REVUE DE STATISTIQUE APPLIQUÉE

F. CHARTIER

Détermination d'une zone de confiance et d'un intervalle de tolérance dans le cas d'une régression linéaire par rapport à une variable

Revue de statistique appliquée, tome 5, n° 3 (1957), p. 121-132

<http://www.numdam.org/item?id=RSA_1957__5_3_121_0>

DÉTERMINATION D'UNE ZONE DE CONFIANCE ET D'UN INTERVALLE DE TOLÉRANCE DANS LE CAS D'UNE RÉGRESSION LINÉAIRE PAR RAPPORT A UNE VARIABLE

par

M. F. CHARTIER

Administrateur à l'Institut National de la Statistique et des Etudes Economiques

Sous réserve de la validité des hypothèses connues (fluctuations aléatoires de la variable expliquée y *distribuées suivant une loi normale de variance constante quelle que soit la valeur de la variable explicative* x*), la méthode des moindres carrés permet d'estimer la valeur moyenne de* y *pour* x *donné et d'avoir sous la forme* y = f(x) *une loi d'estimation moyenne de* y *en fonction de* x.

La loi ainsi obtenue est basée sur un échantillon, le résultat est donc aléatoire et n'est qu'une estimation de la loi réelle (de forme convenablement choisie) décrivant au mieux cette lésion dans la population infinie que l'on peut concevoir.

Il y a donc lieu de préciser, relativement à cette estimation moyenne:

1° *l'incertitude liée à la connaissance de cette ligne d'estimation elle-même,*

2° *les fluctuations possibles de la variable* y *pour* x *donné, autour de cette droite.*

M. CHARTIER étudie ce problème dans le cas particulier d'une régression linéaire, c'est-à-dire d'une relation linéaire descriptive de la variation, en moyenne de y *en fonction* x.

L'un des buts de l'étude de la corrélation qui peut exister entre deux variables est la prévision de la valeur moyenne de l'une d'elles pour une valeur de l'autre.

Voici deux exemples : la Fig. 1 montre que le temps passé pour récolter une tonne de maïs dépend du rendement du champ considéré : il décroît quand le rendement augmente et inversement. On dit qu'il y a corrélation négative . C'est là

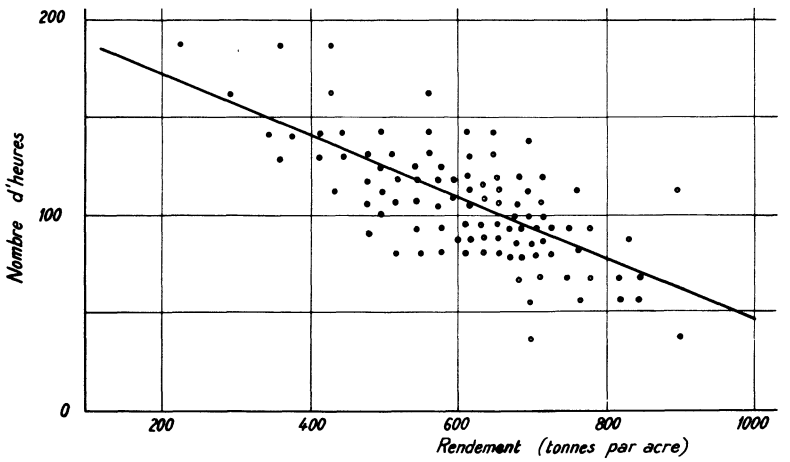

Fig. 1.- Nombre d'heure nécessaires pour récolter une tonne de maïs, en fonction du rendement.

une constatation de bon sens, mais que l'étude de données numériques permet de préciser par la détermination d'une courbe, ici une droite dite **droite de régression** ou mieux **d'estimation** , qui est tracée sur la Fig. 1 et qui permet de prévoir le temps moyen qui sera nécessaire à la récolte d'une tonne pour des champs de rendement donné, même si ce rendement n'a pas été observé au cours de l'établissement des données de base. La Fig. 2 montre comment varient les appréciations visuelles de rendement par un enquêteur qui a visité 37 champs de blé, en fonction du rendement réel de ces champs : les appréciations varient, en moyenne, dans le même sens que les rendements réels - l'enquêteur sera d'autant meilleur que ceci sera plus vrai. On dit alors qu'il y a corrélation positive. On a tracé sur la Fig. 2 une droite qui indique que la loi de variation de l'appréciation moyenne rendement en fonction du rendement réel. On note que cette droite présente une pente un peu plus faible que la diagonale Y = X : l'enquêteur perçoit bien comment un champ se classe parmi les autres, mais il minimise, dans un rapport d'ailleurs constant, le l'ordre de 0,7, l'écart de ce champ a la moyenne. Ainsi, lorsqu'un champ a un rendement réel supérieur de 10 à la moyenne, il ne lui attribue qu'une supériorité de $10 \times 0,7 = 7$ environ. Ayant ainsi "étalonné" notre enquêteur, lorsqu'il rentre d'une visite au cours de laquelle il a apprécié un certain nombre de rendements, nous pouvons déduire de ces appréciations de meilleures estimations du rendement réel.
rendement réel.

