

JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

P. THIONET

**Sur un aspect peu connu de l'analyse des données statistiques
: les modèles de Chernoff**

Journal de la société statistique de Paris, tome 108 (1967), p. 193-204

http://www.numdam.org/item?id=JSFS_1967__108__193_0

© Société de statistique de Paris, 1967, tous droits réservés.

L'accès aux archives de la revue « Journal de la société statistique de Paris » (<http://publications-sfds.math.cnrs.fr/index.php/J-SFdS>) 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/>

IV

SUR UN ASPECT PEU CONNU DE L'ANALYSE DES DONNÉES STATISTIQUES : LES MODÈLES DE CHERNOFF

I. INTRODUCTION

Le statisticien recueille les données en s'efforçant d'éliminer son « équation personnelle »; il doit nous fournir les faits tels qu'ils sont et non pas tels qu'ils devraient être. Mais les données observées (avec toutes les précautions possibles) sont équivalentes à un écheveau de laine qu'on a bien du mal ensuite à démêler. De nouvelles méthodes (encore très vagues), nommées « Item Analysis » en anglais, concernent la présentation de données « élaborées » au lieu de données brutes. Les économistes sont habitués aux traitements qu'on fait subir aux séries temporelles pour en extraire des composantes : tendance (*trend*), cycles, variations saisonnières, résidus. Dans un autre ordre d'idées, l'ajustement sur une distribution statistique de quelque loi de répartition théorique, ou l'analyse d'une distribution statistique considérée comme un mélange en proportions convenables de 2 ou 3 lois théoriques, sont des méthodes classiques de traitement des données.

La méthode que nous allons étudier est beaucoup moins connue et même paraîtra *hérétique* à première vue au statisticien. Cependant quand on a l'expérience d'enquêtes où l'enquêté donne son opinion ou dit ce qu'il sait (sans longue réflexion et sans se reporter à des documents), on n'ignore pas que les réponses (supposées sincères) sont faites toujours un peu *au hasard*; il s'ensuit que certaines statistiques brutes sont certainement aberrantes. On peut se proposer de les corriger.

1^o La première idée est de les confronter avec d'autres données, plus sérieuses. Par exemple la proportion des personnes déclarant avoir leur baccalauréat lors du dernier recensement français paraît à première vue très basse; et on peut la *tester* en demandant au B. U. S. ses propres statistiques. Mais il s'agit de comparer alors l'état de la population des bacheliers à la statistique des *naissances* des bacheliers (B. U. S.) sans avoir la statistique

de *décès* (en négligeant les migrations) : il faudra admettre que la mortalité est la même chez les bacheliers et ailleurs, ce qui n'est pas tout à fait exact : quand ce ne serait que par le fait que les 2 sexes sont très inégalement représentés dans la population d'une part, la sous-population d'autre part (la mortalité masculine est plus forte).

Bien entendu la cause « prestige » doit inciter les gens à se déclarer bacheliers abusivement ; mais la fatigue et la négligence agissent en sens contraire et sans doute plus souvent. Ce travail de réconciliation de statistiques d'origine et de qualité diverses est bien connu des comptables nationaux.

2° Passons au cas où les données entre lesquelles on peut opérer un rapprochement ne sont pas assez comparables. On peut imaginer quelque mécanisme perturbateur, et essayer de voir s'il suffit à expliquer les écarts. La statistique classique connaît très bien ce type d'opération : c'est le *problème des 2 échantillons* par exemple qui est classique dans la théorie des tests, quand on compare des nombres $x_1 x_2 \dots x_n$ et $y_1 y_2 \dots y_n$ dans le cadre d'hypothèses telles que :

$$(H) x_i : \mathcal{N}(\mu, \sigma^2); y_j : \mathcal{N}(\mu', \sigma^2)$$

$$(H^o) x_i, y_j : \mathcal{N}(\mu, \sigma^2)$$

D'ailleurs il s'agit non seulement de choisir entre H^o et H , mais de tirer des données quelque estimation de μ (et μ').

3° A présent supposons qu'on ait affaire à des données manifestement entachées d'erreurs. Recommencer autrement l'enquête statistique aboutirait à d'autres erreurs — ou bien même est souvent exclu (impossibilité ou coût excessif). Les deux solutions orthodoxes : envoyer les questionnaires au pilon — publier les résultats bruts en les accompagnant d'un commentaire destructeur, — peuvent paraître l'une et l'autre détestables. Une idée a fait son chemin : essayer de reconstituer des résultats ayant des chances d'être plus près de la vérité, en imaginant un mécanisme perturbateur, grossier, schématisant les causes d'erreur.

