Statistique et psychologie

Daniel Dugue

STATISTIQUE ET PSYCHOLOGIE

par Daniel DUGUE
Professeur à la Faculté
des Sciences de Caen

---oOo---

Dans ces quelques lignes vont être examinées plusieurs questions qui se trouvent posées par l'application de la Statistique à la Psychologie. Ces questions sont d'ailleurs, les mêmes, quel que soit l'objet pratique de l'application: par exemple la Génétique, la Métallurgie, etc...

Le premier problème concerne le nombre de dimensions de l'être aléatoire soumis à l'Etude. Considérons un individu sur lequel p aptitudes ont été mesurées (avec une précision plus ou moins grande et qui peut varier avec les individus; ce point important sera examiné plus loin). Si cette mesure est répétée pour les N individus examinés les résultats vont constituer un nuage de points dans un espace à p. dimensions. Quelle est la dimension de ce nuage? Ce nombre déterminé, quelle est la signification de chacune des dimensions ? Le problème consiste à ajuster une variété p' (inférieur à p) dimensions à l'ensemble des points obtenus. Inversement on peut se demander si une aptitude n'est pas en fait fonction de plusieurs aptitudes différentes et si l'espace à p dimensions que l'on étudie n'est pas la projection d'un espace à p" (p" > p) dimensions.

Bien entendu la théorie seule peut mettre sur la trace de telles composantes. Les études de factorisation de M. DELAPORTE fondées sur la comparaison des corrélations entre plusieurs grandeurs aident à atteindre ces résultats. De même les études d'arithmétique de variables aléatoires peuvent donner tout au moins théori-

quement des indications. Si l'on connaissait la loi de dispersion théorique de la mesure d'une aptitude on pourrait dans certains cas déterminer si elle est décomposable en une somme ou un produit d'autres variables dont on aurait alors à déterminer le sens expérimental. Ce problème théorique est d'ailleurs loin d'être achevé. De plus il est très souvent discontinu c'est-à-dire si X étant une variable aléatoire de répartition F(x) X est décomposable en une somme ou un produit d'autres variables il n'en est plus de même d'une manière générale pour une variable Y ayant une répartition arbitrairement voisine de F(x). On voit donc à quelles difficultés on se heurte car on ne connaît en général F(x) qu'expérimentalement c'est-à-dire avec une approximation qui peut être très grande mais qui n'est qu'une approximation. La connaissance théorique des répartitions est donc fondamentale pour cette recherche. En particulier il serait bon d'étudier à ce point de vue la répartition des temps de réaction dont la dispersion est très différente d'une répartition normale. Pour employer une image connue ce premier problème se formule ainsi: Quel est le nombre de dimensions de certains vecteurs mentaux (Thurstone : The Vectors of mind). En génétique cela pourrait se traduire ainsi : de combien de gènes dépend l'hérédité de certains caractères (comme la taille). Le rôle de la statistique dans ces deux cas est un rôle de vérification des hypothèses formulées soit par la psychologie soit par la biologie.

Un second problème que l'on peut rencontrer est un problème d'efficacité. Il se pose dans l'exploitation des résultats d'un examen. Nous allons le formuler dans le cas de deux variables. Il reste le même bien que plus compliqué pratiquement si l'on a plus de deux variables.

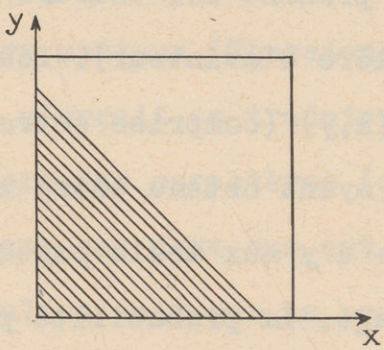
Imaginons que l'on ait deux catégories bien distinctes de tests dans la batterie de sélection du personnel navigant (P.N.)de l'armée de l'air il s'agit de tests écrits et de tests psychomoteurs; dans un examen comme les examens universitaires ou certains concours de recrutement ce serait les épreuves écrites et les épreuves orales ou les épreuves littéraireset les épreuves scientifiques. Appelons x et y les variables aléatoires qui constituent les notes pour chaque groupe de tests: ces deux variables sont évidemment liées. Supposons que l'on ait déterminé la loi de probabilité de l'ensemble soit $\varphi(x, y)\,dx\,dy$ Le résultat de l'examen est lié (où tout au moins on souhaite qu'il soit lié si l'examen est un "bon" examen) à la réussite dans une certaine activité.Pour les batteries de sélection il s'agit de la réussite dans une profession déterminée. Pour les examens universitaires.il pourrait s'agir de la réussite "dans la vie" (ici le terme est plus vague que précédemment). Prenons le cas de la batterie du P.N. l'ensemble des deux variables x, y (tests écrits, tests psychomoteurs) est lié à une variable discrète prenant les valeurs 0 ou 1 (élimination ou réussite dans la carrière d'aviateur). Pour chaque couple x,y il y a donc une quantité f(x,y) (comprise entre 0 et 1) que fixe la proportion des candidats ayant obtenu entre x et x +d x aux tests écrits, entre y et y + d y aux tests psychomoteurs et qui ont réussi à l'école de pilotage. La probabilité pour qu'un candidat ait entre x+ dx aux tests écrits y +dy aux tests psychomoteurs et réussisse à passer son brevet de pilote sera donc $f(x,y)\ \varphi\ (x,y)\,dx\,dy$. Supposons que l'on connaisse à la fois $\varphi(x,y)$ et $f(x,y)$. Cherchons de quelle façon doit se faire l'élimination pour que l'on ait le rendement optimum.