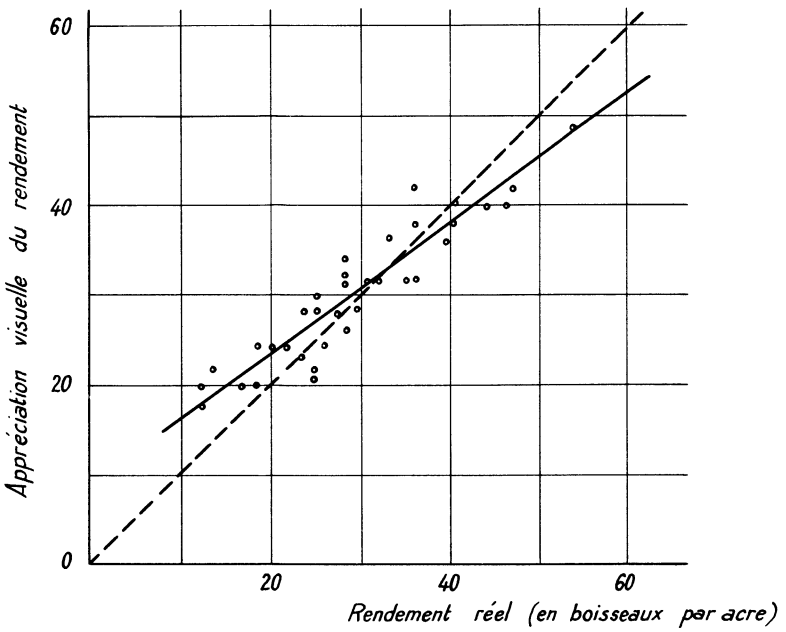

Fig. 2.- Appréciation visuelle du rendement de 37 champs de blé, en fonction du rendement réel.

On pourrait multiplier à l'infini de tels exemples d'étude de corrélation : dans la recherche industrielle, étude des propriétés d'un métal en fonction de sa teneur en certains éléments,... - en économie, étude de la valeur des actions en fonction de leur taux de rapport...

La détermination de la loi de variation de la moyenne de l'une des variables - la variable expliquée - en fonction de l'autre - la variable explicative - ne constitue pas le seul aspect de l'étude de la corrélation entre ces deux variables. Pour une valeur donnée (x) de la variable explicative, du rendement si l'on s'appuie sur le premier exemple ci-dessus, la variable expliquée est susceptible de fluctuations

aléatoires autour de la valeur moyenne f (x) fixée par la droite d'estimation (droite si la liaison est linéaire). On peut alors rechercher un intervalle f (x) - a à f (x) + b tel qu'une proportion donnée, P (= 0,50 ou 0,95 par exemple), des champs de maïs demande un nombre d'heures pour récolter une tonne compris dans cet intervalle. Cet intervalle est généralement appelé **intervalle de tolérance** à P (à 0,50 ou 0,95). Si l'on opère ainsi pour toutes valeurs de x, on obtient deux courbes de part et d'autre de la droite d'estimation qui détermine une **bande de tolérance** à P (à 0,50 ou 0,95), c'est-à-dire une bande telle que, sur l'ensemble des champs il y en aura une proportion P qui nécessiteront un temps à la tonne compris dans cette bande. Le cas le plus simple est celui où les fluctuations aléatoires du temps, pour un rendement donné, sont distribuées suivant une loi normale, de même dispersion (de même écart type) quel que soit le rendement (la variable explicative). En portant 0,67 fois l'écart type de part et d'autre de la droite d'estimation, on a la bande de confiance à 0,50, en le portant 1,96 fois, celle à 0,95.

Mais la détermination de la droite, ou d'une façon plus générale, de la courbe d'estimation et de la bande de tolérance que nous venons d'exposer, suppose que l'on examine l'ensemble des champs de maïs, ou l'ensemble des appréciations de notre observateur, que l'on connaisse toute la **population** comme disent les statisticiens. En fait, il n'en est pas ainsi. On ne dispose que d'un **échantillon d'observations** . Et c' est à partir de cet échantillon que l'on détermine la droite d'estimation (pour des raisons de simplicité, on s'en tient souvent à une droite, à moins que l'allure curviligne du nuage de points ne l'interdise absolument) et la bande de tolérance. On conçoit alors que :

- d'une part, la droite d'estimation obtenue sur un échantillon sera quelque peu différente de celle que l'on aurait obtenue sur un autre, de sorte que la droite déterminée à partir de l'échantillon unique dont on dispose devrait être assortie d'une zone, aléatoire comme elle, mais qui contiendrait avec une probabilité donnée, Π_1, la droite de la population ; cette zone est dite zone **de confiance** (à Π_1, par exemple à 0,90 ou 0,95) ;

- d'autre part, la bande de tolérance à P, déterminée toujours à partir de l'échantillon disponible, contient en fait une proportion P' de la population différente de P. Tout ce que l'on peut espérer, c'est déterminer cette bande de tolérance telle que P' soit au moins égale à P avec une probabilité donnée Π_2.

Il est à peu près intuitif que, sous l'hypothèse énoncée plus haut quant aux fluctuations de la variable expliquée pour une valeur donnée de la variable explicative, **aussi bien la zone de confiance que la bande de tolérance vont s'élargissant lorsque la variable explicative s'écarte de la moyenne des valeurs observées.** Le temps moyen estimé nécessaire à la récolte d'une tonne sera estimé de façon plus précise pour un champ dont le rendement est de 600 tonnes par acre que pour un champ de rendement 1.000 ou seulement 200 (même si la même loi linéaire est valable pour tout l'intervalle de 200 à 1.000). En effet, une erreur aléatoire (due à l'échantillon particulier disponible) sur le coefficient angulaire de la droite ajustée au nuage de points donnera des différences d'ordonnées de cette droite d'autant plus grandes que l'on s'écartera davantage de la valeur moyenne des rendements observés. Il en va de même de la bande de tolérance, bien que le raisonnement qui permet d'établir ce résultat soit plus complexe.