Le plus simple et le plus courant de ces mécanismes supposés est une loi d'erreurs symétriques, d'où l'on conclut que « les erreurs se compensent » et on ne change rien aux résultats du dépouillement, moyennes, proportions pour cent (sinon en les arrondissant pour démystifier le lecteur qui croirait à leur précision). Cette façon de faire est souvent une solution de paresse.

A l'opposé, certain praticien n'hésite pas à torturer les chiffres et à corriger tel ou tel résultat qui lui paraît faux, au risque de ne plus savoir discerner ce qui est *impression subjective*, jugement intéressé, désirs (ou craintes) prises pour des réalités. Il incombe au mathématicien de formaliser le mieux possible les « coups de pouce » à donner aux résultats bruts. Dans ce cadre de pensée entrent les probabilités subjectives et la méthode d'estimation de BAYES.

Ayant posé *a priori* qu'on s'attend à tels résultats avec telle marge d'erreur probable, disposant ensuite des résultats d'une enquête, résultats affectés d'une erreur à leur tour, la *méthode de Bayes* fournit en fait un moyen de redresser les résultats de l'enquête pour tenir compte de l'idée qu'on s'en faisait à l'avance (bien qu'en principe il s'agisse plutôt de rectifier l'idée *a priori* pour tenir compte des faits observés). Elle n'est cependant pas l'objet de cet article.

Nous avons découvert que certains spécialistes de psychométrie s'inspiraient d'une idée tout à fait différente pour redresser les résultats statistiques d'un lot de réponses.

A vrai dire, leur but *final* n'est pas du tout de produire des statistiques mais de *mettre des notes* (scores) à des sujets subissant des sortes d'examen. Pareille tâche n'est pas l'apanage des membres de l'enseignement : Des opérations de remaniement des notes individuelles, pour tenir compte de l'ensemble des notes (de leurs moyennes ou de leurs dispersions trop inégales d'un examinateur à l'autre) sont pratiquées aussi dans l'administration. Mais il s'agit ici d'opérations moins banales; nous les avons découvertes dans deux articles de CHERNOFF, dont d'ailleurs le plus ancien est entièrement repris dans le second. [1] [2]

CHERNOFF dit qu'il étudie une technique proposée par HAMILTON (1950) et discutée par LYERLY (1951); mais il est encore question de HORST (1933) discuté par GULLIKSEN (1950), et aussi de CALANDRA (1941) [Articles parus généralement dans *Psychometrika*].

Remarque générale :

Dans l'emploi des modèles qui suivent, on suppose qu'on a affaire à un nombre élevé de personnes (les sujets) auxquelles des questions sont posées. La probabilité d'obtenir tel ou tel résultat et la proportion des sujets se trouvant dans ce cas sont des nombres bien distincts mais qu'on ne peut supposer très différents l'un de l'autre, en vertu de la loi des grands nombres. En confondant *probabilités* et *proportions*, on se trompe certainement, mais sans doute ne se trompe-t-on pas beaucoup. Pareille assimilation est nécessaire du fait qu'on ne peut mesurer les probabilités, alors que données numériques fournissent des proportions.

II. LE MODÈLE D'ERREUR ET SES IMPLICATIONS STATISTIQUES

1° *Problème le plus simple : Une seule question :*

Soit *une question du type* : Napoléon est-il mort en 1821 ou en 1824?

Soit λ la proportion des personnes bien informées.

Admettons que les autres répondent au *hasard* c'est-à-dire que 1821 ait une chance sur deux d'être choisi. La proportion des réponses correctes sera :

$$p = \lambda + \frac{1 - \lambda}{2} = \frac{1}{2} (1 + \lambda)$$

Si la question est plus ouverte (comme aux jeux radiophoniques) : «indiquez laquelle des dates suivantes est correcte : 1815, 1820, 1821, 1829, 1825, 1830», on admet que la proportion des réponses correctes sera

$$p = \lambda + \frac{1 - \lambda}{r} = \frac{1}{r} + \frac{r - 1}{r} \lambda$$

[1] Herman CHERNOFF, *A new approach to the evaluation of multiple choice questionnaires*. Bull. I. I. S. (Congrès de Paris, 1961, paper n° 86, 6 pages).

