I. LA VARIABILITE AU SEIN DES POPULATIONS

La notion de population recouvre un concept difficilement réductible à une définition unique. Au sens de la génétique, une population représente une entité de reproduction au sein d'une espèce (voir GP et chapitre V). Cette définition est néanmoins vague car elle ne précise ni le type de reproduction, ni le critère permettant d'affecter un individu à une entité plutôt qu'à une autre. Dans ce chapitre, nous considérons la population comme un ensemble d'individus possédant certaines caractéristiques communes : il peut s'agir des pieds de maïs d'une parcelle, d'une colonie d'insectes dans une forêt, des habitants de la commune de Marchastel (Lozère, 38 habitants) ou de la République Populaire de Chine (1,25 milliard d'habitants), etc. Pour décrire la variabilité au sein d'une population, une première solution consiste à fournir le résultat brut de la collecte de données, c'est-à-dire la liste des valeurs numériques mesurées sur tous les individus pour les différents caractères observés. Le volume des données peut être extrêmement important et ne permet pas d'appréhender correctement la situation générale de la population. Aussi, a-t-on recours aux statistiques, dont un des rôles est de synthétiser l'information : le présent chapitre renvoie donc à l'enseignement correspondant.

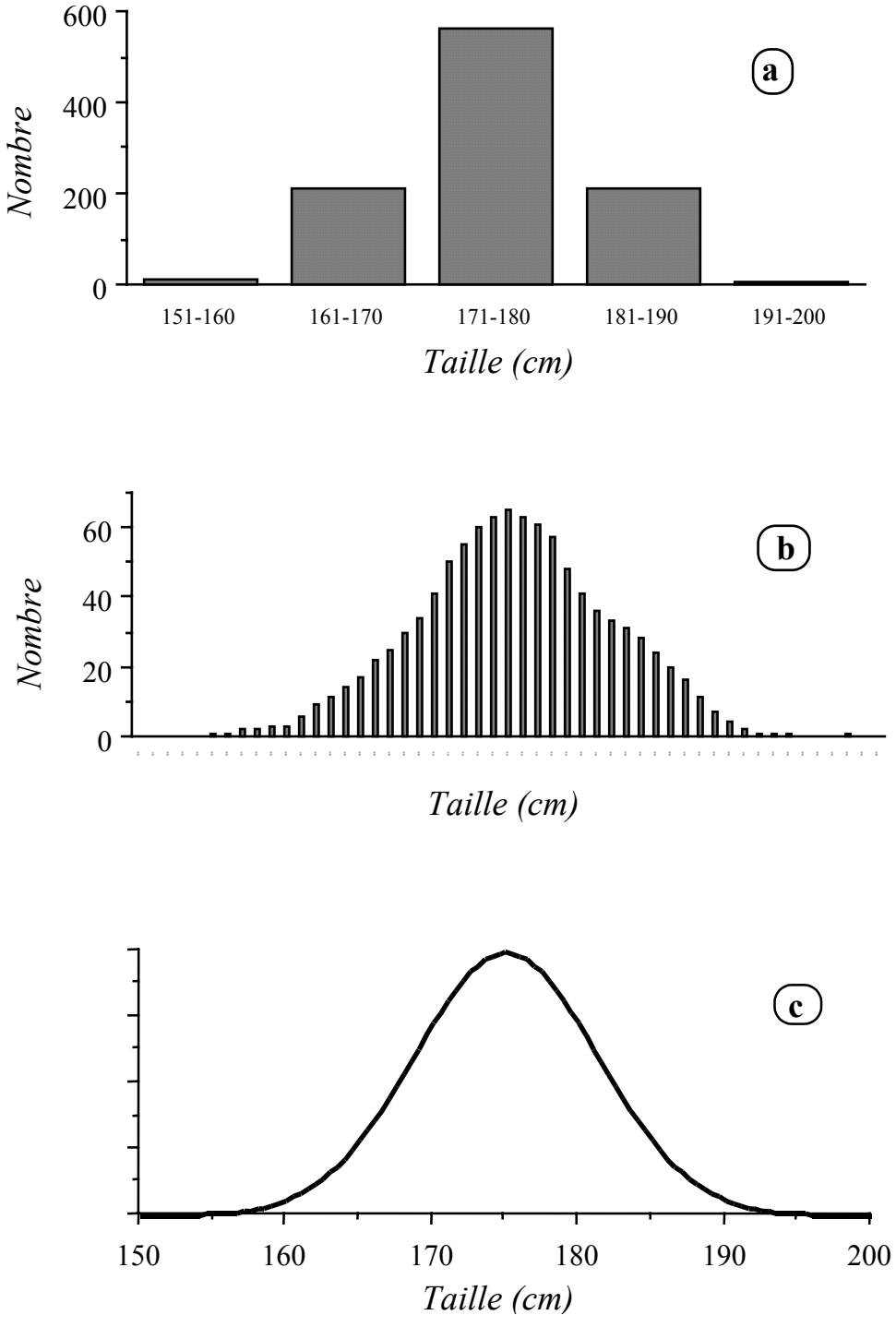
A. Description d'une population pour un caractère