Il résulte des considérations qui précèdent que l'on doit s'arranger, au moment de l'organisation des observations, pour que celles-ci soient réparties dant tout l'intervalle de variation de la variable explicative.

La présente note, inspirée par un article de W. ALLEN WALLIS, "Tolerance Intervals for Linear Regression", paru dans "Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability" a pour objet de montrer,

sur un exemple concret, la méthode à suivre pour la détermination de ces deux éléments : zone de confiance et bande de tolérance.

Les éléments pour traiter le premier problème se trouvent dans la plupart des traités d'Analyse statistique (1). Aussi serons-nous plus bref que pour l'exposé de la solution du second problème, solution qui n'est d'ailleurs qu'approximative.

BASE MATHÉMATIQUE COMMUNE AUX DEUX PROBLÈMES

Soient x la variable explicative et y la variable expliquée. On suppose que, pour x donné, la valeur moyenne de y est :

$$\eta = \alpha + \beta x,$$

fonction linéaire de x, tandis que les fluctuations aléatoires de y autour de cette moyenne suivent une loi normale, d'écart type σ, constant quel que soit x.

A partir d'un échantillon de n observations indépendantes x_i, y_i on estime :

$$\alpha \text{ par } a = \overline{y} - b\,\overline{x}$$

et

$$\beta \text{ par } b = \frac{S\,(x_i - \overline{x})\,(y_i - \overline{y})}{S\,(x_i - \overline{x})^2}$$

où $\overline{x} = S\,x_i / n$ et $\overline{y} = S y_i / n$ (moyenne arithmétique des n observations de x et de y), le signe S désignant une sommation étendue aux n valeurs observées de x et de y.

La moyenne estimée y' de y pour une valeur quelconque x de la variable explicative est :

$$y' = a + b\,x$$

(équation de la droite d'estimation de y en fonction de x), tandis que la variance ou carré de l'écart type σ caractérisant l'amplitude des fluctuations de y autour de sa moyenne pour x donné est estimée par :

$$s^2 = \frac{S\,(y_i - y'_i)^2}{n - 2} = \frac{S\,(y_i - \overline{y})^2 - b^2\,S\,(x_i - \overline{x})^2}{n - 2}$$

où $y'_i = a + b\,x_i$. s est parfois appelé "erreur type" de l'estimation y', appellation assez impropre puisque s ne caractérise pas les fluctuations d'échantillonnage de y', mais celles de y autour de sa valeur moyenne η (ou de son estimation y').

SOLUTION DU PREMIER PROBLÈME : DÉTERMINATION D'UNE ZONE DE CONFIANCE POUR LA DROITE D'ESTIMATION DE Y EN X

L'estimation $y' = \overline{y} + b\,(x - \overline{x})$ de la valeur moyenne de y pour x donné peut s'écrire :

$$y' = S\left[\frac{1}{n} + \frac{x_i - \overline{x}}{S\,(x_i - \overline{x})^2}(x - \overline{x})\right] \cdot y_i$$

soit une forme linéaire des y_i. La variance de chacun des y_i, indépendants entre eux, autour de leur moyenne respective $\eta_i = \alpha + \beta x_i$ est σ^2 ; de sorte que la variance de y' est :

$$\sigma^2_{y'} = S\,k_i^2\,\sigma^2 = \left[\frac{1}{n} + \frac{(x - \overline{x})^2}{S\,(x_i - \overline{x})^2}\right]\sigma^2$$

(1) Voir par exemple : Méthode statistique, par E. Morice et F. Chartier, Deuxième partie Analyse statistique, Chap. IV.

en appelant k_i le coefficient de y_i dans l'expression ci-dessus de y'. On estimera $\sigma^2_{y'}$ par :

$$s^2_{y'} = \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{S\,(x_i - \bar{x})^2} \right] s^2$$

On voit sur l'expression de $\sigma^2_{y'}$, comme sur celle de son estimation $s^2_{y'}$, que la variance de y' croit avec le carré de l'écart de x à $\bar{x}$, le minimum ayant lieu pour $x = \bar{x}$. On peut encore dire que la précision de l'estimation y' de la moyenne η pour x donné, déduite de l'équation de la droite ajustée, équivaut à celle d'une estimation faite à l'aide de :

$$N = \frac{1}{\frac{1}{n} + \frac{(x - \bar{x})^2}{S\,(x_i - \bar{x})^2}} = n\,\frac{S\,(x_i - \bar{x})^2}{S\,(x_i - x)^2}$$

observations de y pour x ayant la valeur considérée. La variance de cette dernière estimation, moyenne de N observations indépendantes serait, en effet,

$$\sigma^2 / N = \sigma^2_{y'}$$

Par exemple, pour les données étudiées au tableau I, comportant 24 observations, la moyenne estimée pour x = 30, pour laquelle $1/N = 0,0428$, a la même précision que si l'on avait fait environ 23 observations de champs dont le rendement soit 30, puis que l'on en ait pris la moyenne arithmétique. En revanche, la moyenne estimée pour x = 50 ($1/N = 0,2687$) n'a qu'une précision du même ordre que la moyenne de 4 observations seulement faites sur des champs de rendement 50.

