

ANNALES DE L'I. H. P.

R.C. GEARY

Tests de la normalité

Annales de l'I. H. P., tome 15, n° 1 (1956), p. 35-65

http://www.numdam.org/item?id=AIHP_1956__15_1_35_0

© Gauthier-Villars, 1956, tous droits réservés.

L'accès aux archives de la revue « Annales de l'I. H. P. » 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/>

Tests de la Normalité

par

R. C. GEARY.

Ce Mémoire traite presque exclusivement des recherches personnelles de l'auteur. Il résume, en effet, plusieurs articles qui se rapprochent assez étroitement et qui ont été publiés durant une longue période. La plupart de ces articles ayant paru dans des journaux de statistique bien connus, je n'ai pas jugé nécessaire de reproduire en détail ici les démonstrations ou les formules (souvent très compliquées). Je n'ai pas donné non plus de références à une bibliographie étendue, celles-ci étant indiquées dans les articles mentionnés à la fin de ce travail.

1. Propriétés d'indépendance de la distribution normale. — Le fait qu'il est possible d'obtenir des distributions exactes (ou, comme dans le cas présent, des approximations serrées plus ou moins contrôlées) de fréquences des diverses fonctions de test pour des échantillons tirés au hasard des univers normaux, dans lesquels les deux paramètres, la moyenne μ et l'écart-type σ , sont inconnus, est dû en grande partie aux propriétés remarquables de l'indépendance stochastique de certaines fonctions symétriques de l'échantillon. Soit x_1, x_2, \dots, x_n l'échantillon normal. Il est bien connu que la moyenne \bar{x} et l'écart-type s sont indépendants stochastiquement. Cette condition nécessaire est aussi suffisante. La démonstration, donnée il y a à peu près vingt ans, supposait l'existence de tous les moments de la population [3]. En 1942, E. Lukacs [15] a démontré le théorème sous l'hypothèse que les moments existent jusqu'au deuxième. Récemment D. Basu et R. G. Laha [10], [14] ont montré que le résultat est vrai sans aucune hypothèse quant aux moments.

Il est de grande importance, du point de vue présent que, pour des échantillons normaux de taille n , une fonction du type

$$(1.1) \quad u = \frac{f(x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x})}{s^k},$$

où la fonction f (ordinairement un polynôme symétrique) est homogène de degré k , soit indépendante stochastiquement de s où

$$(1.2) \quad (n-1)s^2 = \Sigma(x_i - \bar{x})^2.$$

On démontre cette propriété tout d'abord en changeant les variables x en y par une transformation orthogonale [voir, par exemple, formule (3.9), plus loin] dont un terme est

$$(1.3) \quad y_n = \sum \frac{x_i}{n^{\frac{1}{2}}}$$

et, après l'inversion, en substituant en (1.1) qui devient de la forme

$$(1.4) \quad u = \frac{g(y_1, y_2, \dots, y_{n-1})}{s^k},$$

où

$$(1.5) \quad (n-1)s^2 = \sum_{j=1}^{n-1} y_j^2.$$

Par une transformation polaire des variables y_j , de rayon $(n-1)^{\frac{1}{2}}s$ et d'angles $\theta_0, \theta_1, \dots, \theta_{n-3}$ [voir formule (4.3)], u devient

$$(1.6) \quad u = h(\theta_0, \theta_1, \dots, \theta_{n-3}),$$

où la fonction h est indépendante de s . Les deux transformations successives de la fréquence originale

$$(1.7) \quad C \prod_{i=1}^n e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} dx_i$$

montrent que s est indépendant des angles polaires qui sont également indépendants les uns des autres ⁽¹⁾. Par conséquent, la forme de (1.6) montre que u est indépendant de s . Alors

$$(1.8) \quad E(f) = E(us^k) = E(u)E(s^k),$$

⁽¹⁾ On démontre cette propriété incidemment plus loin.

où E indique l'espérance mathématique, d'où

$$(1.9) \quad E(u) = \frac{E(f)}{E(sk)}.$$

Par suite de cette relation, il sera possible de calculer, soit exactement, soit à une approximation serrée, les moments d'un réseau de tests de normalité. Dans la suite nous appellerons la propriété fondamentale (1.9), théorème A.

2. Effet de la non-normalité de l'univers sur les probabilités-types. —

La méthode générale appliquée ici consiste à calculer les valeurs asymptotiquement approximatives des moments inférieurs de quelques tests de signification bien connus et, dans certains cas, à estimer les probabilités à l'aide des moments approximatifs pour des types particuliers des univers non normaux. Ces probabilités sont comparées ensuite avec les valeurs correspondantes pour la théorie normale.

a. Le test z . — On trouve le cas le plus frappant de l'effet de la non-normalité dans la forme la plus simple du test z , destinée à établir l'équivalence probable des variances de la population de deux échantillons indépendants, $x_1, x_2, \dots, x_{n'}$, et $y_1, y_2, \dots, y_{n''}$ respectivement. Posons

$$(2.1) \quad z = \frac{1}{2} \log \frac{(n''-1) \Sigma(x_i - \bar{x})^2}{(n'-1) \Sigma(y_j - \bar{y})^2}.$$

Soit μ_2 le deuxième moment de l'univers et μ_4 le quatrième moment, avec $\beta_2 = \frac{\mu_4}{\mu_2^2}$, il est facile à montrer que la variance de z est approximativement

$$(2.2) \quad \sigma_z^2 = \frac{\beta_2 - 1}{4} \left(\frac{1}{n'} + \frac{1}{n''} \right) = M_2$$

quand n' et n'' sont tellement grands que les termes en n' et n'' de degré inférieur à -1 sont négligeables. Alors, puisque la distribution de fréquences de z tend vers la normalité quand n' et n'' croissent, on peut calculer très facilement à partir de (2.2) pour de grands échantillons les vraies probabilités au-delà des points des probabilités normales. En effet, quand les échantillons sont infiniment grands, l'écart z , à savoir z_0 qui correspond au point de probabilité normale de 2,5 %, par exemple, est

$$(2.3) \quad z_0 = 1,9600(M_2^0)^{\frac{1}{2}},$$

où M_2^0 est la variance quand l'univers est normal, c'est-à-dire quand $\beta_2 = 3$ en (2.2). Si, pourtant, l'univers n'était pas normal et s'il avait une variance M_2 avec $\beta_2 \neq 3$ la vraie probabilité d'un écart plus grand que z_0 en valeur absolue serait, non pas 0,05, mais la probabilité normale relative à l'écart $z_0 M_2^{-\frac{1}{2}}$ (variance unité). Le tableau I montre les vraies probabilités pour diverses valeurs de β_2 .

TABLEAU I.

*L'effet des changements de voussure de l'univers ($\beta_2 - 3$)
pour de grands échantillons sur la probabilité de z .*

β_2 .	Vraie probabilité.	β_2 .	Vraie probabilité.
1,5.....	0,000 089	4.....	0,110
2.....	0,005 6	4,5.....	0,138
2,5.....	0,024	5.....	0,166
3.....	0,050	5,5.....	0,191
3,5.....	0,080	6.....	0,215

Ceci veut dire que quand on *pense* que la probabilité de l'écart est 0,05, la probabilité réelle peut être 0,000 089 si le β_2 de l'univers est 1,5 et 0,215 si $\beta_2 = 6$. Par conséquent si la voussure ($\beta_2 - 3$) de l'univers est négative, l'usage du tableau normal cache des différences significatives, c'est-à-dire entre les variances; si la voussure est positive des différences sont exagérées. A moins que des tests systématiques n'aient établi que la voussure est petite, on ne doit pas se servir du tableau normal de z pour chercher des différences significatives des variances des univers.

Cependant on se sert le plus fréquemment du test z dans l'analyse de la variance et par bonheur dans de telles applications l'effet de la non-normalité est beaucoup moins marqué que dans le cas le plus simple que nous venons de traiter. Supposons qu'on tire k échantillons au hasard du même univers, n_j dans l'échantillon j , avec $\sum n_j = n$. On suppose que n est assez grand pour qu'on puisse négliger les termes en n^{-2} , que k est petit, et que tous les n_j sont du même ordre de grandeur que n , c'est-à-dire que, si

$$(2.4) \quad n_j = \pi_j n, \quad \sum \pi_j = 1,$$

aucun des π_j n'est suffisamment petit pour être négligeable. Dans ce cas

$$(2.5) \quad z = \frac{1}{2} \log \frac{X}{Y},$$

où

$$(2.6) \quad \left\{ \begin{array}{l} (k-1)X = \sum_j n_j (m_{1j} - m_1)^2, \\ (n-k)Y = \sum_j (n_j - 1) m_{2j}, \\ nm_1 = \sum_j n_j m_{1j}, \end{array} \right.$$

avec m_{1j} et m_{2j} , moyenne et variance de l'échantillon j . On trouve que, si $w = \frac{X}{Y}$,

$$(2.7) \quad \left\{ \begin{array}{l} Ew \sim 1 + \frac{2}{n}, \\ Ew^2 \sim \frac{k+1}{k-1} + \frac{1}{n(k-1)^2} [6(k^2-1)(k^2+2k-2-\alpha)(\beta_2-3)], \end{array} \right.$$

où

$$\alpha = \sum_j \pi_j^{-1}.$$

Il est à remarquer que les termes sans n^{-1} dans (2.7) sont indépendants de β_2 et que le troisième moment n'entre pas dans les approximations. La forme de l'expression (2.7) montre que, quand n est grand par rapport à k , l'effet de la voussure de la population est relativement sans importance.