On pourrait imaginer une combinaison linéaire de x et de y et décider que l'on écartera tous ceux pour lesquels la valeur de cette combinaison est inférieure à un nombre donné. Il ne reste plus qu'à fixer les coefficients de cette combinaison. Ce problème assez simple peut être étudié pour un nombre quelconque de variables :

Etant donné $x_1, x_2, \ldots x_p$, résultats de tests dont on connait la liaison interne (coefficients de corrélation les uns avec les autres) et dont on connait la validité (corrélation avec la variable susceptible de prendre les valeurs 0 ou 1 suivant qu'il y a échec ou réussite) il est facile de déterminer $\alpha_1, \alpha_2, \ldots \alpha_p$ tels que la validité de $\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_p x_p$ soit la meilleure. Celà revient à déterminer $\alpha_1, \alpha_2, \ldots \alpha_p$ de telle sorte qu'une forme quadratique en $\alpha_1, \alpha_2, \ldots \alpha_p$ soit maximum.

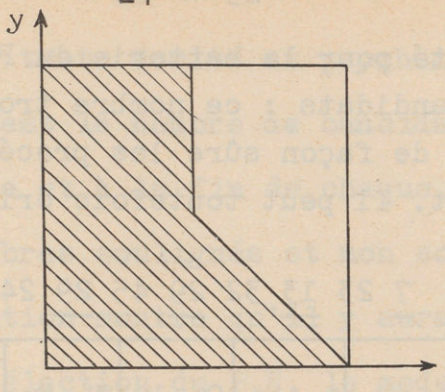
Cette approximation est en général très suffisante quand les tests sont de même nature. Elle ne semble plus l'être quand on a deux groupes de tests manifestement différents (écrits ou psychomoteurs).

Une représentation graphique fera comprendre aisément les différents procédés d'élimination. A l'heure actuelle le procédé employé dans les examens du P.N. est celui de la combinaison linéaire que nous venons de décrire les deux coefficients de x et de y étant égaux: On élimine donc pour $x+y < N$. Ce sera le type I



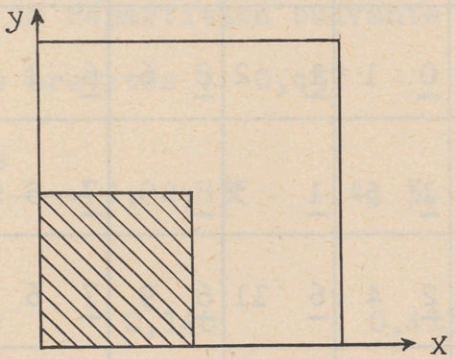
type I

Le procédé des examens universitaires où l'écrit est éliminatoire la somme écrit + Oral l'étant aussi (élimination pour $x < 10$ et $x + y < 20$. (les notes étant "sur 20") entraîne une élimination de la forme suivante (ou type II)



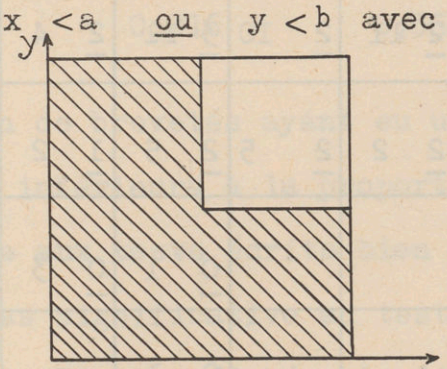
type II

On peut évidemment imaginer d'autres types d'élimination par exemple éliminer tous ceux pour lesquels on a à la fois $x < a$ et $y < b$ ce qui donne le graphique :



type III

ou ceux pour lesquels $x < a$ ou $y < b$ avec le graphique :



type IV

Il est bien évident que si la corrélation entre x et y était voisine de l'unité le nuage de points correspondant à un groupe de candidats se disperserait très peu autour de la bissectrice de l'angle x o y . Dans ces conditions tous ces types d'élimination reviendraient au même.

Mais en général la corrélation entre x et y est assez différente de 1 (pour la batterie du P.N. $\tau_{xy} = 0{,}387$) On peut naturellement imaginer d'autres formes pour le domaine d'élimination. Avant d'exposer la théorie de cette question examinons le

graphique de validité pour la batterie du P.N. Ce graphique porte seulement sur 309 candidats : ce nombre trop faible encore ne permet pas de modifier de façon sûre les procédés d'élimination en vigueur actuellement. Il peut toutefois orienter les recherches futures.

tests écrits	1	2	3	4	5	6	7	8	9	
9					0 1	3 0	5 1	3 1	2 0	13 3
8				1 1	2 2	3 3	5 1	2 2	3 0	16 9
7			0 1	1 2	8 6	6 1	12 2	4 0	1 0	32 12
6		1 1	1 5	1 3	8 9	7 5	3 5	5 2	0 1	26 31
5	0 3	0 7	2 4	6 11	6 9	7 6	3 4	2 2		26 46
4	0 2	0 4	2 11	2 10	3 11	2 4	4 1	1 0	0 1	14 44
3	0 2	0 3	2 2	2 5	2 5	1 2	0 1			7 20
2	0 2	0 2			0 1	0 3		1 0		1 8
1					0 1					0 1
	135 174									

tests psychomoteurs

- En abscisse ont été portées les notes en tests psychomoteurs (de 1 à 9) et en ordonnées les notes aux tests écrits (de 1 à 9) - Dans chaque carré le chiffre de gauche souligné représente le nombre de candidats brevetés après l'école de pilotage ayant eu la

note correspondante aux tests écrits et psychomoteurs. Le chiffre de droite non souligné est le nombre de candidats éliminés. Au-dessus de chaque colonne et à la fin de chaque ligne ont été portées les sommes des nombres soulignés et non soulignés. Un examen rapide de cette répartition montre qu'il y aurait sans doute intérêt à adopter pour la sélection du P.N. le mode d'élimination des examens universitaires (type II), les tests psychomoteurs jouant le rôle de l'écrit et les tests écrits jouant le rôle de l'oral.

