REVUE DE STATISTIQUE APPLIQUÉE

E. MORICE

Les méthodes d'analyse de la variance

Revue de statistique appliquée, tome 3, nº 2 (1955), p. 65-82
<http://www.numdam.org/item?id=RSA_1955__3_2_65_0>

© Société française de statistique, 1955, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

LES MÉTHODES D'ANALYSE DE LA VARIANCE

par

E. MORICE

Inspecteur Général à l'Institut National de la Statistique et des Études Économiques

L'apparente simplicité avec laquelle s'appliquent les méthodes classiques d'analyse de la variance à des schémas élémentaires de comparaisons d'échantillons ou d'études de validité de liaisons entre variables, n'est pas sans inconvénients.

Faute de s'être assuré du fait que les hypothèses de base sont vérifiées, ou tout au moins acceptables, en première approximation, ou encore par suite d'une erreur d'appréciation des paramètres (« degrés de liberté ») qui entrent en jeu dans les tests utilisés, on risque d'aboutir à des conclusions erronées ou sans rapport réel avec la question posée.

Sans présenter une théorie complète de la question, qui dépasserait de beaucoup le cadre d'un article de revue, il n'est sans doute pas dénué d'intérêt d'essayer de préciser brièvement les hypothèses de base et les conditions de validité de l'emploi des tests statistiques utilisés.

I. - DISTRIBUTION DE χ^2

On donne ce nom à la distribution d'une variable aléatoire, généralement désignée par la lettre grecque χ^2, dont la loi de probabilité élémentaire est définie par :

$$p(\chi^2)\, d(\chi^2) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\nu/2)} (\chi^2)^{\frac{\nu}{2}-1} e^{-\frac{\chi^2}{2}} d(\chi^2)$$

$$0 < \chi^2 < \infty \ .$$

Cette loi, en plus de la variable aléatoire χ^2 dépend donc d'un paramètre ν appelé nombre de degré de liberté de la distribution.

La distribution de χ^2 a été mise en tables qui, le plus généralement, donnent les fractiles supérieurs de la distribution, c'est-à-dire, les valeurs χ_0^2, telles que

$$p\left[\chi^2 \geqslant \chi_0^2\right] = \int_{\chi_0^2}^{\infty} p(\chi^2)\ d(\chi^2) = P \ ,$$

pour diverses valeurs de P, en fonction du nombre ν de degrés de liberté.

Ainsi pour $\nu = 10$, $P = 0,05$, on lit directement dans la table :

$$\left[P\ (\chi^2 \geqslant 18,31)\ ,\ \nu = 10\right] = 0,05$$

Pour $\nu = 10$, il y a cinq chances sur 100 que χ^2 dépasse la valeur $\chi_0^2 = 18,31$.

On peut résumer ainsi qu'il suit les propriétés essentielles de cette distribution :

1 - Si $u_1, u_2, \ldots, u_n$ sont n observations **indépendantes** constituant un échantillon tiré d'une population **normale** de moyenne nulle et de variance égale à l'unité

les variables aléatoires $\chi^2 = \sum_{i=1}^{n} u_i^2$ correspondant aux divers échantillons que l'on pourrait concevoir à partir de cette population sont distribuées comme χ^2 avec :

$$\nu = n \text{ degrés de liberté}$$

Il en résulte immédiatement que si $x_1, \ldots, x_n$ sont dans les mêmes hypothèses (indépendance et normalité), n observations constituant un échantillon tiré d'une population de moyenne m et de variance σ^2, les variables :

$$\chi^2 = \frac{\sum (x_i - m)^2}{\sigma^2}$$

sont distribuées comme χ^2 avec :

$$\nu = n \text{ degrés de liberté .}$$

2 - Il arrive assez généralement en pratique que la moyenne m de la population soit inconnue et estimée par la moyenne $\overline{x}$ de l'échantillon.

Dans ces conditions, et sous réserve des mêmes hypothèses que ci-dessus, on démontre encore que la quantité :

$$\chi^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{\sigma^2}$$

est distribuée comme χ^2, mais avec

$$\nu = n - 1 \text{ degrés de liberté}$$

Sous cette forme, la distribution de χ^2 permettra de traiter les deux problèmes suivants :

(a) - Si la variance de la population est connue, ou si l'on fait une hypothèse sur la valeur σ^2 de cette variance, déterminer un intervalle de probabilité $1 - P$, pour la quantité $\sum (x - \overline{x})^2$ et par conséquent pour la variance de l'échantillon.

$$\sigma_e^2 = \frac{\sum (x - \overline{x})^2}{n}$$

Ainsi, par exemple, pour $P = 0,05 = 0,025 + 0,025$, on lira dans la table, pour $n = 12$, c'est-à-dire : $\nu = 11$

$$P(\chi^2 \geqslant 21,9) = 0,025$$
$$P(\chi^2 \geqslant 3,82) = 0,975$$

donc

$$P(3,82 \leqslant \chi^2 \leqslant 21,9) = 0,95$$

Si, par hypothèse (population bien connue en vertu d'observations antérieures) on a :

$$\sigma^2 = 0,18$$

l'inégalité précédente s'écrit :

$$P\left(3,82 \leqslant \frac{\sum (x-\overline{x})^2}{0,18} \leqslant 21,9\right) = 0,95$$

c'est-à-dire

$$P\left((3,82 \times 0,18 \leqslant \sum (x - \overline{x})^2 \leqslant 21,9 \times 0,18\right) = 0,95$$

Si toutes les hypothèses précédentes sont fondées (indépendance des observations, population normale de variance $\sigma^2 = 0,18$), il n'y a donc que cinq chances sur cent pour que, relativement à l'échantillon observé, la somme des carrés des écarts $\sum (x - \overline{x})^2$ soit extérieure à l'intervalle :

(0,688 ; au lieu de, 3,942)

c'est-à-dire, cinq chances sur cent pour que la variance σ_e^2 de l'échantillon soit extérieure à l'intervalle

(0,057 ; au lieu de, 0,328)

(b) - Si la variance σ^2 de la population est inconnue et si l'on ne possède que l'échantillon observé $x_1, \ldots x_n$, on pourra déterminer un intervalle de confiance à $1 - P$ pour cette variance σ^2.

Ainsi, par exemple, toujours pour $P = 0,05 = 0,025 + 0,025$, si un échantillon de $n = 10$ observations a donné :

$$\sum (x - \bar{x})^2 = 36,10$$

On aura encore ($\nu = 9$)

$$P\left[2,70 \leqslant \frac{36,9}{\sigma^2} \leqslant 19\right] = 0,95$$

La résolution, par rapport à σ^2, de l'inégalité ci-dessus, donnera pour σ^2 l'intervalle de confiance (1,94 ; 13,7) à 0,95.