b. Le test t . — Supposons que la moyenne de la population soit zéro. Alors

$$(2.8) \quad t = n^{\frac{1}{2}} \frac{\bar{x}}{s}.$$

Dans [8] des développements allant jusqu'aux termes en n^{-2} (c'est-à-dire ne tenant aucun compte ni de $n^{-\frac{5}{2}}$ ni des puissances moins élevées) sont données pour les six premiers cumulants de t en termes de $\lambda_m = \frac{\lambda'_m}{\lambda_2^{\frac{m}{2}}}$, où les λ'_m sont les cumulants de l'univers, c'est-à-dire $\lambda'_m = 0$ pour $m > 2$ pour des univers normaux. On ne reproduit pas ici ces formules, qui sont compliquées. On se sert des six développements pour déduire des approximations des probabilités de t pour des échantillons

de 10 de six univers non normaux particuliers, qui sont caractérisés par des valeurs différentes de

$$\beta_1^{\frac{1}{2}} = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} \quad \text{et} \quad \beta_2 = \frac{\mu_4}{\mu_2^2}.$$

Les probabilités approximatives qui correspondent au point de la probabilité normale 0,025 pour t , c'est-à-dire $t = -2,262$ sont données dans le tableau II.

TABLEAU II.

Probabilité de t inférieur à $-2,262$ pour des échantillons de 10 (pour sept univers).

Univers.	$\lambda_3 = \beta_1^{\frac{1}{2}}$.	$\lambda_4 = \beta_2 - 3$.	Probabilité.
Normal.....	0	0	0,025
2.....	0	1	0,024
3.....	1/2	0	0,041
4.....	$1/\sqrt{2}$	1/2	0,047
5.....	1	0	0,072?
6.....	1	1	0,086?
7.....	1/2	1/2	0,043

Comme on l'a proposé dans des travaux antérieurs (W. S. Gosset, 1908; R. C. Geary, 1936) le tableau montre que la distorsion n'est pas sensible pour les univers symétriques, c'est-à-dire pour ceux où $\beta_1 = 0$. Il n'est pas douteux, cependant, que les probabilités dans le tableau-type de t pour des univers normaux peuvent différer sérieusement des vraies probabilités quand les univers sont très asymétriques. Il s'agit jusqu'ici des probabilités « d'un côté » : la distorsion n'est pas importante, en général, même pour des univers asymétriques, dans les régions $|t| > t_0$, où t_0 est le point de la probabilité de l'univers normal.

Quand on se sert de t comme test de signification de la différence entre deux moyennes, la fonction t a la forme

$$(2.9) \quad t = \frac{(\bar{x}' - \bar{x}'')(n' + n'' - 2)^{\frac{1}{2}}}{[(n' - 1)s'^2 + (n'' - 1)s''^2]^{\frac{1}{2}}} \left(\frac{n' n''}{n' + n''} \right)^{\frac{1}{2}}$$

pour des échantillons de n' et n'' , avec \bar{x}' et \bar{x}'' , moyennes et s'^2 et s''^2 , variances des échantillons. Quand on suppose que les moyennes des deux univers sont égales on a obtenu en [8] les expressions pour les

quatre premiers cumulants pour des univers non normaux caractérisés par leurs cumulants. A partir des valeurs approximatives des cumulants de t , on a estimé les vraies probabilités sur quatre exemples pour $t \leq \tau$, où τ est la valeur négative du tableau normal pour la probabilité 0,025 de t . Les résultats sont donnés dans le tableau III.

TABLEAU III.

Exemple.	n' .	n'' .	λ_3' .	λ_3'' .	λ_4' .	λ_4'' .	Probabilité.
1.....	12	4	1	-1	1	-1	0,045
2.....	18	6	1	-1	1	-1	0,041
3.....	7	4	$1/\sqrt{2}$	$1/\sqrt{2}$	1/2	1/2	0,027
4.....	10	6	1	0	1	0	0,036

L'exemple 3 suggère que si les univers dont on a tiré les deux échantillons sont les mêmes, les distorsions ne sont pas importantes.

Il faut remarquer que les probabilités données dans le tableau III (ainsi que dans le tableau II) ne sont que grossièrement approchées, les échantillons sont beaucoup trop petits pour que les résultats prétendent à l'exactitude. Ces exemples ont seulement pour but de montrer que la probabilité réelle peut être très différente des valeurs classiques pour de petits échantillons.

3. Réseau de tests de la normalité $a(c)$. — Nous considérons maintenant les propriétés d'un réseau de tests symétriques de la normalité $a(c)$ où

$$(3.1) \quad a(c) = \frac{\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|^c}{\left[\frac{1}{n} \sum (x_i - \bar{x})^2 \right]^{\frac{1}{2}c}},$$

avec $\bar{x} = \sum \frac{x_i}{n}$ et c est réel. Dans [8] on a montré que si c est non négatif et si $\mu_{|k|} = E|x|^k$, où $k = \max(2c, 4)$, existe, alors la fréquence de $a(c)$ tend vers la normalité au fur et à mesure que n croît. La démonstration de ce théorème comprend deux étapes :

(i) $a_1(c)$, qui est la valeur de $a(c)$ avec $\bar{x} = 0$ et la moyenne de l'univers zéro, tend vers la normalité;

(ii) $a_1(c)$ tend en probabilité vers $a(c)$,

dans les conditions indiquées. La démonstration de (i), à son tour, implique le développement du dénominateur limité à trois termes (y compris le reste) selon le théorème de Taylor. On soupçonne que le théorème est vrai, même pour une classe limitée de valeurs négatives de c . En effet, la démonstration donnée dans [8] est assez compliquée et maladroite : en la simplifiant on trouverait peut-être un théorème plus fort.

On remarquera que $a(4) = b_2$, le test le plus familier de la vousure, qui est une estimation cohérente de β_2 de l'univers est égal à 3 quand l'univers est normal. Aussi $a(1) = \alpha$, dont la distribution de la fréquence est considérée dans [4] où l'on donne un tableau des points inférieurs et supérieurs de la probabilité de 1, 5 et 10 % pour des échantillons de 11 et au-dessus. Le test α tend en probabilité vers

$$\left(\frac{2}{\pi}\right)^{\frac{1}{2}} = 0,797\ 884\ 560\ 8\dots$$

Par suite du théorème A, le calcul des moments de $a(c)$ pour des échantillons normaux n'offre aucune difficulté théorique. En effet, le $k^{\text{ième}}$ moment M'_k est

$$(3.2) \quad M'_k[a(c)] = E[a(c)]^k = \frac{E\left(\frac{1}{n} \sum |x_i - \bar{x}|^c\right)^k}{E\left[\frac{1}{n} \sum (x_i - \bar{x})^2\right]^{\frac{1}{2}ck}}$$

La valeur exacte du dénominateur est bien connue :

$$(3.3) \quad E\left[\frac{1}{n} \sum (x_i - \bar{x})^2\right]^{\frac{1}{2}ck} = \frac{\left(\frac{2}{n}\right)^{\frac{1}{2}ck} \left(\frac{n+ck-3}{2}\right)!}{\left(\frac{n-3}{2}\right)!}$$

La valeur du numérateur de (3.2) s'obtient en développant $(\sum |z_i|^c)^k$, avec $z_i = x_i - \bar{x}$, par le théorème multinomial et en calculant l'espérance mathématique de chaque terme, en se servant de la répartition combinée des fréquences des z_i , qui, pour quatre valeurs (z_1, z_2, z_3, z_4), par exemple, est

$$(3.4) \quad \frac{1}{(2\pi)^2} \left(\frac{n}{n-4}\right)^{\frac{1}{2}} \exp\left[-\frac{n-3}{2(n-4)}(z_1^2 + z_2^2 + z_3^2 + z_4^2) - \frac{1}{n-4}(z_1 z_2 + z_1 z_3 + \dots)\right] dz_1 dz_2 dz_3 dz_4.$$

Naturellement cette expression représente aussi la fréquence de n'importe quel groupe de quatre termes de (z_1, z_2, \dots, z_n) . Les fréquences correspondantes pour une, deux et trois variables z sont données aussi en [4], formule (11). Il est évident qu'il ne faut que ces quatre distributions pour calculer les quatre premiers moments de $a(c)$. Les moments sont donnés dans [8] approximativement sous la forme de développements en $\frac{1}{n}$ à plusieurs termes. C'est seulement le premier moment qu'on peut calculer exactement à partir des fonctions connues. Cette valeur est

$$(3.5) \quad M'_1(a(c)) = \frac{\pi^{-\frac{1}{2}} (n-1)^{\frac{1}{2}c} \left(\frac{c-1}{2}\right)! \left(\frac{n-3}{2}\right)!}{\left(\frac{n+c-3}{2}\right)!}.$$

Pour $c=0$ et $c=2$, on trouve naturellement que $M'_1=1$ parce que dans de tels cas $a(c)$ est égal à l'unité pour chaque échantillon. Pour toute autre valeur positive de c (et peut-être pour certaines valeurs négatives) on peut construire un test de la normalité, au moins théoriquement. Pour un tel test il faut une distribution de fréquences pour des échantillons normaux pour chaque valeur de n . On n'a pas encore trouvé possible d'établir une distribution *exacte* pour aucun des membres du réseau de test $a(c)$, sauf seulement pour de très petits échantillons qui ne sont pas utiles dans la pratique.

