

BULLETIN

DU

Musée royal d'Histoire  
naturelle de Belgique

Tome XXIV, n° 11,  
Bruxelles, avril 1948.

MEDEDEELINGEN

VAN HET

Koninklijk Natuurhistorisch  
Museum van België

Deel XXIV, n° 11.  
Brussel, April 1948.

ERREURS DE MESURE ET D'ÉCHANTILLONNAGE  
DANS LES GRAPHIQUES  
DE LA CROISSANCE RELATIVE,

par Elisabeth DEFRISE-GUSSENHOVEN (Bruxelles).

Le traitement quantitatif des phénomènes est un progrès décisif en biologie. Il a comme conséquence l'usage des mathématiques, dont aucun biologiste ne conteste plus l'utilité. Cependant, notons qu'en biologie une formule mathématique ne doit jamais constituer par elle-même un but de recherche, mais seulement l'un des moyens pour mieux décrire les faits : une formule ou une courbe doit toujours avoir un sens biologique.

Les graphiques sont une des formes les plus fréquentes d'étude quantitative des phénomènes biologiques. Ils donnent une représentation commode des données numériques et, parmi les multiples avantages qu'ils présentent, je ne veux citer que celui-ci : souvent l'examen, même rapide, d'un graphique, permet de trouver une loi qu'il serait bien long d'établir par le calcul. Supposons, par exemple, que l'on se demande si, au cours de la croissance d'un enfant, il existe une relation linéaire entre sa taille et la longueur de son bras, c'est-à-dire, désignant par  $x_t$  sa taille et par  $y_t$  la longueur de son bras à l'âge  $t$ , si l'on a, entre  $x_t$  et  $y_t$ , une relation linéaire telle que  $y_t = ax_t + b$ , vérifiée à chaque âge. D'assez longs calculs, notamment par l'application de la méthode des moindres carrés, permettent de voir s'il existe une telle relation. Mais il est beaucoup plus simple d'employer la méthode graphique. On considère  $x_t$  et  $y_t$

comme les coordonnées d'un point dans un plan rapporté à des axes rectangulaires et l'on vérifie, à la règle, si les points correspondant aux âges successifs de l'enfant, sont approximativement alignés (fig. I). Et s'ils le sont (c'est-à-dire si la loi linéaire se trouve vérifiée), c'est alors seulement que l'on calculera, par la méthode des moindres carrés, les constantes  $a$  et  $b$  de l'équation de la droite  $y_t = ax_t + b$ , qui passe par les différents points.

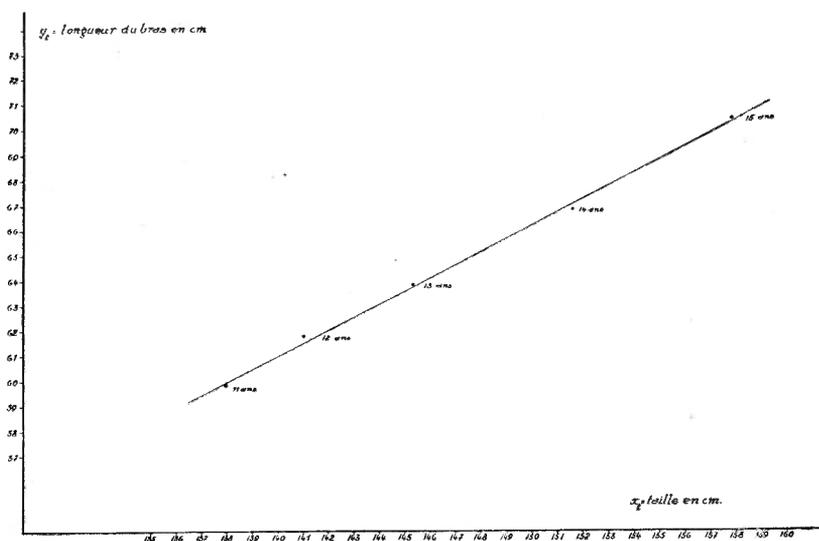


Fig. I. — Les points représentés ont comme abscisse la taille et comme ordonnée la longueur du bras d'un enfant, mesurées d'année en année. Il est facile de vérifier à la règle si ces points sont sensiblement alignés, c'est-à-dire si, au cours de la croissance, il existe une loi linéaire entre la taille et le bras de cet enfant.

La méthode graphique est parfois même indispensable. Je pense notamment à certaines courbes donnant la vitesse de croissance relative chez l'homme, qui révèlent, à la puberté, une double poussée de croissance très caractéristique (1). Ces courbes n'ont pratiquement pas d'équation calculable, de sorte

(1) Cette particularité de la croissance pubertaire a été mise en évidence par F. TWIESSELMANN, en vue d'un travail en cours d'impression, Mémoires du Musée royal d'Histoire naturelle, 1948.

que pour les étudier, il faut bien avoir recours à un diagramme.

Cependant, sous peine de conduire à des conclusions fausses, l'usage des graphiques requiert certaines précautions :

1° Tout d'abord, la condition suivante doit être satisfaite : une loi énoncée pour des grandeurs doit être indépendante du choix des unités utilisées pour mesurer ces grandeurs (2).

Reprenons l'exemple envisagé plus haut. Si  $x$  et  $y$  sont les mesures de la taille et du bras avec une certaine unité (le cm., par exemple), les mesures  $x'$ ,  $y'$  des mêmes grandeurs avec une autre unité  $l$  fois plus petite (le mm., par exemple) sont :

$$x' = lx, \quad y' = ly$$

Une loi linéaire liant bras et taille a un sens indépendant du choix des unités, car si l'on a

$$y = ax + b,$$

on a aussi

$$y' = ax' + lb.$$

De plus, on voit que non seulement la forme linéaire de la relation est conservée, mais aussi le coefficient  $a$ , qui mesure la pente de la droite représentative de la loi.

Ce coefficient  $a$ , autrement dit la pente de cette droite, a donc aussi un sens *biologique*.