Avec les hypothèses déjà mentionnées, on démontre que la variable aléatoire

$$t = (y' - \eta) \cdot \frac{\sqrt{N}}{s}$$

suit une loi de Student de moyenne nulle, avec $\nu = n - 2$ degrés de liberté, laquelle se confond pratiquement avec une loi normale (0,1) si $\nu > 30$.

Le seuil de confiance Π étant fixé, c'est-à-dire aussi la valeur t_Π de la variable de Student ou normale telle qu'il y ait une probabilité Π que $|t| < t_\Pi$, on peut dire qu'il y a une probabilité Π que :

$$y' - t_\Pi \frac{s}{\sqrt{N}} < \eta < y' + t_\Pi \frac{s}{\sqrt{N}}$$

La détermination des deux termes extrêmes de cette double inégalité pour toute valeur de x définit la zone de confiance à Π (par exemple à 0,95) pour la droite d'estimation de y en x.

Il existe des tables de la loi de Student donnant t_Π pour diverses valeurs de Π et de ν.

APPLICATION : Reprenons les appréciations visuelles de rendement de 37 champs de blé déjà citées, mais, afin de rendre plus sensible l'élargissement de la zone de confiance quand x s'écarte de $\bar{x}$, ne conservons que 24 observations par élimination systématique de 1 observation sur 3 après les avoir rangées dans l'ordre croissant des valeurs de x (et si nécessaire de y), ce qui ne modifie pas sensiblement l'allure générale du nuage de points, ni l'équation de la droite d'estimation de y en fonction de x.

On trouve au tableau I les 24 couples (x , y) et les moyennes et sommes de carrés qui s'en déduisent. La droite d'estimation ajustée, $y' = 9,684 + 0,707\,x$, est représentée sur la Fig. 3. Son équation permet d'estimer la moyenne y' autour de laquelle se dispersent les observations y, avec un écart type estimé $s = 3,13$.

Pour calculer les limites de l'intervalle $y' \pm t_\Pi \cdot s/\sqrt{N}$ recouvrant avec une probabilité $\Pi = 0,95$ la moyenne vraie η, inconnue, on a utilisé la valeur du t de Student, soit 2,07. Cette valeur est voisine du t normal, 1,96.

On voit que la demi-amplitude de l'intervalle $t_\pi\, s/\sqrt{N}$, croît comme $1/\sqrt{N}$ quand x s'écarte de $\bar{x} = 28,50$, son minimum, pour $x = \bar{x}$ étant de 1,32. Les limites inférieures et supérieures de la zone de confiance figurent dans les deux dernières colonnes au bas du tableau I et sont représentées sur la Fig. 3.

TABLEAU I -

Détermination de la zone de confiance à 0,95 pour une droite d'estimation.

Appréciations visuelles de rendement (y) en fonction du rendement réel (x). Les n observations :

x	y	x	y	x	y	x	y
12	20	24	21	28	32	36	38
13	22	24	22	28	34	36	42
17	20	24	30	29	30	41	38
18	24	25	24	29	32	41	40
22	24	26	28	32	32	46	40
23	23	28	26	35	32	47	42

$$n = 24 \quad \bar{x} = 28,500 \quad \bar{y} = 29,833$$

$$S\,(x - \bar{x})^2 = 2.036 \quad S\,(x - \bar{x})\,(y - \bar{y}) = 1439 \quad S\,(y - \bar{y})^2 = 1233$$

$$b = \frac{S\,(x - \bar{x})\,(y - \bar{y})}{S\,(x - \bar{x})^2} = 0,707 \quad a = \bar{y} - b\,\bar{x} = 9,684$$

Droite d'estimation ajustée = $\underline{y' = 9,684 + 0,707\,x}$

$$S\,(y - y')^2 = S\,(y - \bar{y})^2 - b^2\,S\,(x - \bar{x}) = 215$$

Ecart type de y pour x donné : $s = \sqrt{\frac{S\,(y - y')^2}{n - 2}} = 3,13$

Pour $\Pi = 0,95$ et $\nu = n - 2 = 22$ degrés de liberté, $t_\Pi = 2,07$

x	$\frac{1}{24} + \frac{(x - 28,50)^2}{2036}$ $= 1/N$	$\frac{2,07 \times 3,13}{\sqrt{N}}$ $= t_\Pi\, s/\sqrt{N}$	$9,68 + 0,707\,x$ $= y'$	$y' - t_\Pi\, s/\sqrt{N}$	$y' + t_\Pi\, s/\sqrt{N}$
0	0,4407	4,30	9,68	5,38	13,98
10	0,2098	2,97	16,75	13,78	19,72
20	0,0762	1,79	23,82	22,03	25,61
28,50	0,0417	1,32	29,83	28,51	31,15
30	0,0428	1,34	30,89	29,35	32,23
40	0,1066	2,11	37,96	35,85	40,07
50	0,2687	3,36	45,03	41,67	48,39
60	0,5291	4,71	52,10	47,39	56,81

