Pour un enseignement de la statistique dans le premier cycle

par P.L. HENNEQUIN, Professeur à l'Université de Clermont

DEUXIEME PARTIE (*)

III — *Publication des résultats*

Le plus souvent, l'enquêteur s'engage à garder confidentielle l'application recensement R. Par contre, il est intéressant de diffuser P_Ω et il suffit pour cela de publier sa restriction à $\mathcal{A}_S$ ensemble des alternatives élémentaires. Il suffit même de publier S_R et la restriction de P_Ω à $\mathcal{A}_{S_R}$.

Bien entendu, la table de P_Ω contient card S_R lignes et peut être volumineuse. On se contente souvent d'en publier des extraits (par exemple la valeur des P_Ω (A_j^i) mais on perd alors de l'information.

Pour déterminer P_Ω et S_R nous allons décrire ici deux algorithmes de dépouillement en employant une terminologie relative à un ordinateur. Mais ces algorithmes s'appliquent très bien, si card Ω et card S ne sont pas trop grands, à un dépouillement qui peut être effectué à la main par les élèves d'une classe.

Premier algorithme : Commençons par numéroter les σ éléments de S (si tous les r_i sont égaux il est naturel d'utiliser le système de base r_i) ; soit $s(\omega)$ le nombre affecté à la réponse contenue dans le formulaire ω . Supposons que nous disposions d'un ordinateur dont la mémoire est divisée en emplacements, repérés par une adresse.

Le nombre d'emplacements est supposé supérieur à σ et chaque emplacement peut contenir un nombre compris entre 0 et K avec $K \geqslant$ card Ω. Au départ, chaque emplacement contient le nombre 0. Nous lisons successivement les formulaires remplis, après les avoir convertis par exemple en cartes perforées et à la lecture de ω nous augmentons d'une unité le contenu de l'emplacement de mémoire, d'adresse $s(\omega)$.

(*) Voir la première partie dans le Bulletin 290, page 535 et suivantes.

Si nous travaillons à la main, nous divisons une feuille de papier en deux colonnes et en σ lignes. Dans la première colonne, nous inscrivons, à raison d'un par ligne, les éléments de S ou leur code. Pour chaque ω, nous cochons un bâton dans la ligne $s(\omega)$. Une fois tous les questionnaires lus, nous comptons le nombre de bâtons de chaque ligne.

Nous obtenons ainsi les σ nombres card Ω_s et les $P_\Omega(A_s)$ s'en déduisent par division par card Ω.

Si τ désigne le temps nécessaire pour lire un questionnaire et ajouter 1 au contenu d'une mémoire et τ' le temps d'une division par card Ω, la durée du dépouillement est égale à

$$\tau \text{ card } \Omega + \tau' \text{ card } R(\Omega)$$

(car en fait seules interviennent les divisions d'un nombre *non nul* par card Ω).

Il peut se faire que card Ω soit beaucoup plus petit que card S et que ce premier algorithme conduise à introduire un grand nombre d'emplacements mémoires ou de lignes, qui, en fait, seront inutilisés.

Dans ce cas, il est préférable de procéder comme suit.

Deuxième algorithme : un élément de S étant représenté par une suite $\{q_i, 1 \leqslant i \leqslant n\}$, ordonnons S par l'ordre lexicographique. Par une première lecture de la réponse $q_n(\omega)$ à la dernière question, nous rangeons les formulaires remplis suivant l'ordre des $q_n(\omega)$ croissants. Par une seconde lecture, nous rangeons la pile ainsi obtenue suivant l'ordre des q_{n-1} croissants et ainsi de suite, au bout de n opérations, les $s(\omega)$ sont rangés dans l'ordre croissant. Si ce tri sur une question qui comporte r_i réponses possibles nécessite un temps $t\, r_i$, le temps total pour ce tri est θ card Ω où $\theta = t \sum_{i=1}^{n} r_i$ (à la main θ est voisin de τ). On reprend alors les $s(\omega)$ ainsi ordonnés, soient $\{s(\omega_i)\}$. On inscrit $s(\omega_1)$ puis le nombre de fois p_1 où il figure (consécutivement) dans la suite $\{s(\omega_i)\}$, $s(\omega_{p_1+1})$ et le nombre de fois où il figure dans la suite, etc... On obtient ainsi l'ensemble $R(\Omega)$ et, en divisant par card Ω, la restriction de P_Ω à $\mathcal{A}_R$.

Si θ' désigne le temps nécessaire pour la lecture d'un élément de la suite ordonnée $\{s(\omega_i)\}$ et sa comparaison avec le précédent, le dépouillement suivant ce deuxième algorithme dure

$$(\theta + \theta') \text{ card } \Omega + \tau' \text{ card } R(\Omega)$$

La comparaison des temps de calcul dépend du matériel utilisé. A la main, le second algorithme est un peu plus long mais c'est le seul applicable si card S est infini (cf. § IV) ou même très grand devant card Ω.