1. La distribution

Une représentation graphique permet de rendre compte de la manière dont les valeurs numériques se répartissent dans la gamme de variation observée : c'est la distribution. Il s'agit d'un graphe où en abscisse se trouvent les valeurs numériques et en ordonnée la fréquence (ou le nombre) des individus que l'on trouve avec une valeur donnée ou dans un intervalle donné. La figure 1 représente la distribution de la taille de mille étudiants de l'université de Harvard. On voit que l'allure de la distribution change selon la précision avec laquelle a été faite la mesure, ou, ce qui revient au même, la largeur des classes que l'on constitue pour réaliser le graphe. Si le pas de classe adopté est de 10 cm, les étudiants se répartissent en cinq classes (Figure 1.a). Si l'on peut mesurer à 1 cm près, les classes se subdivisent (Figure 1.b). Si l'on poursuit le processus, en affinant les mesures et en supposant que l'on peut mesurer un très grand nombre d'individus, on tend vers une distribution continue (Figure 1.c).

Figure 1. Distribution de la taille de 1 000 étudiants de sexe masculin de l'université de Harvard (Etats-Unis), selon que l'on constitue des classes de 10 cm (a) ou de 1 cm (b), et comparaison avec la courbe de la loi normale (c).

Source : Castle, 1916



La figure 1 nous indique également que la distribution de la taille, dans la population observée, se rapproche, dans sa forme, de la courbe « en cloche » caractéristique de la loi normale (voir Stat).

Ainsi, comme on l'a déjà évoqué dans l'introduction, les caractères quantitatifs présentent une variation continue. La figure 2 montre quelques exemples de variation observée pour des caractères d'importance agronomique ou zootechnique. Même lorsque le caractère mesuré est par nature discontinu, car représentant la somme d'un nombre d'objets distincts (on parle dans ce cas de caractères méristiques : nombre de grains sur un épi de maïs, nombre de soies abdominales chez la drosophile, nombre de jeunes dans une portée de truie, etc.), le nombre de classes observées peut être élevé, et l'on considère la variation de ce type de caractère comme continue (voir figure 2.d).

La distribution normale est une distribution très fréquemment rencontrée pour un grand nombre de caractères dans toutes les espèces (voir figure 2). De ce fait, l'analyse statistique des caractères auxquels on s'intéresse en génétique quantitative est souvent facilitée. Toutefois, les caractères pour lesquels on ne peut pas admettre la normalité de la distribution nécessitent un traitement approprié. Par exemple, une transformation mathématique des données par des fonctions telles que le logarithme, la racine, etc. permet souvent de normaliser les distributions.

2. La moyenne

La distribution reste néanmoins peu facile à manipuler, et ne donne pas immédiatement une valeur représentative du caractère dans la population : si nous demandons à un obstétricien combien pèse un bébé à la naissance en France, il ne nous présentera pas une courbe de distribution, mais nous répondra, « autour de 3,2 kg ». La moyenne, ou espérance, est le concept statistique le plus utilisé pour donner un ordre de grandeur d'un caractère. Cependant, on peut parfois être amené à donner le mode (valeur la plus fréquemment observée) ou la médiane (valeur autour de laquelle l'effectif de la population se répartit équitablement). En cas de distribution normale, ces trois paramètres sont égaux.