3 - Il arrive souvent que l'on soit conduit à estimer une variance s^2 à partir d'une expression $Q = \sum L_i^2$, dans laquelle les quantités L_i sont des formes linéaires non indépendantes dépendant des N observations liées par k relations linéaires (on dira alors que Q est une forme quadratique de rang N - k).

Dans ce cas, si $\frac{Q}{N-k}$ est une estimation correcte de la variance envisagée σ^2 la variable aléatoire $\frac{Q}{\sigma^2}$ est distribuée comme χ^2 avec $\nu = n - k$ degré de liberté.

Ainsi, par exemple, si on considère k échantillons de $n_1, .., n_i, .., n_k$ observations indépendantes provenant de populations normales de même variance σ^2, chaque expression de la forme :

$$\frac{\sum_j (x_{ij} - \bar{x}_i)}{n_i - 1} \qquad j = 1, 2, \ldots, n_k$$

indépendante de toute hypothèse sur les moyennes de ces diverses populations, est une estimation correcte de σ^2, c'est-à-dire telle que :

$$E\left[\frac{\sum (x_{ij} - \bar{x}_i)^2}{n_i - 1}\right] = \sigma^2 .$$

Il est évident que si l'hypothèse de l'unicité de variance est valable, on peut avoir une meilleure estimation de σ^2 en utilisant l'information globale fournie, non par un seul, mais par l'ensemble des échantillons.

On démontre effectivement que l'expression

$$s^2 = \frac{Q}{N-k} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{N-k} , \qquad (N = n_1 + \ldots + n_k)$$

est une estimation correcte de σ^2, basée sur $\nu = N - k$ degrés de liberté. On a, en effet, N expressions $L_{ij} = x_{ij} - \bar{x}_i$, mais elles sont liées par k relations, car dans chaque échantillon, on a :

$$\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) = 0 .$$

Dans ces conditions, la variable aléatoire

$$\frac{Q}{\sigma^2} = \frac{\sum \sum (x_{ij} - \bar{x}_i)^2}{\sigma^2}$$

est distribuée comme χ^2 avec $\nu = N - k$ degrés de liberté.

Les propriétés fondamentales de la distribution de χ^2 sont résumées par les deux théorèmes suivants :

4 - Théorème d'addition (FISHER)

Si k expressions indépendantes $A_1, A_2, \ldots, A_k$ sont individuellement distribuées comme χ^2 avec des degrés de liberté en nombres respectifs $\nu_1, \nu_2, \ldots, \nu_k$, leur somme :

$$A = A_1 + A_2 + \ldots A_k$$

est aussi distribuée comme χ^2 avec un nombre de degrés de liberté

$$\nu = \nu_1 + \nu_2 + \ldots \nu_k$$

Ainsi, par exemple, dans le cas précédent, chaque expression

$$A_i = \frac{Q_i}{\sigma^2} = \frac{\sum_j (x_{ij} - \bar{x}_i)^2}{\sigma^2}$$

étant distribuée comme χ^2 avec $\nu_i = n_i - 1$ degrés de liberté, leur somme

$$A = \frac{\sum_i \sum_j (x_{ij} - \bar{x}_i)^2}{\sigma^2}$$

est distribuée comme χ^2 avec $\nu = \nu_1 + \ldots \nu_k = N - k$ degrés de liberté.

5 - Théorèmes de partition (COCHRAN - FISHER)

(a) - Si $A_1, A_2, \ldots A_k$ sont des quantités respectivement distribuées comme χ^2, avec $\nu_1, \nu_2, \ldots, \nu_k$ degrés de liberté, et si leur somme $A = \sum_{i=1}^{k} A_i$ est aussi distribuée comme χ^2, avec :

$$\nu = \nu_1 + \nu_2 + \ldots \nu_k$$

degré de libertés, les variables A_i sont indépendantes en probabilité.

(b) - Si $A, A_1, A_2, \ldots A_k$ sont respectivement distribuées comme χ^2 avec $\nu, \nu_1, \ldots, \nu_k$ degrés de liberté, si $A_1, \ldots A_k$ sont indépendantes et si l'on a :

$$A = A_1 + \ldots A_k + B, \text{ avec } \sum \nu_i < \nu$$

B est aussi distribué comme χ^2 avec $\nu' = \nu - \sum \nu_i$ degrés de liberté.

II. - DISTRIBUTION DE F (Fisher-Snedecor)

On donne ce nom à la distribution d'une variable aléatoire F, dont la loi de probabilité élémentaire est définie par :

$$p(F)\, dF = \frac{\Gamma\left(\frac{\nu_1 + \nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \nu_1^{\frac{\nu_1}{2}} \nu^{\frac{\nu_2}{2}} \frac{F^{\frac{\nu_1}{2}-1}}{(\nu_2 + \nu_1 F)^{\frac{\nu_1+\nu_2}{2}}}\, dF .$$

Cette loi, en plus de la variable aléatoire F, dépend donc de deux paramètres ν_1 et ν_2, appelés encore degrés de liberté, dont le sens sera précisé ci-après.

La distribution de F est liée à la distribution de χ^2 par la propriété fondamentale suivante :

Si A' et A'' sont deux variables aléatoires indépendantes, respectivement distribuées comme χ^2 avec ν' et ν'' degrés de liberté, le rapport :

$$F = \frac{\frac{A'}{\nu'}}{\frac{A''}{\nu''}}$$

est distribué comme F avec ν' et ν'' degrés de liberté.

La distribution de F a été mise en tables qui, le plus généralement, donnent les fractiles supérieurs de la distribution, c'est-à-dire les valeurs F_0 de F, telles que

$$P(F \geqslant F_0) = P ,$$

pour diverses valeurs de P, en fonction de ν_1 et ν_2.

Etant donné que le rapport :

$$F' = \frac{1}{F} = \frac{\frac{A''}{\nu''}}{\frac{A'}{\nu'}}$$

est aussi distribué comme F avec cette fois, ν'' et ν' degrés de liberté, les tables sont, en général, réduites aux valeurs de F supérieures à l'unité.

En effet, α étant un nombre positif supérieur à l'unité, on a :

$$P\left(F \leqslant \frac{1}{\alpha}, \nu', \nu''\right) = P\left(F \geqslant \alpha, \nu'', \nu'\right)$$

Il suffira d'intervenir les paramètres ν' et ν'' d'entrée dans les tables, dans lesquelles ν_1 désigne le nombre de degrés de liberté relatif au **plus grand** des deux termes.