On a en effet la répartition suivante :

- Proportion générale de brevetés : 0,437

- Proportion de brevetés ayant eu les notes :⟶ 1, 2, 3 | 4, 5, 6 | 7, 8, 9

aux tests ↓ :

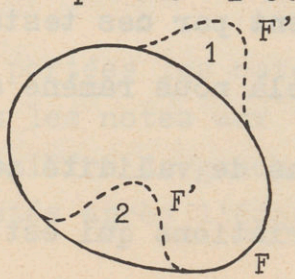
	1, 2, 3	4, 5, 6	7, 8, 9
psychomoteurs :	0,140	0,413	0,700
écrits :	0,216	0,353	0,718

La proportion de brevetés ayant eu une faible note aux tests psychomoteurs est inférieure à la proportion de brevetés ayant eu une faible note aux tests écrits bien qu'à vrai dire cette différence ne soit pas significative au test de Student. Pour les 3notes moyennes et les3 notes élevées la différence entre les deux catégories de tests est encore moins significative. Sous réserve d'un contrôle futur il semble donc qu'il y ait lieu de préférer pour les candidats ayant de faibles notes aux tests psychomoteurs la sélection uniquement par ces tests à la sélection par l'ensemble de la batterie. Celà nous ramène au type II.

Toutes ces questions de validité posent un problème de maximum lié de calcul des variations qui est du même ordre que

celui résolu par M. Jersy NEYMANN au moyen du smooth test dans ses études sur la vérification des hypothèses. Ici intervient le fait que l'on doit prendre un nombre déterminé de candidats qui correspond à la capacité de l'école de pilotage. On a N candidats pilotes parmi lesquels des candidats inconnus dont on sait le nombre B sont brevetables.

On désire en choisir n fixe de telle sorte que la proportion de brevetés soit maximum. Dans ce problème n'intervient donc qu'une sorte d'erreur : le risque d'accepter un candidat qui ne pourra pas obtenir son brevet; on ne tient pas compte du risque d'éliminer un candidat qui pourrait être un bon pilote car l'école ne peut recevoir qu'un nombre limité d'élèves. Cela conduit à tracer dans le plan des x y une courbe fermée F telle que $\iint_F \varphi(x,y)\,dx\,dy$ soit une quantité donnée ici $\frac{n}{N}$ de manière que le nombre de candidats retenus soit n et telle que $\iint_F f(x,y)\,\varphi(x,y)\,dx\,dy$ soit maximum. En employant la méthode des multiplicateurs de Lagrange il faut trouver cette courbe de telle sorte que $\iint_F \left[f(x,y) - \lambda\right]\varphi(x,y)\,dx\,dy$ soit maximum, λ étant pris tel que $\iint_F \varphi(x,y)\,dx\,dy$ soit égal à la valeur donnée. Si $f(x,y)$ est telle que pour tous les points ξ, η intérieurs à la courbe $f(x,y) = \lambda$ $f(\xi,\eta) > \lambda$ et pour les points extérieurs $f(\xi,\eta) < \lambda$ la courbe F sera la courbe $f(x,y) = \lambda$. En effet soit F en trait plein et considérons la modification F' qui consiste à ajouter la partie 1 et à retrancher la partie 2.



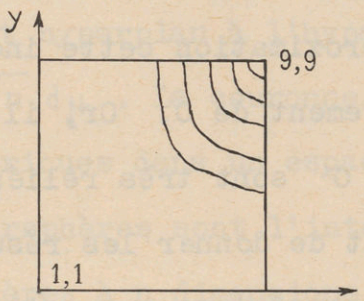
On doit avoir $\iint_1 \varphi(x,y)\,dx\,dy = \iint_2 \varphi(x,y)\,dx\,dy$

D'après l'hypothèse sur $f(x,y)$, $f(x,y) - \lambda$ est positif dans 2 et négatif dans 1, on en déduit :

$$\iint_F \left[f(x,y) - \lambda\right] \varphi(x,y)\,dx\,dy > \iint_{F'} \left[f(x,y) - \lambda\right] \varphi(x,y)\,dx\,dy$$

et F fournit bien la réponse. $f(x,y)$ étant donné la validité des tests écrits et psychomoteurs est une fonction ayant son maximum au point 9.9 et décroîtra quand x ou y fixe, y ou x décroît de 9 à 1.

Les courbes de niveau auront donc la forme suivante et



les domaines s'obtiendront en joignant à ces courbes des portions de côtés de carré. Tout revient donc à estimer avec le maximum de précision les courbes de niveau et par conséquent les deux fonctions théoriques φ et f. On sait par le théorème de Borel Cantelli que l'histogramme des fréquences converge presque certainement vers la surface théorique mais les nombres que nous avons jusqu'à présent sont trop faibles pour permettre même un commencement d'estimation puisque l'on a au plus 17 candidats dans une même case. Ce qui correspond à un écart-type sur les fréquences de l'ordre de 0,125 dans les cas plus précis.