[2] Hermann CHERNOFF, *The scoring of multiple choice questionnaires*. A. M. S., 33-2, June 1962, pp. 375-393.

r désignant le nombre des options supposées également probables pour l'ignorant. Il s'ensuit (aux écarts aléatoires près) qu'on peut estimer λ connaissant p par les formules

$$\lambda = 2p - 1 \quad \text{si } r = 2.$$

et plus généralement

$$\lambda = \frac{rp - 1}{r - 1}$$

sous réserve qu'on ait

$$p > 1/r$$

Le tableau suivant permet de passer des résultats statistiques bruts p aux résultats rectifiés λ (compte tenu du modèle d'erreur).

Valeurs de λ : proportion de personnes bien informées								
Valeur de p (réponses exactes)	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
$r = 2$	—	—	—	—	0,20	0,40	0,60	0,80
3	—	—	0,10	0,25	0,40	0,55	0,70	0,85
4	—	0,07	0,20	0,33	0,47	0,60	0,73	0,87
5	—	0,125	0,250	0,375	0,500	0,625	0,750	0,875
6	0,04	0,16	0,28	0,40	0,52	0,64	0,76	0,88

2° Cas d'un couple de questions

Supposons que les sujets aient à répondre à 2 questions, dont chacune est du type : r réponses prévues (également probables en cas d'ignorance).

On désigne par p_{11} la proportion des sujets répondant correctement aux 2 questions

p_{10} la proportion des sujets répondant correctement à la 1^{re} question seule

p_{01} la proportion des sujets répondant correctement à la 2^e question seule

enfin : $p_{00} = 1 - p_{11} - p_{10} - p_{01}$ la proportion des sujets ne répondant correctement à aucune des 2 questions.

Les p sont connues (après les examens). On les assimile à des probabilités.

On désigne par λ_{11} λ_{10} λ_{01} et λ_{00} les proportions de sujets *connaissant* la réponse correcte aux 2 questions, à la 1^{re} question seule, à la 2^e question seule, enfin à aucune des questions. Les λ sont inconnus. On les assimile aussi à des probabilités.

Imaginons en outre qu'il y ait indépendance entre les tirages au sort (mentaux) qu'effectuent les sujets ignorants, quand ils sont doublement ignorants (ce qui suppose qu'il n'existe aucun souci de *cohérence* entre leurs réponses aux 2 questions).

Alors on a (en confondant proportions avec probabilités) :

$$\left| \begin{array}{l} p_{11} = \lambda_{11} + \frac{\lambda_{10}}{r} + \frac{\lambda_{01}}{r} + \frac{\lambda_{00}}{r^2} \\ p_{10} = \lambda_{10} \left(\frac{r-1}{r} \right) + \lambda_{00} \frac{r-1}{r^2} \\ p_{01} = \lambda_{01} \left(\frac{r-1}{r} \right) + \lambda_{00} \frac{r-1}{r^2} \\ p_{00} = \lambda_{00} \left(\frac{r-1}{r} \right)^2 \end{array} \right| \quad (1)$$

d'où :

$$\left| \begin{array}{l} \lambda_{00} = \left(\frac{r}{r-1} \right)^2 p_{00} \\ \lambda_{10} = \frac{r}{r-1} p_{10} - \frac{r}{(r-1)^2} p_{00} \\ \lambda_{01} = \frac{r}{r-1} p_{01} - \frac{r}{(r-1)^2} p_{00} \\ \lambda_{11} = p_{11} - \frac{1}{r-1} (p_{10} + p_{01}) + \frac{1}{(r-1)^2} p_{00} \end{array} \right| \quad (2)$$

D'où l'on tire les estimations des λ en fonction des fréquences observées pour les p .

Vérification :

$$\begin{aligned} \lambda_{00} + \lambda_{01} + \lambda_{10} + \lambda_{11} &= p_{11} + \left(\frac{r}{r-1} - \frac{1}{r-1} \right) p_{10} + \left(\frac{r}{r-1} - \frac{1}{r-1} \right) p_{01} + \\ &\quad \frac{r^2 - 2r + 1}{(r-1)^2} p_{00} = p_{11} + p_{10} + p_{01} + p_{00} = 1 \end{aligned}$$

Discussion :

On a omis de tenir compte des écarts aléatoires (probabilités et proportions étant confondues dans ce calcul). On s'expose donc à trouver certains des λ_i négatifs, ou encore plus grands que 1, sans que le modèle soit nécessairement aberrant.