III. - APPLICATION AUX PROBLÈMES DE COMPARAISON DE VARIANCES

1. - Comparaison de deux variances

Considérons deux échantillons :

E_1 $(x_1 \ldots x_{n_1})$: Echantillon de n_1 observations indépendantes de moyenne $\bar{x}$ provenant d'une population normale de variance inconnue σ_1^2;

E_2 $(y_1 \ldots y_{n_2})$: Echantillon de n observations indépendantes de moyenne $\bar{y}$ provenant d'une population normale de variance inconnue σ_2^2;

et proposons-nous de rechercher si l'hypothèse d'une variance commune σ^2 est acceptable.

$$H_0 \longrightarrow \sigma_1^2 = \sigma_2^2 = \sigma^2$$

Dans cette hypothèse, les quantités :

$$s_1^2 = \frac{\sum (x - \bar{x})^2}{n_1 - 1} \qquad \text{et} \qquad s_2^2 = \frac{\sum (y - \bar{y})^2}{n_2 - 1}$$

sont des estimations correctes de σ^2.

De plus, les quantités :

$$A_1 = \frac{\sum (x - \bar{x}_1)^2}{\sigma^2} \qquad A_2 = \frac{\sum (x - \bar{x}_2)^2}{\sigma^2}$$

sont alors distribuées comme χ^2 avec $\nu_1 = n_1 - 1$; $\nu_2 = n_2 - 1$.

Par conséquent, le rapport :

$$F = \frac{\frac{A_1}{\nu_1}}{\frac{A_2}{\nu_2}} = \frac{s_1^2}{s_2^2}$$

est distribué comme $F(\nu_1, \nu_2)$.

S'il résulte de l'examen de la table de F que la probabilité d'obtenir pour ce rapport une valeur supérieure ou égale (ou inférieure ou égale) à celle fournie par l'observation, est trop petite, on sera conduit à rejeter l'hypothèse H_0.

<u>Exemple</u>

$n_1 = 13 \qquad s_1^2 = 0,07 \qquad \nu_1 = 12$

$n_2 = 25 \qquad s_2^2 = 0,05 \qquad \nu_2 = 24$

F (observé) $= \frac{s_1^2}{s_2^2} = 1,40$

La table montre que pour $\nu_1 = 12$, $\nu_2 = 24$, la probabilité d'obtenir $F \geqslant 1,40$ est supérieure à 0,05.

$P = 0,01$	0,05	
$F \geqslant 3,03$	2,18	1,40

Au niveau considéré, on n'a donc pas de raison suffisante de conclure que σ_1^2 est supérieure à σ_2^2.

Avec les mêmes données, on peut se proposer de déterminer l'intervalle (F_1, F_2) qui a une probabilité donnée, par exemple $1 - P = 0,98$, de contenir le rapport observé dans l'hypothèse de l'égalité des variances.

On lira encore dans la table que :

$$P = 0,01 \text{ pour que } \frac{s_1^2(12)}{s_2^2(24)} > 3,03$$

$$P = 0,01 \text{ pour que } \frac{s_2^2(24)}{s_1^2(12)} > 3,78$$

c'est-à-dire :

$$\frac{s_1^2(12)}{s_2^2(24)} < \frac{1}{3,78}$$

d'où en définitive

$$P = 0,98 \text{ pour que } \frac{1}{3,78} < \frac{s_1^2}{s_2^2} < 3,03$$

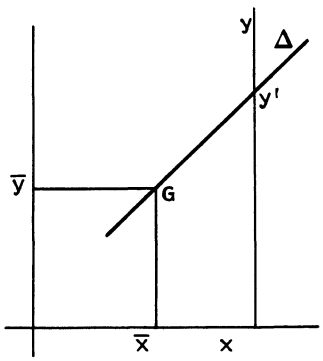
Le rapport $F = 1,40$ est intérieur à cet intervalle : au niveau considéré, il n'y a pas lieu de conclure à l'inégalité des variances.

2. - Test de signification d'un coefficient de corrélation (ou de régression) linéaire

Soit $y' = \bar{y} + b(x - \bar{x})$ (1)

avec

$$b = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{\sum y(x-\bar{x})}{\sum (x-\bar{x})^2} \quad \begin{matrix}(1)\\(2)\end{matrix}$$



l'équation d'estimation linéaire de y en fonction de x obtenue par la méthode des moindres carrés. Admettons comme hypothèse acceptable que la distribution de y dans la population est normale de variance σ_y^2 et que les distributions liées de y pour x donné sont aussi normales de variance commune : $\sigma_{y.x}^2 = \sigma_y^2 (1 - r^2)$,

le coefficient de corrélation r étant défini par la relation :

$$r^2 = \frac{\sum (y-\bar{y})^2 - \sum (y-y')^2}{\sum (y-\bar{y})^2} = 1 - \frac{\sum (y-y')^2}{\sum (y-\bar{y})^2}$$

d'où on déduit aisément, compte tenu de (1), la définition classique :

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \cdot \sum (y-\bar{y})^2}}$$

tous les calculs étant faits à partir des n couples observés (x, y) de l'échantillon dont on dispose.

A partir de l'identité :

$$y - \bar{y} = (y - y') + (y' - \bar{y}) ,$$

et compte tenu des relations (1) et (2), on en déduit la relation ;

$$\sum (y - \bar{y})^2 = \sum (y - y')^2 + \sum (y' - \bar{y})^2 ,$$

de la forme précédemment envisagée : $Q = Q_1 + Q_2$

On remarquera d'abord que les trois formes quadratiques Q, Q_1, Q_2, satisfont aux conditions suivantes :

a) $Q = \sum (y - \bar{y})^2$ est une forme quadratique de rang $\nu = n - 1$, les n formes linéaires $y - \bar{y}$ étant liées par la relation

$$\sum (y - \bar{y}) = 0$$

(b) $Q_1 = \sum (y' - \bar{y})^2$ est une somme de carrés de n formes linéaires qui se met immédiatement sous la forme

$$Q = b^2 \sum (x - \bar{x})^2 = u^2$$

avec

$$u = \frac{x_1 - \bar{x}}{\sqrt{\sum (x-\bar{x})^2}} y_1 + \ldots + \frac{x_n - \bar{x}}{\sqrt{\sum (x-\bar{x})^2}} y_n ,$$

c'est-à-dire une forme quadratique ne faisant intervenir qu'une seule combinaison linéaire des y.

Q_2 est donc une forme quadratique de rang $\nu_1 = 1$.

