4. DE QUELQUES "ADJUVANTS" AU CALCUL STATISTIQUE

par Michel JULLIEN (psycho-sociologue)

-:-:-:-:-:-:-

On rencontre de plus en plus de chercheurs en Sciences Humaines conscients de la nécessité de recourir au calcul statistique dès qu'ils veulent démonter une structure de relations ou établir un pronostic : la statistique est la seule voie qui puisse permettre aux Sciences du "fluctuant" (Sciences de l'homme, biologie et toutes les disciplines où les facteurs sont pratiquement impossibles à maîtriser ou à dénombrer) de se prétendre sciences expérimentales ; elle est pour ces sciences ce que sont les ballons, éprouvettes et balances pour la chimie : elle ne remplace pas le chercheur et la construction théorique, mais en est le support nécessaire. Les sciences de la matière ont pu progresser grâce à de multiples petites trouvailles techniques (appareils ou méthodes), fort humbles, mais sans lesquelles les théories n'auraient pu sortir du stade des hypothèses - A l'inverse, dans les sciences humaines, la théorie précède et déborde trop souvent les moyens méthodologiques de "l'asseoir" ; ces moyens méthodologiques existent pourtant et sont parfois connus des chercheurs ; mais ces chercheurs sont facilement rebutés par l'aspect ingrat des calculs nécessaires : ceux-ci procèdent souvent par itérations successives, et leurs résultats semblent souvent peu "payants" en regard du temps et de la tension intellectuelle qui leur sont consacrés.

C'est dans l'espoir de persuader mes collègues qu'il n'est pas tellement redoutable de se lancer dans un véritable traitement statistique de l'information, que je rappelle ou propose ci-après quelques "trucs".

1°) Utilisation d'abaques.

A l'inverse des "tables", ils ont le mérite de permettre des estimations rapides et d'indiquer si le calcul vaut d'être effectué en détail. Par ailleurs ils peuvent être confiés, après entraînement, à un personnel subalterne d'un niveau B.E.P.C., ce qui permet d'envisager d'importantes "campagnes de calcul".

A. Pour ma part je trouve très utile ceux qui se trouvent dans le "Morice et Chartier- Méthodes statistiques Tome II - INSEE 1954" - On y trouve des abaques sur :

1°) Loi de Poisson (fluctuations d'échantillonnage de fréquences faibles).

2°) Loi binomiale (Intervalle de confiance d'une fréquence).

Ces deux abaques éviteraient à bien des chercheurs de se lancer dans de profonds raisonnements - sur des différences entre "pourcentages -. Ces raisonnements font un bel effet mais hélas il arrive que la différence constatée n'ait aucune signification, eu égard à l'échantillon utilisé.

3°) Test d'homogénéité de la variance (préalable à l'étude de la différence entre moyennes).

4°) Analyse de séquences (probabilité du nombre de séquences en fonction des effectifs totaux et partiels).

5°) Somme des carrés des différences successives (utile pour l'étude de la variation d'une moyenne dans une population à variance constante, sur petits échantillons).

B. Par ailleurs j'avais signalé dans mon rapport annuel 1961, l'existence d'abaques publiés par le Bulletin du CERP (tome X 1961 n° 1 - "De quelques principes de nomographie").

Des abaques de grand modèle, plus pratiques, sont en vente au CERP, 13, rue Paul Chautard Paris XVe - D'une précision suffisante pour éviter la majorité des calculs, ils sont fondés sur l'épreuve du χ^2 .

Ils indiquent :

n° 1 et 2 : N (effectif de l'échantillon), au-dessus duquel la différence entre deux distributions est significative ou non [tableau à (2 x 2) cases].

n° 3 et 4 - Valeur numérique de χ^2 pour tableau à (2 x 2) cases - Valeur absolue de Phi (un des coefficients de contingence).

n° 5 : comme 1, pour tableau à (2 x c) cases -

n° 6 : comme 1, pour tableau à (1 lignes x c colonnes) cases.
n° 7 : valeur numérique de χ^2: comparaison fréquence expérimentale à fréquence théorique 0,50. Comparaison de deux fréquences dans des échantillons appariés.
n° 8 : comme 7, mais pour fréquence théorique quelconque.