3^o Amélioration du modèle : Introduction de l'ignorance partielle

Au lieu de s'en tenir à 2 états possibles : connaissance-ignorance totale, on aurait intérêt à se rapprocher de la réalité en introduisant dans le modèle une ignorance graduelle. Chernoff (p. 393) nous informe qu'il est en train d'étudier un modèle continu, où la probabilité qu'un sujet connaisse la réponse à la i ème question est

$$\lambda_i = 1 - e^{-\theta t_i}$$

où t_i est un indicateur de la facilité de la question i ,

et où θ est un indicateur de l'aptitude du sujet à répondre juste.

Il a été procédé dans cette voie par Lord et Allan Birnbaum; nous ignorons le résultat de ces recherches. Revenons à un modèle discret d'ignorance; celui de Horst (p. 382).

4^o Modèle pour une seule question, à r réponses possibles (Hypothèses : Relation d'ordre entre les réponses).

On suppose qu'il existe un *ordre* dans les r réponses possibles, depuis la vraie jusqu'à la plus absurde. Si elles étaient ainsi rangées, tout le monde choisirait la première; mais elles sont présentées en désordre. Le niveau de connaissances d'un individu consiste dans la liste de réponses possibles, amputée de celles qu'il sait être fausses.

Soit λ_i la proportion de la population qui tire au sort sa réponse entre les réponses possibles nos 1, 2, 3... i . Les sujets parfaitement renseignés choisissent la bonne réponse :

la réponse 1; ils sont en proportion λ_1 ; à l'opposé, les sujets totalement ignorants (en proportion λ_r) n'éliminent aucune réponse possible de leurs choix, ils répondent n'importe quoi (avec probabilité $1/r$).

La probabilité qu'un sujet ait ainsi choisi la réponse j est

$$p_j = \frac{\lambda_j}{j} + \frac{\lambda_{j+1}}{j+1} + \dots + \frac{\lambda_r}{r} \quad (1)$$

Les proportions $\hat{p}$ étant substituées aux probabilités p (le nombre de sujets interrogés étant supposé grand), on en tire l'estimation des λ_j à partir de :

$$\lambda_j = j(p_j - p_{j+1}); \text{ en particulier } \lambda_1 = p_1 - p_2. \quad (2)$$

5° *Modèle plus souple : Hypothèse : Relation d'ordre partiel entre les réponses possibles* (ce que Chernoff appelle *complex orderings*)

1. On suppose que les connaissances d'un sujet peuvent être représentées par un ensemble S de réponses qui lui semblent possibles, parmi lesquelles :

l'une est la réponse qu'il donne, tirée au sort dans S (avec d'égales probabilités);

une est nécessairement la vraie (celle tirée au sort ou une autre).

On voit que, comme précédemment, on n'accepte pas qu'un sujet puisse laisser la vraie réponse en dehors des réponses lui semblant possibles.

2. On s'organise pour *estimer* les probabilités λ_i ... qu'un sujet (pris au hasard) ait ses connaissances représentées par l'ensemble S_i .

Exemple : $S_1 = \{a\}, S_2 = \{a, b\}, S_3 = \{a, c\}, S_4 = \{a, b, c, d\}$

$$\left| \begin{array}{l} p_a = \lambda_1 + \frac{1}{2}(\lambda_1 + \lambda_3) + \frac{1}{4}\lambda_4 \\ p_b = \frac{1}{2}\lambda_2 + \frac{1}{4}\lambda_4 \\ p_c = \frac{1}{2}\lambda_3 + \frac{1}{4}\lambda_4 \\ p_d = \frac{1}{4}\lambda_4 \end{array} \right| \quad (1) \quad \left| \begin{array}{l} \lambda_1 = p_a - p_b - p_c + p_d \\ \lambda_2 = 2(p_b - p_d) \\ \lambda_3 = 2(p_c - p_d) \\ \lambda_4 = 4p_d \end{array} \right| \quad (2)$$

Les p sont estimés par les fréquences $\hat{p}$ correspondantes, résultat brut des observations. Les λ sont estimés par les formules (2) où les $\hat{p}$ sont substituées aux p . Il est clair qu'avec une organisation défectueuse on n'arriverait plus à estimer les λ : Le modèle serait *non identifiable* (Le modèle est identifiable s'il consiste en $r-1$ équations linéaires à $r-1$ inconnues λ linéairement indépendantes avec un déterminant non nul) ⁽¹⁾.