Supposons maintenant que l'on recherche une relation entre la taille et le poids d'un enfant aux différents âges. Si  $x$  et  $p$  mesurent la taille et le poids avec des unités déterminées (cm. et kgr., par exemple), les mesures  $x'$  et  $p'$  de ces mêmes grandeurs avec d'autres unités (mm. et gr., par exemple) seront :

$$x' = lx, \quad p' = mp.$$

Une loi linéaire entre taille et poids peut avoir un sens, car si

$$x = ap + b, \quad x' = \frac{la}{m} p' + lb.$$

Mais ici, si la linéarité est conservée, le coefficient  $a$  ne l'est pas, de sorte que ce coefficient, c'est-à-dire la pente de la droite représentative de la loi, ne peut avoir de sens biologique.

(2) Cette remarque se rattache aux considérations de dimensions en physique et d'invariance en mathématique.

De la même façon, on voit immédiatement qu'une relation de la forme  $y = kx^\alpha$  peut exprimer une loi biologique, que  $x$  et  $y$  soient des tailles, des poids, ou d'autres mensurations; l'exposant  $\alpha$  a un sens intrinsèque, mais non le coefficient  $k$ .

Ces considérations peuvent être utiles dans deux directions: à la fois empêcher de rechercher en vain un sens biologique à des grandeurs mathématiques qui ne peuvent à priori en avoir aucun, et attirer l'attention sur des grandeurs mathématiques pour lesquelles on peut chercher une interprétation biologique.

2° Passons à une deuxième précaution essentielle.

Dans la recherche graphique d'une courbe exprimant une loi suivie par des grandeurs biologiques, on tâche de faire passer une courbe d'une certaine espèce (par exemple, une droite, une courbe d'équation  $y = kx^\alpha, \dots$ ) par des points représentatifs. Ainsi, dans l'exemple évoqué au début, on a tâché de faire passer une droite par les points représentatifs de la stature et de la longueur du bras de l'enfant aux différents âges.

Bien entendu, cette courbe ne passe jamais qu'*approximativement* par les différents points. Le tout est de juger si ces points sont assez voisins de la courbe tracée. Or, si l'on fait le dessin à une autre échelle, les distances des points à la courbe vont évidemment toutes être multipliées par un même facteur. De sorte que, suivant que l'on fait le dessin à petite ou à grande échelle, on sera tenté de dire que telle courbe est satisfaisante ou non, ce qui est absurde. Les écarts « acceptables » entre les points et la courbe doivent donc être jugés, non pas suivant un critère absolu de petite distance *géométrique* sur le graphique, mais suivant un critère lié aux phénomènes biologiques et aux procédés de mesure et qui donnera une distance *géométrique* variant avec l'échelle.

Dans ce qui suit, nous nous proposons d'indiquer un tel critère, en envisageant deux cas bien distincts: A. loi relative à *un individu*; B. loi relative à des *moyennes* pour une catégorie d'individus.

Nous ne citons que les hypothèses essentielles et les résultats, en reportant à la fin une esquisse de la suite des calculs. D'autre part, si tous nos exemples sont choisis dans la croissance relative, nos résultats s'appliquent évidemment aussi dans d'autres cas.

## A. LOI RELATIVE A UN INDIVIDU — ERREURS DE MESURE.

1<sup>er</sup> cas (fig. II).

Pour chaque grandeur on ne fait qu'une mesure.

Supposons qu'il s'agisse de trouver une loi de croissance relative entre la taille et le bras d'un enfant.

Mesurons sa taille à un âge donné: soit  $x$  la valeur lue sur la toise. Evaluons, d'après la précision de la toise, une limite supérieure de l'erreur  $\epsilon_x$  commise et, sur un graphique, portons en abscisses la valeur  $x$  et les deux quantités  $x - \epsilon_x$  et  $x + \epsilon_x$ , entre lesquelles on est sûr de trouver la « vraie » taille de l'enfant (3).

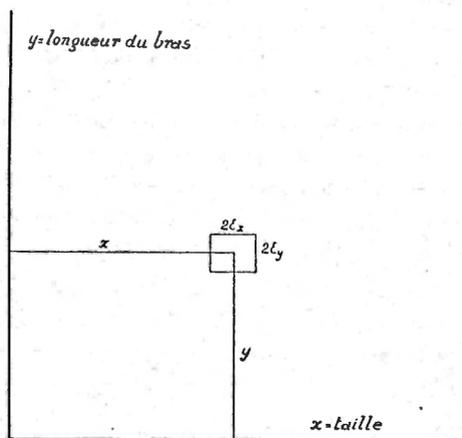


Fig. II. — Erreurs de mesure. — Pour chaque grandeur on ne fait, qu'une mesure. La « vraie » taille du sujet est comprise entre  $x - \epsilon_x$  et  $x + \epsilon_x$  et la « vraie » longueur du bras entre  $y - \epsilon_y$  et  $y + \epsilon_y$ . Le rectangle contient donc le point image des « vraies » longueurs de la taille et du bras à l'âge considéré. A chaque âge, on aura un rectangle pareil. Une courbe, image d'une loi de croissance relative taille/bras sera acceptable pourvu qu'elle passe à l'intérieur de chaque rectangle.

(3) Il n'est peut-être pas inutile de rappeler cette remarque que la taille d'un enfant (même en négligeant les difficultés spécifiques inhérentes aux grandeurs biologiques), comme la longueur d'une règle, qui seraient des nombres exactement définis, existent seulement en tant que concepts.

De même, mesurons son bras et portons en ordonnées la valeur  $y$  lue sur la toise ainsi que les quantités  $y - \epsilon_y$  et  $y + \epsilon_y$  entre lesquelles se trouve sûrement la « vraie » longueur du bras.

Le point  $[x, y]$ , résultat des mesures, est ainsi entouré d'un rectangle de côtés  $2\epsilon_x$  et  $2\epsilon_y$ , qui contient le point représentatif des longueurs de la taille et du bras à l'âge considéré.

