	2 896	8 4096	2 1152	7072
42	0 0	0 0	6 3528	4 2688	2 1512	7728
otal	2304	6248	14112	11840	5508	40012

Dans chaque case intérieure vous remarquez l'existence de deux nombres l'un en petit caractères : c'est l'effectif n_{ij}

l'autre en gros caractères : c'est le produit $n_{ij}X_iY_j$

Par exemple, on a obtenu 880 en faisant le produit de la valeur de X (égale à 20) qui est sur la ligne par la valeur de Y (égale à 11) qui est sur la colonne, le résultat obtenu est multiplié par la valeur de l'effectif (égal à 4) qui se trouve dans cette case.
quand on fait la somme des nombres on obtient dans la dernière ligne les totaux sur les colonnes et dans la dernière colonne les totaux sur les lignes.
La somme des nombres sur la dernière ligne est égale à la somme des nombres sur la dernière colonne, c'est le total des totaux. C'est

$$\sum_{i=1}^{k}\sum_{j=1}^{p}n_{ij}X_iY_j$$

Il suffit de diviser ce nombre par N et d'ôter du résultat le produit des moyennes de $\overline{X}$ et de $\overline{Y}$ pour obtenir la valeur de la covariance ; C'est-à-dire

$$Cov(X;Y)=\frac{40012}{150}-(20{,}04.12{,}693)\approx 12{,}379$$

Remarque :

Pour la vérification des calculs il est plus simple de recourir à la procédure suivante :
- faire la somme des totaux de la dernière ligne,
- faire la somme des totaux de la dernière colonne,
- comparer les deux résultats.

Ils doivent être nécessairement égaux.
Si jamais ils ne le sont pas il y a eu erreur. Il faut alors la chercher et la corriger.

V.F – Coefficient de corrélation linéaire

Avant d'introduire ce paramètre nous allons d'abord essayer de justifier son existence, montrer son utilité et voir à quel besoin il répond.

Nous avons étudié la covariance au paragraphe précédent. Nous avons pris soin de vérifier, qu'effectivement, elle peut déceler une liaison entre deux variables que nous appellerons **corrélation** (nous n'utiliserons pas le mot **relation** car ce n'est pas, comme en analyse, une relation fonctionnelle rigide qui lie les deux variables), nous avons vu qu'elle peut aussi indiquer le sens de cette liaison.

→ Nous avons même évoqué le terme de **force de cette liaison**.

- Lorsque les points se trouvent presque exclusivement dans les zones I et III la covariance est positive et a tendance à être grande (en valeur absolue).
- Lorsque les points se trouvent presque exclusivement dans les zones II et IV la covariance est négative et a tendance à être grande (en valeur absolue).

Mais quand on dit que la covariance est grande, elle est grande à quelle point ?
A partir de quelle valeur peut-on dire qu'il y a une forte covariance entre les deux variables ?
Soit, par exemple, deux couples de variables (X ; Y) et (Z ; T). Supposons que leurs covariances respectives soient

- $Cov(X;Y)=527{,}2$ et $Cov(Z;T)=0{,}45$

Ainsi, $\Delta = b^2 - 4.a.c \le 0$

C'est-à-dire $$\Delta = 4.\mathrm{Cov}^2(X,Y) - 4.\sigma_X^2.\sigma_Y^2 \le 0$$

autrement dit

$$\mathrm{Cov}^2(X,Y) \le \sigma_X^2.\sigma_Y^2$$

Ce qui conduit à

$$|\mathrm{Cov}(X,Y)| \le \sigma_X.\sigma_Y$$

Finalement, en conclusion, la covariance est toujours, en valeur absolue, inférieure ou égale au produit des écarts-types.

Pour répondre à la question du début, nous pouvons donc dire que la covariance ne peut pas augmenter indéfiniment mais elle a une borne. La valeur maximale qu'elle peut atteindre est égale au produit des écarts-types des variables X et Y. Plus elle s'approche de cette valeur plus la liaison entre X et Y est forte.
Nous pouvons utiliser cette idée pour créer un paramètre qui rend compte de la force de la liaison : **Le coefficient de corrélation linéaire**.
C'est le rapport de la covariance au produit des écarts-types, c'est-à-dire

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

Remarque :

1. Puisque $|Cov(X,Y)| \le \sigma_X.\sigma_Y$

Alors

$$\boxed{-1 \le \rho \le +1}$$

C'est une propriété importante de ce coefficient.

2. ρ est appelé coefficient de corrélation linéaire. On peut se demander pourquoi le mot **linéaire** ?