2°) Utilisation des machines à calculer.

A travers les nombreux contacts que j'ai eus à ce sujet, j'ai constaté que nombreux sont ceux qui ne savent pas exploiter pleinement les possibilités de leur machine, et par suite hésitent à entreprendre des calculs intéressants - Les audacieux se bornent aux éternels calculs de pourcentage (en ignorant parfois l'usage bien commode de l'inverse de l'effectif total comme multiplicateur constant !).

Dès que l'on dépasse le calcul de ses propres dépenses ménagères et le contrôle de son compte bancaire (utiles mais quand même accessoires dans une activité de recherche), l'utilisation d'une machine doit être programmée, "un peu" comme pour un ordinateur (toute proportion gardée, bien entendu !). Un programme de calcul consiste à analyser les opérations successives prévues, déterminer les éventualités et leurs probabilités d'apparition, standardiser les séquences en fonction des possibilités de la machine - Les notices accompagnant les machines ne donnent bien souvent que des programmes de types commerciaux (tenue de stock, comptabilité, calculs d'intérêt), leurs auteurs supposant peut-être que les chercheurs scientifiques sont normalement dotés de moyens de calculs plus puissants (!).

En pratique il est préférable de concevoir un programme pour chaque campagne de calcul : certaines particularités du calcul envisagé permettent parfois de "sauter" quelques maillons de la chaîne, et de gagner du temps. Un programme bien fait doit permettre...

1°) de rendre la succession des opérations automatique au point que l'opérateur n'a plus à réfléchir sur chaque phase : d'où fatigue moindre pour le chercheur et possibilité de confier les calculs à un personnel moins qualifié...

2°) de gagner quelques secondes sur chaque opération élémentaire : ceci devient très avantageux dès que le nombre d'opérations est de l'ordre de 10 000, ce qui arrive facilement (par exemple : 12 000 opérations pour l'analyse des contingences dans 28 tableaux à environ 7 lignes x 8 colonnes).

Je propose ci-après, à titre d'exemples, quelques programmes conçus pour des machines que je connais : imprimantes à mémoire (c'est le minimum admissible pour des calculs de recherche), (Lagomarsino Division, Olivetti Divisumma 24), et "à mémoire" et double totalisateur (c'est ce que nous devrions normalement avoir), (Olivetti Tetractys) - Il existe d'autres machines non imprimantes intéressantes (du type Facit à trois compteurs, transfert et mémoire - ou Fridden à clavier multiple, extraction de racines carrées), mais je n'ai pas eu l'occasion de travailler beaucoup sur elles ; les principes resteraient probablement les mêmes.

Symboles utilisés (ordre d'appuyer sur la touche correspondante)

C - inscription au clavier de ...

IM -"- en mémoire (réalisée à l'enfoncement d'une autre touche)

M sortie de mémoire (------- id. -------)

X entrée comme multiplicande

= -"- -"- multiplicateur

= Neg. -"- -"- -"- négatif

D -"- -"- dividende

÷ -"- -"- diviseur (le quotient reste en mémoire automatiquement)

T total sur totalisateur ordinaire (noir ou vert selon les diverses Tetractys)

S total partiel du totalisateur ordinaire

* total sur totalisateur annexe (bleu ou noir - respectivement sur Tetractys)

◇ total partiel annexe

A / N.A. } les produits sont ou non donnés automatiquement

0 ⟶ A transfert automatique du totalisateur ordinaire au totalisateur annexe (Tetractys)

TR transfert automatique d'un produit comme multiplicande de l'opération suivante

. / .. / ... } nombres de 0 qu'il y a lieu de placer après le dernier chiffre inscrit.