Ainsi, les points  $[x, y]$ , résultats des mesures aux différents âges, seront tous entourés d'un rectangle et la courbe donnant la loi de croissance relative entre la taille et le bras, devra passer à l'intérieur de chaque rectangle, sans passer nécessairement par son centre.

On sent immédiatement les inconvénients de cette méthode. Comment évaluer  $\epsilon_x$ ? Il faut choisir  $\epsilon_x$  assez grand pour être certain que la taille se trouve entre  $x - \epsilon_x$  et  $x + \epsilon_x$  et, en même temps assez petit pour que la loi cherchée soit précise. Deux opérateurs différents donneront à  $\epsilon_x$  et à  $\epsilon_y$  des valeurs différentes : il n'y a aucun moyen univoque de les déterminer.

C'est pour remédier à ces inconvénients que l'on mesure *plusieurs fois* chaque grandeur, en appliquant aux résultats obtenus le calcul des probabilités, comme nous allons le voir.

2<sup>e</sup> cas (fig. III).

Pour chaque grandeur, on fait *plusieurs mesures* (4).

En mesurant plusieurs fois la taille de l'enfant, on trouve des valeurs légèrement différentes, si les graduations de la toise sont suffisamment serrées.

Voici 20 lectures obtenues en mesurant 20 fois la taille d'un même sujet avec une toise graduée en mm. :

1643 mm.	2 fois
1644 mm.	1 fois
1645 mm.	0 fois
1646 mm.	2 fois
1647 mm.	5 fois
1648 mm.	5 fois
1649 mm.	4 fois
1650 mm.	0 fois
1651 mm.	1 fois

(4) FELDMAN, W. M., 1935, *Biomathematics*. (London, p. 387.)

Les erreurs commises (5) sont dues à une accumulation de causes à effets très petits et suivant de très près la loi de GAUSS.

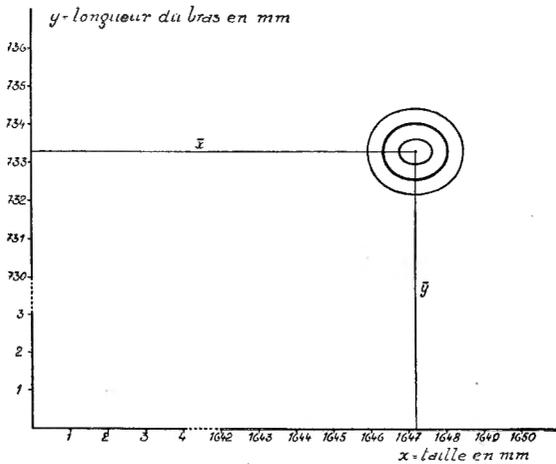


Fig. III. — Erreurs de mesure. — On fait *plusieurs mesures* pour chaque grandeur. Le centre des ellipses a comme abscisse la moyenne  $\bar{x}$  des mesures de la taille et comme ordonnée la moyenne  $\bar{y}$  des mesures du bras. La seconde ellipse contient, avec une probabilité de 91 %, le point image des « vraies » longueurs de la taille et du bras du sujet à l'âge donné. A chaque âge, on aura une ellipse pareille, qui devra être traversée par toute courbe représentant une loi de croissance relative taille/bras chez cet enfant.

Si l'on désigne par  $\bar{x}$  la moyenne arithmétique des  $n$  mesures et par  $\sigma_x$  leur dispersion, on sait que la longueur  $X$  de la taille se trouve

$$\text{entre } \bar{x} - \frac{\sigma_x}{\sqrt{n}} \text{ et } \bar{x} + \frac{\sigma_x}{\sqrt{n}}, \text{ avec une probabilité de } 68,3 \%,$$

$$\text{entre } \bar{x} - \frac{2 \sigma_x}{\sqrt{n}} \text{ et } \bar{x} + \frac{2 \sigma_x}{\sqrt{n}}, \text{ avec une probabilité de } 95,4 \%,$$

(5) Nous supposons ici qu'il n'y a pas d'erreurs systématiques, dont nous parlerons dans une note ultérieure.

entre  $\bar{x} - \frac{3 \sigma_x}{\sqrt{n}}$  et  $\bar{x} + \frac{3 \sigma_x}{\sqrt{n}}$ , avec une probabilité de 99,7 %, une quasi-certitude.

Ces formules appliquées aux chiffres donnés plus haut montrent que la « vraie » taille du sujet mesuré est comprise entre 1646,75 mm. et 1647,65 mm. avec une probabilité de 68,3 %, entre 1646,30 mm. et 1648,10 mm. avec une probabilité de 95,4 %, entre 1645,85 mm. et 1648,55 mm. avec une probabilité de 99,7 %.

Nous arrivons ainsi à une évaluation de l'erreur bien plus précise qu'en prenant une seule mesure de la taille.

Opérons de la même façon pour le bras. Voici les 20 mesures obtenues :

731 mm.	2 fois
732 mm.	6 fois
733 mm.	3 fois
734 mm.	5 fois
735 mm.	2 fois
736 mm.	0 fois
737 mm.	2 fois

Désignons par  $\bar{y}$  la moyenne arithmétique de ces quantités et par  $\sigma_y$  leur dispersion.

En effectuant les calculs, on trouve que la longueur Y du bras est comprise

entre 732,98 mm. et 733,72 mm. avec une probabilité de 68,3 %, entre 732,61 mm. et 734,09 mm. avec une probabilité de 95,4 %, entre 732,24 mm. et 734,46 mm. avec une probabilité de 99,7 %.

D'une manière générale, supposons que l'on fait  $n$  mesures de la taille et  $n'$  mesures du bras (dans notre exemple,  $n$  et  $n'$  valent 20, mais ils peuvent aussi ne pas être égaux).