1. On pourra se reporter à VAJDA, *Mathematical Programming* (1961), p. 57, § 4.6, *Trees, Triangularity*, et § 4.7, *Incidence matrix*. La matrice d'incidence M met en évidence un arbre; mais ce pourrait ne pas être le cas avec un autre exemple.

$$\begin{array}{c} a \ b \ c \ d \\ S_1 \\ S_2 \\ S_3 \\ S_4 \end{array} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} = M$$

6° *Modèle à plusieurs questions* (et non plus à plusieurs réponses) : Étude d'un cas très simple.

Supposons une suite de s questions, telles que tout sujet qui connaît la réponse à la $i^{\text{ème}}$ question connaît également la réponse aux questions $(i-1)$ $(i-2)$...

Supposons qu'il ne connaisse pas la réponse à la question $(i+1)$; alors il répond au hasard parmi les r choix qui lui sont offerts.

Admettons qu'il y ait indépendance entre ses réponses (aléatoires) aux questions $(i+1)$ $(i+2)$...

On dira qu'un sujet appartient à la classe i s'il connaît la réponse correcte à la question i , mais non à la réponse $i+1$. Soit $\lambda_1 \lambda_2 \dots \lambda_t \dots \lambda_s$ les proportions de sujets appartenant à ces s classes (et λ_0 la proportion de ceux qui ne connaissent aucune réponse). Les réponses du sujet ont des probabilités $p_0 p_1 p_2 \dots p_j \dots$ d'être exactes j fois.

Les proportions correspondantes (statistique des réponses) estiment $p_0 p_1 \dots p_j$. Les relations (1) donnent les p_j (λ_t), on les résout suivant (2) donnant les λ_t (p_j).

Relations (1)

p_0 = probabilité de réponses fausses aux s questions (c'est-à-dire à la 1^{re})

$$= \lambda_0 \left(1 - \frac{1}{r}\right) = \left(\frac{r-1}{r}\right) \lambda_0$$

p_1 = probabilité de répondre juste à la 1^{re} et faux aux questions 2, 3, ... (c'est-à-dire à la 2^e)

$$= \lambda_0 \frac{1}{r} \left(1 - \frac{1}{r}\right) + \lambda_1 \left(1 - \frac{1}{r}\right) = \frac{r-1}{2} (\lambda_0 + \lambda_1 r)$$

$$p_2 = \lambda_0 \frac{1}{r^2} \left(1 - \frac{1}{r}\right) + \lambda_1 \frac{1}{r} \left(1 - \frac{1}{r}\right) + \lambda_2 \left(1 - \frac{1}{r}\right)$$

$$p_j = \frac{(r-1)}{r^{j+1}} \sum_{t=0}^j \lambda_t r^t$$

Relations (2)

$$\lambda_0 = \frac{r}{r-1} p_0$$

$$\lambda_1 = \frac{r}{r-1} p_1 - \frac{\lambda_0}{r}$$

$$\lambda_1 = \frac{r}{r-1} p_1 - \frac{1}{r-1} p_0 = \frac{rp_1 - p_0}{r-1}$$

$$\lambda_2 = \frac{r}{r-1} p_2 - \frac{1}{r} - \frac{\lambda_0}{r^2} = \frac{rp_2 - p_1}{r-1}$$

$$\lambda_j = (r-1)^{-1} (r p_j - p_{j-1})$$

Discussion : $\lambda_j > 0 \quad \Leftrightarrow \quad \underline{p_{j-1} < r p_j} \quad (3)$

Le modèle ne s'applique pas si les données $\hat{p}$ ne vérifient pas les inégalités (3).

Remarque : Dans le cas précédent les psychologues diraient que les questions posées apprécient des connaissances qu'on sait *repérer*; ils parleraient d'*échelle*.