Une unité d'action sur la machine est signalée par un tiret.

1er programme - Calcul d'une moyenne - sur données groupées - par changement d'origine et d'unité.

X_o = nouvelle origine

i = intervalle de classe

$X' = \frac{X - X_o}{i}$ (nouvelle variable)

N = effectif total de la distribution

Formule $m = \frac{NXo + i\ (\Sigma\ x')}{N}$

Machine (sur N.A.)

- C x' 1 + ou -
- C x' 2 + ou -

.....

C.....x' k..... pour les k classes

- IM T
- C i..... (2 zéros de plus que le nombre de chiffres de N) si T < 0 = Neg. si T > 0 =
- C N..... (autant de zéros que i) x
- C X_o =
- C N ÷

- Lecture de m à deux décimales (selon le nombre de zéros attribués à i et N)
- T (sortie du reste de la dernière division - machine à zéro).

2e programme - Calcul de l'Ecart-type (σ) - sur données groupées - par changement d'origine et d'unité.

Formule : $\sigma = \sqrt{\frac{i^2 \left(\Sigma\ x'^2 - \frac{(\Sigma\ x')^2}{N}\right)}{N - 1}}$

Machine (sur N.A.)

- C $\Sigma x'$ x
- C $\Sigma x'$ =
- C N ÷
- T (sortie du reste - totalisateur purgé)
- C x'^2_1 + } les x'^2 sont en général facile à obtenir : (η x') x'
- C x'^2_2 +

.....

C..... x'^2_k (pour k classes) +

- M - = (rappel du précédent quotient)

- IM m
- C i^2 =
- C N-1 ÷

▫ Lecture Variance (σ^2)

- T (sortie du reste-purge)

▫ La lecture de σ peut se faire au moyen de tables de carrés, ou par le moyen de tables d'extraction par coefficient de division, en poursuivant sans s'arrêter le calcul ainsi

- C Nombre le plus proche de σ^2 +
  (lu sur la table)
- M + (rappel du dernier quotient)
- C Coefficient correspondant ÷
  lu sur la table)

▫ Lecture σ

- T (purge du dernier reste)

3e programme - Calcul du t de student (homogénéité des variances admises)

Ce programme est conçu pour Tetractys (2 totalisateurs). Sur machines plus simples il y aurait lieu de "lire" quelques résultats pour les réintroduire ultérieurement par clavier.

Formule : $|t| = \dfrac{(M_A - M_B)}{\sqrt{\sigma^2 \left(\frac{1}{N_A} + \frac{1}{N_B}\right)}}$

Machine (sur A)

- C 1..... D
- C N_A ÷
- M + (totalisateur annexe)
- C 1..... D
- C N_B ÷
- M + (totalisateur annexe)
- IM *

} peut-être effectué plus rapidement au moyen de tables des inverses (surtout pour machines plus simples)

(Machine sur N.A.) (attention !)

- C σ^2 =
- S (carré du dénominateur de la formule)

□ Lecture de S pour extraction de racines par tables ou calcul immédiat par coefficient de division ainsi :

- C Nombre le plus proche de S +
- C Coefficient correspondant ÷
- T (reste purgé)

□ Lecture et conservation du quotient

- C m (la plus élevée) +
- C m (la plus faible) -
- C quotient précédent ou $\sqrt{S}$ ÷

□ Lecture de |t|

T (reste purgé)

Si on a le courage de s'y lancer, on s'aperçoit vite que ces programmes sont simples. Bien entendu il faut avoir compris la signification du calcul entrepris (ce qui n'est pas toujours le cas chez les anciens littéraires que nous sommes pour la plupart).

*

* *

J'ai mis au point d'autres programmes (limites de confiance de fréquences, coefficients de contingence, analyse de variance) pour Tetractys. Je crois inutile de les publier dans le détail avant de savoir si cela intéresse quelqu'un, mais serais heureux de correspondre éventuellement avec des collègues à ce sujet. Je me bornerai, en attendant, à en exposer les principes:

1°) <u>sur le calcul des fréquences</u>.