Traçons, sur un graphique, avec le point  $[\bar{x}, \bar{y}]$  comme centre, trois ellipses concentriques, dont les axes sont parallèles aux axes coordonnés et ont respectivement comme longueur

$$\frac{6 \sigma_x}{\sqrt{n}} \quad \text{et} \quad \frac{6 \sigma_y}{\sqrt{n'}}$$

$$\frac{4 \sigma_x}{\sqrt{n}} \quad \text{et} \quad \frac{4 \sigma_y}{\sqrt{n'}}$$

$$\frac{2 \sigma_x}{\sqrt{n}} \quad \text{et} \quad \frac{2 \sigma_y}{\sqrt{n'}}$$

Comme les mesures de la taille et du bras à un âge donné sont *indépendantes*, il n'y a entre elles aucune corrélation, de sorte que l'ellipse extérieure contient, avec une quasi-certitude,

$99,7 \times \frac{99,7}{100} = 99,3 \%$ , le point  $[X, Y]$ , représentatif des longueurs de la taille et du bras.

La seconde ellipse contient le point  $[X, Y]$  avec une probabilité de  $95,4 \times \frac{95,4}{100} = 91,0 \%$  et l'ellipse intérieure avec une

probabilité de  $68,3 \times \frac{68,3}{100} = 46,6 \%$ .

Dans chaque problème particulier, on choisira l'ellipse selon le degré de certitude désiré.

Dans l'étude de la croissance relative, il me semble que la seconde ellipse donne une probabilité suffisante.

Il en résulte que l'on peut procéder comme suit. A chaque âge, on mesure la taille et le bras un certain nombre  $n$  et  $n'$  de fois; on calcule les moyennes arithmétiques  $\bar{x}$  et  $\bar{y}$  et les dispersions  $\sigma_x$  et  $\sigma_y$  des deux séries de mesures; on marque le point  $[\bar{x}, \bar{y}]$  sur le graphique et on l'entoure d'une ellipse d'axes

$$\frac{4 \sigma_x}{\sqrt{n}} \quad \text{et} \quad \frac{4 \sigma_y}{\sqrt{n'}}$$

Cela donne, pour notre exemple, le point de coordonnées  $[1647,20 \text{ mm.}, 733,35 \text{ mm.}]$  entouré d'une ellipse très petite dont les axes ont 1,80 mm. et 1,48 mm.

On a ainsi, pour chaque âge, un point, image des moyennes des mesures, entouré d'une ellipse. La courbe donnant la loi cherchée doit passer à l'intérieur de toutes les ellipses, puisque celles-ci contiennent, avec une probabilité de 91,0 %, les points représentatifs des longueurs de la taille et du bras.

On remarquera que si les nombres  $n$  et  $n'$  des mesures augmentent, les longueurs des axes des ellipses diminuent. Cependant, en pratique, on n'arrivera jamais à réduire l'ellipse à un point en augmentant indéfiniment le nombre de mesures. En effet, outre le travail excessif que cela demanderait, il faut se

dire que les chances de commettre des erreurs systématiques augmentent lorsque l'on accroît le nombre des mesures. Pour chaque sorte de grandeur, c'est l'expérience qui indiquera l'ordre de grandeur optimum pour le nombre de mesures. En biologie, il me semble qu'en prenant de 15 à 20 mesures par grandeur, on obtient une bonne précision.

B. PASSONS MAINTENANT AU CAS OÙ L'ON ÉTUDIE UNE LOI SE RAPPORTANT A DES MOYENNES POUR PLUSIEURS INDIVIDUS.

On dispose, par exemple, de plusieurs groupes d'enfants d'âges différents, mais provenant d'un milieu homogène. Il serait préférable d'avoir un seul groupe et de le suivre pendant plusieurs années. Chez les hommes, cela présente des difficultés manifestes, mais certains animaux à croissance rapide pourraient fort bien être mesurés à intervalles réguliers jusqu'à la fin de leur développement.

Prenons un premier groupe de  $n$  enfants, d'âge  $t$ . Mesurons la taille de chaque enfant et calculons la moyenne arithmétique  $\bar{x}$  de toutes les tailles trouvées.

Il suffit de faire une seule mesure par enfant. En effet, bien que chaque taille soit entachée d'une erreur de mesure, ces erreurs se compensent très sensiblement lorsqu'on fait la moyenne, pourvu qu'il y ait assez d'enfants et que chaque taille ait été mesurée avec la même précision (6).  $\bar{x}$  sera donc pratiquement égal à la moyenne arithmétique des « vraies » tailles des enfants.

Par conséquent, chaque fois que nous ferons des moyennes de grandeurs suffisamment nombreuses mesurées avec la même précision, nous n'aurons plus à nous occuper des erreurs de mesures.

Mais ces moyennes sont sujettes à un autre genre d'erreur : *les erreurs d'échantillonnage*.

En effet, revenons à nos  $n$  enfants d'âge  $t$ . Ils constituent un *échantillon* de l'ensemble plus vaste de tous les enfants du

(6) « Précision » a ici son sens mathématique. La paramètre de précision d'une série de mesures est égal à  $\frac{1}{2\sigma^2}$  et est d'autant plus grand que la dispersion des mesures est plus faible. DELTHEIL, R., 1930, *Erreurs et moindres carrés*. (Paris.)

même âge vivant dans les mêmes conditions qu'eux. Ce qui nous intéresse, c'est la taille moyenne de *la population*, et non pas la taille moyenne de l'échantillon : les lois que nous cherchons doivent se rapporter aux moyennes de cette population et non pas à celles de l'échantillon.

En termes de probabilités, nous pouvons déduire certains renseignements sur la structure du groupe complet à partir de données relatives à l'échantillon.