Nous avons vu que $|\mathrm{Cov}(X,Y)| \le \sigma_X.\sigma_Y$

Quand est ce que cette inégalité devient-elle égalité ? C'est-à-dire

$$|\mathrm{Cov}(X,Y)| = \sigma_X.\sigma_Y$$

Réponse : C'est lorsque l'expression (M) est nulle.
(M) ne peut être nulle que si tous ses termes sont nuls. Autrement dit si, pour tout i

$$R.(X_i - \overline{X}) - (Y_i - \overline{Y}) = 0$$

Cette condition signifie que tous les points sont sur une même droite d'équation

$$Y_i - \overline{Y} = R.(X_i - \overline{X})$$

Peut-on dire, à la vue de ces valeurs, que la covariance entre X et Y est plus forte que la covariance entre Z et T ?
Est-ce parce que 527,2 est plus grand que 0,45 qu'on peut tirer cette conclusion ?
Mais si on vous disait, par exemple, que X et Y ont été mesurés en mètres et que Z et T ont été mesurés en millimètres cette conclusion reste-t-elle valable ?
Reprenons la formule de la covariance

$$Cov(X;Y) = \left[\frac{1}{N}.\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij} X_i Y_j\right] - \overline{X}_M.\overline{Y}_M$$

Nous voyons sur cette formule que la covariance est en relation directe avec les unités dans lesquelles ont été mesurées X et Y. Si on change donc ces unités la covariance change de valeur, et parfois considérablement.

Qu'est-ce qu'il y a alors lieu de faire dans cette situation, ?

Pour cela, considérons la quantité suivante :

$$\frac{1}{N}\sum_{i=1}^{k}\left[R.(X_i - \overline{X}) - (Y_i - \overline{Y})\right]^2 \qquad (M)$$

Cette quantité est nécessairement positive (ou nulle) car c'est une somme de carrés.
Développant cette expression nous obtenons

$$R^2\left[\frac{1}{N}\sum_{i=1}^{k}(X_i - \overline{X})^2\right] - 2.R\left[\frac{1}{N}\sum_{i=1}^{k}(X_i - \overline{X}).(Y_i - \overline{Y})\right] + \left[\frac{1}{N}\sum_{i=1}^{k}(Y_i - \overline{Y})^2\right]$$

Nous y reconnaissons les formules des variances de X et de Y et celle de leur covariance, on peut donc l'écrire, simplement

$$R^2.\mathrm{Var}(X) - 2.R.\mathrm{Cov}(X,Y) + \mathrm{Var}(Y)$$

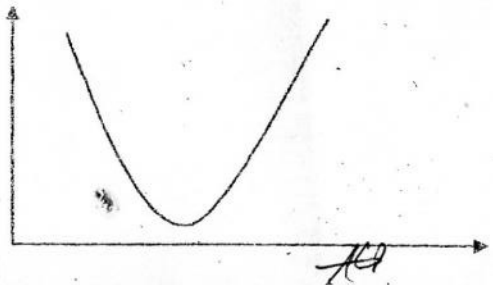
Qu'on peut écrire aussi

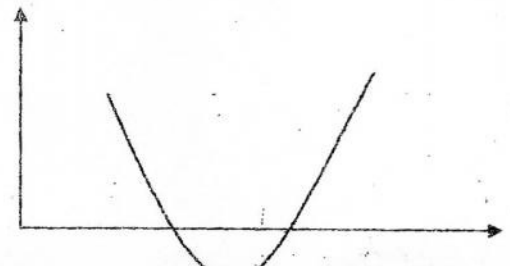
$$\sigma_X^2.R^2 - 2.\mathrm{Cov}(X,Y).R + \sigma_Y^2$$

C'est est un trinôme du second degré en R de la forme

$$a.R^2 - b.R + c$$

Nous avons vu que ce trinôme est nécessairement positif (ou nul) quelle que soit la valeur de R, pour qu'il en soit ainsi il faut que son déterminant soit négatif ou nul. (car c'est une parabole qui se situe au dessus de l'axe des abscisses, Il ne faut jamais qu'elle descende en dessous, c'est-à-dire qu'il ne faut jamais que le trinôme ait deux solutions).





Ainsi, nous aurons $|\text{Cov}(X,Y)| = \sigma_X \cdot \sigma_Y$ ou, ce qui revient au même, $|\rho| = 1$ lorsque tous les points se trouvent exactement sur une droite, et

- si $\rho = +1$ la pente de la droite est positive, (X et Y varient dans le même sens)

- si $\rho = -1$ la pente de la droite est négative (X et Y varient dans le sens contraire).

- Quand ρ est proche de 1 nous comprenons que les points du nuage sont proches de la droite.
- Quand ρ est proche de zéro, le nuage de points a une forme arrondie et cela veut dire qu'il n'y a pas une liaison linéaire entre les variables X et Y.

ρ ne sert donc qu'à montrer l'existence d'une liaison linéaire entre les variables. Si ρ n'est pas proche de 1 cela veut seulement dire qu'il ne peut y avoir de liaison linéaire entre les variables X et Y mais il est toujours possible qu'elles ont une autre forme de liaison (non linéaire).

Exemple de calcul :

Reprenons notre exemple du début du chapitre. Nous avons

$$\rho = \frac{\text{Cov}(X,Y)}{\sigma_X . \sigma_Y} = \frac{12,379}{9,36.3,60} \approx 0,367$$

Nous déduisons de ce résultat que, pour les fraisiers, il n'y a pas de relation linéaire entre le nombre de leurs feuilles et le nombre de leurs fruits.