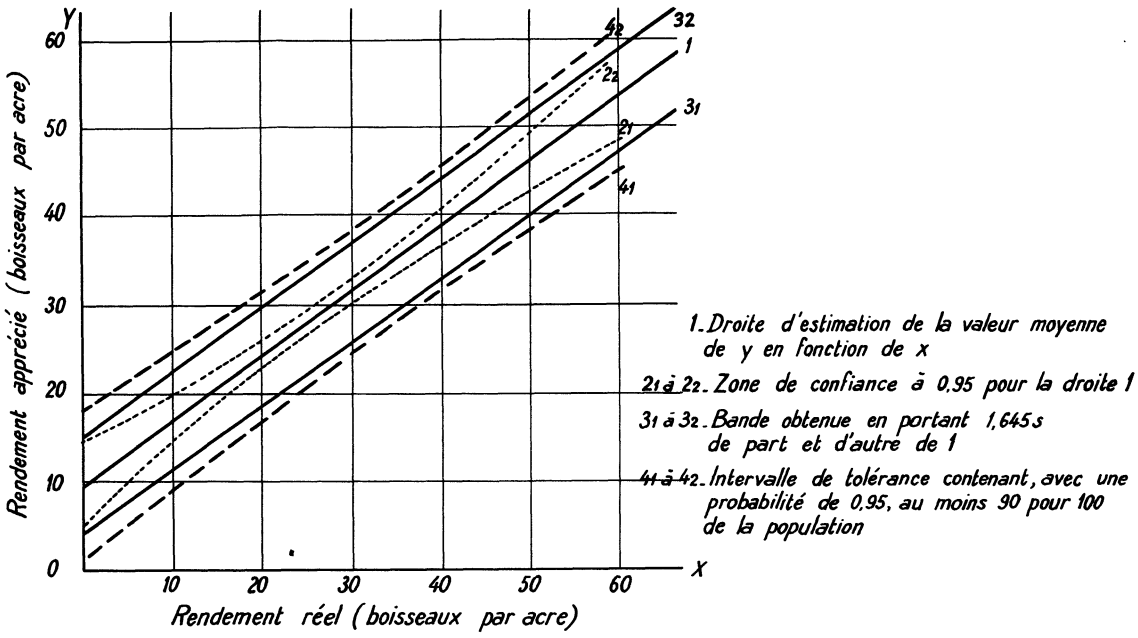

Fig. 3.- Etude d'un échantillon de 24 observations (cf. tableaux I et II).

SOLUTION DU SECOND PROBLÈME : DÉTERMINATION D'UN INTERVALLE DE TOLÉRANCE

(contenant une proportion de la population au moins égale à P avec une probabilité II, P et II étant donnés).

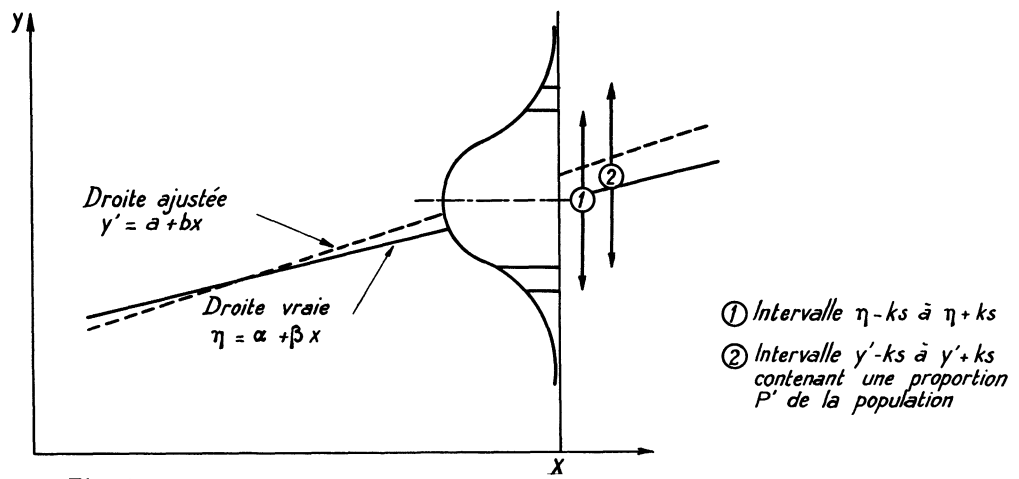

Fig. 4.- Variation de la proportion P' de la population dans l'intervalle y' - ks à y' + ks en fonction de y' et de s

Pour une valeur donnée de x, la variable y est distribuée normalement autour de la moyenne η ($= \alpha + \beta x$), avec l'écart type σ (Fig. 4). L'échantillon nous donne la moyenne estimée y' ($= a + b x$) et l'écart type estimé s. Le problème consiste à trouver une constante k telle qu'il y ait une probabilité II que la proportion aléatoire de la population comprise dans l'intervalle aléatoire $y' \pm ks$ soit au moins égale à P, ceci pour toutes valeur de la variable x.

On voit sur la Fig. 4 que, pourvu que l'intervalle soit assez grand (limites se trouvant dans les ailes de la courbe normale), c'est-à-dire si P est assez grand, les fluctuations aléatoires de y' modifient peu la proportion P' de la population effectivement incluse entre $y' \pm ks$. Il y a, en effet, à peu près compensation entre ce qui est perdu à la limite inférieure et ce qui est gagné à la limite supérieure si y' est supérieur à η et inversement si y' est inférieur à η (ne pas oublier que y' est une variable normale, centrée sur η, d'écart type $\sigma/\sqrt{N}$, qui ne s'écarte pas beaucoup de σ tant que N est grand.