Par exemple, bien que nous ne puissions pas calculer la moyenne arithmétique  $\bar{X}$  des tailles du groupe complet, nous savons que cette valeur est comprise

entre  $\bar{x} - \frac{\sigma_x}{\sqrt{n}}$  et  $\bar{x} + \frac{\sigma_x}{\sqrt{n}}$ , avec une probabilité de 68,3 %,

entre  $\bar{x} - \frac{2 \sigma_x}{\sqrt{n}}$  et  $\bar{x} + \frac{2 \sigma_x}{\sqrt{n}}$ , avec une probabilité de 95,4 %,

entre  $\bar{x} - \frac{3 \sigma_x}{\sqrt{n}}$  et  $\bar{x} + \frac{3 \sigma_x}{\sqrt{n}}$ , avec une probabilité de 99,7 %,

si  $\bar{x}$  désigne la moyenne arithmétique des tailles de l'échantillon,  $\sigma_x$  leur dispersion et  $n$  leur nombre.

De même, la moyenne  $\bar{Y}$  des longueurs du bras du groupe total est comprise

entre  $\bar{y} - \frac{\sigma_y}{\sqrt{n}}$  et  $\bar{y} + \frac{\sigma_y}{\sqrt{n}}$ , avec une probabilité de 68,3 %,

entre  $\bar{y} - \frac{2 \sigma_y}{\sqrt{n}}$  et  $\bar{y} + \frac{2 \sigma_y}{\sqrt{n}}$ , avec une probabilité de 95,4 %,

entre  $\bar{y} - \frac{3 \sigma_y}{\sqrt{n}}$  et  $\bar{y} + \frac{3 \sigma_y}{\sqrt{n}}$ , avec une probabilité de 99,7 %,

si  $\bar{y}$  désigne la moyenne des longueurs du bras de l'échantillon et  $\sigma_y$  leur dispersion.

Le point  $[\bar{x}, \bar{y}]$ , image des moyennes de l'échantillon d'âge  $t$ ,

se trouve ainsi entouré de rectangles d'erreur qui contiennent avec une probabilité connue le point  $[\bar{X}, \bar{Y}]$  représentatif des moyennes de la population qui, seule, nous intéresse. Nous allons tâcher de réduire ces zones d'erreur.

Dans l'échantillon envisagé, il y a une certaine *corrélation* entre les longueurs de la taille et du bras: plus la taille d'un sujet est élevée, plus il a de chances d'avoir aussi un long bras. Calculons le coefficient de corrélation  $r$  dans l'échantillon d'âge  $t$ ; nous supposons qu'il est égal à celui du groupe total d'âge  $t$ , ce qui est le cas lorsque l'échantillon comprend un nombre suffisant de sujets.

Faisons encore les hypothèses suivantes: les variables taille et bras sont distribuées *normalement* dans le groupe total et la corrélation taille-bras  $y$  est aussi *normale*. Ces hypothèses, vérifiées dans de nombreux cas, permettent d'appliquer la théorie générale de l'échantillonnage, qui conduit aux résultats suivants.

Entourons le point  $[\bar{x}, \bar{y}]$  de trois ellipses concentriques, dites d'égale probabilité, ayant comme équations

$$\frac{(x - \bar{x})^2}{\sigma_x^2} - \frac{2r (x - \bar{x})(y - \bar{y})}{\sigma_x \sigma_y} + \frac{(y - \bar{y})^2}{\sigma_y^2} = \text{constante C.} \quad (\text{fig. IV})$$

L'ellipse extérieure, pour  $C = \frac{9}{n} (1 - r^2)$ , contient le point  $[\bar{X}, \bar{Y}]$ , image des moyennes du groupe total, avec une probabilité de  $99,7 \times \frac{99,7}{100} = 99,3 \%$ .

La seconde ellipse, pour  $C = \frac{4}{n} (1 - r^2)$ , contient le point  $[\bar{X}, \bar{Y}]$  avec une probabilité de  $95,4 \times \frac{95,4}{100} = 91,0 \%$ .

L'ellipse intérieure, pour  $C = \frac{1}{n} (1 - r^2)$ , contient  $[\bar{X}, \bar{Y}]$  avec une probabilité de  $68,3 \times \frac{68,3}{100} = 46,6 \%$ .

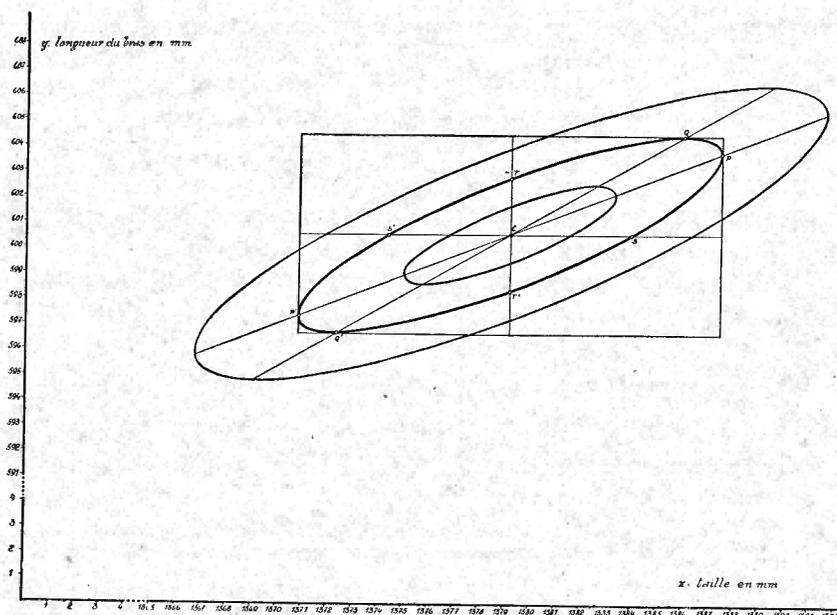


Fig. IV. — Erreurs d'échantillonnage. Ellipses concernant le groupe des garçons de Pro Juventute âgés de 11 ans. — Le point C a comme coordonnées la moyenne  $\bar{x}$  des tailles et la moyenne  $\bar{y}$  des longueurs du bras. Les droites TT' et PP' d'une part, SS' et QQ' d'autre part, sont des couples de diamètres conjugués pour les 3 ellipses. La seconde ellipse contient avec une probabilité de 91 % le point  $[\bar{X}, \bar{Y}]$ , représentatif des moyennes de la population totale des garçons de 11 ans.