Nous allons maintenant donner un critère permettant d'examiner si une répartition de notes est normale ou non. Il nous a

été suggéré par l'étude sur différents tests de deux séries de variables aléatoires : la moyenne et l'écart-type de groupes de sujets.

A l'heure actuelle l'Institut de Psychotechnique et Biométrie de l'Université d'Alger a dans ses archives des résultats d'examens portant sur 2.140 sujets pour un certain nombre de tests différents. Ces 2.140 personnes ont été groupées en 10 séries différentes et dans chacune de ces dix séries la moyenne et l'écart-type ont été calculés pour chacun des tests. Si les notes étaient distribuées suivant la loi normale de GAUSS-LAPLACE ces deux variables seraient indépendantes et le coefficient de corrélation qui mesure en première approximation cette indépendance ne devrait pas s'écarter significativement de 0. Or, il n'en est rien et dans certains cas μ et σ sont très reliés par une corrélation en général négative. Avant de donner les résultats numériques auxquels nous arrivons nous allons faire une rapide étude théorique du problème. Supposons n résultats expérimentaux indépendants et obéissant à la même loi normale de moyenne m et d'écart-type s. La loi de probabilité de l'ensemble est :

$$\frac{1}{(2\pi)^{\frac{n}{2}} s^n} \; e^{\frac{-\sum (x_i - m)^2}{2 s^2}} \; dx_1 \ldots\ldots dx_n$$

ou $$\frac{1}{(2\pi)^{\frac{n}{2}} s^n} \; e^{-\frac{1}{2 s^2}\left[n \sigma^2 + n (\mu - m)^2\right]} \; dx_1 \ldots\ldots dx_n$$

si l'on pose $\mu = \frac{\sum x_i}{n}$ et $\sigma^2 = \frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2$

La loi de répartition des deux statistiques σ et μ sera :

$$\frac{1}{(2\pi)^{n/2}\,\delta^n}\; e^{-\frac{n\sigma^2}{2\delta^2}}\; e^{-\frac{n(\mu-m)^2}{2\delta^2}}\; dV$$

dV étant l'extension du domaine de l'espace à n dimensions compris entre les quatre surfaces statistiques $\frac{\Sigma x_i}{n} = \mu$ et $\frac{\Sigma x_i}{n} = \mu + d\mu$

d'une part et $\frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2 = \sigma^2$ et $\frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2 = (\sigma + d\sigma)^2$

d'autre part

Ce volume est celui d'un cylindre ayant pour base l'équivalent d'une couronne circulaire de l'hyperplan $\frac{\Sigma x_i}{n} = \mu$ et pour hauteur la distance de cet hyperplan à l'hyperplan $\frac{\Sigma x_i}{n} = \mu + d\mu$ Cette hauteur sera donc $\sqrt{n}\, d\mu$. La couronne sera comprise entre deux hypersphères concentriques dans un espace euclidien à (n-1) dimensions. Ces deux hypersphères sont l'intersection avec le plan $\frac{\Sigma x_i}{n}$ des deux hypersphères à n dimensions.

$\Sigma x_i^2 = n\sigma^2 + n\mu^2$ et $\Sigma x_i^2 = n(\sigma + d\sigma)^2 + n\mu^2$

Le rayon de ces deux hypersphères sera donc $\sqrt{n}\,\sigma$ et $\sqrt{n}(\sigma + d\sigma$

Si $A_{n-1}\, R^{n-1}$ est l'extension d'une hypersphère de rayon R dans un espace à n-1 dimensions l'extension de la couronne sera :

$$(n-1)\, A_{n-1}\, R^{n-2}\, dR$$

et dans le cas particulier $(n-1)\, A_{n-1}\, n^{\frac{n-1}{2}}\, \sigma^{n-2}\, d\sigma$

La loi des deux variables μ et σ sera donc :

$$\frac{(n-1)\, An{-}1\, n^{n/2}}{(2\pi)^{n/2}\;\; \delta^n}\; e^{-\frac{n\sigma^2}{2\delta^2}}\; e^{-\frac{n(\mu-m)^2}{2\delta^2}}\; \sigma^{n-2}\, d\sigma\, d\mu$$

La quantité A_{n-1} mesure de l'ensemble intérieur à une hypersphère

de rayon 1 dans l'espace à (n-1) dimensions se calcule aisément à l'aide des coordonnées polaires (voir Borel - Introduction Géométrique à quelques théories physiques).

On est conduit à deux expressions différentes suivant que n est pair ou impair

Si n est pair $A_{n-1} = \frac{2^{n-1} \pi \frac{n}{2} \left(\frac{n-2}{2}\right)!}{(n-1)!}$

Si n est impair $A_{n-1} = \frac{2}{n-1} \frac{\pi \frac{n-1}{2}}{\left(\frac{n-3}{2}\right)!}$

L'expression à laquelle on aboutit montre bien que quel que soit n les statistiques σ et μ sont indépendantes puisque la loi de l'ensemble est de la forme $f(\sigma)\, g(\mu)\, d\sigma\, d\mu$. Ce résultat est dû à Mr. R.A. FISHER, Mr. GEARY et LU KACZ ont établi la réciproque de ce théorème en montrant que si σ et μ sont indépendants la variable parente est normale. C'est la démonstration de Mr. LU KACZ que je résume ci-dessous. Elle a été publiée dans les "Annals of Mathematical Statistics" (Mars 1942).