IV — *Questionnaires ouverts*

IV.1 — Jusqu'à maintenant nous avons supposé que la personne interrogée devait, pour chaque question Q_i, choisir une réponse dans une liste explicite E_i à r_i éléments. Or il n'est pas toujours commode de décrire à l'avance l'ensemble des réponses possibles à une question et il vaut mieux, dans ce cas, considérer cet ensemble comme une partie d'un ensemble infini (N, Z, Q ou R).

Il en est ainsi pour les questions :

Combien avez-vous d'enfants ? de timbres postes dans votre collection ? Quel est votre bénéfice annuel (éventuellement négatif) ?

Quelle est votre date de naissance ?

Si l'on pose simultanément p telles questions, leur ensemble équivaut à une question unique à valeurs dans R^p, ainsi "quelles sont vos mensurations ? ".

Des espaces plus complexes sont nécessaires pour repérer les réponses aux questions :

où logez-vous ? où prenez-vous vos vacances ? quelles sont vos performances sportives ? quelles sont vos aptitudes mathématiques ?

Chaque fois qu'un questionnaire comprend au moins une telle question, nous dirons qu'il est *ouvert*.

Soit donc $\mathcal{X}$ l'espace dans lequel nous plongeons l'ensemble des réponses possibles à une question Q. Le recensement permet de définir une application X_Q de Ω dans $\mathcal{X}$. Comme pour R, le statisticien ne s'intéresse pas directement à X_Q mais à

$$P_{\Omega}^{X_Q}(B) = f_{\Omega} \circ X_Q^{-1}(B) = \frac{\text{card}\,\{\omega \,/\, X_Q(\omega) \in B\}}{\text{card}\,\Omega}$$

au moins pour certaines parties B de $\mathcal{X}$.

Comme ici Ω est fini, il en est de même de $X_Q(\Omega)$ et on peut définir $P_\Omega^{X_Q}$ pour toutes les parties de $\mathfrak{X}$: soit b_i, $1 \leqslant i \leqslant k$, les éléments de $X_Q(\Omega)$; on a

$$P_\Omega^{X_Q}(B) = P_\Omega^{X_Q}(B \cap X_Q(\Omega)) = \sum_{b_i \in B} P_\Omega^{X_Q}(\{b_i\}) \qquad (1)$$

avec

$$P_\Omega^{X_Q}(\{b_i\}) = \frac{\text{card}\,\{\omega \,/\, X_Q(\omega) = b_i\}}{\text{card}\,\Omega} = p_i$$

La double suite $\{(b_i, p_i), 1 \leqslant i \leqslant k\}$ constitue la *distribution* de X et (1) permet de calculer $P_\Omega^{X_Q}(B)$ pour tout $B \subset \mathfrak{X}$ à partir de cette distribution. On définit ainsi une probabilité sur $\mathcal{P}(\mathfrak{X})$ (qui n'est autre que l'image par X de la probabilité uniforme sur Ω).

Pour un questionnaire ouvert, on appellera application recensement l'application R de Ω dans $\prod_{i=1}^{n} \mathfrak{X}_i$ définie par les n applications X_{Q_i} correspondant aux n questions, suivant

$$R(\omega) = \{X_{Q_1}(\omega), X_{Q_2}(\omega), \dots X_{Q_n}(\omega)\}.$$

La donnée de R permet de définir une probabilité P_Ω^R sur $(\prod_{i=1}^{n} \mathfrak{X}_i)$: $P_\Omega^R = f_\Omega \circ R^{-1}$.

Une fois le recensement effectué, il est possible qu'on s'intéresse à des applications X de Ω dans $\mathfrak{X}$ ne correspondant pas à une question figurant dans le questionnaire. Par exemple l'application "quel est le montant de vos impôts ? ".

A cette application X correspond, comme plus haut, une probabilité sur $\mathcal{P}(\mathfrak{X})$ définie par $P_\Omega^X = f_\Omega \circ X^{-1}$. Mais il est possible de calculer $P_\Omega^X(B)$ sans connaître f_Ω, si l'on connaît P_Ω^R et si $X^{-1}(B)$ appartient au sous-anneau $\mathcal{B}_R$ de $\mathcal{P}(\Omega)$ engendré par les $R^{-1}(s)$.

En effet, si $X^{-1}(B) = C$ et si $C = R^{-1}(D)$ avec $D \subset \prod_{i=1}^{n} \mathfrak{X}_i$,

$$P_\Omega^X(B) = f_\Omega \circ X^{-1}(B) = f_\Omega(C) = f_\Omega(R^{-1}(D)) = f_\Omega \circ R^{-1}(D) = P_\Omega^R(D)$$

Définition

Nous dirons qu'une application X de Ω est *mesurable* sur $(\Omega, \mathcal{B}_R)$ (sous entendu dans $(\mathfrak{X}, \mathcal{P}(\mathfrak{X}))$) si $\forall$ $B \in \mathcal{P}(\mathfrak{X})$, $X^{-1}(B) \in \mathcal{B}_R$.