Figure 2. Quelques distributions observées pour des caractères d'importance agronomique ou zootechnique.

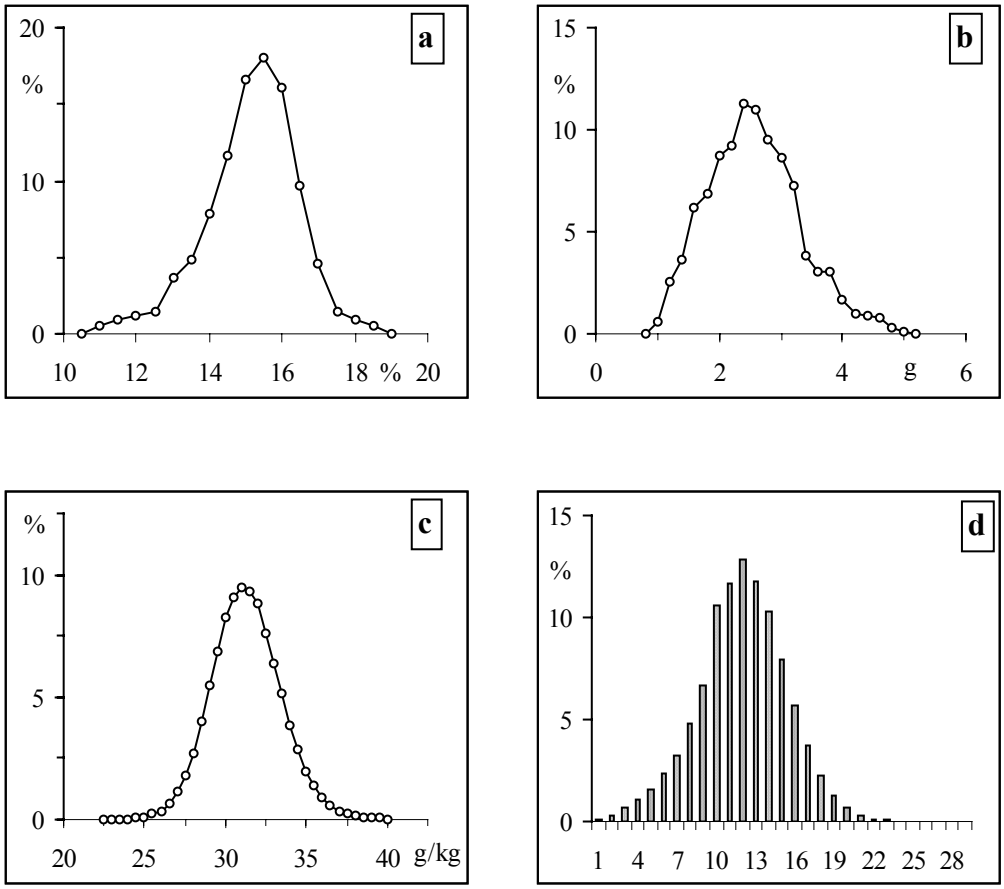
N = nombre total d'individus ayant fait l'objet d'une mesure

a : Teneur en saccharose (%) dans la racine de betterave à sucre. Variété du début du XXème siècle. (*N* = 42 997 ; Source : De Vries, 1909).

b : Poids de grain (g) dans un épi de blé (*N* = 790 ; Source : INA P-G, 2000).

c : Teneur en protéines du lait de vache (g/kg), calculée sur l'ensemble de la lactation. Race Montbéliarde, contrôle de performances en ferme en 1988 (*N* = 251 705 ; Source : FNOCL, 1989).

d : Taille de portée (nombre de jeunes nés par portée) chez la truie. Race Large-White, contrôle de performances en ferme de 1990 à 2000. (*N* = 724 123 ; Source : ITP-INRA, 2000).



3. La variance

Pour décrire l'amplitude de la dispersion autour de la moyenne, on utilise le plus souvent la variance, qui est la moyenne des carrés des écarts à la moyenne (voir le mémento statistique en fin de ce chapitre) et s'exprime dans le carré de l'unité du caractère mesuré. L'écart-type, quant à lui, est égal à la racine carrée de la variance et s'exprime dans l'unité du caractère.