Soit $\varphi(u,v)$ la fonction caractéristique de la loi liée de μ et σ^2, $F(x)$ étant la loi de probabilité totale de la variable parente, on a :

$$\varphi(u,v) = \int e^{i(u\mu + v\sigma^2)} \, dF(x1) \ldots dF(xn)$$

L'indépendance de μ et σ et par conséquent de μ et σ^2 entraîne :

$$\varphi(u,v) = \varphi(u,o)\ \varphi(o,v)$$

Donc :

$$\varphi'_v(u,0) = \varphi(u,0)\ \varphi'_v(0,0)$$

$\varphi(u,o) = \theta\left(\frac{u}{n}\right)^n$; θ étant la fonction caractéristique de la

variable parente

$$\varphi'_v(u,0) = i\int \sigma^2 e^{iu\mu} \qquad dF(x_1)\text{-------}\ dF(x_n)$$

$$\varphi'_v(0,0) = i\int \sigma^2 \qquad dF(x_1)\text{-------}\ dF(x_n)$$

Ces dérivations impliquent que l'on fait l'hypothèse de l'existence du second moment de la variable parente, donc du premier moment m, et de l'écart-type s.

$$\text{Or } \sigma^2 = \frac{\sum_1^n x_i^2}{n} - \mu^2 = \frac{n-1}{n^2}\sum_1^n x_i^2 - \frac{1}{n^2}\sum_{i\neq j} x_i x_j$$

$$\text{Donc : } \varphi'_v(u,o) = i\int \left[\frac{n-1}{n^2}\sum_1^n x_i^2 - \frac{1}{n^2}\sum_{i\neq j} x_i x_j\right] e^{i\frac{u}{n}\sum x_i}\, dF(x_1)\ldots\ldots dF(xn)$$

$$\text{Or : } \int \sum x_i^2\ e^{i\frac{u}{n}\sum x_i}\, dF(x_1)\ldots\ldots dF(x_n) = \sum_1^n \int x_i^2\ e^{i\frac{u}{n}(x_i+\ldots+x_n)}\, dF(x_1)\ldots dF(x_n)$$

$$= n\left[\int x^2 e^{i\frac{u}{n}x}\, dF(x)\right]\left[\int e^{i\frac{u}{n}x}\, dF(x)\right]^{n-1} = -n\,\theta''_{u^2}\left(\frac{u}{n}\right)\left[\theta\left(\frac{u}{n}\right)\right]^{n-1}$$

$$\text{et : } \int \sum_{i\neq j} x_i x_j\ e^{i\frac{u}{n}(x_1+\ldots+x_n)}\, dF(x_1)\ldots\ldots dF(x_n) = \sum \int x_i x_j\ e^{i\frac{u}{n}(x_1+\ldots+x_n)}\, dF(x_1)\ldots\ldots dF(x_n)$$

$$= n(n-1)\left[\int x_1 x_2\ e^{i\frac{u}{n}(x_1+x_2)}\, dF(x_1)\, dF(x_2)\right]\left[\theta\left(\frac{u}{n}\right)\right]^{n-2} = -n(n-1)\left[\theta'_u\left(\frac{u}{n}\right)\right]^2\left[\theta\left(\frac{u}{n}\right)\right]^{n-2}$$

$$\text{Donc : } \varphi'_v(u,o) = -i\frac{n}{n}\frac{1}{n}\,\theta''_{u^2}\left(\frac{u}{n}\right)\left[\theta\left(\frac{u}{n}\right)\right]^{n-1} + i\frac{n-1}{n}\left[\theta'_u\left(\frac{u}{n}\right)\right]^2\left[\theta\left(\frac{u}{n}\right)\right]^{n-2}$$

$$\int \sigma^2\, dF(x_1)\ldots\ldots dF(x_n) = \frac{n-1}{n}\,s^2$$

Donc en remplaçant $\frac{u}{n}$ par t : $-\theta\theta''_{t^2} + \theta'^2_t = s^2\theta^2$

ou

$$\frac{d}{dt}\frac{\theta'}{\theta} = -\delta^2 \; ; \; \frac{\theta'}{\theta} = -\delta^2 t + im \; ; \; \log\theta = -\frac{\delta^2}{2} t^2 + imt$$

et θ est bien la fonction caractéristique d'une variable normale. Pour que la variable parente soit normale il est donc nécessaire et suffisant que σ et μ soient indépendants. Cette indépendance pourrait être testée selon les méthodes exposées par Wassily Hoeffding dans son article "A non Parametric test of Independance" (Annals of Mathematical Statistics - Vol XIX N° 4 - Décembre 1948). Nous nous contenterons ici d'utiliser le coefficient de corrélation de Pearson. Si le coefficient de corrélation de σ et μ est significativement différent de 0 , σ et μ ne peuvent être indépendants et on peut en conclure la non normalité de la variable parente. Mais si σ et μ ont un coefficient de corrélation ne différent pas significativement de zéro il peut se faire que σ et μ soient liés et on ne peut conclure que la variable parente est sûrement normale. La nullité significative de $r_{\sigma\mu}$ est donc un critère nécessaire et non suffisant de normalité.