Chacune des trois ellipses a comme centre le point  $[\bar{x}, \bar{y}]$  et comme couples de diamètres conjugués (7) :

d'une part, la droite de régression  $y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$   
 et la droite  $x = \bar{x}$  ;

d'autre part, l'autre droite de régression  $\bar{x} - x = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$   
 et la droite  $y = \bar{y}$ .

(7) Deux diamètres d'une ellipse sont conjugués lorsque chacun joint le milieu des cordes parallèles à l'autre.

Comme dans le cas des erreurs de mesures, nous choisirons la seconde ellipse, d'équation

$$\frac{(x - \bar{x})^2}{\sigma_x^2} - \frac{2r(x - \bar{x})(y - \bar{y})}{\sigma_x \sigma_y} + \frac{(y - \bar{y})^2}{\sigma_y^2} = \frac{4}{n} (1 - r^2)$$

qui contient le point  $[\bar{X}, \bar{Y}]$  avec une probabilité élevée (91,0 %).

Voici quelques propriétés de cette ellipse. Elle est tangente aux quatre côtés du rectangle de centre  $[\bar{x}, \bar{y}]$  et dont les

côtés ont  $\frac{4 \sigma_x}{\sqrt{n}}$  et  $\frac{4 \sigma_y}{\sqrt{n}}$ .

Les points de contact sont

$$P \equiv \left[ \bar{x} + \frac{2 \sigma_x}{\sqrt{n}}, \bar{y} + \frac{2r \sigma_y}{\sqrt{n}} \right]$$

$$P' \equiv \left[ \bar{x} - \frac{2 \sigma_x}{\sqrt{n}}, \bar{y} - \frac{2r \sigma_y}{\sqrt{n}} \right]$$

$$Q \equiv \left[ \bar{x} + \frac{2r \sigma_x}{\sqrt{n}}, \bar{y} + \frac{2 \sigma_y}{\sqrt{n}} \right]$$

$$Q' \equiv \left[ \bar{x} - \frac{2r \sigma_x}{\sqrt{n}}, \bar{y} - \frac{2 \sigma_y}{\sqrt{n}} \right]$$

L'ellipse coupe la droite parallèle à l'axe  $ox$  ( $y = \bar{y}$ ) aux points

$$S \equiv \left[ \bar{x} + \frac{2 \sigma_x \sqrt{1 - r^2}}{\sqrt{n}}, \bar{y} \right]$$

$$S' \equiv \left[ \bar{x} - \frac{2 \sigma_x \sqrt{1 - r^2}}{\sqrt{n}}, \bar{y} \right]$$

Elle coupe la droite parallèle à l'axe  $oy$  ( $x = \bar{x}$ ) aux points

$$T \equiv \left[ \bar{x}, \bar{y} + \frac{2 \sigma_y \sqrt{1 - r^2}}{\sqrt{n}} \right]$$

$$T' \equiv \left[ \bar{x}, \bar{y} - \frac{2 \sigma_y \sqrt{1 - r^2}}{\sqrt{n}} \right]$$

Ces points permettent facilement de tracer l'ellipse.

Pour chaque échantillon d'âge donné, nous entourons d'une telle ellipse le point connu  $[\bar{x}, \bar{y}]$ , image des moyennes (fig. V).

La loi cherchée, liant au cours de la croissance les moyennes de la taille et du bras, sera représentée par une courbe assujettie à passer à l'intérieur de chaque ellipse, sans devoir passer nécessairement par chaque centre.

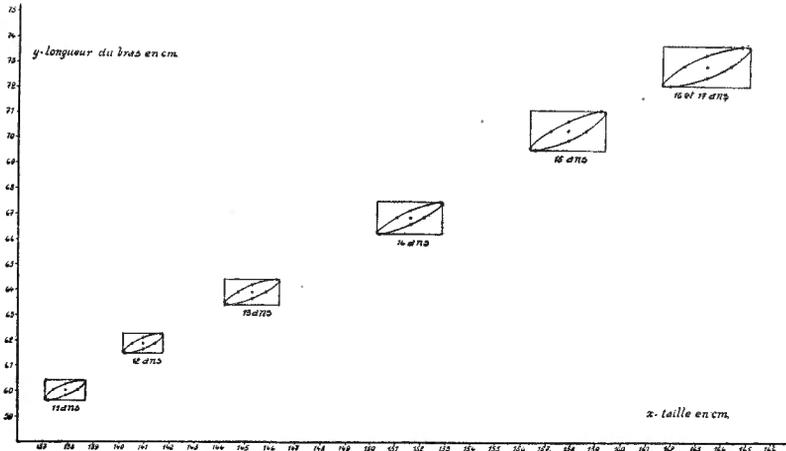


Fig. V. — Garçons de Pro Juventute groupés par âge. Ellipses d'erreurs dues à l'échantillonnage. — La première ellipse, par exemple, est relative aux garçons de 11 ans. Son centre a comme abscisse la moyenne  $\bar{x}$  des tailles de l'échantillon connu et comme ordonnée la moyenne  $\bar{y}$  des longueurs du bras. Elle contient, avec une probabilité de 91 %, le point inconnu représentant les moyennes de la taille et du bras de la population totale des garçons de 11 ans. Une courbe d'une loi de croissance relative taille-bras, sera acceptable pourvu qu'elle passe à l'intérieur de chaque ellipse. Voir, figure IV, un tracé détaillé de l'ellipse concernant les garçons de 11 ans.