A titre d'illustration, le tableau 1 donne les paramètres des distributions reportées aux figures 1 et 2. Une propriété intéressante de la loi normale est que l'on peut facilement calculer la proportion d'individus dont la valeur se situe au-delà (ou en deçà) d'un certain seuil par rapport à la moyenne ; des tables donnent ces proportions pour une loi normale centrée réduite, c'est-à-dire dont la moyenne est nulle et la variance est égale à 1 (voir Stat). Ainsi, il est bien connu que, dans le cas d'une distribution normale, 95 % des valeurs se situent dans un intervalle allant de -1,96 à + 1,96 écarts-types de part et d'autre de la moyenne. Si l'on préfère, on peut également dire que seulement 2,5 % des valeurs se situent à plus de 1,96 écarts-types au-delà de la moyenne, et réciproquement en deçà de la moyenne. A titre d'exemple, la distribution de la teneur en protéines du lait étant rigoureusement ajustée à une distribution normale (figure 2.c), les chiffres du tableau 1 nous indiquent que 95 % des vaches Montbéliardes contrôlées en 1988 ont eu une valeur de taux protéique comprise entre 27,6 et 35,4 g/kg, ce qui représente un bon aperçu du champ de variation du caractère dans la population considérée.

Tableau 1. Caractéristiques des échantillons pour lesquels la distribution des valeurs mesurées a été donnée (cf. figures 1 et 2).

Caractère	Nombre d'individus mesurés	Moyenne	Ecart-type
Taille (cm)	1 000	175	6,3
Teneur en saccharose (%)	42 997	15,2	1,2
Poids de grain par épi (g)	790	2,58	0,75
Teneur en protéines du lait (g/kg)	251 705	31,5	2,0
Taille de portée	724 123	12,0	3,4

B. Description d'une population pour deux caractères

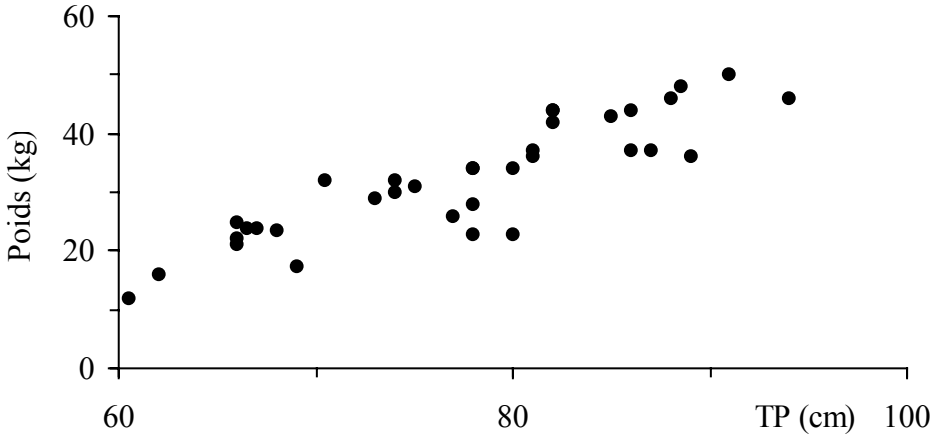
Lorsque l'on étudie une population, il est rare que l'on ne s'intéresse qu'à un seul caractère mesurable sur les individus qui la composent. En amélioration des plantes ou des animaux, ce n'est quasiment jamais le cas. Il est donc important de disposer d'outils permettant, au minimum, de décrire une population de façon bidimensionnelle.

1. La distribution à deux dimensions

Pour rendre compte graphiquement d'une distribution à deux dimensions, on construit un nuage de points : chaque individu est représenté sur le graphe par un point dont les coordonnées sont égales aux valeurs respectives pour les deux caractères étudiés. La figure 3 montre une construction de ce type, à partir de données recueillies sur une espèce d'ongulé sauvage, le bouquetin des Alpes, *Capra ibex ibex*. La forme du nuage de points suggère une liaison statistique entre les deux caractères mesurés, liaison qu'il est possible de quantifier.

Figure 3. Relation entre le tour de poitrine (TP) et le poids dans un échantillon de 35 mâles de bouquetin des Alpes (*Capra ibex ibex*).