Il est plus simple d'étudier ici la quantité $\frac{\Sigma\sigma_i\mu_i}{\rho} - \frac{\Sigma\sigma_i}{\rho}\frac{\Sigma\mu_i}{\rho} = \theta$ (p étant le nombre de groupes de variables ici 10) de valeur moyenne nulle dans le cas de normalité. Dans cette même publication M. Choudhury donne la loi de répartition de $\Sigma\sigma_i\mu_i$ mais en supposant que m est nul. Nous ne chercherons pas ici de test non paramétrique car il est difficile d'établir la loi de θ mais nous calculerons son écart-type et on appliquera ensuite le théorème de Tchebycheff.

$$\text{Variance de } \theta = E\left[\frac{\Sigma\sigma_i\mu_i}{\rho} - \frac{\Sigma\sigma_j}{\rho}\frac{\Sigma\mu_k}{\rho}\right]^2$$

$$= \frac{1}{\rho^2} E\left(\Sigma\sigma_i\mu_i\right)^2 - \frac{2}{\rho^3} E\left[\Sigma\sigma_i\mu_i \, \Sigma\sigma_j \, \Sigma\mu_k\right] + \frac{1}{\rho^4} E\left[\Sigma\sigma_j \, \Sigma\mu_k\right]^2$$

$$E\left[\Sigma\sigma_i\mu_i\right]^2 = E\left[\Sigma\left(\sigma_i\mu_i\right)^2 + \sum_{j\neq k}\sigma_j\mu_j\sigma_k\mu_k\right]$$

$$= \rho E\left[(\sigma\mu)^2\right] + \rho(\rho-1)E(\sigma_j\mu_j\sigma_k\mu_k) = \rho E(\sigma^2)E(\mu^2) + \rho(\rho-1)\left[E(\sigma)\right]^2\left[E(\mu)\right]^2$$

$$E\left[\Sigma\sigma_i\mu_i\,\Sigma\sigma_j\,\Sigma\mu_k\right] = E\left[\Sigma\sigma_i\mu_i\left[\Sigma\sigma_j\mu_j + \sum_{j\neq k}\sigma_j\mu_k\right]\right]$$

$$= E\left[\left[\Sigma\sigma_i\mu_i\right]^2 + \Sigma\sigma_i\mu_i\sum_{j\neq k}\sigma_j\mu_k\right]$$

$$= E\left[\left(\Sigma\sigma_i\mu_i\right)^2 + \sum_{i\neq k}\sigma_i^2\mu_i\mu_k + \sum_{i\neq j}\sigma_i\sigma_j\mu_i^2 + \sum_{\substack{i\neq j\\ j\neq k\\ k\neq i}}\sigma_i\sigma_j\mu_i\mu_k\right]$$

$$= \rho E(\sigma^2)E(\mu^2) + \rho(\rho-1)\left[E(\sigma)\right]^2\left[E(\mu)\right]^2 + \rho(\rho-1)E(\sigma^2)\left[E(\mu)\right]^2$$

$$+ \rho(\rho-1)\left[E(\sigma)\right]^2E(\mu^2) + \rho(\rho-1)(\rho-2)\left[E(\sigma)\right]^2\left[E(\mu)\right]^2$$

$$E\left[\Sigma\sigma_j\,\Sigma\mu_k\right]^2 = E\left[\left(\Sigma\sigma_j\right)^2\right]E\left[\left(\Sigma\mu_k\right)^2\right]$$

$$= E\left[\Sigma\sigma_i^2 + \sum_{j\neq k}\sigma_j\sigma_k\right]E\left[\Sigma\mu_i^2 + \sum_{j\neq k}\mu_k\right]$$

$$= \left[\rho E(\sigma_i^2) + \rho(\rho-1)\left[E(\sigma)\right]^2\right]\left[\rho E(\mu^2) - \rho(\rho-1)\left[E(\mu)\right]^2\right]$$

On déduit de ces égalités

Variance de $\theta = \dfrac{\rho-1}{\rho^2}$ variance de σ variance de μ

$= \dfrac{\rho-1}{\rho}$ $\dfrac{s^2}{2n}$ $\dfrac{s^2}{n}$

Si n est le nombre de variables dans chacun des ρ groupes.

Ici s est une quantité inconnue et nous serons amené à l'estimer faute de le connaître : nous adopterons la valeur moyenne de p quan-

tités σ soit $\frac{\Sigma \sigma_i}{\rho}$. Cette estimation n'est pas celle de maximum de likelihood, qui pour l'ensemble des données que nous avons serait

$$\sqrt{\frac{\Sigma \sigma_i^2}{\rho} + \frac{\Sigma \mu_i^2}{\rho} - \left(\frac{\Sigma \mu_i}{\rho}\right)^2}$$

Mais l'écart-type de ces deux estimations est le même $\sqrt{\frac{\delta}{2 n \rho}}$ et par conséquent on peut sans inconvénient de perte d'information choisir le plus simple à calculer :

En fait on effectuera donc le quotient :

$$\frac{\frac{\Sigma \sigma_i \mu_i}{\rho} - \frac{\Sigma \sigma_j}{\rho} \frac{\Sigma \mu_k}{\rho}}{\frac{\sqrt{\rho - 1}}{\sqrt{2}\,\rho} \left(\frac{\Sigma \sigma_j}{\rho}\right)^2 \frac{1}{n}}$$

dont on ignore la loi de distribution. Il est bien évident que cette loi ne dépendra ni de s ni de m. Nous serons obligé de nous contenter en première approximation de la majoration assez grossière de Tchebycheff, c'est-à-dire que nous considérons que la variable aléatoire $\frac{\Sigma \sigma_i \mu_i}{\rho} - \frac{\Sigma \sigma_j}{\rho} \frac{\Sigma \mu_k}{\rho}$ s'écartera significativement de 0 si le rapport est supérieur à 4, en valeur absolue.

Voici les résultats obtenus sur les 18 tests que nous avons soumis à ce critère :

.....