7^o *Autre exemple :* Deux examens (d'espagnol) ont lieu l'un avant, l'autre après un stage de formation (de 6 mois). On admet qu'en gros tous les étudiants savent plus d'espagnol après qu'avant. On pourrait leur poser la même question à 6 mois de distance. On n'a pas osé le faire de peur que certains se souviennent de la réponse qu'ils ont faite 6 mois plus tôt. On procède comme suit.

On suppose qu'il existe un *ordre naturel* de choix $1\ 2\ \dots\ r$ parmi les r réponses possibles. Un sujet qui répond 1 (c'est-à-dire juste) peut appartenir à $S_1 = \{1\}$, à $S_2 = \{1, 2\}$, à $S_3 = \{1, 2, 3\} \dots$ ou $S_r = \{1, 2 \dots r\}$. L'état de sa science (inconnu) au 1^{er} examen est défini par l'indice h et au 2^e examen par l'indice i de l'ensemble S auquel il appartient à ces instants.

Ses réponses sont (de même) représentées par deux indices j (1^{er} examen) et k (2^e examen); avec $j \leq h \leq r; k \leq i \leq r$.

On connaît $(j\ k)$. On ne connaît pas le $(h\ i)$ correspondant. On a alors des relations de la forme :

$$\text{Relations (1) :} \quad p(j\ k) = \sum \lambda(h\ i) c(h\ i)$$

$$\text{Relations (2) :} \quad \lambda(h\ i) = \sum p(j\ k) d(j\ k)$$

Le premier $\sum$ s'entend $j \leq h; k \leq i$, et le second $h \leq j; i \leq k$. Restent à définir les 2 matrices $[c]$ et $[d]$ de coefficients.

Exemple :

Faisons-le sur l'exemple (artificiel) que donne CHERNOFF. On suppose $r = 3$

Matrice des $p(j\ k)$ (Données)

	$k = 1$	2	3	marge
$j = 1$	0,340	0,130	0,035	0,505
2	0,180	0,100	0,025	0,305
2	0,100	0,070	0,020	0,190
	0,620	0,300	0,080	1,000

Matrice des $\lambda(h\ i)$ (résultat des calculs)

	$i = 1$	2	3	
$h = 1$	0,130	0,040	0,030	200
2	100	100	030	230
3	090	300	180	570
	0,320	0,440	0,240	1,000

Les marges et les relations 1.2 :

Les sujets de classe $h = 1$ répondent $j = 1$ nécessairement

Ceux de classe $h = 2$ répondent $j = 1$ ou 2 par moitié

Ceux de classe $h = 3$ répondent $j = 1, 2, 3$ par tiers

$$0,505 = 0,200 + \frac{1}{2} 0,230 + \frac{1}{3} 0,570$$

$$0,305 = \frac{1}{2} 0,230 + \frac{1}{3} 0,570$$

$$\text{Enfin} \quad 0,190 = \frac{1}{3} 0,570$$

De même pour les sujets de classe $i = 1; i = 2; i = 3$.

Ainsi les *marges* du tableau $\lambda(h, i)$ sont calculables par la technique du § 4 ci-dessus (question unique). Mais *il n'y a pas indépendance* entre h et i ; et c'est bien normal s'il s'agit de 2 questions sur la même discipline. Nous constatons (marges du tableau λ) que les réponses sont bien meilleures au 2^e examen, sans que (pourtant) le triangle nord est du tableau soit garni de zéros (ces 10 % sont ceux qui ont désappris).

Intérieur des tableaux : Le plus simple est de décrire en *détail* le passage de la matrice des λ à celle des p (Relations (1)) :

Matrice (p_{jk}) obtenue à partir de la matrice λ_{hk} :

	colonne $k = 1$	$k = 2$	$k = 3$
ligne $j = 1$	$130 + \frac{40}{2} + \frac{30}{3}$ $+ \frac{100}{2} + \frac{100}{4} + \frac{30}{6}$ $+ \frac{90}{3} + \frac{300}{6} + \frac{180}{9}$	$\frac{40}{2} + \frac{30}{3}$ $+ \frac{100}{4} + \frac{30}{6}$ $+ \frac{300}{6} + \frac{180}{9}$	$\frac{30}{3}$ $+ \frac{30}{6}$ $+ \frac{180}{9}$
ligne $j = 2$	$\frac{100}{2} + \frac{100}{4} + \frac{30}{6}$ $+ \frac{90}{3} + \frac{300}{6} + \frac{180}{9}$	$\frac{100}{4} + \frac{30}{6}$ $+ \frac{300}{6} + \frac{180}{9}$	$\frac{30}{6}$ $+ \frac{180}{9}$
ligne $j = 3$	$\frac{90}{3} + \frac{300}{6} + \frac{180}{9}$	$\frac{300}{6} + \frac{180}{9}$	$\frac{180}{9}$

d'où : Coefficients de la relation 1 : $c(h, i) = 1/hi$

On trouve ensuite :