(c) $Q_2 = \sum (y - y')^2$ est une forme quadratique de rang $\nu_2 = n - 2$, car les n formes linéaires $y - y'$ sont liées par deux relations indépendantes :

$$\sum (y - y') = 0$$

$$\sum (x - \bar{x})(y - y') = \sum (x-\bar{x})(y-\bar{y}) - b \sum (x - \bar{x})^2 = 0 .$$

De plus, on a :

$$\nu = \nu_1 + \nu_2$$

Supposons maintenant qu'en plus des hypothèses précisées ci-dessus, on veuille tester l'hypothèse d'une corrélation nulle

$$r = 0 \ , \quad (\text{ou} \quad b_{yx} = 0).$$

<u>Dans ce cas</u> la variance conditionnelle σ^2_{yx} et la variance σ^2_y sont égales et les trois expressions

$$s^2 = \frac{1}{n-1} \sum (y - \bar{y})^2 = \frac{Q}{n-1}$$

$$s_1^2 = \sum (y' - \bar{y})^2 = Q_1$$

$$s_2^2 = \frac{1}{n-2} \sum (y - y')^2 = \frac{Q_2}{n-2}$$

sont des estimations correctes de cette variance σ^2.

On a alors : $E(s^2) = E(s_1^2) = E(s_2^2) = \sigma^2$.

De plus, en raison du théorème de Cochran, Q_1 et Q_2 sont indépendantes et le rapport

$$F = \frac{Q_1}{\frac{Q_2}{n-2}} = \frac{Q_1}{\frac{1}{n-2} Q_2}$$

est distribué comme F avec $\nu_1 = 1$ et $\nu_2 = n - 2$ avec :

$$E\left[Q_1\right] = E\left[\frac{Q_2}{n-2}\right] = \sigma^2$$

autour d'une valeur moyenne voisine de l'unité $\left(\frac{\nu_2}{\nu_2 - 2}\right)$.

Par contre, si la régression (ou la corrélation) est significativement différente de zéro, $\frac{Q_2}{n-2}$ est une estimation de $\sigma^2_{y.x} < \sigma^2_y$

De plus :

$$\begin{aligned} E\left[Q_1\right] &= E\left[Q\right] - E\left[Q_2\right] \\ &= (n-1)\,\sigma^2_y - (n-2)\,\sigma^2_{y.x} \end{aligned}$$

Etant donné que :

$$\sigma^2_y > \sigma^2_{yx}$$

on a alors :

$$E\left[Q_1\right] > \sigma^2_{y.x}$$

et le rapport F sera alors, en moyenne, supérieur à l'unité.

Il en résulte que rechercher si la corrélation est significative, c'est rechercher si le rapport F est significativement supérieur à l'unité, ce qui pourra être testé à l'aide de la table de la distribution de F.

On remarquera, pour parler un langage simple, que ceci revient à chercher (compte tenu des hypothèses initiales sur la normalité et l'indépendance) si la variance **expliquée** par la régression :

$\sum (y' - \bar{y})^2$ est significativement supérieure à la variance **résiduelle** :

$$\frac{\sum (y - y')^2}{n-2}$$

Ces résultats sont généralement présentés sous forme d'un tableau appelé tableau d'analyse de la variance :

Source de variation	Somme des carrés	Degrés de liberté	Variance moyenne
Droite de régression	$\sum (y'-\bar{y})^2 = b^2\sum (x-\bar{x})^2 = r^2\sum (y-\bar{y})^2$	1	$s_1^2 = r^2\sum (y-\bar{y})^2$
Résiduelle	$\sum (y-y')^2 = (1-r^2)\sum (y-\bar{y})^2$	$n-2$	$s_2^2 = \frac{(1-r^2)\sum(y-\bar{y})^2}{n-2}$
totale	$\sum (y-\bar{y})^2$	$n-1$	

On remarquera que l'on a :

$$F = \frac{s_1^2}{s_2^2} = \frac{r^2}{1-r^2}(n-2) \qquad \nu_1 = 1 \quad \nu_2 = n-2$$

3. - Test de l'homogénéité d'un groupe de k moyennes (analyse de variance : cas d'un seul facteur de variabilité)

	Échantillons			
	(1) ...	(i) ...	(k)	
		x_{i1}		
		x_{ij}		
		$x_{i.n_i}$		
Moyennes	$\bar{x}_1$	$\bar{x}_i$	$\bar{x}_k$	$\bar{x}$

Considérons un groupe de k échantillons d'observation x_{ij} ($j^{ème}$ valeur de l'échantillon i) Soient :

$\bar{x}_1 \ldots \bar{x}_i \ldots \bar{x}_k$ les moyennes de ces échantillons d'effectifs ; $n \ldots n \ldots n_k$ et $\bar{x}$ leur moyenne générale :

$$\bar{x} = \frac{\sum n_i x_i}{\sum n_i} = \frac{\sum n_i x_i}{N}$$

$$N = \sum n_i$$

On a tout d'abord la relation :

$$\sum_i \sum_j (x_{ij} - \bar{x})^2 = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

de la forme

$$Q = Q_1 + Q_2$$

Dans tout ce qui suit, nous supposerons d'abord que chaque échantillon est constitué par n observations **indépendantes** provenant d'une population **normale** et que toutes ces populations normales ont **même variance** σ^2 et nous nous proposerons de tester l'hypothèse d'une même moyenne m pour ces populations.

Si cette hypothèse est vérifiée :

1. $Q = \sum \sum (x_{ij} - \bar{x})^2$ est une forme quadratique de rang $\nu = N - 1$.

2. $Q_1 = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$ est une somme de carrés de k formes linéaires

$$L_i = \sqrt{n_i}\,(\bar{x}_i - \bar{x}) \qquad i = 1, 2, \ldots, k,$$

liées par une relation $\sum_{i=1}^{k} \sqrt{n_i}\, L_i = 0$.

donc Q_1 est une forme quadratique de rang $\nu_1 = k - 1$.

3. $Q_2 = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2$ est une somme de carrés de N formes linéaires

$$L_{ij} = x_{ij} - \bar{x}_i$$

liées par k relations indépendantes :

$$\sum_{j=1}^{n_i} L_{ij} = 0 \qquad\qquad i = 1, 2, \ldots, k$$

donc Q_2 est une forme quadratique de rang $\nu_2 = N - k$ et on a $\nu = \nu_1 + \nu_2$.

De plus, dans l'hypothèse d'une même moyenne m et d'une même variance les quantités :

$$s^2 = \frac{Q}{n-1} \; ; \qquad s_1^2 = \frac{Q_1}{k-1} \; ; \qquad s_2^2 = \frac{Q_2}{N-k} \; ;$$

sont trois estimations correctes de cette variance σ^2 (on peut démontrer directement que l'on a

$$E(s^2) = E(s_1^2) = E(s_2^2) = \sigma^2 \;) .$$

Dans ces conditions, en application du théorème de Cochran et de la définition de la distribution de F :

I. Les quantités $\dfrac{Q}{\sigma^2}$; $\dfrac{Q_1}{\sigma^2}$; $\dfrac{Q_2}{\sigma^2}$

sont distribuées comme χ^2 avec : $\nu = N - 1$; $\nu_1 = k - 1$; $\nu_2 = N - k$.