Tests.	$r_{\sigma\mu}$	$\frac{\Sigma\sigma_i\mu_i}{\rho} - \frac{\Sigma\sigma_i \Sigma\mu_i}{\rho^2}$	$\frac{\sqrt{\rho-1}}{\sqrt{2}\,\rho}\ \frac{\delta^r}{n}$	
I	- 0,703	- 1,170	0,105	non gaussien
II	- 0,000245	- 0,00014	0,076	
III	+ 0,400	+ 0,133	0,023	non gaussien
IV	- 0,506	- 0,283	0,046	non gaussien
V	- 0,235	- 0,421	0,059	non gaussien
VI	- 0,233	- 1,260	0,657	
VII	- 0,831	- 0,881	0,094	non gaussien
VIII	- 0,392	- 2,540	0,337	non gaussien
IX	- 0,105	- 0,030	0,030	
X	+ 0,286	+ 0,0740	0,014	non gaussien
XI	- 0,484	- 0,024	0,006	non gaussien
XII	+ 0,009	+ 0,003	0,015	
XIII	- 0,459	- 4,636	0,665	non gaussien
XIV	+ 0,558	+16,868	4,669	
XV	- 0,848	-18,024	1,012	non gaussien
XVI	+ 0,258	+13,830	9,570	
XVII	- 0,335	- 0,970	0,097	non gaussien
XVIII	- 0,654	-38,823	16,653	

Les six derniers tests XIII à XVIII sont des tests psychomoteurs les autres étant des tests écrits. Parmi les 11 coefficients de corrélation différant de 0 d'une manière significative on remarque que deux seulement sont positifs et les autres négatifs. En cherchant à interpréter cet écart à zéro des coefficients $r_{\sigma\mu}$ on peut noter que cela signifie que pour les tests dont le $r_{\sigma\mu}$ est négatif σ diminue "en moyenne" quand μ augmente. Donc les échantillons de résultats à forte valeur moyenne sont moins dispersés que dans le cas de la loi de Gauss. Pour ces lois de dispersion les grandes valeurs sont moins probables que dans la

loi normale alors qu'au contraire pour les tests à $\tau_{\sigma\mu}$ positifs μ et σ croissent en même temps "en moyenne" et les grandes valeurs sont plus probables que dans le cas de la loi normale. Le cas du test XV est à cet égard particulièrement significatif. La note du test est égale à 400-t, t étant un certain temps de réaction. Si l'on avait fait l'étude sur la valeur t on aurait donc trouvé $\tau_{\sigma\mu}$ égal à +0,848 : ce qui confirme le fait bien connu que dans la distribution des temps de réaction les grandes valeurs sont beaucoup plus probables que dans le cas de la loi normale (Darmois - Statistiques et Applications p. 104).

Le fait que l'on soit obligé de rejeter l'hypothèse de la distribution normale dans la majorité des cas peut s'expliquer par l'intervention de plusieurs causes.

Tout d'abord admettons que l'aptitude que l'on étudie est une variable aléatoire normale d'une valeur moyenne et d'un écart-type bien déterminé (comme par exemple la taille dans une population homogène). Mais ce que l'on connaît n'est pas l'aptitude elle-même mais le résultat d'une mesure par conséquent un nombre entaché d'une erreur obéissant à la loi de Gauss. L'écart-type de cette seconde loi peut varier d'un individu à l'autre. Le résultat aléatoire que l'on obtient est donc la somme d'une variable aléatoire fixe et d'une variable aléatoire qui dépend de l'individu examiné: la note est donc déterminée au moyen d'une chaîne à deux éléments. La loi élémentaire φ (Z) de distribution de cette note Z est égale à l'intégrale.

$$\frac{1}{2\pi\sigma}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left[\frac{(x-m)^2}{\sigma^2}+\frac{(z-x)^2}{\sigma^2(x)}\right]}\frac{dx}{\sigma(x)}$$

m et σ sont la moyenne et l'écart-type dans la première population et σ (x) l'écart-type de l'erreur avec laquelle on mesure la quantité x. Il n'est donc pas étonnant que le résultat final ne soit pas gaussien.

Il serait intéressant d'étudier la loi φ (Z) dans le cas de certaines formes particulièrement simples par exemple dans le cas linéaire ou σ (X) $= A_X + B$. Quelle que soit la précision du test on ne peut éliminer cette seconde variable. On sait que l'on peut avoir une mesure de cette précision ou plus exactement de cette fidèlité en partageant le test en deux parties et en étudiant la corrélation des notes obtenues par un même individu à la première et à la deuxième partie. Une deuxième cause de non normalité des résultats est le fait que l'aptitude elle-même peut ne pas être normalement distribuée (c'est le cas du temps de réaction dont on a parlé tout à l'heure) comme par exemple le poids d'une population homogène et celà nous ramène à la question du nombre de dimensions d'une aptitude.

Le cas du test XVII montre que certaines distributions non normales peuvent donner un $\tau_{\sigma\mu}$ nul ou ne s'écartant pas de zéro d'une manière significative puisque le quotient de