On montrera plus loin que, pour des échantillons infiniment grands, $a(4)$ est le test le plus efficace dans un champ d'univers symétriques. Par conséquent, la déduction, nécessairement par des méthodes empiriques, de la distribution de $a(4)$ pour des échantillons normaux est un problème important qui n'a pas été résolu. En se servant de la formule convenable de K. Pearson et des valeurs exactes des quatre premiers moments donnés par R. A. Fisher, E. S. Pearson a calculé un tableau de fréquences pour $a(4)$ pour $n \geq 100$: on ne peut pas se fier aux probabilités calculées par ces méthodes, pour des échantillons plus petits. La difficulté principale est que la tendance à la normalité de la distribution de $a(4)$ lorsque n croît est extrêmement lente ce qui rend impossible l'emploi de l'approximation de Gram-Charlier, par exemple. Pour $n=100$, $B_2^{\frac{1}{2}}=1,27$ et $B_2=6,75$ pour $a(4)$, très différent des valeurs normales 0 et 3 respectivement. De l'autre côté c'est principa-

lement le voisinage de la normalité dans la distribution de $\alpha(1)$ qui a rendu possible de calculer le tableau des fréquences :

pour $n = 11$,

$$B_1^{\frac{1}{2}} = -0,41, \quad B_2 = 3,14;$$

pour $n = 101$,

$$B_1^{\frac{1}{2}} = -0,17, \quad B_2 = 3,05.$$

On montrera plus loin que les probabilités calculées de $\alpha(1)$ pour $n \geq 11$ sont des approximations très serrées.

Pour encourager des chercheurs à préparer un tableau de la probabilité de $\alpha(4)$ pour des échantillons de petite taille et de taille moyenne on a fourni les valeurs exactes pour les sept premiers moments de $\alpha(4)$. On donne des références aux calculs des six premiers moments dans [6] où l'on trouvera aussi le septième moment. La méthode de [6] se sert de la propriété, qui est facile à démontrer, que, pour des échantillons d'un univers normal, de moyenne zéro et de variance unité,

$$(3.6) \quad E(x_1 - \bar{x})^{a_1} (x_2 - \bar{x})^{a_2} \dots (x_p - \bar{x})^{a_p}$$

s'obtiendra comme le coefficient de $t_1^{a_1} t_2^{a_2} \dots t_p^{a_p}$ dans le développement de

$$(3.7) \quad \frac{a_1! a_2! \dots a_p!}{k! 2^k} \left[t_1^2 + t_2^2 + \dots + t_p^2 - \frac{(t_1 + t_2 + \dots + t_p)^2}{n} \right]^k,$$

où $2k = \Sigma a_i$. Il n'est pas difficile de déterminer les coefficients quand k est petit. Cependant pour calculer le septième moment, $k = 14$, il fallait développer une technique, même une arithmétique spéciale. Le résultat définitif ([6], formule (3.9)) est donné sous la forme du quotient d'un polynôme de puissance 13 en n par $(n+1)(n+3) \dots (n+25)$. La plupart des coefficients dans le numérateur sont des nombres astronomiquement grands.

Il est assez évident que, pour des échantillons de taille modérée, la distribution de fréquences de $\alpha(c)$ doit être très semblable à celle de $\alpha_1(c)$ où

$$(3.8) \quad a_1(c) = \frac{\frac{1}{n} \Sigma |x_i|^c}{\left(\frac{1}{n} \Sigma x_i^2 \right)^{\frac{c}{2}}},$$

quand la moyenne de l'univers est zéro. On a proposé [2] cette forme du test de la normalité en prenant pour n la taille de l'échantillon

diminuée de l'unité et $c = 1$. Le procédé était le suivant : Soit, comme précédemment, l'échantillon normal, x_1, x_2, \dots, x_n dans lequel les observations sont indépendantes, ce qui implique, en particulier, qu'elles sont arrangées au hasard. On introduit les variables y données par la transformation orthogonale d'Helmert :

$$(3.9) \quad \left\{ \begin{array}{l} y_1 = \frac{x_1 - x_2}{2^{\frac{1}{2}}}, \\ y_2 = \frac{x_1 + x_2 - 2x_3}{6^{\frac{1}{2}}}, \\ \dots\dots\dots \\ y_{n-1} = \frac{x_1 + x_2 + \dots + x_{n-1} - (n-1)x_n}{[n(n-1)]^{\frac{1}{2}}}, \\ y_n = \frac{x_1 + x_2 + \dots + x_n}{n^{\frac{1}{2}}}. \end{array} \right.$$

Les y_i sont alors des variables normales indépendantes mais maintenant y_1, y_2, \dots, y_{n-1} ont une moyenne d'univers égale à zéro. Par conséquent on peut appliquer le test $a_1(c)$ aux $(n - 1)$ variables transformées. L'objection principale à ce test est que la valeur de $a_1(c)$ dépend de l'ordre particulier de l'échantillon original, afin qu'il y ait en général $n!$ valeurs de $a_1(c)$ pour chaque échantillon. Malheureusement il n'existe pas de transformation orthogonale symétrique sauf pour $n = 4$.

Pour établir que, pour des échantillons de taille modérée la distribution de $a(c, n)$ doit être assez semblable à celle de $a_1(c, n - 1)$, on a dressé le tableau IV qui compare les valeurs des paramètres principaux de la fréquence pour chacune des fonctions de test pour $c = 1$ et $c = 4$ et pour $n = 26$ et $n = 51$.

TABLEAU IV.

Les paramètres de la fréquence de $a_1(c, n - 1)$ et $a(c, n)$

Test..	pour $c = 1$ et 4 et $n = 26$ et 51 .							
	$n = 26$		$n = 51$					
	Moyenne.	Écart type.	$B_1^{\frac{1}{2}}$	B_2	Moyenne.	Écart type.	$B_1^{\frac{1}{2}}$	B_2
$a_1(1, n - 1) \dots$	0,8059	{ 0,0410	— 0,33	3,16	0,8019	{ 0,0295	— 0,24	3,10
$a(1, n) \dots\dots$								
$a_1(4, n - 1) \dots$	2,7778	{ 0,7411	1,79	9,21	2,8846	{ 0,5997	1,59	8,50
$a(4, n) \dots\dots$								

Il est assez évident qu'à partir du tableau IV, pour une envergure assez étendue des valeurs de c et pour des échantillons de taille modérée pour lesquels nous pourrions espérer qu'un test de la normalité serait effectif, la distribution de $a_1(c)$ est très proche de celle de $a(c)$. L'effet de cette remarque est que le calcul des moments et la dérivation (nécessairement approximative) des fréquences est beaucoup plus simple dans le cas de $a_1(c)$ que de $a(c)$. En particulier, le calcul d'un tableau des probabilités de $a_1(4)$ pour des tailles de 25 et plus vaut la peine d'être établi.

Des tests qui sont des fonctions exactes l'un de l'autre pour chaque échantillon sont évidemment également efficaces. Par conséquent, il sera intéressant d'étudier la corrélation entre $a_1(c)$ et $a_1(c')$, $c \neq c'$, qui devrait être ± 1 si la relation fonctionnelle était exactement linéaire. On a donné dans [8] des formules pour la valeur approchée du coefficient de corrélation même quand la forme de l'univers est tout à fait générale. En particulier, pour des échantillons normaux infiniment grands le coefficient est $-\frac{1}{[12(\pi-3)]^{\frac{1}{2}}} = -0,767$ pour $c=1$, $c'=4$, et pour $c=2,2$ et $c'=4$ le coefficient est $+0,887$.

4. Distributions de fréquences. — Les tests $a(c)$ [et $a_1(c)$] qu'on a examinés jusqu'ici, bien que désignés comme tests de la normalité sont en réalité des tests de la voussure. Ils ne sont pas efficaces comme tests d'asymétrie. Naturellement si l'échantillon se montre asymétrique (en probabilité) par suite de quelque test, il est non normal. Aussi, par conséquent, les tests de l'asymétrie de l'univers ne sont guère moins importants que les tests de la voussure. D'une manière analogue aux tests de la voussure nous pouvons considérer le champ des tests d'asymétrie

$$(4.1) \quad g(c) = \frac{\frac{1}{n}(-\Sigma' |x_i - \bar{x}|^c + \Sigma'' |x_i - \bar{x}|^c)}{\left[\frac{1}{n} \Sigma (x_i - \bar{x})^2 \right]^{\frac{c}{2}}}, \quad c > 0,$$

où Σ' s'étend aux observations x_i inférieures à la moyenne \bar{x} et Σ'' s'étend au reste de l'échantillon. Puisque le théorème A s'applique à $g(c)$ pour des échantillons normaux, il n'y a pas de difficulté théorique à trouver les moments inférieurs de $g(c)$: le moment d'ordre k est le quotient du

moment du numérateur par le moment du dénominateur de l'ordre k .