Source : Toïgo, 1998



2. La corrélation

Le coefficient de corrélation (de Pearson) est le paramètre usuel pour quantifier l'association entre deux variables. Ce coefficient se calcule comme étant le rapport de la covariance entre les deux variables au produit de leurs écarts-types respectifs ; la covariance est l'espérance du coproduit des deux variables centrées (voir mémento statistique). Le coefficient de corrélation entre deux variables X et Y est sans dimension et varie de -1 à +1. Il mesure le ***degré de liaison linéaire*** entre les deux variables mais n'implique pas toujours une relation de cause à effet. Si la corrélation est positive, les valeurs élevées de Y sont préférentiellement associées à des valeurs élevées de X. En cas de corrélation négative, ce sont les valeurs faibles de Y qui sont associées préférentiellement aux valeurs élevées de X. Quand le coefficient de corrélation est élevé en valeur absolue, l'association est rigoureuse et le nuage de points est resserré autour d'une droite. *A contrario*, une valeur proche de zéro indique que la liaison linéaire est imparfaite, une valeur de zéro signifiant l'indépendance linéaire entre les deux variables.

Lorsque deux variables suivent conjointement une loi normale, leur coefficient de corrélation est le meilleur indicateur de leur association statistique. A l'inverse, deux variables peuvent avoir entre elles une liaison forte mais non linéaire (par exemple X sur l'ensemble des réels et $Y = X^2$) et être en corrélation nulle, car comme indiqué plus haut, la corrélation mesure une liaison qui est de type linéaire. Lorsque l'on ne connaît pas *a priori* le type de liaison entre deux variables, l'observation de leur distribution bidimensionnelle est indispensable et se restreindre au seul calcul du coefficient de corrélation pour juger de leur liaison peut conduire à de grossières erreurs.

3. La régression linéaire

L'existence d'une corrélation linéaire entre deux variables permet de prédire la valeur d'une variable à partir de la valeur prise par l'autre variable. En effet, connaissant X, il est possible de prédire Y par $\hat{Y}$, selon l'équation suivante :

$$\hat{Y} = a + bX$$

Le coefficient de régression (*b*) représente la pente de la droite, c'est-à-dire la quantité selon laquelle, ***en moyenne***, la variable *Y* varie lorsque la variable *X* croît d'une unité. Quant au coefficient *a*, il représente simplement l'ordonnée de la droite à l'origine (c.à.d. le zéro de la variable *X*). Les coefficients *a* et *b* de cette droite de régression sont déterminés de façon à minimiser la variance d'erreur entre les valeurs prédites par l'équation ci-dessus ($\hat{Y}$) et les valeurs réelles de *Y* sur l'échantillon considéré (méthode des moindres carrés, voir Stat). Plus la corrélation entre les deux variables est élevée, plus la prédiction de l'une à partir de l'autre est précise (plus $\hat{Y}$ est proche de *Y*).

L'intérêt de la technique de régression linéaire est de permettre une prédiction de la valeur d'un individu pour un caractère qui nous intéresse alors que la mesure en est difficile ou très coûteuse, voire impossible dans les conditions où l'on se trouve. L'objectif des mesures faites chez le bouquetin des Alpes et présentées à la figure 3 était justement de fournir un prédicteur simple du poids des animaux afin de permettre le suivi régulier des populations, notamment de la croissance des animaux. On conçoit en effet que, dans l'habitat naturel du bouquetin, le transport d'une bascule soit quelque peu malaisé. Un mètre-ruban est bien moins encombrant et d'un maniement très facile. Les prédictions que l'on peut faire du poids des animaux à partir de la simple meure de leur tour de poitrine sont jugées suffisamment fiables pour l'objectif que l'on s'est assigné : les erreurs de prédiction sont de faible ampleur et elles n'ont que peu de conséquences pratiques.

La régression fait l'objet de très nombreuses applications, dans des domaines très variés. Notamment, il sera largement discuté, dans le cadre des enseignements correspondants, de l'usage extrêmement courant que l'on fait de la régression linéaire en amélioration des plantes et en amélioration des animaux, pour prédire ce que l'on ne peut en général pas observer directement, la valeur génétique pour un caractère donné, à partir de ce que l'on voit, le phénotype.