2. Les quantités Q_1 et Q_2 sont indépendantes.

3. Le rapport

$$F = \frac{\dfrac{Q_1}{k-1}}{\dfrac{Q_2}{N-k}} = \frac{s_1^2}{s_2^2} = \frac{\dfrac{1}{k-1} \sum n_i (\bar{x}_i - \bar{x})^2}{\dfrac{1}{N-k} \sum\sum (x_{ij} - \bar{x}_i)^2}$$

est distribué comme F avec $\nu_1 = k - 1$, $\nu_2 = N - k$ autour d'une valeur moyenne voisine de l'unité $\left(\dfrac{\nu_2}{\nu_2 - 2}\right)$.

Par contre, si les k populations ont des moyennes différentes, $m_1, \ldots m_i, \ldots, m_k$, la variance estimée s_2^2 sera encore telle que

$$E(s_2^2) = \sigma^2 ,$$

mais la variance estimée dans s_1^2 proviendra de deux sources indépendantes :

1. La part provenant de la fluctuation des x_{ij} dans chaque échantillon et qui intervient dans la variance des $\bar{x}_i$ pour $\dfrac{\sigma^2}{n_i}$, c'est-à-dire pour σ^2 dans $\dfrac{Q_1}{k-1}$.

2. La part provenant de la dispersion propre des m_i .

Ces deux composantes s'ajoutant, on aura encore :

$$E(s_1^2) > \sigma^2$$

et le rapport

$$F = \frac{s_1^2}{s_2^2}$$

aura tendance à être en moyenne supérieur à l'unité.

Le test F permettra donc de voir si le rapport F observé est tel qu'il n'y a qu'une probabilité très petite que cette valeur soit atteinte du seul fait des fluctuations aléatoires, c'est-à-dire si - pour un niveau de confiance donné - il y a lieu de considérer que, dans leur ensemble, les différences constatées entre (E_1), ... (E_k) peuvent être valablement attribuées, non aux fluctuations d'échantillonnage, mais à des différences systématiques.

Le tableau d'analyse de la variance se présentera comme suit :

Source de variation	Somme des carrés	Degrés de liberté	Variances	F
Entre échantillons (Effet de E)	$Q_1 = \sum_i n_i (\bar{x}_i - \bar{x})^2$	k-1	$s_1^2 = \frac{Q_1}{k-1}$	$\frac{s_1^2}{s_2^2}$
Résiduelle	$Q_2 = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2$	N-k	$s_2^2 = \frac{Q_2}{N-k}$	
Total	$Q = \sum_i \sum_j (x_{ij} - \bar{x})^2$	N-1	$s^2 = \frac{Q}{N-1}$	

Pour l'exécution des calculs, on remarquera que si l'on désigne par :

X la somme de toutes les observations

X_i la somme des n_i observations de l'échantillon E_i

On a :

$$Q = \sum_i \sum_j (x_{ij} - \bar{x})^2 = \sum_i \sum_j x_{ij}^2 - X\bar{x}$$

$$Q_1 = \sum_i n_i (\bar{x}_i - \bar{x})^2 = \sum_i X_i x_i - X\bar{x}$$

(Les calculs peuvent d'ailleurs encore être simplifiés en remplaçant tous les x_{ij} par leurs différences x_{ij} - a, comptées à partir d'une origine arbitraire).

Le dernier terme $Q_2 = \sum \sum (x_{ij} - \bar{x}_i)^2$ pourra être calculé par différence.

Il est important de remarquer que, en plus des hypothèses de **normalité et d'indépendance** déjà posées au début de cette étude, les raisonnements faits supposent de plus **l'unicité de variance** dans les k populations échantillonnées.

Pratiquement, cette hypothèse est assez généralement plausible, les différences entre les populations, P_1, ..., P_k soumises à la comparaison ne portant pas sur la nature même des phénomènes, mais sur des modifications relativement légères ayant pour effet de déplacer la courbe de distribution sans la modifier.

Néanmoins, cette hypothèse qui peut d'ailleurs être testée (Test L_1 de Neyman et Pearson ou test de Bartlett) n'est pas obligatoirement vérifiée.

4. - Etude simultanée de deux facteurs A et B susceptibles de diverses modalités ($A_1 \cdots A_k$),($B_1 \cdots B_n$) .

Les observations x_{ij} se présentent alors sous la forme suivante :

		Facteur A					
		A_1	...	A_i	...	A_k	*Moyennes*
Facteur B	B_1	x_{11}		x_{i1}		x_{k1}	r_1
	⋮						
	B_j	.		x_{ij}			r_j
	⋮						
	B_n						r_n
Moyennes		c_1	...	c_i	...	c_k	$\bar{x}$

$\bar{x}$ étant la moyenne générale des nk observations, on peut écrire :

$$x_{ij} - \bar{x} = (r_j - \bar{x}) + (c_i - \bar{x}) + (x_{ij} - r_j - c_i + \bar{x})$$

d'où l'on déduit encore :

$$\sum\sum (x_{ij} - \bar{x})^2 = k\sum_j (r_j - \bar{x})^2 + n\sum_i (c_i - \bar{x})^2 + \sum\sum d^2$$

avec :

$$(d = x_{ij} - r_j - c_i + \bar{x})$$

de la forme :

$$Q = Q_1 + Q_2 + Q_3 .$$

Les mêmes raisonnements que précédemment montrent encore que ces formes quadratiques sont de rangs respectifs :

$$\nu = nk - 1 ; \qquad \nu_1 = n - 1 ; \qquad \nu_2 = k - 1 ; \qquad \nu_3 = (n-1)(k-1)$$

avec

$$\nu = \nu_1 + \nu_2 + \nu_3 ;$$

si, de plus, on fait les mêmes hypothèses de base (indépendance, normalité, unicité de variance σ^2) l'hypothèse à tester de l'absence de l'effet de la cause A (même moyenne c aux fluctuations aléatoires près pour les diverses colonnes), conduit à conclure que le rapport F

$$F = \frac{\dfrac{Q_2}{k-1}}{\dfrac{Q_3}{(n-1)(k-1)}} = \frac{\dfrac{n}{k-1}\sum_i (c_i - \bar{x})^2}{\dfrac{1}{(n-1)(k-1)}\sum\sum d^2}$$

est distribué comme F avec $\nu_1 = k - 1$, $\nu_2 = (n - 1)(k - 1)$ (avec la même hypothèse pour l'effet B).