On notera que $g(3) = \frac{m_3}{m_2^{\frac{3}{2}}}, = b_1^{\frac{1}{2}}$, test bien connu d'asymétrie. C'est

une grande simplification que, pour des échantillons normaux, la distribution de fréquences de $g(c)$ soit symétrique. Dans cette partie on considère les distributions de fréquence $a_1(1)$, $a(1)$ et $g(3)$.

Il sera peut-être convenable de commencer cette partie par quelques remarques générales au sujet des distributions de fréquences des fonctions des échantillons normaux, par exemple t , s et z . Tandis que les distributions de fréquences ou de la probabilité sont fondamentales dans la théorie des sondages, en pratique les fonctions de l'échantillon pour lesquelles on connaît les distributions *exactes* de fréquences sont malheureusement peu nombreuses, même si l'on postule un univers normal, et même si l'on considère comme « connue » une fréquence sous la forme d'une intégrale multiple pourvu que l'ordre de l'intégrale soit indépendant de la taille n de l'échantillon : nous pouvons ordinairement écrire la probabilité sous la forme d'une intégrale d'ordre n , mais on ne peut regarder une telle expression comme une solution utile. Quand on connaît les moments des fonctions de l'échantillon, dans la pratique on a recours aux distributions approximatives en utilisant, par exemple, la « série différentielle » de l'École scandinave (dont le développement de Gram-Charlier est un cas particulier) ou le système de Karl Pearson. Dans de tels procédés, cependant, ce n'est que très rarement qu'on peut découvrir les limites serrées des erreurs de l'approximation, à défaut des restes approximatifs ou pour tous les points ou pour quelques points de la probabilité de la fonction de l'échantillon. Quant aux fonctions de test $a_1(1)$ et $g(3)$ on peut trouver un intérêt plus général dans les méthodes dont l'on se sert pour déterminer les distributions approximatives, parce qu'elles tiennent compte, non seulement des moments normaux des tests mais aussi de ce qu'on peut appeler *l'anatomie* des distributions. La méthode, en effet, comprend les cas suivants :

(i) Établir une formule intégrale d'itération pour les ordonnées de la distribution.

(ii) Trouver les limites du test et les points intermédiaires (appelés *chaîtons*) entre lesquels la distribution a des formes fonctionnelles différentes.

(iii) Établir la forme de la distribution près des limites de l'envergure.

(iv) En connaissant la distribution exacte pour les valeurs les plus petites de n [= 2 dans le cas de $a_1(1)$ et = 3 dans le cas de $g(3)$], employer l'itération intégrale [voir (i)] pour estimer les ordonnées pour des valeurs successives de n [jusqu'à $n=10$ pour $a_1(1)$ et $n=8$ pour $g(3)$].

(v) Continuer l'itération jusqu'à ce que les distributions soient étroitement semblables à l'approximation de Gram-Charlier.

(vi) Se servir de l'approximation de Gram-Charlier pour toutes les tailles plus grandes que la dernière valeur dans l'itération.

La méthode implique que si une formule générale (ici celle de Gram-Charlier) donne une approximation serrée de la distribution réelle pour une valeur particulière de $n = v$, l'approximation sera aussi bonne, au moins, pour $n > v$. Il est évident que la méthode ne fournit pas une formule *exacte* pour la distribution comme le fait, par exemple, l'expression bien connue de la distribution de t pour des échantillons normaux. La méthode est, pourtant, beaucoup plus précise que la façon ordinaire d'ajuster les courbes. Peut-être pourrait-on décrire la méthode comme « quasi exacte ».

L'application de cette méthode à $a_1(1)$ et à $g(3)$ est facilitée considérablement par le fait que dans les deux cas il est évident que, même pour des valeurs petites de n , l'expression de Gram-Charlier doit fournir une bonne approximation puisque les semi-invariants d'ordre 3 et plus sont très petits. Malheureusement il en sera tout autrement dans le cas de la fonction de test b_2 [= $a(4)$] que l'auteur a suggéré comme un problème à aborder. Par bonheur ou par bon jugement, pourtant, il sera peut-être possible, de trouver une forme convenable pour une fonction de répartition [= $f(b_2)$] relative à des tailles modérées de l'échantillon : ce serait peut-être un test de l'efficacité de la fonction de la baser sur les six premiers moments et de comparer ensuite la valeur du septième moment, calculé à partir de la fonction avec la valeur exacte. On pourrait sans aucun doute déterminer les ordonnées de la distribution de proche en proche en commençant avec la distribution connue pour $n=4$ (trouvée par Mc Kay), bien que les calculs soient considérables. On

qui correspond à la région polaire

$$\frac{\pi}{2} \geq \varphi_i \geq 0 \quad (i = 0, 1, 2, \dots, n-2).$$

De (4.3) et (4.5)

$$(4.6) \quad n^{\frac{1}{2}} \omega_n = c_{n-2} + s_{n-2}(c_{n-2} + s_{n-3}(c_{n-4} + \dots)) = c_{n-2} + (n-1)^{\frac{1}{2}} \omega_{n-1} s_{n-2}$$

qui est le premier pas de l'itération. Puisque les angles polaires sont tous indépendants pour des échantillons normaux, la probabilité jointe de ω_{n-1} et φ_{n-2} est

$$(4.7) \quad C_{n-2} s_{n-2}^{n-2} d\varphi_{n-2} f(\omega_{n-1}) d\omega_{n-1},$$

où $f(\omega_{n-1})$ représente la fréquence de ω_{n-1} et pour la constante C_{n-2} ,

$$C_{n-2} \int_0^{\frac{\pi}{2}} s_{n-2}^{n-2} d\varphi_{n-2} = 1.$$

Ensuite, de (4.6), quand φ_{n-2} est fixe,

$$n^{\frac{1}{2}} d\omega_n = (n-1)^{\frac{1}{2}} s_{n-2} d\omega_{n-1},$$

et, en substituant, pour $d\omega_{n-1}$, l'élément (4.7) devient

$$C_{n-2} \left(\frac{n}{n-1} \right)^{\frac{1}{2}} s_{n-2}^{n-2} d\varphi_{n-2} f(\omega_{n-1}) d\omega_n.$$

Finalement, la distribution de fréquence de ω_n se dérive de celle de ω_{n-1} par la formule d'itération intégrale suivante :

$$(4.8) \quad f(\omega_n) = C_{n-2} \left(\frac{n}{n-1} \right)^{\frac{1}{2}} \int_0^{\cos(\alpha-\beta)} s_{n-2}^{n-2} f(\omega_{n-1}) dc_{n-2},$$

avec, de (4.6),

$$\omega_{n-1} = \frac{n^{\frac{1}{2}} \omega_n - c_{n-2}}{(n-1)^{\frac{1}{2}} s_{n-2}}$$

et avec

$$\cos \alpha = n^{-\frac{1}{2}}, \quad \cos \beta = \omega_n.$$

Il est facile de montrer (voir [2], p. 320) que $f(\omega_n)$ a une forme

fonctionnellement différente dans chacun des intervalles

$$\left(\frac{p}{n}\right)^{\frac{1}{2}} \geq \omega_n \geq \left(\frac{p-1}{n}\right)^{\frac{1}{2}} \quad (p = 2, 3, \dots, n).$$

La fréquence est zéro en dehors des intervalles (c'est-à-dire pour $\omega_n \leq n^{-\frac{1}{2}}$, $\omega_n \geq 1$). On appelle les points $\omega_n = \left(\frac{p}{n}\right)^{\frac{1}{2}}$ les *chaînon*s.

Quant aux autres propriétés générales (c'est-à-dire pour toutes les valeurs de n) de la distribution on trouve :

(i) Près de la limite inférieure de l'envergure, c'est-à-dire $\omega_n = n^{-\frac{1}{2}}$, la forme de la distribution est

$$(4.10) \quad \frac{2^{n-1}}{\pi^{\frac{n}{2}}} n^{\frac{n+1}{2}} \frac{\left(\frac{1}{2}n-1\right)!}{(n-2)!} \left(\omega_n - \frac{1}{n^{\frac{1}{2}}}\right)^{n-2} d\omega_n;$$

(ii) Dans l'intervalle supérieur $1 \geq \omega_n \geq \left(\frac{n-1}{n}\right)^{\frac{1}{2}}$ la distribution est

$$(4.11) \quad A(1 - \omega_n^2)^{\frac{n-3}{2}} d\omega_n,$$

où

$$(4.12) \quad A \int_0^{\pi} \sin^{n-2} \varphi d\varphi = 2^n;$$

(iii) Dans le second intervalle $\left(\frac{n-1}{n}\right)^{\frac{1}{2}} \geq \omega_n \geq \left(\frac{n-2}{n}\right)^{\frac{1}{2}}$ la distribution est

$$(4.13) \quad \left[\frac{2^{n-1} n(n-2)}{\pi(n-1)^{\frac{1}{2}}} \int_0^{\cos \varphi} \left(1 - \frac{x^2}{n-1}\right)^{\frac{n-4}{2}} dx - \frac{n-2}{2} A \right] (1 - \omega_n^2)^{\frac{n-3}{2}} d\omega_n,$$

où $\omega_n = \cos \varphi$ et A est donné par (4.12).