$$\frac{\Sigma_\rho \sigma_i \mu_i}{\Sigma_\rho \sigma_i \; \Sigma_\rho \mu_i}$$

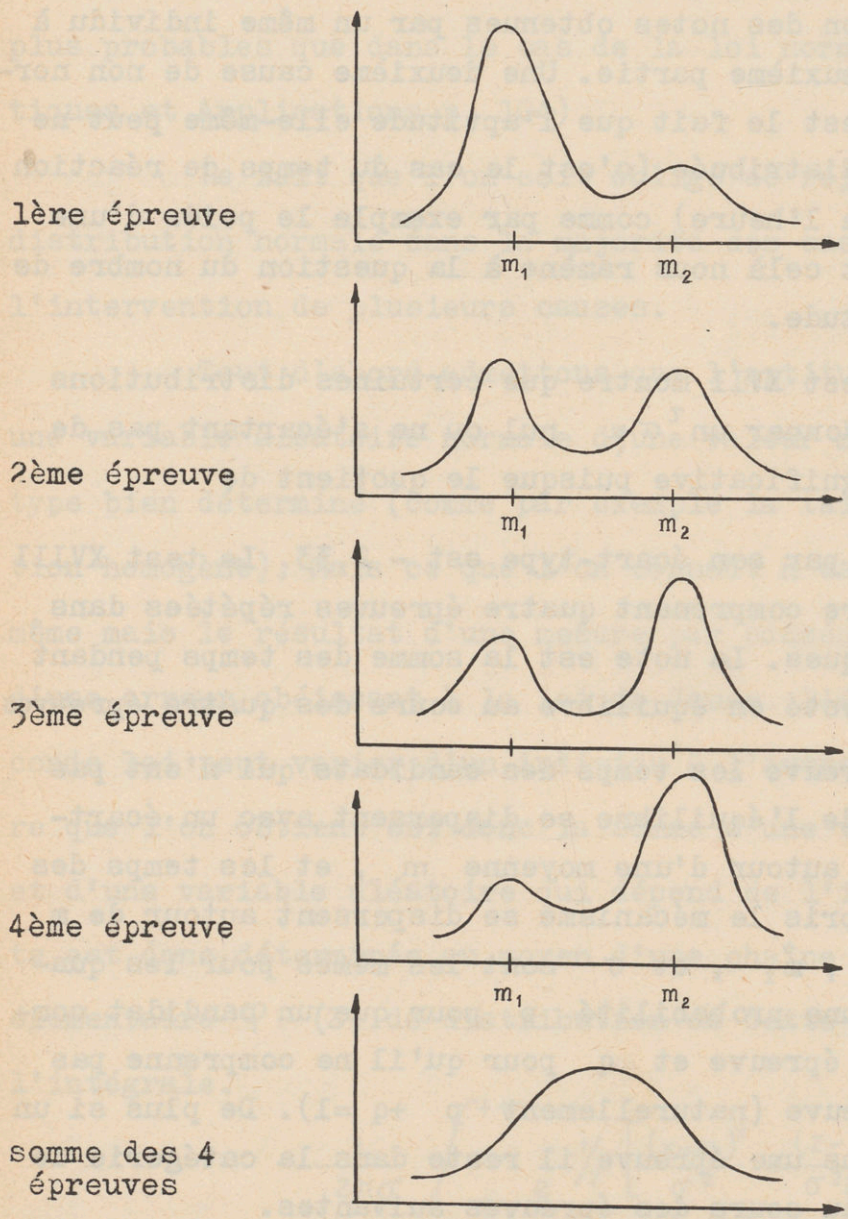
par son écart-type est - 2,33. Le test XVIII est un test d'équilibre comprenant quatre épreuves répétées dans des conditions identiques. La note est la somme des temps pendant lequel le sujet est resté en équilibre au cours des quatre épreuves.

A chaque épreuve les temps des candidats qui n'ont pas compris le mécanisme de l'équilibre se dispersent avec un écart-type assez faible σ autour d'une moyenne m , et les temps des candidats qui ont compris le mécanisme se dispersent autour de m avec le même σ ; m_1 , m_2 , et σ sont les mêmes pour les quatre épreuves. Il y a une probabilité p pour que un candidat comprenne au cours d'une épreuve et q pour qu'il ne comprenne pas au cours de cette épreuve (naturellement $p + q = 1$). De plus si un candidat a compris dans une épreuve il reste dans la catégorie de ceux qui ont compris au cours des épreuves suivantes.

De ces hypothèses on déduit la répartition des probabili-

tés suivantes q^4 ne comprennent au cours d'aucune épreuve
$q^3 p$ comprennent à la 4ème épreuve
$q^2 p$ comprennent à la 3ème épreuve
$q\ p$ comprennent à la 2ème épreuve
p comprennent à la 1ère épreuve

La valeur moyenne de la somme des temps au cours des quatre épreuves est donc $4\ m_1$, pour la 1ère catégorie, $3\ m_1 + m_2$ pour la 2ème catégorie, $2\ m_1 + 2\ m_2$ pour la 3ème catégorie , $m_1 + 3\ m_2$ pour la 4ème et $4\ m_2$ pour la cinquième. L'écart-type est égal à $\sqrt{4}\ \sigma$ pour toutes les catégories



La note du test est donc la somme d'une variable aléatoire d'écart-type 2σ et de valeur moyenne nulle et d'une variable discontinue prenant la valeur $4m$ avec la probabilité q^4, $3m_1 + m_2$ avec la probabilité $q^3 p$, $2m_1 + 2m_2$ avec la probabilité $q^2 p$, $m_1 + 3m_2$ avec la probabilité $q\,p$, et $4m_2$ avec la probabilité p. Sa fonction caractéristique est :

$$e^{-2\sigma^2 t^2}\left[q^4 e^{i4m_1 t} + q^3 p\, e^{i(3m_1+m_2)t} + q^2 p\, e^{i(2m_1+2m_2)t} + q p\, e^{i(m_1+3m_2)t} + p\, e^{i4m_2 t}\right]$$

Chaque épreuve a une répartition très différente de la répartition normale (courbe bimodale) mais on voit que la somme des résultats se distingue difficilement d'une variable de Gauss en employant le critère que nous avons étudié.