En revanche, toute fluctuation aléatoire de s a un effet sensible sur P', car il n'y a plus glissement de l'intervalle avec compensation approximative aux deux extrémités, mais accroissement (si s augmente) ou diminution (si s diminue) de son amplitude et par suite de P'.

On montre (1) que, pour la détermination pratique de k, on peut considérer $y' - \eta$ comme ayant la valeur moyenne $\sigma / \sqrt{N}$ (c'est-à-dire précisément l'écart type de y').

On doit alors déterminer k tel qu'il y ait une probabilité Π (par exemple 0,95) que la proportion de la population comprise entre $\eta + \sigma / \sqrt{N} \pm ks$ soit au moins égale à P, c'est-à-dire que l'on ait avec une probabilité Π :

$$\frac{1}{\sqrt{2\Pi}} \int_{\eta + \frac{\sigma}{\sqrt{N}} - ks}^{\eta + \frac{\sigma}{\sqrt{N}} + ks} e^{-\frac{(y-\eta)^2}{2\sigma^2}} \frac{dy}{\sigma} \geqslant P$$

La réalisation de cette inégalité dépend évidemment de la valeur de la variable aléatoire s.

Soit k_p la valeur de k telle que:

$$\frac{1}{\sqrt{2\Pi}} \int_{\eta + \frac{\sigma}{\sqrt{N}} - k_p\sigma}^{\eta + \frac{\sigma}{\sqrt{N}} + k_p\sigma} e^{-\frac{(y-\eta)^2}{2\sigma}} \frac{dy}{\sigma} = P$$

(Il existe une seule valeur k_p satisfaisant à cette condition pour P donné car l'intégrale définie du premier membre croît de façon continue de 0 à 1 quand k varie de 0 à l'infini).

La condition énoncée précédemment pour la détermination de k équivaut à dire qu'il faut qu'il y ait une probabilité Π que

$$k\, s \geqslant k_p\, \sigma.$$

Or, on connaît la loi de s : la quantité $\nu s^2 / \sigma^2$ suit une loi de χ^2 à $\nu = n - 2$ degré de liberté. Soit alors χ^2_Π la valeur de χ^2 qui est dépassée avec une probabilité Π. On doit prendre :

$$k \geqslant k_p \Big/ \sqrt{\frac{\chi^2_\Pi}{\nu}} = \frac{k_p \sqrt{\nu}}{\chi_\Pi}$$

(1) Voir le principe de la démonstration en Appendice.

DÉTERMINATION PRATIQUE DE Kp :

Dans l'équation qui définit k , on peut supposer $\eta = 0$ et $\sigma = 1$, de sorte que l'équation devient :

$$\frac{1}{\sqrt{2\Pi}} \int_{\frac{1}{\sqrt{N}} - k}^{\frac{1}{\sqrt{N}} + k} e^{-y^2/2} \, dy = P$$

Pour N donné, les tables usuelles de la fonction intégrale de la loi normale

$$F(t) = \frac{1}{\sqrt{2\Pi}} \int_{-\infty}^{t} e^{-x^2/2} \, dx$$

permettrait de trouver une valeur approchée de k_p. On montre que l'on a approximativement, pourvu que $N \geqslant 1$:

$$k_p = t_p \left(1 + \frac{1}{2N} - \frac{2t_P^2 - 3}{24 N^2}\right)$$

où t_P est la valeur de la variable normale telle que $F(t_p) - F(-t_p) = P$, valeur donnée par les tables usuelles. Indiquons seulement ici les valeurs vraisemblablement les plus courantes de t_p et de $(2t_p^2 - 3)/24$:

P	t_p	$\frac{2t_P^2 - 3}{24}$	P	t
0,50	0,674	- 0,087	0,00	0,000
0,80	1,282	+ 0,012	0,60	0,842
0,90	1,645	0,101		
0,95	1,960	0,195		
0,98	2,326	0,326	0,96	2,054
0,99	2,576	0,428		

(Les valeurs pour P = 0,00, 0,60, 0,96 sont utiles dans la détermination de χ_Π ci-après).

DÉTERMINATION PRATIQUE DE χ_Π :

Les tables de R.A. FISHER donnent, pour Π et ν fixés, la valeur de χ_Π^2 (et non de χ_Π) ; mais les valeurs de Π que l'on trouve dans les tables usuelles sont 0,90, 0,80, 0,70, 0,50..., ce qui est parfois insuffisant. Onpeut alors utiliser l'approximation :

$$\chi_\Pi^2 = \nu \left(1 - \frac{2}{9\nu} - t_{2\Pi - 1} \sqrt{\frac{2}{9\nu}}\right)^3$$

où $t_{2\Pi-1}$ est la valeur de la variable normale t défini comme t_p ci-dessus (Si $\Pi < 1/2$, remplacer $- t_{2\Pi-1}$ par $+ t_{1-2\Pi}$).