Quant aux propriétés particulières, la distribution a été déterminée complètement pour $n = 2, 3, 4$ (voir [2]). Pour $5 \leq n \leq 10$ on a calculé les ordonnées de la distribution par quadrature en utilisant la formule d'itération (4.8) pour des valeurs de ω_n généralement aux intervalles de 0,03. A chaque étape ($n = 5, 6, \dots$) la fréquence totale était estimée, les petites déviations à partir de l'unité étant distribuées par ajustement des ordonnées. Pour $n = 10$ on a aussi calculé les fré-

quences en employant la formule de Gram-Charlier. On a trouvé que les deux séries étaient en concordance excellente (voir diagramme I). Le

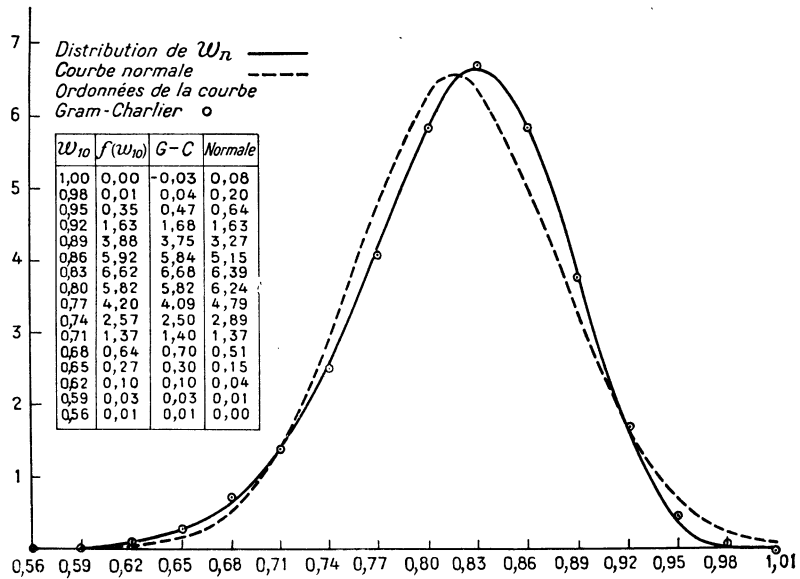


DIAGRAMME I.
Distribution de fréquences de w_n .

tableau V compare les valeurs des points 1 et 5 % de w_n selon les deux systèmes.

TABLEAU V.

Points de la probabilité de w_n , « réels » et de Gram-Charlier.

	Point 1 %.		Point 5 %.	
	Supérieur.	Inférieur.	Supérieur.	Inférieur.
« Réels ».....	0,941	0,656	0,911	0,710
Gram-Charlier (à deux termes).....	0,942	0,656	0,912	0,708

La ressemblance étroite des constantes de la distribution pour $a_1(1, n-1)$ et $a(1, n)$ (voir tableau IV) semble justifier la supposition qu'on pourrait aussi se servir de la formule de Gram-Charlier pour construire un tableau des points de la probabilité pour $n \geq 11$. On a dressé, par conséquent, dans [4] et [5] un tableau qui montre les points 1, 5 et 10 % supérieurs et inférieurs de la probabilité pour diverses valeurs de n jusqu'à $n = 1000$, muni de graphiques pour déterminer la non-normalité probable pour toutes les tailles de l'échantillon.

La distribution des fréquences de $b_1^{\frac{1}{2}}$ pour des échantillons normaux. — La méthode de la dérivation de la distribution de $b_1^{\frac{1}{2}}$ est essentiellement semblable à la méthode pour $a_1(1)$ bien que le calcul soit beaucoup plus compliqué. En mettant $b_1^{\frac{1}{2}} = t_n$, nous avons

$$(4.14) \quad t_n = \frac{n^{\frac{1}{2}} \Sigma (x_i - \bar{x})^2}{[\Sigma (x_i - \bar{x})^2]^{\frac{3}{2}}}.$$

Si l'on applique une transformation orthogonale d'Helmert (3.8) aux x_i on peut exprimer (4.14) en termes des $(n-1)$ variables y_1, y_2, \dots, y_{n-1} auxquelles on applique ensuite une transformation polaire, cette fois pourtant en $(n-1)$ variables c'est-à-dire avec $(n-2)$ angles polaires $\varphi_0, \varphi_1, \dots, \varphi_{n-3}$. On trouve l'itération

$$(4.15) \quad f(t_n) = \left(\frac{n-1}{\pi n} \right)^{\frac{1}{2}} \frac{\frac{1}{2}(n-3)!}{\frac{1}{2}(n-4)!} \int dx (1-x^2)^{\frac{1}{2}(n-7)} f(t_{n-1}),$$

où

$$(4.16) \quad n^{\frac{1}{2}} t_{n-1} = \frac{(n-1)^{\frac{1}{2}} t_n - 3x + (n+1)x^3}{(1-x^2)^{\frac{3}{2}}}$$

dans l'intégrale.

Dans (4.15) l'intégration s'étend à toutes les valeurs de x ($-1 \leq x \leq +1$) qui donnent des valeurs de $f(t_{n-1})$ différentes de zéro. On trouve les chaînons

$$(4.17) \quad t_n = k r_n = \frac{n-2k}{[k(n-k)]^{\frac{1}{2}}} \quad (k=1, 2, \dots, n-1).$$

La fréquence est nulle pour $t_n \geq \frac{n-2}{(n-1)^{\frac{1}{2}}}$. Les limites de l'envergure correspondent respectivement à $k=1$ et $k=n-1$. De (4.17), il est évident que l'origine est un chaînon seulement quand l'effectif de l'échantillon est un nombre pair.

Puisque la distribution de t_n est symétrique autour de l'origine, il faut considérer seulement les fréquences pour $t_n \geq 0$. Comme dans le cas de $a_1(1)$, nous trouvons que la distribution de t_n a une forme fonctionnelle différente dans chaque intervalle entre les chaînons. Dans

le cas de t_n on s'est décidé d'avoir égard au fait que l'itération intégrale successive (c'est-à-dire quand n croît) fournit un degré d'aplanissement qui s'accroît progressivement : autrement on ne pourrait pas espérer trouver une seule fonction continue permettant de se rapprocher efficacement de la vraie fréquence. Pour exprimer mathématiquement ce caractère, nous considérons la notion d'*ordre de contact* aux chaînons des diverses fonctions selon la définition suivante :

DÉFINITION. — Deux fonctions ont un ordre de contact k^{ν_n} au chaînon k^r_n si les valeurs des fonctions et celles de leurs premières ($k^{\nu_{n-1}}$) dérivées sont bornées et si elles prennent la même valeur au point chaînon. On peut montrer sans difficulté que $k^{\nu_n} = k-1^{\nu_{n-1}} + 1$.

Il sera convenable de dresser (dans le tableau suivant) pour les tailles les plus petites les valeurs des chaînons et leurs ordres de contact. On ne montre que les chaînons positifs. Les ordres de contact k^{ν_n} résulteront d'une proposition donnée plus loin, qui montre les valeurs réelles des fréquences auprès de la limite de l'envergure. On notera l'aplanissement non décroissant dans la direction de $t_n = 0$.

TABLEAU VI.

Valeurs de k^r_n et k^{ν_n} pour $n = 3$ à 8 inclusivement.

n .	1 ^{er} chaînon.		2 ^o chaînon.		3 ^o chaînon.		4 ^o chaînon.	
	1^r_n .	1^{ν_n} .	2^r_n .	2^{ν_n} .	3^r_n .	3^{ν_n} .	4^r_n .	4^{ν_n} .
3.....	$1/\sqrt{2}$	0	-	-	-	-	-	-
4.....	$2/\sqrt{3}$	0	0	0	-	-	-	-
5.....	$3/2$	1	$1/\sqrt{6}$	1	-	-	-	-
6.....	$4/\sqrt{5}$	1	$1/\sqrt{2}$	2	0	2	-	-
7.....	$5/\sqrt{6}$	2	$3/\sqrt{10}$	2	$1/\sqrt{12}$	3	-	-
8.....	$6/\sqrt{7}$	2	$2/\sqrt{3}$	3	$2/\sqrt{15}$	3	0	4

Près des limites $\left[t_n = \pm \frac{n-2}{(n-1)^{\frac{1}{2}}} \right]$ la fréquence est

$$(4.18) \quad f(t_n) = \frac{1}{3\pi^{\frac{1}{2}}} \frac{\frac{1}{2}(n-3)!}{\frac{1}{2}(n-4)!} \frac{(n-1)^{\frac{1}{2}(n-3)}}{(3n \cdot n-2)^{\frac{1}{2}(n-4)}} \left(\frac{n-2}{n-1} - t_n^2 \right)^{\frac{1}{2}(n-4)},$$

de sorte que, pour $n = 3$, la fréquence soit

$$(4.19) \quad f(t_3) = \frac{1}{\pi} \left(\frac{1}{2} - t_3^2 \right)^{-\frac{1}{2}}.$$

Pour $n = 4$, la fréquence est approximativement

$$(4.20) \quad f(t_4) = 0,285222 - 0,366466 \log_{10} |t_4| - 0,009178 |t_4|^{\frac{1}{2}} + 0,031359 |t_4|.$$

Les deux premiers coefficients de (4.20) sont exacts. Les deux autres coefficients sont obtenus par les conditions :

$$(i) \quad f_4 \left(\frac{2}{3^{\frac{1}{2}}} \right) = \frac{3^{\frac{1}{2}}}{2} [\text{de (4.18)}];$$

(ii) la fréquence totale est l'unité.