En effet, si ces hypothèses sont réalisées, les quantités :

$$s^2 = \frac{Q}{nk-1} \qquad s_r^2 = \frac{Q_1}{n-1}$$

$$s_c^2 = \frac{Q_2}{k-1} \qquad s_e^2 = \frac{Q_3}{(n-1)(k-1)}$$

sont quatre estimations correctes de σ^2 :

$$E(s^2) = E(s_c^2) = E(s_r^2) = E(s_e^2) = \sigma^2$$

De plus, les quantités Q_1, Q_2, Q_3 sont indépendantes. Par contre, l'existence de l'effet A sera caractérisée par le fait que F est significativement supérieur à l'unité, car dans ce cas on aura encore :

$$E(s_e^2) = \sigma^2$$

mais pour la même raison que ci-dessus :

$$E(s_c^2) = \sigma^2 + \lambda^2 .$$

De même, l'existence de l'effet B sera caractérisée par le fait que le rapport

$$F = \frac{\dfrac{k}{n-1}\sum_j (r_j - \bar{x})^2}{\dfrac{1}{(n-1)(k-1)}\sum\sum d^2} , \qquad \begin{array}{l} \nu_1 = n-1 \\ \nu_2 = (n-1)(k-1) \end{array}$$

est significativement supérieur à l'unité.

Le tableau d'analyse de la variance se présentera comme ci-dessous :

Source de variation	Somme des carrés	degrés de liberté	Variance	F
Entre les lignes (effet de B)	$Q_1 = k \sum_j (r_j - \bar{x})^2$	n - 1	$s_r^2 = \frac{Q_1}{n-1}$	$\frac{s_r^2}{s_e^2}$
Entre les colonnes (effet de A)	$Q_2 = n \sum_i (c_i - \bar{x})^2$	k - 1	$s_c^2 = \frac{Q_2}{k-1}$	$\frac{s_c^2}{s_e^2}$
Résiduelle	$Q_3 = \sum\sum d^2$	(n-1) (k-1)	$s_e^2 = \frac{Q_3}{(n-1)(k-1)}$	
Total	$Q = \sum\sum (x_{ij} - \bar{x})^2$	nk - 1	$s^2 = \frac{Q}{nk-1}$	

L'éxécution des calculs se fait sans difficulté : en effet, si on désigne par :

X la somme de toutes les observations

C_i la somme de la colonne A_i

R_j la somme de la ligne B_j

on aura encore :

$$\sum\sum (x_{ij} - \bar{x})^2 = \sum\sum x_{ij}^2 - X\bar{x}$$

$$k \sum_j (r_j - \bar{x})^2 = \sum_j R_j r_j - X\bar{x}$$

$$n \sum_i (c_i - \bar{x})^2 = \sum_i C_i c_i - X\bar{x} ,$$

Le dernier terme $\sum\sum d^2$ étant calculé par différence.

Mais l'extension à ce dernier cas du raisonnement fait dans le cas de l'analyse pour un seul facteur de variabilité suppose implicitement (en raison de l'hypothèse d'indépendance) l'indépendance des effets éventuels des causes A et B, c'est-à-dire l'addivité des effets dus à la variation de ces causes (absence d'interaction).

Le modèle mathématique utilisé est en fait :

$$x_{ij} = m + \alpha_i + \beta_j + \varepsilon_{ij}$$

Dans cette relation, α_i et β_j caractérisent respectivement l'influence des causes A_i et B_j avec

$$\sum \alpha_i = 0 \qquad \sum \beta_j = 0$$

et ε_{ij} les variations aléatoires normalement et indépendamment distribuées autour de zéro avec la variance σ^2.

IV. - INFLUENCE DE LA NON-RÉALISATION DES HYPOTHÈSES DE BASE DE L'ANALYSE DE VARIANCE

Rappelons que ces hypothèses sont les suivantes :

1. Normalité de la distribution des erreurs expérimentales
2. Indépendance des erreurs expérimentales
3. Additivité des effets des causes sous contrôle.

4. Constance de la variance des erreurs expérimentales quelle que soit l'importance des effets de ces causes.

Prises dans leur ensemble, ces hypothèses semblent impliquer une sévère restriction des types de problèmes auxquels les techniques d'analyse de variance peuvent s'appliquer. Il y a lieu de rechercher dans quelle mesure la validité du test est affectée lorsque ces hypothèses ne sont pas vérifiées.

1. - Non normalité

Une étude mathématique de l'influence de la non normalité sur la validité du test F est évidemment très difficile.

E.S. PEARSON a étudié expérimentalement la distribution de F à partir de six exemples différents de population non-normales. L'étude empirique ainsi réalisée est insuffisante pour fixer avec précision les seuils à 5 % et 1% des distributions correspondantes · néanmoins, il semble en résulter que le test F classique puisse être appliqué sans erreur grave aux échantillons provenant de distributions telles que celles que l'on rencontre généralement.

COCHRAN estime que les valeurs de F données par la table pour les seuils 5% et 1% correspondent, en fait, respectivement à des seuils compris entre 4% et 7% d'une part, 0,5% et 2% d'autre part.

En général, la non normalité a pour effet de faire paraître le résultat plus significatif qu'il ne l'est en réalité.

Dans la pratique, il pourra arriver que l'on ait à priori, en raison d'observations antérieures suffisamment nombreuses, quelque opinion valable sur la normalité approximative des distributions dans les populations envisagées.

S'il n'en est pas ainsi, on ne peut guère envisager de tester cette normalité à l'aide des échantillons observés, en général beaucoup trop peu importants.

Si, en raison même de la nature des phénomènes étudiés, il y a quelque raison à priori de suspecter la non normalité, on peut quelquefois envisager un changement de variable capable d'améliorer une approximation vers la normalité.

Une méthode d'analyse n'impliquant aucune hypothèse relativement à la normalité des populations a été envisagée par PITMAN et WELCH (1).

2 - Non indépendance entre les erreurs

S'il existe une corrélation ρ entre les erreurs ou fluctuations ε_i relatives, par exemple, aux k observations d'un même facteur, B_j par exemple, l'espérance mathématique de la variance estimée dans la méthode d'analyse de la variance n'est plus égale à σ^2.

Il est théoriquement possible de calculer cette espérance mathématique qui dépend de ρ, mais étant donné l'impossibilité d'estimer ρ à partir de l'échantillon il est difficile de préciser l'influence de l'existence de telles corrélations sur la validité du test F.

Il est en tout cas nécessaire de prendre toutes précautions pour éviter autant que possible leur existence.