Cette approximation est satisfaisante pour $\nu \geqslant 3$. Pour $\nu = 1$ et 2 il est plus simple d'utiliser les valeurs exactes qui sont respectivement :

$$\chi_\Pi^2 = t_{1-\Pi}^2 \quad \text{et} \quad \chi_\Pi^2 = - 2 \log_e \Pi = - 0{,}8686 \log_{10} \Pi$$

On peut encore utiliser l'approximation :

$$\chi_\Pi^2 = \frac{1}{2} \left(\sqrt{2\nu - 1} - t_{2\Pi - 1}\right)^2$$

plus simple, mais moins précise, au moins tant que ν n'atteint pas une trentaine.

On trouve, au tableau ci-après, la valeur de $\frac{\sqrt{\nu}}{\chi_\Pi}$ pour les valeurs indiquées de Π et ν :

	Π = 0,50	0,80	0,90	0,95	0,98	0,99
ν = 1	1,484	3,947	7,956	15,95	39,90	79,81
2	1,201	2,117	3,079	4,406	7,036	9,975
3	1,126	1,728	2,266	2,989	4,027	5,108
4	1,092	1,557	1,939	2,372	3,053	3,670
6	1,060	1,398	1,647	1,916	2,300	2,623
8	1,044	1,319	1,514	1,711	1,984	2,205
12	1,029	1,240	1,380	1,515	1,695	1,833
20	1,018	1,171	1,268	1,357	1,471	1,556
30	1,011	1,133	1,207	1,274	1,356	1,417
40	1,008	1,112	1,173	1,229	1,295	1,342
60	1,005	1,087	1,135	1,179	1,231	1,265
120	1,003	1,058	1,090	1,118	1,153	1,175

Ces valeurs ont été obtenues à partir de la table de χ^2 de Fisher, les valeurs de χ^2 y étant lues avec 3 décimales. Il est donc probable que la troisième décimale donnée ci-dessus est inexacte. Toutefois il est à penser que l'erreur absolue est de l'ordre de quelques millièmes seulement.

APPLICATION : Utilisons encore les 24 couples d'observations du tableau I (rendement réel x, et applications visuelles y). Fixons P à 0,90 et Π à 0,95, c'est à-dire déterminons la bande de tolérance contenant, avec une probabilité Π = 0,95 au moins une proportion P = 0,90 (90 %) de la totalité des observations que l'enquêteur pourrait exécuter dans les mêmes conditions.

On trouve au tableau II les phases successives de cette détermination.

D'abord le calcul de k_p, ou plutôt de k_p / t_p. Il suffira ensuite pour obtenir ks d'écrire $ks = (k_p/t_p)\left[t_p \left(\sqrt{\nu} / \chi_\Pi\right) s\right]$, le facteur entre crochets ayant une valeur constante.

La valeur de $\sqrt{\nu} / \chi_\Pi$ ne figure pas dans la table donnée ci-dessus pour ν = 22, mais seulement pour ν = 20 et ν = 30 (avec Π = 0,95). Par interpolation, il vient :

$$1,357 + (1,274 - 1,357) \cdot \frac{2}{10} = 1,340 .$$

La formule approchée :

$$\frac{\nu}{\chi^2_\Pi} = \frac{1}{\left(1 - \frac{2}{9\nu} - t_{2\Pi-1} \sqrt{\frac{2}{9\nu}}\right)^3} \quad \text{avec } t_{2\Pi-1} = t_{0,90} = 1,645$$

donne $\nu/\chi^2_\Pi = 1,7836$ et $\sqrt{\nu}/\chi_\Pi = 1,335$, valeur ne différent que de 0,005, soit 4°/$_{oo}$ environ de celle obtenue par interpolation. On a utilisé 1,335.

On calcule ensuite $t_p \left(\sqrt{\nu} / \chi_\Pi\right) s$, soit 6,874. D'où les valeurs de ks = 6,874 (k_p / t_p), puis celles des limites supérieure et inférieure de la zone de tolérance $y' \pm ks$.

La Fig. 3 représente les deux courbes limitant cette zone. On voit qu'elle va s'élargissant aux deux extrémités, quand x s'écarte de $\bar{x}$ = 28,50. Elle s'élargit toutefois relativement moins que l'intervalle de confiance servant à l'estimation de la droite $\eta = \alpha + \beta x$. En revanche, la zone de confiance ainsi déterminée est

sensiblement plus large que la bande de largeur constante obtenue en portant parallèlement à l'axe des y, de part et d'autre de la droite ajustée y' = a + bx, soit :

$$t_p \; s = 1,645 \; s = 5,15, \; t_p \text{ variable normale}$$

Soit : $t_p \; s = 1,717 \; s = 5,37$, t_p variable de Student pour $\nu = 22$ d. l. ce qui, en toute rigueur, est incorrect d'une manière comme de l'autre. C'est ce qu'on voit sur la Fig. 3 où l'on a représenté les deux droites $y' \pm 5,15$.

Détermination de la bande de tolérance contenant, avec la probabilité $\Pi = 0,95$, une proportion au moins égale à $P = 0,90$ de la population.

TABLEAU II

Détermination de la bande de tolérance contenant, avec la probabilité $\Pi = 0,95$, une proportion au moins égale à $P = 0,90$ de la population.

Données du Tableau I.