Il apparaîtra de la comparaison suivante que l'expression (4.20) est presque exacte :

TABLEAU VII.

Comparaison des valeurs des moments, réelles et calculées, pour $n = 4$.

Moment.	Valeur réelle.	Valeur calculée de la formule (4.20).
μ_2	0,342 857	0,342 930
μ_4	0,258 941	0,258 979
μ_6	0,240 503	0,240 263
μ_8	0,245 940	0,245 949

On peut ajouter que quand on s'est servi expérimentalement de la forme $A |t_4| + B t_4^2$ pour les deux derniers termes de $f(t_4)$, la comparaison des moments était excellente, mais pour chacune des quatre comparaisons la formule (4.20) était plus proche de la vérité.

Pour $n = 5$ à 8 exclusivement la fréquence s'obtient comme suit. Les ordonnées des fréquences (et toujours la fréquence à $t_n = 0$) s'obtiennent par des quadratures successives en employant la formule d'itération (4.15). On a supposé que les fonctions dans les intervalles entre les chaînons sont des polynômes, en passant par les ordonnées des chaînons déjà déterminées, ayant égard à l'ordre de contact aux chaînons et à la forme de la fonction aux limites de l'envergure. Les coefficients des polynômes sont déterminés en égalisant les moments, jusqu'aux degrés nécessaires, aux moments des fonctions. Pour $n \geq 8$ on s'est servi d'une approximation de Gram-Charlier (moyenne zéro et variance

unité) sous la forme

$$(4.21) \quad \begin{cases} f(\omega) = \exp \left[\frac{\lambda_4}{4! \lambda_2^2} \left(\frac{d}{d\omega} \right)^4 + \frac{\lambda_6}{6! \lambda_2^3} \left(\frac{d}{d\omega} \right)^6 + \frac{\lambda_8}{8! \lambda_2^4} \left(\frac{d}{d\omega} \right)^8 \right] \Phi(\omega), \\ \Phi(\omega) = (2\pi)^{-\frac{1}{2}} \exp \left(-\frac{1}{2} \omega^2 \right), \end{cases}$$

où les λ sont les semi-invariants de $t_n = b_1^{\frac{1}{2}}$. On a calculé les points de la probabilité 0,10, 0,05, 0,025, 0,01 et 0,001 pour $n = 8$ en utilisant la technique de Cornish-Fisher et aussi les probabilités « réelles » au-delà de ces points calculées par le procédé d'itération. Les résultats sont les suivants :

Probabilité de Gram-Charlier.....	0,10	0,05	0,025	0,01	0,001
Probabilité « réelle »...	0,096 855	0,050 459	0,026 825	0,011 504	0,001 155

Quand on considère que le procédé d'itération introduit ses propres erreurs cumulatives et que la courbe « réelle » est une synthèse des fonctions qui changent de forme à chaque chaînon, la concordance montrée doit être regardée comme satisfaisante. Il n'est guère douteux que, pour $n \geq 8$, on peut employer avec confiance la formule (4.21) pour construire un tableau des probabilités de $t_n = b_1^{\frac{1}{2}}$. Le diagramme II montre l'évolution de la courbe de fréquences de $n = 3$ à $n = 8$ inclusivement. Nous montrons dans chaque cas la courbe normale pour le même écart-type. On remarquera le changement frappant dans la forme des courbes entre $n = 5$ et $n = 6$.

5. Efficacité relative des divers tests de la normalité. — Considérons en premier lieu le cas des échantillons indéfiniment grands. Le problème est de déterminer la valeur de c qui rend le test $\alpha(c)$ [voir (3.1)] le plus sensible aux déviations de la normalité de l'univers. Par conséquent il nous faut envisager un champ d'autres univers non normaux. Puisque nous avons à identifier la voussure (et non pas l'asymétrie), ce champ doit être symétrique. La forme la plus simple du champ symétrique est celui de Gram-Charlier :

$$(5.1) \quad (2\pi)^{-\frac{1}{2}} \left[1 + \frac{\lambda_4}{24} \left(\frac{d}{dx} \right)^4 \right] \exp \left(-\frac{x^2}{2} \right).$$

On suppose, sans manque de généralité, que la variance de l'univers

est égale à l'unité. L'univers normal est naturellement un membre du champ : on le trouve pour $\lambda_4 = 0$.

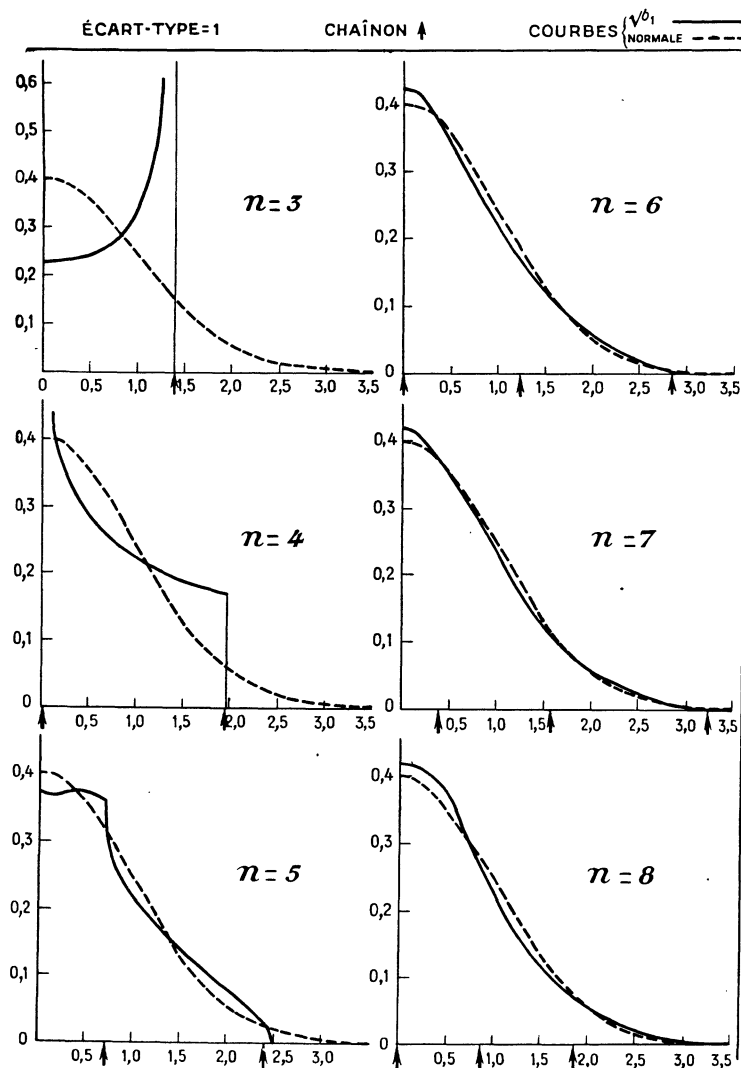


DIAGRAMME II.

L'évolution d'une fréquence : distribution de $\sqrt{b_1}$ pour des échantillons normaux (côté positif).

La méthode générale consiste à déterminer la moyenne $M_1(c, \lambda_4)$ de $a(c)$ pour la distribution (S. 1) et à calculer le *discriminant symé-*

trique $\rho(c)$ où

$$(5.2) \quad \frac{\lambda_4 \rho(c)}{4!} = \frac{M_1(c, \lambda_4) - M_1(c, 0)}{\sigma(c) n^{\frac{1}{2}}},$$

où $\sigma(c)$ est l'écart-type de $a(c)$ pour des échantillons normaux. En retenant seulement les termes jusqu'à ceux en $n^{-\frac{1}{2}}$, nous trouvons

$$(5.3) \quad M_1(c, \lambda_4) = \left(\frac{c-1}{2}\right)! \left(\frac{2^c}{\pi}\right)^{\frac{1}{2}} \left(1 + \frac{\lambda_4}{2^4} c \overline{c-2}\right),$$

$$(5.4) \quad \sigma(c) = \left(\frac{2^c}{\pi n}\right)^{\frac{1}{2}} \left[\left(\frac{2c-1}{2}\right)! \pi^{\frac{1}{2}} - \left(\frac{c-1}{2}\right)! \frac{c^2+2}{2} \right]^{\frac{1}{2}}.$$

De (5.2), (5.3) et (5.4) on trouve le discriminant

$$(5.5) \quad \rho(c) = c(c-2) \left[\frac{\left(\frac{2c-1}{2}\right)! \pi^{\frac{1}{2}}}{\left(\frac{c-1}{2}\right)!^2} - \frac{c^2+2}{2} \right]^{-\frac{1}{2}}.$$

En différentiant $\rho(c)$ nous trouvons $\rho'(c) = 0$ pour $c = 4$, quand $a(c) = b_2$. La raison de ce résultat remarquable est, sans doute, la forme que nous avons donnée au champ des univers considérés. Du point de vue mathématique le résultat n'est pas évident, parce que le calcul de la dérivée de $\rho(c)$ est assez compliqué. Du point de vue pratique la valeur de ce résultat dépend naturellement de ce qu'on croit être l'efficacité du développement de Gram-Charlier (5.1) pour représenter des univers symétriques. Le tableau VIII montre qu'il y a peu de différence dans l'efficacité du champ de tests $a(c)$ pour l'envergure $2 \leq c \leq 5$ environ

TABLEAU VIII.