Coefficients de la relation 2 : $\lambda(h, i) = hi[p(h, i) - p(h+1, i) - p(h, i+1) + p(h+1, i+1)]$

Si l'on n'avait pas tenu compte de la 2^e question on aurait eu (cf. II 1)

$$\min V = \frac{\lambda(p - \lambda)}{p} \quad \text{avec } p = p_{11} + p_{10}; z + t = \lambda_{11} + \lambda_{10} = \lambda$$

On peut montrer qu'on a :

$$\min W < \min V$$

grâce à la combinaison des 2 questions.

Annexe.

Montrons qu'il en est bien ainsi.

Les 3 paraboles d'équation $\lambda(p - \lambda) = u$, $z(p_{11} - z) = v$, $t(p_{10} - t) = w$ ont pour sommets $(\lambda = \frac{p}{2}, u = p^2/4)$; $(z = \frac{p_{11}}{2}, v = \frac{p_{11}^2}{4})$; $(t = \frac{p_{10}}{2}, w = \frac{p_{10}^2}{4})$

Donc :

$$\text{Max min } V = \frac{p}{4}, \text{ pour } \lambda = p/2$$

$$\text{Max min } W = \frac{p_{11}}{4} + \frac{p_{10}}{4} = \frac{p}{4}, \text{ pour } z + t = \frac{p_{11} + p_{10}}{2} = \frac{p}{2}$$

Reste à voir que $\min W < \min V$ dans les autres cas. C'est l'inégalité peu connue

$$\frac{z(a-z)}{a} + \frac{t(b-t)}{b} \leq \frac{(z+t)(a+b-z-t)}{a+b} \quad (1)$$

$$\text{ou } \frac{z^2}{a} + \frac{t^2}{b} \geq \frac{(z+t)^2}{a+b} \quad (2)$$

$$\text{ou } z^2 + \frac{b}{a}z^2 + t^2 + \frac{a}{b}t^2 \geq z^2 + t^2 + z3t \quad (3)$$

$$\text{ou } \left(\sqrt{\frac{b}{a}}z - \sqrt{\frac{a}{b}}t \right)^2 \geq 0 \quad (4)$$

IV RÉSUMÉ ET CONCLUSION

Après avoir rappelé que le problème éternel du statisticien est de faire de bonnes statistiques avec des données médiocres, et que l'emploi de modèles à erreurs aléatoires n'est pas une innovation en statistique, on a présenté des modèles trop peu connus des statisticiens et que CHERNOFF étudie dans deux articles parus en 1961 et 1962.

On a surtout développé (partie II) les mécanismes par lesquels on peut ainsi passer de statistiques brutes (proportions p d'individus classés suivant leurs réponses) à des statistiques rectifiées proportions λ d'individus classés suivant leur état réel). Les questions sont d'un type tel qu'une partie des réponses soit faite au hasard (dans l'ignorance totale ou partielle où se trouve l'enquête).

On n'a que très sommairement (partie III) indiqué la façon d'apporter des corrections aux notes qui sont attribuées aux sujets interrogés, lorsqu'il s'agit d'un examen sur le niveau de connaissances et non d'une enquête statistique.

En conclusion, on peut se demander si les réponses à certaines enquêtes (par sondage ou non) de la statistique ne gagneraient pas quelquefois à être ainsi « redressées » dans le cadre d'un modèle à erreurs aléatoires. Il faudrait bien se garder, d'ailleurs d'y voir une panacée, car l'hypothèse de base (toutes les réponses sont équi-probables dans l'ignorance) est certainement inexacte, surtout si l'ordre des réponses possibles n'est pas permuté d'un questionnaire à l'autre pour en éliminer le rôle.

P. THIONET

*
* * *