X est mesurable sur $(\Omega, \mathcal{B}_R)$ si et seulement si il existe au moins une application Y de $\prod_{i=1}^{n} \mathfrak{X}_i$ dans $\mathfrak{X}$ telle que $X = Y \circ R$ autrement dit si X peut se calculer à partir de R. Ainsi le questionnaire "déclaration de revenus" permet de calculer le montant des impôts sur le revenu.

IV.2 — Pour poursuivre l'étude d'une application mesurable X, il est nécessaire de structurer $\mathfrak{X}$. Les deux structures les plus usuelles sont celles d'espace *ordonné* et celle d'espace *vectoriel*. Le cas particulier $\mathfrak{X} = \mathbf{R}$ auquel on se limite souvent est à la fois totalement ordonné et vectoriel ; il nous semble préférable de le traiter après les cas plus généraux pour bien séparer le rôle joué par les deux structures.

IV.3 — Si $\mathfrak{X}$ est ordonné par $\prec$ on associera à X sa *fonction de répartition*, application F croissante de $\mathfrak{X}$ dans $[0, 1]$ définie par

$$F(x) = \sum_{b_i \prec x} p_i .$$

Si $\mathfrak{X}$ et $\mathfrak{X}'$ sont *totalement* ordonnés et si h est une application *strictement* croissante de C dans C', la fonction de répartition F' de $h \circ X$ est telle que

$$F' \circ h = F$$

En fait, la fonction de répartition n'apporte, dans notre cas où Ω est fini (donc aussi $X(\Omega)$) aucune information supplémentaire par rapport à la distribution de X. Si card $X(\Omega)$ est grand, il sera nécessaire de résumer l'information par le calcul de *valeurs centrales*.

Si $\mathfrak{X}$ est *totalement ordonné*, on pourra définir des médianes, et, plus généralement des *quantiles* de X : si α est un réel compris entre 0 et 1, on appelle quantile d'ordre α tout élément x de $\mathcal{X}$ tel que

$$\sum_{b_i < x} p_i \geqslant \alpha \quad \text{et} \quad \sum_{x < b_i} p_i \geqslant 1 - \alpha$$

Il existe toujours au moins un b_j qui satisfait cette double inégalité. On appelle médiane un quantile d'ordre $\frac{1}{2}$.

Si h est une application *croissante* de $\mathcal{X}$ dans $\mathcal{X}'$ totalement ordonné et si q_α est un quantile d'ordre α de X, h (q_α) est un quantile d'ordre α et h o X.

IV.4 — Si $\mathcal{X}$ est vectoriel, on peut associer à X sa moyenne sur Ω :

$$E\,X = \frac{\sum_{\omega \in \Omega} X(\omega)}{\text{card}\,\Omega} = \sum_i p_i b_i$$

On définit ainsi une application *linéaire* de $\mathcal{X}^\Omega$ dans $\mathcal{X}$. Si $\mathcal{X}$ est *en outre* ordonné par un ordre $>$ compatible avec la structure vectorielle, $X > 0$ implique $E\,X > 0$.

IV.5 — Terminons par un exemple où les notions de moyennes et de médianes peuvent s'introduire dans une leçon de géographie politique sur la notion de *commune* en France.

Le recensement de 1969 a permis de classer les communes de France par ordre de population croissante suivant le tableau page suivante (où Paris est considérée comme une commune).

Si l'on envoie à *chaque commune* (par exemple au Maire) un questionnaire sur la population de la commune, on obtiendra comme

moyenne : $\frac{50.10^6}{37.708} \simeq 1\,325$

médiane : $\simeq$ 400

Mais si l'on pose cette fois la question à *toute la population*, on obtiendra comme

moyenne : $\simeq$ 190 000

médiane : $\simeq$ 4 500

(cette dernière quantité est appelée parfois population *médiale* d'une commune française).