$0 < c < 2$.	Discriminant $\rho(c)$.	$2 < c$.	Discriminant $\rho(c)$.
+ 0.....	-2,334	2 + 0.....	4,460
0,1.....	-2,541	2,1.....	4,508
0,2.....	-2,725	2,5.....	4,666
0,5.....	-3,188	3,0.....	4,801
0,7.....	-3,441	3,9.....	4,898
1,0.....	-3,758	4,0.....	4,900
1,1.....	-3,851	4,1.....	4,898
1,5.....	-4,166	5,0.....	4,818
1,9.....	-4,405	6,0.....	4,602
2 - 0.....	-4,460	7,0.....	4,288
		8,0.....	3,906

pour les grands échantillons. Plus la valeur de c est petite, plus il est facile de calculer les fréquences empiriques de $a(c)$, quand la taille n est donnée. Cette analyse montre qu'il serait utile de construire un tableau de la probabilité pour $a(c, n)$ pour $c = 2, 1$, par exemple. Nous donnons dans [8] la plupart des formules dont on a besoin.

D'une manière analogue nous pouvons considérer l'efficacité du réseau de tests d'asymétrie $g(c)$ [la formule (4.1)] à l'égard du champ des univers

$$(5.6) \quad (2\pi)^{-\frac{1}{2}} \left[1 + \frac{\lambda_3}{6} (x^3 - 3x) \right] \exp\left(-\frac{x^2}{2}\right),$$

c'est-à-dire le Gram-Charlier à deux termes pour la variance de l'univers égale à l'unité. Dans ce cas le *discriminant asymétrique* $r(c)$ est

$$(5.7) \quad r(c) = (c-1) \left[\left(\frac{2c-1}{2} \right)! \left(\frac{c}{2} \right)!^{-2} \frac{\pi^{\frac{1}{2}}}{2} - 1 \right]^{-\frac{1}{2}}.$$

Cette expression prend une valeur maximum pour $c = 3$ [c'est-à-dire $r'(3) = 0$] quand le test est le $b_1^{\frac{1}{2}}$ bien connu. Les valeurs de $r(c)$ pour quatre valeurs de c sont les suivantes :

c .	$r(c)$.	c .	$r(c)$.
2.....	2,370	4.....	2,389
3.....	2,450	5.....	2,236

La marge en faveur de $g(c)$ pour $c = 3$, en comparaison des autres tests dans l'envergure $2 \leq c \leq 5$ est, par conséquent, petite.

Les puissances des tests de la normalité. — L'analyse précédente de l'efficacité relative des tests de la normalité ne s'applique qu'aux échantillons infiniment grands. Pour comparer les efficacités pour une valeur donnée de n nous devons avoir recours à la théorie des fonctions de puissance de Neyman-Pearson. Il est bien connu que dans la pratique cette théorie implique le calcul :

(i) des points de la probabilité du test dans l'hypothèse normale selon des probabilités définies (0,05, 0,01, etc.);

(ii) de la probabilité, sur telle ou telle hypothèse non normale pour l'univers, du test au-delà des points normaux de la probabilité. On préfère le test qui donne la probabilité la plus élevée.

Nous donnons dans [8] les calculs qui sont formidables et approximatifs. Les deux tests que nous comparons sont $a(1)$, parce que pour ce test nous connaissons les probabilités pour tout n (voir [4] et [5]), et $a(4)$ parce que ce test est le plus efficace par rapport au Gram-Charlier (§. 1) dont on se sert aussi pour l'analyse des fonctions de puissance. On utilise seulement deux tailles de l'échantillon à savoir $n = 100$ et $n = 500$ et une seule valeur de λ_4 , à savoir $\lambda_4 = \frac{1}{2}$. Comme autre concession à la simplicité on utilise $a_1(c)$ [voir (4.5)] au lieu de $a(c)$. Il fallait utiliser des valeurs assez élevées de n et une valeur basse de λ_4 pour donner une mesure de vraisemblance à l'approximation des distributions des probabilités de $a_1(4)$ pour $n = 100$, que nous avons déterminés à partir de la formule-type IV de Karl Pearson (basée sur les développements asymptotiques des quatre premiers moments) comme suit :

Distributions approximatives des fréquences de $a_1(4)$ pour $n = 100$:

$$(5.8) \quad \text{Normal : } \lambda_4 = 0 \quad \left\{ \begin{array}{l} C \cos^{11,33500} e^{13,015430} dx, \\ \text{tg } \theta = \frac{x - 1,873387}{0,765849}, \\ \log_{10} C = \bar{3},2644596; \end{array} \right.$$

$$(5.9) \quad \lambda_4 = \frac{1}{2} \quad \left\{ \begin{array}{l} C \cos^{6,00960} e^{2,31280} dx, \\ \text{tg } \theta = \frac{x - 2,8522}{0,9062}, \\ \log_{10} C = \bar{1},7499974. \end{array} \right.$$

En raison de la valeur élevée de l'effectif, nous avons considéré qu'il serait assez exact de se servir de l'approximation de Gram-Charlier dans le cas $n = 500$, par laquelle on pourrait utiliser les formules de Cornish-Fisher pour trouver les points de la probabilité. Nous avons appliqué certaines vérifications pour établir que la formule de Gram-Charlier pour $a_1(4)$, $n = 500$ est suffisante. Tous les calculs pour $a_1(1)$ sont basés sur les fréquences de Gram-Charlier. Nous donnons les puissances des deux tests pour les deux tailles de l'échantillon dans le tableau IX.

On peut expliquer le tableau en se référant au premier nombre dans la colonne (4) : 0,0648 est la probabilité estimée de trouver $a_1(4) \geq 4,3836$ [voir la colonne (2)] pour $n = 100$ quand la fréquence de l'univers est actuellement (§. 1) avec $\lambda_4 = \frac{1}{2}$.

TABLEAU IX.

Puissance estimée de $a_1(c)$ pour distinguer (5.1) pour $\lambda_4 = \frac{1}{2}$ de la distribution normale ($\lambda_4 = 0$) à quatre niveaux de la probabilité. Échantillons de 100 et 500.

Probabilité, théorie normale. (1)	Point de la probabilité normale.		Puissance de la fréquence (5.1) avec $\lambda_4 = \frac{1}{2}$.	
	$c = 4$ (supérieur). (2)	$c = 1$ (inférieur). (3)	$c = 4$. — (4)	$c = 1$. — (5)
$n = 100$.				
0,01.....	4,3836	0,7482	0,0648	0,0695
0,05.....	3,7744	0,7642	0,1995	0,1979
0,10.....	3,5195	0,7725	0,3163	0,3037
0,20.....	3,3110	0,7824	0,4525	0,4597
$n = 500$.				
0,005.....	3,7062	0,773167	0,1934	0,2067
0,01.....	3,6094	0,775684	0,2920	0,2790
0,05.....	3,3766	0,782482	0,5955	0,5196
0,10.....	3,2695	0,786058	0,7392	0,6509

L'analyse précédente nous a permis de conclure assez fermement que pour des échantillons infiniment grands, $a(4)$ est préférable à $a(1)$. Le tableau IX montre que cette conclusion est assujettie à une restriction importante dans son application générale : pour des échantillons de taille moyenne (c'est-à-dire $n = 100$ à peu près) $a(4)$ ne montre aucune supériorité. Même pour des échantillons plus grands ($n = 500$) la supériorité de $a(4)$ n'est pas aussi marquée qu'on aurait pu le croire. Naturellement ces conclusions sont très expérimentales, parce qu'elles sont basées sur un seul univers différent de l'univers normal et sur des échantillons de deux tailles ($n = 100$ et $n = 500$) seulement. Il est désirable que ces calculs soient étendus.

Cette analyse soulève un problème d'une importance générale : est-ce que les conclusions quant à l'efficacité relative des tests de signification qui sont vraies pour des échantillons infiniment grands — et de telles conclusions sont basées ordinairement sur de tels échantillons — sont nécessairement vraies pour des échantillons de taille moyenne, c'est-à-dire de l'espèce des échantillons qu'on rencontre dans la pratique? Quant à la réponse, il y a une seule certitude : le problème ne sera pas abordé ici.

6. Conclusion : Stade des tests de la normalité. — Nous avons considéré les problèmes suivants :

- (i) l'effet de la non-normalité de l'univers sur les probabilités, basées sur l'hypothèse que l'univers est normal, pour certains tests de signification;
- (ii) le calcul des probabilités des tests de normalité $\alpha(1)$ et $b_1^{\frac{1}{2}}$;
- (iii) l'efficacité relative de divers tests des réseaux $\alpha(c)$ et $g(c)$.