3. - Non additivité

La non additivité des effets, par exemple, l'hypothèse d'un modèle de la forme

$$x_{ij} = m\,\alpha_i\,\beta_j\,(1+\varepsilon_{ij})$$

a pour effet d'augmenter la variance résiduelle qui intervient dans le dénominateur de F, à moins que cette variance résiduelle ne soit petite par rapport aux effets des causes, cet accroissement de la variance résiduelle pourra, en général, être négligé.

---

(1) - PITMAN - The analysis of variance test - Biometrika 29 - 1937 - p. 322-335.

WELCH - On the z Test in randomized blocks and latin Squares - Biometrika 29 - 1937 - p. 21. 52.

4. - Non uniformité de la variance

La non uniformité de la variance a aussi pour effet de réduire la sensibilité du test.

Dans le cas de l'analyse de variance relative à un seul facteur de variabilité, il est, en général, possible de tester l'homogénéité de variance dans les diverses populations P_1, ..., P_i, ... P_k, en utilisant, par exemple, le texte L_1 de NEYMAN et PEARSON, basé sur la distribution de :

$$L_1 = \frac{s_G^2}{s_A^2}$$

avec, dans le cas de k échantillons de même effectif n :

$$s_G^2 = (s_1^2 \cdot s_2^2 \cdot \ldots \cdot s_k^2)^{\frac{1}{k}}$$

$$s_A^2 = \frac{1}{k} \sum_1^k s_i^2$$

$$s_i^2 = \frac{1}{n} \sum_{j=1}^n (x_{ij}^2 - \bar{x}_i^2)$$

Ce rapport L_1 doit tendre vers l'unité dans l'hypothèse de l'homogénéité des variances · s'il est significativement inférieur à l'unité il y a hétérogénéité dans les variances et on ne peut plus utiliser le test F.

Les tables établies pour

$$n_1 = \ldots = n_i \ldots = n_k = n$$

peuvent encore être utilisées quand n_i est variable, en prenant pour n la moyenne arithmétique des n_i, à condition qu'aucun des n_i ne soit inférieur à 15 ou 20.

5. - Changements de variables en vue de stabiliser les variances

Lorsque la variable x a une distribution dans laquelle la variance est fonction de la moyenne (loi binomiale, loi de Poisson, ...) il est évident que les différences significatives entre moyennes d'échantillons sont incompatibles avec l'hypothèse d'une variance uniforme.

Dans de tels cas, il est souvent possible d'envisager un changement de variable $y = f(x)$ tel que la variance des y sera pratiquement indépendante des moyennes.

(a) - <u>Cas d'une distribution binomiale</u> - Si x est une fréquence, par exemple proportion des pièces défectueuses dans un échantillon de n observations provenant d'une population dans laquelle cette fréquence est p, la variance de x est

$$\frac{p(1-p)}{n}$$

si on considère la variable transformée

$$y = \text{arc sin} \sqrt{x}$$

on montre que sa variance est approximativement $\frac{1}{4n}$, si y est exprimé en radians, ou $\frac{821}{n}$, si y est exprimé en degrés.

Cette transformation est efficace, sauf pour les valeurs extrêmes de x. BARTLETT a suggéré de compter les rapports 0/n et n/n, respectivement pour 1/4 n et (n - 1/4)/n.

Cette transformation améliore l'approximation de p vers la normalité, d'autant plus médiocre à l'origine que p était plus voisin de zéro (ou un).

Plus généralement, si la variance est liée à la moyenne m par une relation empirique de la forme

$$\sigma_x^2 = k\,m\,(1-m)$$

la même transformation donne lieu à une variance stabilisée approximativement égale à 0,25 k.

(b) - Cas d'une distribution de Poisson - Si x est une variable distribuée suivant une loi de Poisson de moyenne m, sa variance est aussi égale à m.

Dans ce cas, la variable transformée :

$$y = \sqrt{x}$$

ou mieux encore

$$y = \sqrt{x + 1/2}$$

a pour $m > 3$ une variance sensiblement constante approximativement égale à 0,25.

Plus généralement, si la variance est liée à la moyenne par une relation de la forme

$$\sigma_x^2 = \lambda^2 m$$

la même transformation donnera lieu à une variance stabilisée approximativement égale à $0,25\ \lambda^2$.

(c) - Cas de populations dont l'écart-type est sensiblement proportionnel à la moyenne (coefficient de variation constant : c)

Dans ce cas, la variable transformée

$$y = \log_{10} x$$

ou

$$y = \log_{10} (x + 1)$$

(si x peut prendre la valeur zéro)
a une variance sensiblement constante et égale à $0,189\ c^2$ (ou à c^2 si l'on a pris $y = \mathrm{Log}_e\ x$).

On notera de plus qu'un tel changement de variable transforme les effets multiplicatifs en effets additifs.

V. - RISQUES D'ERREURS

Il ne faut pas oublier, de plus, que tout test statistique conduit, non pas à une conclusion absolue (telle chose est vraie, ou telle chose n'est pas vraie), mais à une conclusion considérée, dans les conditions de l'expérimentation, comme plus vraisemblable que la conclusion contraire ou qu'une autre conclusion.

Mais la conclusion à laquelle on s'arrête - et en vertu de laquelle on prend une décision - peut se trouver fausse elle comporte essentiellement deux risques d'erreur.

Si l'on désigne par H l'hypothèse à tester, formulée de manière précise, appliquer à ce problème un test déterminé, c'est accepter de courir deux risques :

- Risque de première espèce ou risque de rejeter l'hypothèse H alors qu'elle est vraie;
- Risque de seconde espèce ou risque d'accepter l'hypothèse H alors qu'elle est fausse.

Soient α et β les probabilités caractérisant respectivement le premier et le second risque, probabilités que l'on souhaitera généralement être petites, sinon très petites (appréciation technique ou même subjective des inconvénients ou dangers de prendre une dé cision basée sur une conclusion fausse).

On remarquera que dans ce qui précède, le premier problème est formulé de manière précise : on définit l'hypothèse H et on se donne α qui peut être regardé comme la proportion des cas ou l'on rejettera à tort l'hypothèse H, si l'on effectue de nombreux tests dans des conditions analogues.

Mais il n'en est pas de même du second problème, en général plus complexe: l'hypothèse testée H pouvant être fausse de bien des manières. (Dans un domaine simplement numérique on remarquera la différence qui existe entre les conditions

$$A = a \qquad A \neq a$$

a étant un nombre donné).

Pour préciser numériquement ce qu'il faut entendre par risque de seconde espèce, il faut pouvoir remplacer la définition toute négative "hypothèse fausse" par une définition précise d'une hypothèse alternative H' contre laquelle on teste l'hypothèse H.