Rappelons simplement ici l'utilité de se servir de l'inverse de l'effectif total comme multiplicateur constant de chacun des effectifs partiels. Dans certaines machines (Tetractys en particulier), tout quotient reste en mémoire et peut donc intervenir directement comme multiplicateur dans la suite des autres opérations.

De plus il est intéressant de reporter dans le totalisateur annexe toutes les fréquences calculées, et contrôler ainsi immédiatement la suffisante précision des calculs : si le total s'écarte exagérément de 1 (ou de 100 si l'on tient aux %), il y aura lieu de faire un calcul plus précis de l'inverse (source majeure de la divergence).

Ainsi l'on calculera $\frac{1}{N}$... - ce numérateur affecté d'un nombre suffisant de 0 pour la précision désirée (il est bon d'obtenir au moins 3 décimales significatives à l'inverse), on "purgera" le reste, et l'on poursuivra sans désemparer la multiplication de chacun des effectifs partiels.

A ce propos, je me permets de placer ici une remarque qui me "démange" depuis que je lis des rapports de sciences humaines : combien de décimales doit-on affecter à un "pourcentage" quelconque (ce qui revient à choisir entre "pour cent", "pour mille", ou "pour dix-mille", etc...) ?

Deux problèmes se posent : soit que l'on veuille se servir des % dans une argumentation (donc prouver quelque chose à partir de différences constatées), soit que l'on désire seulement présenter des résultats.

a) Dans le premier cas l'élément dominant est l'intervalle de confiance des fréquences calculées (qui, soit dit en passant, ne dépend que partiellement de l'importance de l'échantillon, mais ceci est un autre problème). Les seules preuves légitimes ne peuvent être fournies que par un calcul adéquat de limites de confiance ou par un "Khi Carré" ; pour le reste peu importe le nombre de décimales choisi.

b) Dans le deuxième cas le but est de présenter sous une forme simplifiée (tableau de fréquences), des résultats complexes et ceci sans perdre "de l'information" mais aussi sans paraître fournir plus d'information que l'on en a trouvée. Rappelons que 12,3 % ou 123 $^{o}/_{oo}$ sont strictement équivalents et que c'est par une prétention mensongère au sérieux et à l'honnêteté scientifique que l'on adopte souvent la première écriture.

Si N = 900, il est légitime de présenter ses résultats au millième près (0,123 ou 123 $^{o}/_{oo}$), car ce millième est "presque" un individu réel. Il serait dommage par contre d'en rester au centième (0,12 ou 12 %), car ce centième représente presque 10 individus.

Si N = 150, le millième est moins légitime car il représente une plus petite fraction d'individu, mais se contenter de "pour cent" sans décimales serait faire disparaître des individus peut-être significatifs (1 % représentant plus d'un individu).

Si N - 90 présenter des résultats sous la forme 12,5 %, tendrait à faire croire au lecteur que l'on est capable de distinguer entre 12,3 % et 12,4 %, ce qui n'est nullement en accord avec l'observation brute : le ",3" et le ",4" ne représentent même pas un dixième d'individu

Si N = 8 on devra heroïquement abandonner le sacro-saint pourcentage, et s'exprimer en "pour dix" ou à la rigueur en 10, 20, 30... 70 %, en tout cas ne pas donner négligemment au lecteur l'impression d'un potentiel bien plus grand d'information.

En définitive la meilleure écriture, dans la présentation des tableaux de fréquences semble être le rapport brut (entre 0 et 1) sous la forme 0,2 (de N = 1 à 9) ou 0,25 (de N = 10 à 99) ou 0,255 (de N = 100 à 999) (ou mieux l'écriture anglo-saxonne .2 .25 .255 etc...).

0,2 ou .2 signifiera automatiquement que l'échantillon de référence n'est pas supérieur à 10 (et donnera par la même la possibilité de faire une estimation des limites de confiance).