A titre d'exemple, voici les chiffres se rapportant à un groupe de 223 enfants de 11 ans (8) (cf. fig. IV) :

$$\begin{aligned}\bar{x} &= 137,89 \text{ cms}; & \sigma_x &= 6,21 \text{ cms}; & r &= 0,82 \\ \bar{y} &= 60,06 \text{ cms}; & \sigma_y &= 2,94 \text{ cms}; & n &= 223.\end{aligned}$$

Le calcul des quantités

$$\begin{aligned}& \frac{2 \sigma_x}{\sqrt{n}}, & \frac{2 \sigma_y}{\sqrt{n}} \\ & \frac{2r \sigma_x}{\sqrt{n}}, & \frac{2r \sigma_y}{\sqrt{n}} \\ & \frac{2 \sigma_x \sqrt{1-r^2}}{\sqrt{n}}, & \frac{2 \sigma_y \sqrt{1-r^2}}{\sqrt{n}}\end{aligned}$$

permet de déterminer les coordonnées des points P, P', Q, Q', S, S', T, T' grâce auxquels on peut tracer l'ellipse.

#### CONCLUSIONS.

Les mathématiques peuvent rendre de précieux services en biologie; mais, notamment dans les graphiques, il ne faut pas négliger certaines précautions. Lorsqu'on représente deux grandeurs, résultats d'expérience, de mesures ou moyennes de mesures, par un point d'un plan, une question essentielle est celle-ci: *quelle est la zone d'incertitude qui entoure ce point?* Une courbe, assujettie à passer par le point, doit passer à l'intérieur de cette zone.

Nous avons résolu la question dans les cas suivants :

1° Le point représente deux mesures faites sur un seul sujet; il est alors entouré d'un rectangle défini par la limite supérieure de l'erreur sur chaque mesure.

2° Le point représente encore deux quantités relatives à un seul sujet, mais pour chacune on fait plusieurs mesures. La

(8) Il s'agit des enfants de Pro Juventute. Leurs mesures, relevées par F. TWIESSELMANN et G. GERKENS, se trouvent à la section d'Anthropologie et de Préhistoire du Musée royal d'Histoire naturelle.

zone d'incertitude est délimitée par une ellipse, dont les dimensions dépendent de la précision des mesures.

3° Le point est l'image des moyennes de deux grandeurs mesurées chez tous les sujets d'un échantillon, qui est tiré d'un univers homogène relativement à ces grandeurs. Moyennant *certaines hypothèses* sur la normalité des variables, on peut délimiter la zone d'incertitude par une ellipse, dont les dimensions dépendent des erreurs dues à l'échantillonnage et de la corrélation éventuelle qui existe entre les deux variables.

Si maintenant l'on cherche une loi liant deux quantités qui varient, soit avec le temps, soit avec la température, soit avec la nourriture, etc., c'est-à-dire si l'on cherche une courbe passant par les différents points d'un graphique, on a un critère précis: elle doit passer à l'intérieur de la zone d'incertitude dont chaque point est entouré.

Parmi plusieurs courbes satisfaisant à cette condition, on choisira de préférence celle qui

1° a une équation simple,

2° passe le plus près des centres des rectangles ou des ellipses.

Signalons toutefois que si une loi absolument rigoureuse liait les vraies valeurs dont les points du graphique ne sont qu'une approximation, il n'y aurait que des chances *très faibles* pour que la courbe, image de la loi, passe précisément par les centres. Autrement dit, une courbe qui passerait exactement par les centres ne doit pas faire illusion sur la rigueur de la loi qu'elle représente (9).

Ces résultats trouvent une large application dans l'étude de la croissance relative, où les points des graphiques varient avec le temps. Qu'il s'agisse d'hommes, d'animaux ou de plantes, il est essentiel de tenir compte des erreurs de mesures, lorsqu'on suit le développement d'un seul sujet, ou des erreurs d'échantillonnage, lorsqu'on étudie la croissance d'un groupe. Dans l'exemple que nous avons choisi pour les erreurs de mesure, celles-ci étaient faibles, la taille d'un homme étant une mesure assez facile à prendre. Mais pour d'autres mesures, on pourrait trouver une ellipse d'erreur plus grande. Quant aux ellipses d'erreurs dues à l'échantillonnage, elles sont presque toujours

(9) RISSER, R., 1932, *Application de la statistique à la démographie et à la biologie*. (Paris, pp. 248 et 249.)

importantes à considérer et il est bon de les marquer sur tous les graphiques concernant la croissance relative. Par exemple, lorsque celle-ci semble présenter un changement de la constante de croissance, les ellipses d'erreur permettent de se rendre compte si la courbe change de pente d'une façon brusque ou continue.

Notons que, si l'on est amené à choisir la loi d'allométrie  $y = kx^z$ , on a avantage à passer aux coordonnées logarithmiques. Cela ne présente aucune difficulté. Les ellipses dont nous avons parlé deviennent, sur le nouveau graphique, des courbes fermées entourant les points  $[\log. \bar{x}, \log. \bar{y}]$ . Il est facile de tracer ces courbes, puisque l'on en connaît huit points et les tangentes en quatre de ces points. Le point

$P \equiv \left[ \bar{x} + \frac{2 \sigma_x}{\sqrt{n}}, \bar{y} + \frac{2 r \sigma_y}{\sqrt{n}} \right]$  devient, par exemple, le point de

coordonnées  $\log. \left( \bar{x} + \frac{2 \sigma_x}{\sqrt{n}} \right), \log. \left( \bar{y} + \frac{2 r \sigma_y}{\sqrt{n}} \right)$ .

Les ellipses d'erreur peuvent aussi servir en systématique, lorsqu'on examine des graphiques de corrélation entre deux caractères pour des espèces différentes. Chaque espèce est figurée par un nuage de points dont chacun est l'image des deux caractères chez un individu et par un point central, image des moyennes des deux caractères. Si l'on entoure chaque point central d'une ellipse d'erreur, on peut tirer certaines conclusions des positions relatives de ces ellipses. Lorsqu'elles sont extérieures les unes aux autres, la divergence entre les moyennes n'est pas due aux erreurs d'échantillonnage, mais à des différences inhérentes aux espèces examinées (10). Si deux ellipses ont une partie commune, il y a une certaine probabilité pour que la divergence des centres soit due uniquement aux erreurs d'échantillonnage; mais le calcul de cette probabilité étant assez compliqué, on a d'abord recours à d'autres critères, tels la comparaison des moyennes par le calcul de l'erreur moyenne sur la différence des moyennes et la comparaison des coefficients de corrélation.