Quant à (i), nous avons montré que les vraies probabilités peuvent différer considérablement des probabilités calculées dans l'hypothèse que l'univers est normal. La seule exception des tests qui ont été examinés est le cas important de l'analyse de la variance. On devrait répéter que la divergence considérable que nous avons trouvée dans le cas du test t s'applique seulement à la probabilité « d'un côté » : la probabilité de $|t| \geq t_0$ n'est pas beaucoup influencée par la non-normalité de l'univers.

Quant à (ii), des tableaux des probabilités sont disponibles pour des échantillons de toutes les tailles plus grandes que 10 pour $\alpha(1)$, qui est un test de la voussure, et pour $b_1^{\frac{1}{2}}$ qui est un test de l'asymétrie. Quant à (iii), nous avons de bonnes raisons pour penser que tous les deux sont efficaces, $\alpha(1)$ aussi efficace que $\alpha(4)$ pour des échantillons de tailles modérées, et $b_1^{\frac{1}{2}}$ [et $\alpha(4)$] le plus efficace pour des échantillons infiniment grands dans un champ étendu des hypothèses alternatives.

Il faut admettre, cependant, que les meilleurs tests de la normalité ne sont pas très sensibles. Ce n'est que très exceptionnellement dans la pratique qu'on peut être assuré de la non-normalité d'un petit échantillon. Il serait peut-être prudent de soupçonner la non-normalité quand la valeur du test est inférieure au point de la probabilité 10 % (au lieu de 5 %, etc.).

Mais même quand nous sommes assez sûrs que l'univers n'est pas normal, il reste la question : « Qu'est-ce qu'on peut faire dans ce cas ? » La réponse, naturellement, dépendra des circonstances. Si le tableau de la probabilité du test sous l'hypothèse normale (par exemple le test t d'un côté) montre un haut degré de signification (par exemple la probabilité $< 0,001$) on peut conclure avec confiance que, quel que soit l'univers, la signification existe encore : il ne nous faut pas connaître

les probabilités exactes mais seulement que l'hypothèse nulle n'est pas soutenable. Pourtant, quand la probabilité est proche de la limite de la décision (par exemple 0,05 à peu près) dans le cas d'un échantillon probablement non normal, il est difficile de savoir que faire. Il est vrai qu'on pourrait supposer une forme plausible pour la distribution non normale de l'univers et calculer la distribution approximative du test, mais un tel procédé n'est pas à recommander dans la pratique parce qu'il prend trop de temps. Une inférence quant à la non normalité n'est pas d'ordinaire un but en lui-même mais une étape dans le jugement de la signification basée sur un test particulier.

Bien qu'il y ait une bibliographie considérable relative à l'effet de la non-normalité de l'univers sur les tests bien connus et que des travaux continuent à paraître, le problème principal, à savoir de déterminer à partir de l'échantillon donné la correction à appliquer aux probabilités normales, n'a pas été résolu, et une solution semble improbable. Il suffira d'établir une limite supérieure de la probabilité pourvu que cette limite soit assez forte. A une certaine époque, l'auteur pensait qu'une limite pourrait être déterminée à partir des théorèmes du type Bienaymé-Tchebychef mais il a abandonné cette recherche en lisant le quatrième chapitre bien connu du travail de M. Fréchet [13].

M. Fréchet a montré, en effet, que, pour les nombreux théorèmes de ce type, la limite supérieure de la probabilité pourrait être réalisée dans chaque cas sous des conditions fixées d'avance. Pour des buts pratiques, la limite de Bienaymé-Tchebychef est beaucoup trop faible. Par exemple, pour un écart de trois fois l'écart-type, Bienaymé-Tchebychef donne une limite supérieure de la probabilité de $1/9$, tandis que par suite de la théorie normale la probabilité est $1/370$. Diverses extensions du théorème original de Bienaymé-Tchebychef améliorent la limite supérieure mais pas au point d'apporter une limite raisonnablement proche de ce qu'on soupçonne devoir être la vraie valeur. De plus, pour se servir d'une théorie du type Bienaymé-Tchebychef, il faut connaître la variance où les moments d'ordre plus élevés qui sont en général inconnus. On pourrait, sans doute, estimer ces moments à partir de l'échantillon lui-même mais, à moins que l'échantillon ne soit très grand (et il ne l'est jamais dans la pratique), les erreurs du sondage des estimations des moments sont considérables. En dehors des difficultés des calculs, à cause de ces erreurs il est inutile, en particulier, de

rapprocher les probabilités au moyen des courbes de Pearson, par exemple.

Pour éviter la difficulté de faire des inférences à partir des échantillons au moyen de divers tests sans connaître la distribution, peut-être la technique la plus utile est-elle la méthode des *permutations*, due à R. A. Fisher (*voir*, par exemple, [12]), et implicite dans beaucoup de ses travaux. Cette méthode a été développée considérablement par E. J. G. Pitman [16] et récemment par G. E. P. Box et S. L. Anderson [11]. Comme vous le savez, la méthode est implicite dans la détermination de la signification du tableau de contingence 2×2 en employant l'algèbre des combinaisons quand les nombres sont petits ou χ^2 (un degré de liberté) quand les nombres sont élevés. Un autre exemple (donné par Fisher) est de déterminer s'il y a une différence significative entre m valeurs du type A et n valeurs du type B (par exemple les tailles des Français et des Anglais). On mesure les moyennes des deux types. Ensuite on mélange les échantillons et l'on calcule les moyennes pour chacun des $\binom{m+n}{m}$ échantillons de m de l'univers borné et on les range par ordre de grandeur. L'ordre de la moyenne dans cette suite détermine la probabilité de l'hypothèse nulle. Si m et n ne sont pas petits, il serait préférable, du point de vue pratique, d'ajuster une courbe (par exemple de Pearson) aux $\binom{m+n}{m}$ moyennes. On remarquera que, selon cette méthode, on peut calculer les moments *exacts*, et les formules sont maniables pour les quatre ou cinq premiers moments. Alternativement, on pourrait tirer au hasard un petit nombre p (10-20 par exemple) des échantillons et juger la signification *approximative* par l'ordre de la valeur de la moyenne du type A par rapport aux p moyennes.

Il est évident qu'on pourrait appliquer la théorie et la pratique de la méthode des permutations dans presque tous les cas où l'on utilise d'habitude des tests de signification sous l'hypothèse que la distribution de l'univers est normale. Actuellement, une grande partie de la théorie est établie. Il nous faut simplement un traité systématique et uniforme sur le sujet. Un tel travail devrait contenir un examen d'ordre philosophique de la notion générale (*voir* en particulier, l'article de G. A. Barnard [9] qui traite du cas du tableau 2×2).

BIBLIOGRAPHIE.

A. — *Articles de R. C. Geary.*

(Voir les bibliographies des autres auteurs dans les articles cités, particulièrement dans [8].)

- [1] *A general expression for the moments of certain symmetrical functions of normal samples* (*Biometrika*, t. 25, 1933, p. 184).
- [2] *The ratio of the mean deviation to the standard deviation as a test of normality* (*Biometrika*, t. 27, 1935, p. 310).
- [3] *The distribution of Student's ratio for non-normal samples* [*J. Roy. Stat. Soc.* (Supplement), t. 3, 1936, p. 178].
- [4] *Moments of the ratio of the mean deviation to the standard deviation for normal samples* (*Biometrika*, t. 28, 1936, p. 295).
- [5] (avec E. S. PEARSON) *Tests of Normality*, 1938.
- [6] (avec J. P. G. WORLLEDGE) *On the computation of universal moments of tests of statistical normality derived from samples drawn at random from a normal universe. Application to the calculation of the seventh moment of b_2* (*Biometrika*, t. 34, 1947, p. 98).
- [7] *The frequency distribution of $b_1^{\frac{1}{2}}$ for samples of all sizes drawn at random from a normal population* (*Biometrika*, t. 34, 1947, p. 68).
- [8] *Testing for normality* (*Biometrika*, t. 34, 1947, p. 209).

B. — *Autres travaux mentionnés dans le texte.*

- [9] G. A. BARNARD, *Significance tests for 2×2 tables* (*Biometrika*, t. 34, 1947, p. 123).
- [10] D. BASU et R. G. LAHA, *On some characteristics of the normal distribution* (*Sankhya*, t. 13, 1954, p. 359).
- [11] G. E. P. BOX et S. L. ANDERSON, *Permutation theory in the derivation of robust criteria and the study of departure from assumption* (*J. Roy. Stat. Soc.*, 1954).
- [12] R. A. FISHER, *Design of Experiments*, 5^e édit., 1949, p. 44.
- [13] M. FRÉCHET, *Généralités sur les probabilités. Variables aléatoires*, chap. IV, 1937.
- [14] R. G. LAHA, *On an extension of Geary's theorem* (*Biometrika*, t. 40, 1953, p. 228).
- [15] E. LUKACS, *A characteristic of the normal distribution* (*Ann. Math. Stat.*, t. 13, 1942, p. 91).
- [16] E. J. G. PITMAN. I. *Significance tests which may be applied to samples from any population* [*J. Roy. Stat. Soc.* (Supplement), t. 4, 1937, p. 119]; II. *The correlation coefficient test* (*ibid.*, t. 4, 1937, p. 225); III. *The analysis of variance test* (*Biometrika*, t. 29, 1938, p. 322).