Pour $P = 0,90$ on a $t_p = 1,645$ et $(2 t_p^2 - 3)/24 = 0,101$

Pour $\Pi = 0,95$ et $\nu = 22$, $\sqrt{\nu} / \chi_\Pi = 1,335$. D'autre part, $s = 3,13$ (1)

x	$\frac{1}{N}$ (1)	$1 + \frac{1}{2N} - \frac{2t_p^2 - 3}{24 N^2} = k_p / t_p$	$k_p \frac{\sqrt{\nu}}{\chi_\Pi} s = ks$	y' (1)	y' - ks	y' + ks
0	0,4407	1,2008	8,25	9,68	1,43	17,93
10	0,2098	1,1005	7,56	16,75	9,19	24,31
20	0,0762	1,0375	7,13	23,82	16,69	30,95
28,50	0,0417	1,0207	7,02	29,83	22,81	36,85
30	0,0428	1,0212	7,02	30,89	23,87	37,91
40	0,1066	1,0522	7,23	37,96	30,73	45,19
50	0,2687	1,1271	7,75	45,03	37,28	52,78
60	0,5291	1,2363	8,50	52,10	43,60	60,60

APPENDICE

Sur la valeur moyenne de la différence $y' - \eta$.

HYPOTHÈSES : Pour x donné, la population est normalement distribuée par rapport à y, avec la moyenne η et la variance σ^2, deux quantités inconnues pour lesquelles on dispose des estimations indépendantes respectives :

y', de la loi normale $(\eta, \sigma / \sqrt{N})$.

s^2, distribué comme $\sigma^2 \chi^2 / \nu$ pour ν degrés de liberté.

La proportion P' de la population comprise dans l'intervalle y' - ks à y' + ks est:

$$P' = \frac{1}{\sqrt{2\Pi}} \int_{y'-ks}^{y'+ks} e^{-\frac{(y-\eta)^2}{2\sigma^2}} \frac{dy}{\sigma}$$

C'est une variable aléatoire comme y' et s (k est une constante) qui est au moins égale à P si le point de coordonnées y', s se trouve dans une certaine région R_o du plan des y', s schématisée sur la Fig. 5.

Cette région R_o admet comme axe de symétrie la verticale $y' = \eta$ et elle est limitée inférieurement par une courbe dont la concavité est tournée vers le haut

(1) D'après le tableau 1.

(grandes valeurs de s) : pour deux valeurs opposées mais de même module de $y' - \eta$, il faut que s ait au moins une certaine valeur, la même que $y' - \eta$ soit positif ou négatif, mais d'autant plus grande que le module de $y' - \eta$ est plus grand.

Désignons par F (P) la probabilité totale que P' soit au moins égale à P et par F (P; z) la probabilité conditionnelle que P' soit au moins égale à P, sachant que $y' - \eta = z$. La loi de probabilité élémentaire de $y' - \eta = z$ est

$$\frac{1}{\sqrt{2\Pi}} e^{-Nz^2/2\sigma^2} \sqrt{N}\, dz/\sigma .$$

On a :

$$F(P) = \int_{z=-\infty}^{z=\infty} \frac{1}{\sqrt{2\Pi}} e^{-\frac{Nz^2}{2\sigma^2}} \sqrt{N}\, \frac{dz}{\sigma} \times F(P ; z)$$

$$= E \left[F (P ; z) \right]$$

en désignant par le symbole E l'espérance mathématique.

Le développement en série de Taylor de F (P ; z) s'écrit, compte tenu de ce que F (P ; z) est une fonction symétrique de z:

$$F (P ; z) = F (P ; o) + \frac{z^2}{2!} \frac{\partial^2 F}{\partial z^2}(P ; o) + \frac{z^4}{4!} \frac{\partial^4 F}{\partial z^4}(P ; 0) + \ldots .$$

En prenant l'espérance mathématique de ce développement, comme :

$$E (z^2) = \sigma^2/N \quad \text{et} \quad E (z^4) = 3\, \sigma^4/N^2,$$

on obtient :

$$F (P) = E \left[F (P ; z) \right]$$

$$= F (P ; o) + \frac{\sigma^2}{2N} \frac{\partial^2 F}{\partial z^2}(P ; o) + \frac{3\, \sigma^4}{4!\, N^2} \frac{\partial^4 F}{\partial z^4}(P ; o) + \ldots .$$

Si l'on rapproche le développement ci-dessus de F (P ; z) et le développement ainsi obtenu F (P), pour F (P), on voit que l'on a approximativement :

$$F (P) \simeq F (P ; \sigma / \sqrt{N}),$$

la différence entre les deux développements n'apparaissant que pour les termes en $1/N^2$, $1/N^4$, ... qui sont petits si N est grand.

Ayant ainsi fixé l'un des deux paramètres dont dépend F, à savoir y', on est ramené à un problème à une seule variable s dont on a exposé la solution plus haut.

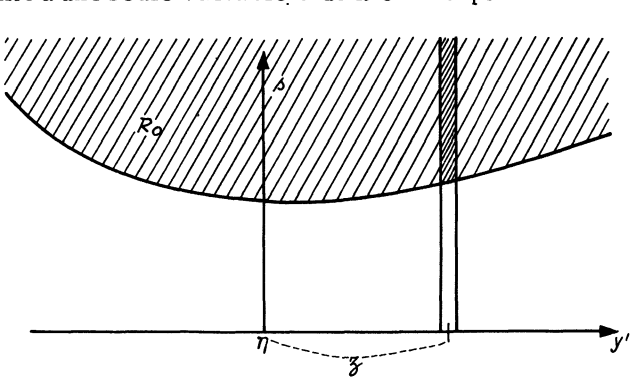

Fig. 5