Examinons ce qui se passe lorsqu'on emploie le test F dans les problèmes envisagés ci-dessus.

Considérons, par exemple, le problème traité au III-3 (Test de l'homogénéité d'un groupe de k moyennes).

Supposons qu'on utilise le test F au niveau $\alpha = 0,05$; compte tenu des nombres de degrés de liberté : $\nu_1 = k - 1$, $\nu_2 = N - k$, la table de la distribution de F nous donne une valeur critique :

$$F_{0,05}(\nu_1, \nu_2)$$

Par exemple pour k = 5, n = 10.

$$F_{0,05} = 2,57$$

D'autre part, l'ensemble des observations utilisées nous donne une valeur expérimentale :

$$F = \frac{s_1^2}{s_2^2}$$

Que nous donne la règle envisagée ?

Si $1 < F < 2,57$ nous constatons que l'hétérogénéité expérimentalement constatée - dans le sens prévu dans le cas d'une hétérogénéité réelle - est telle qu'elle a plus de 5 chances sur 100 d'être effectivement constatée dans le cas où les échantillons observés proviendraient de populations (normales et de même variance) dans lesquelles cette hétérogénéité n'existerait pas.

Ceci veut dire qu'en répétant l'observation sur un grand nombre d'ensembles analogues provenant de telles populations, on trouverait environ 5% de ces ensembles donnant lieu à une valeur expérimentale de F supérieure à la valeur critique choisie $F_{0,05} = 2,57$.

Le risque de rejeter l'hypothèse H (homogénéité) alors qu'elle est vraie est donc $\alpha = 0,05$, mais le risque de deuxième espèce : accepter l'hypothèse de l'homogénéité, alors qu'elle n'existe pas, n'est pas pris en considération.

D'un point de vue pratique, ceci se conçoit aisément ; si l'on compare deux méthodes, ou deux machines, c'est en vue de remplacer éventuellement la moins bonne par la meilleure. Si l'on étudie l'homogénéité des moyennes d'un groupe d'échantillons, c'est en vue de prendre les mesures nécessaires pour assurer cette homogénéité dans le sens favorable au but poursuivi.

Dans un cas comme dans l'autre, ceci peut conduire à prendre des mesures coûteuses (remplacement d'une machine par une autre, d'un procédé de fabrication par un autre,...), il importe donc de ne prendre ces décisions que si elles sont vraiment utiles.

Par contre, si tout se passe à peu près comme si l'homogénéité existait - il est certain d'ailleurs que l'homogénéité parfaite n'existe pas - il n'y a, en général, pas d'inconvénient à l'admettre.

D'une part, l'augmentation des effectifs des échantillons (on sait de quelle façon en dépend la précision des estimations de s_1^2 et s_2^2, c'est-à-dire leurs intervalles de confiance), et d'autre part, la répétition des groupes d'observations, permettront à cet égard d'améliorer ou de confirmer la validité des conclusions.

Il est encore un autre point sur lequel il paraît nécessaire d'attirer l'attention.

L'homogénéité - sous réserve de la validité des hypothèses de base (normalité, unicité de variance) - est caractérisée par le fait que F ne doit pas trop s'écarter de 1, en deçà ou au delà et la table F permet, moyennant permutation des nombres de degrés de liberté, de définir un intervalle considéré comme compatible avec

cette homogénéité, par exemple :

$$\frac{1}{F_{0,05}(\nu_2 , \nu_1)} \quad \text{à} \quad F_{0,05}(\nu_1 , \nu_2)$$

Mais l'hypothèse de l'hétérogénéité, dans le problème envisagé, doit conduire à une valeur de F tendant à être significativement <u>supérieure</u> à l'unité et la table F suffira pour répondre à la question si $F > 1$. Par exemple, elle nous permettra de dire : il y a moins d'une chance sur mille que dans l'hypothèse de l'homogénéité le rapport F puisse, en raison des seules fluctuations d'échantillonnage, atteindre une telle valeur, il est donc raisonnable de soupçonner l'existence d'une cause systématique qui, si elle existe dans le cadre de l'existence des hypothèses de base, ne saurait être que l'hétérogénéité des moyennes.

Par contre, si la valeur numérique de F, fournie par l'expérience, est notablement inférieure à l'unité, la seule réponse de la table F est qu'un tel résultat a une faible (ou très faible) probabilité d'être compatible avec l'hypothèse $s_1^2 = s_2^2$, si s_1^2 et s_2^2 sont des estimations statistiquement indépendantes d'une même variance. Mais on ne pourra rien en conclure en ce qui concerne les moyennes, puisque - compte tenu des hypothèses de base - une alternative est seule possible :

Homogénéité $\longrightarrow$ F généralement voisin de 1.

Hétérogénéité $\longrightarrow$ F tend à s'éloigner de 1, mais en devenant plus grand.

Ceci doit donc conduire à suspecter la validité d'emploi du test, c'est-à-dire la validité de cet ensemble d'hypothèses de base et par conséquent à examiner attentivement ces hypothèses.

VI. - CONCLUSION

Les considérations qui précèdent n'ont pas pour but de minimiser l'importance des méthodes d'analyse de la variance, qui constituent un puissant moyen d'investigation, mais elles montrent que, même dans des cas très élémentaires tels que ceux qui ont été envisagés au paragraphe III, l'emploi de ces méthodes ne saurait se réduire à la manipulation arithmétique d'un tableau de chiffres.

Quelques connaissances statistiques, bien sûr, mais aussi un esprit critique averti, une parfaite connaissance technique du domaine étudié et des conditions dans lesquelles les observations ont été faites, sont des conditions nécessaires pour que l'analyse statistique apporte quelque lumière sur des problèmes de comparaison que l'arithmétique - élémentaire ou savante - ne saurait résoudre seule.

BIBLIOGRAPHIE

- H. CRAMER - Mathematical Methods of Statistics
  Princeton University Press - 1946
- H.B. MANN - Analysis and Design of Experiments
  Dover Publications Inc. - New-York - 1949
- KEMPTHORNE - The design and analysis of experiment
  Chapmann and Hall - Londres - 1952
- W.G. COCHRAN - Some consequences when the assumptions for the analysis of variance ar not satisfied
  Biometrics - vol. III (1947) - p. 22-38
- CHURCHILL EISENHART - The assumptions underlying the analysis of variance
  Biometrics - vol'. III (1947) - p. 1-21
- VESSEREAU - Méthodes statistiques en biologie et agronomie
  Baillière et fils - Paris - 1948
- VESSEREAU - Cours professé au stage de formation statistique générale (Centre de formation des Ingénieurs)