Je tiens à remercier M. GLIBERT qui a bien voulu lire mon manuscrit et à exprimer ma reconnaissance au docteur

(10) Il est bien entendu que les échantillons doivent avoir un nombre suffisant de sujets.

F. TWIESELNANN qui m'a constamment guidé de ses précieux conseils; sans son aide, je n'aurais pas écrit ce petit article.

ESQUISSE DE LA SUITE DES CALCULS.

Voici comment on trouve l'équation des ellipses d'égale probabilité dans le cas des erreurs dues à l'échantillonnage.

Soit P la population dont on a tiré l'échantillon E. En gardant les notations précédentes, appelons  $\bar{X}$ ,  $\bar{Y}$  les moyennes de la taille et du bras de la population. Appelons  $\bar{x}$ ,  $\bar{y}$ ,  $\sigma_x$ ,  $\sigma_y$ ,  $r$  les moyennes, les dispersions et la corrélation de la taille et du bras dans l'échantillon E. Supposons que l'on tire de P d'autres échantillons de  $n$  sujets et appelons  $\bar{x}_i$ ,  $\bar{y}_i$  les moyennes d'un tel échantillon  $E_i$  ( $i = 1, 2, \dots, i, \dots, k$ ).

1° Si  $n$  est assez grand,  $r$  est aussi le coefficient de corrélation taille/bras dans la population P.

2° Si les variables taille et bras sont distribuées normalement dans P et si la corrélation qui les lie est elle aussi normale,

$\frac{\sigma_x}{\sqrt{n}}$  et  $\frac{\sigma_y}{\sqrt{n}}$  sont les dispersions de  $\bar{x}_i$  et  $\bar{y}_i$  autour de  $\bar{X}$  et  $\bar{Y}$  et

le coefficient de corrélation liant  $\bar{x}_i$  et  $\bar{y}_i$  est égal à  $r$  (11).

3° Il s'ensuit que la surface normale de corrélation entre  $\bar{x}_i$  et  $\bar{y}_i$  a comme équation (12)

$$(*) z = \frac{k}{2\pi \frac{\sigma_x \cdot \sigma_y}{n} \sqrt{1-r^2}} e^{-\frac{n}{2(1-r^2)} \left\{ \frac{(x - \bar{X})^2}{\sigma_x^2} - \frac{2r(x - \bar{X})(y - \bar{Y})}{\sigma_x \cdot \sigma_y} + \frac{(y - \bar{Y})^2}{\sigma_y^2} \right\}}$$

Elle donne en chaque point  $\left. \begin{array}{l} x = \bar{x}_i \\ y = \bar{y}_i \end{array} \right\}$  la fréquence  $z_i$  de ce point.

Elle est coupée par des plans parallèles au plan  $xy$  suivant des ellipses d'égale probabilité, dont les projections sur le plan  $xy$  sont des ellipses d'équation

(11) KENDALL, M. G., 1945, *The Advanced Theory of Statistics*. (London, I, p. 340.)

(12) YULE, G. U. and KENDALL, M. G., 1946, *An Introduction to the Theory of Statistics*. (London.)

$$\frac{(x - \bar{X})^2}{\sigma_x^2} - \frac{2r(x - \bar{X})(y - \bar{Y})}{\sigma_x \cdot \sigma_y} + \frac{(y - \bar{Y})^2}{\sigma_y^2} = \text{constante C.}$$

Ces ellipses ont comme couples de diamètres conjugués la droite de régression  $y - \bar{Y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{X})$  avec la droite  $x = \bar{X}$  et l'autre droite de régression  $x - \bar{X} = r \frac{\sigma_x}{\sigma_y} (y - \bar{Y})$  avec la droite  $y = \bar{Y}$ .

Appelons  $\xi$  l'ellipse pour laquelle C prend la valeur  $\frac{4}{n} (1 - r^2)$ .

Elle est inscrite dans le rectangle de centre  $[\bar{X}, \bar{Y}]$  et de côtés  $\frac{4\sigma_x}{\sqrt{n}}$  et  $\frac{4\sigma_y}{\sqrt{n}}$  et contient  $95,4 \times \frac{95,4}{100} = 91,0 \%$  des points  $[\bar{x}_i, \bar{y}_i]$ . Elle contient donc le point  $[\bar{x}, \bar{y}]$  avec une probabilité de  $91,0 \%$ .

Inversement, l'ellipse d'équation

$$\frac{(x - \bar{x})^2}{\sigma_x^2} - \frac{2r(x - \bar{x})(y - \bar{y})}{\sigma_x \sigma_y} + \frac{(y - \bar{y})^2}{\sigma_y^2} = \frac{4}{n} (1 - r^2),$$

égale à l'ellipse  $\xi$ , mais de centre connu  $[\bar{x}, \bar{y}]$ , contient le point inconnu  $[\bar{X}, \bar{Y}]$  avec une probabilité de  $91,0 \%$ .

On passe facilement de ce cas à celui des erreurs de mesures en faisant  $r = 0$  dans l'équation (\*) et en indiquant que les dispersions de  $\bar{x}_i$  et  $\bar{y}_i$  autour de X et Y sont  $\frac{\sigma_x}{\sqrt{n}}$  et  $\frac{\sigma_y}{\sqrt{n}}$ .

On trouve alors l'équation suivante

$$\frac{n(x - \bar{x})^2}{\sigma_x^2} + \frac{n'(y - \bar{y})^2}{\sigma_y^2} = 4$$

qui est celle d'une ellipse de centre  $[\bar{x}, \bar{y}]$ , d'axes parallèles aux axes coordonnés et de longueur  $\frac{4\sigma_x}{\sqrt{n}}$  et  $\frac{4\sigma_y}{\sqrt{n}}$ . Elle contient le point  $[X, Y]$  avec une probabilité de  $91,0 \%$ .