0,20 indiquera une fréquence équivalente mais sur échantillon de 10 à 99 individus.

0,02 ou .02 indiquera le dixième de la fréquence précédente, mais sur échantillon équivalent.

0,020 indiquera la même fréquence sur échantillon de 100 à 999.

0,001 représentera le 1 pour mille précisions habituellement réservées aux démographes (sans justification d'ailleurs car il peut leur arriver de travailler sur des villages de moins de 100 habitants, ou plus souvent sur des villes de 9 000 habitants - auquel car le "pour dix mille" semble plus "informant" au niveau de la présentation des résultats bien entendu).

2°) Limites de confiance sur une fréquence.

Les procédés sont variés et appellent la réalisation des programmes adaptés. Rappelons aussi qu'il est toujours préférable d'utiliser des abaques à échelle suffisamment grande.

Cependant pour le cas où il est souhaitable de faire un calcul plus précis et si l'on utilise la formule $f \pm \sqrt{\frac{pq}{N}}$ voici quelques indications :

a) Calculer d'abord l'expression sous radical en choisissant un nombre de décimales adéquat (attention à l'extraction finale de la racine qui réduira de moitié ce nombre).

- Produit p q passé en mémoire ou sur totalisateur annexe.
- Division par N.

Si l'on extrait la racine par la méthode des coefficients de division, on conservera automatiquement le dernier quotient en mémoire (après purge du reste), on entrera f puis on fera M - pour avoir la limite inférieure et M + M + (2 fois) pour avoir la limite supérieure.

3°) Analyse de Variance.

Il y aura évidemment rupture entre les calculs des "sommes de carrés" et l'analyse elle-même. Les premières devront être prélevées au cours du programme "calcul de l'Ecart-type" (après la huitième "palier", non compris les itérations des x'^2_k).

- on cherchera évidemment à obtenir le F de Snédécor aussi directement que possible.
- on sera obligé (du moins sur une Tetractys) de noter à part le dénominateur de F (après calcul de la somme des sommes de carrés et division par N-k) (k = nombre de groupes).
- on calculera ensuite le numérateur (variance des moyennes), (que l'on mettra en mémoire ou sur totalisateur annexe) selon l'ordre suivant : différences, pondérations, somme des différences, pondérés, quotient par (k-1).
- on "entrera" à nouveau le dénominateur pour obtenir F.

La principale difficulté est de conserver une attitude constante en ce qui concerne le nombre de décimales, lequel devra être impérativement fixé à l'avance, et prévu de façon à palier le manque de souplesse de ces machines à ce sujet. Si ces précautions sont prises la lecture de F se fera directement.

Les coefficients de contingence sont trop nombreux pour être abordés ici : j'ai fait un programme pour le coefficient de liaison Tetrachorique (Lt) et pour Phi (à partir du Khi Carré) dont les utilisations ne sont pas universelles. Chaque type de problème demande donc à être étudié.

L'important, dans la réalisation d'un programme, est de bien étudier les éventualités possibles (en particulier le signe de certains résultats) et... de l'écrire. C'est à ce stade que l'on s'aperçoit souvent des contradictions ou que l'on découvre de nouveaux cheminements plus courts.

Nota. Au dernier SICOB (Salon du Matériel de Bureau) ont été présentés de petits ordinateurs pour la recherche, lents mais souples et ne nécessitant pas l'intervention d'un programmeur.
Un chercheur en économie de l'ORSTOM s'est déjà intéressé à l'IBM 1 130.
J'ai découvert au SICOB l'EMD 800 (produit par Dassault), qui à partir de 60 000 F est déjà assez complet pour les calculs effectués dans les centres Outre-Mer. Leur usage nous permettrait de varier et tester nos hypothèses sans être rebutés par la longueur des calculs : en quelques dizaines de secondes par exemple l'EMD 800 fournit un "t" de Student.