MEMENTO STATISTIQUE

(inspiré de Minvielle, 1990)

Paramètre	Population de taille infinie		Estimateur sur un échantillon de taille *n*		Champ de variation
	Symbole	Définition	Symbole	Formule de calcul	
Moyenne	μ, μ_X	$\mathrm{E}(X)$	$\overline{X}$, $\hat{\mu}$	$\frac{1}{n}\sum X$	- ∞ , + ∞
Variance	σ^2, σ_X^2 V_X, $\mathrm{Var}(X)$	$\mathrm{E}\left(\left[X-\mathrm{E}(X)\right]^2\right)$ $=\mathrm{E}\left(X^2\right)-\left[\mathrm{E}(X)\right]^2$	s_X^2, $\hat{\sigma}_X^2$ $\hat{V}_X$	$\frac{1}{n-1}\left[\left(\sum X^2\right)-\left(\sum X\right)^2/n\right]$ $=\frac{1}{n-1}\left[\left(\sum X^2\right)-n\left(\overline{X}\right)^2\right]$	0 , + ∞
Covariance	σ_{XY} $\mathrm{Cov}(X,Y)$	$\mathrm{E}\left(\left[X-\mathrm{E}(X)\right]\left[Y-\mathrm{E}(Y)\right]\right)$ $=\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)$	s_{XY}	$\frac{1}{n-1}\left[\left(\sum XY\right)-\left(\sum X\right)\left(\sum Y\right)/n\right]$	- ∞ , + ∞
Corrélation	ρ_{XY} $r(X,Y)$	$\frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}$	$r(X,Y)$ $\hat{r}(X,Y)$	$\frac{s_{XY}}{s_X s_Y}=\frac{\left[\left(\sum XY\right)-\left(\sum X\right)\left(\sum Y\right)/n\right]}{\sqrt{\left[\left(\sum X^2\right)-\left(\left(\sum X\right)^2\right)/n\right]\left[\left(\sum Y^2\right)-\left(\left(\sum Y\right)^2\right)/n\right]}}$	- 1 , + 1
Régression	$\beta_{Y/X}$ $b_{Y/X}$	$\frac{\mathrm{Cov}(X,Y)}{\sigma_X^2}$	$b_{Y/X}$ $\hat{b}_{Y/X}$	$\frac{s_{XY}}{s_X^2}=\frac{\left[\left(\sum XY\right)-\left(\sum X\right)\left(\sum Y\right)/n\right]}{\left[\left(\sum X^2\right)-\left(\left(\sum X\right)^2\right)/n\right]}$	- ∞ , + ∞

EXERCICES

Le tableau ci-dessous donne les valeurs numériques ayant servi à l'établissement du graphe relatif à la liaison entre le tour de poitrine (TP) et le poids chez 35 mâles de bouquetin des Alpes (Toïgo, 1998 ; cf. figure 3 dans le chapitre I). A partir des éléments partiels de calcul qui sont fournis, calculer :

- la moyenne et l'écart-type de chacun des caractères.
- les coefficient de corrélation entre les deux caractères.
- l'équation de la droite de régression permettant de prédire le poids en fonction du tour de poitrine.

n°	TP (cm)	Poids (kg)	n°	TP (cm)	Poids (kg)
1	60,5	12,0	19	80,0	34,0
2	62,0	16,0	20	78,0	34,0
3	69,0	17,5	21	78,0	34,0
4	66,0	21,0	22	89,0	36,0
5	66,0	22,0	23	81,0	36,0
6	80,0	23,0	24	81,0	37,0
7	78,0	23,0	25	86,0	37,0
8	68,0	23,5	26	87,0	37,0
9	67,0	24,0	27	82,0	42,0
10	66,5	24,0	28	85,0	43,0
11	66,0	25,0	29	82,0	44,0
12	77,0	26,0	30	86,0	44,0
13	78,0	28,0	31	82,0	44,0
14	73,0	29,0	32	88,0	46,0
15	74,0	30,0	33	94,0	46,0
16	75,0	31,0	34	88,5	48,0
17	74,0	32,0	35	91,0	50,0
18	70,5	32,0			

	TP	Poids	TP x Poids
Somme	27 09,0	1 131,0	-
Somme des carrés	212 304,0	39 852,5	-
Somme des co-produits	-	-	90 132,5