Nombre d'habitants	Nombre de communes	Pourcentage de la population
de 1 à 100	3 877	0,5
101 à 200	7 514	2,2
201 à 500	12 616	7,9
501 à 1 000	6 824	9,2
1 001 à 1 500	2 484	5,8
1 501 à 2 000	1 134	3,9
2 001 à 2 500	671	3
2 501 à 3 500	727	4,2
3 501 à 5 000	548	4,3
5 001 à 10 000	642	8,6
10 001 à 20 000	345	9,3
20 001 à 30 000	139	6,9
30 001 à 50 000	98	7,4
50 001 à 100 000	60	7,8
100 001 à 200 000	28	6,8
200 001 à 300 000	4	2
320 000	1	1,4
370 000	1	
530 000	1	1,1
890 000	1	1,8
2 590 000	1	5,2
TOTAL	37 708	100

V — *Probabilités conditionnelles. Indépendance.*

V.1 — Soit A_β une alternative non négligeable, c'est-à-dire telle que $P_\Omega\,(A_\beta) \neq 0$ et Ω_β la partie correspondante de Ω :

$$\Omega_\beta = R^{-1} \circ I_2^{-1} \circ I_1^{-1}\,(A_\beta)$$

Restreignons le dépouillement à Ω_β ; on a :

$$P_{\Omega_\beta}\,(A_\alpha) = f_{\Omega_\beta}\,(\Omega_\alpha) = \frac{\text{card}\,(\Omega_\alpha \cap \Omega_\beta)}{\text{card}\,\Omega_\beta} = \frac{P_\Omega\,(A_\alpha \wedge A_\beta)}{P_\Omega\,(A_\beta)}$$

On définit ainsi une nouvelle fonction sur $\mathcal{A}$ que nous noterons par abréviation P_β et que nous appellerons probabilité conditionnée par A_β ; elle est comme P_Ω positive, additive et telle que $P_\beta(A_U) = P_\beta(A_\beta) = 1$. Sa restriction à $\mathcal{A}_\beta$ est donc aussi une probabilité. On a $P_\Omega = P_U$.

Elle possède en outre les propriétés suivantes :

Si les Ω_{α_i} forment une partition de Ω on dira que les A_{α_i} forment une partition de A_U, c'est-à-dire que

$$A_{\alpha_i} \neq A_O,\ i \neq j \Rightarrow A_{\alpha_i} \wedge A_{\alpha_j} = A_O,\ \bigvee_i A_{\alpha_i} = A_U$$

Si en outre $\forall\ i$, $P_\Omega(A_{\alpha_i}) \neq 0$, alors

$$\forall A \in \mathcal{A},\ P_\Omega(A) = \sum_i P_{\alpha_i}(A)\, P_\Omega(A_{\alpha_i}) \qquad (1)$$

(Règle des probabilités composées).

Cette règle permet de fractionner une enquête en enquêtes relatives chacune à une partie de la population (fraction d'un pays en régions ou communes).

Si, en outre, $P_\Omega(A_\alpha) \neq 0$,

$$P_\alpha(A_{\alpha_i}) = \frac{P_{\alpha_i}(A_\alpha)\, P_\Omega(A_{\alpha_i})}{\sum_j P_{\alpha_j}(A_\alpha)\, P_\Omega(A_{\alpha_j})} \qquad (2) \quad \text{(Formule de Bayes)}$$

La formule de Bayes permet par exemple à une compagnie d'assurances automobile d'évaluer, à partir des déclarations d'accident de ses clients pendant une année, la probabilité qu'un accident de voiture ait été provoqué par un conducteur d'âge i quand on connaît le nombre des clients d'âge i et la probabilité pour qu'un client d'âge i ait un accident comme conducteur.

V.2 — *Indépendance de deux alternatives*

L'alternative A_α est dite ***indépendante de l'alternative*** A_β (relativement à Ω et P ou encore à P_Ω), où $P_\Omega(A_\beta) \neq 0$, si

$$P_\beta(A_\alpha) = P_\Omega(A_\alpha) \qquad (1)$$

autrement dit si la probabilité de A_α est la même dans Ω et dans Ω_β.

Cette relation équivaut à

$$P_\Omega (A_\alpha \wedge A_\beta) = P_\Omega (A_\alpha) P_\Omega (A_\beta) \qquad (2) \qquad \text{et} \quad P_\Omega (A_\beta) \neq 0$$

Si $P_\Omega (A_\beta) = 0$, a fortiori $P_\Omega (A_\alpha \wedge A_\beta) = 0$ donc (2) est satisfaite.
De plus (2) est symétrique en α et β .

(2) définit *l'indépendance de* A_α et A_β .

Si $P_\Omega (A_\alpha) \neq 0$ et $P_\Omega (A_\beta) \neq 0$, (2) implique donc à la fois que A_α est indépendante de A_β et A_β indépendante de A_α .

Si (A_α, A_β) sont indépendantes, il en est de même de $(\overline{A_\alpha}, A_\beta)$, $(A_\alpha, \overline{A_\beta})$, $(\overline{A_\alpha}, \overline{A_\beta})$, en effet

$$P_\Omega (\overline{A_\alpha} \wedge A_\beta) = P_\Omega (A_\beta) - P_\Omega (A_\alpha \wedge A_\beta) = (1 - P_\Omega (A_\alpha)) P_\Omega (A_\beta)$$
$$= P_\Omega (\overline{A_\alpha}) P_\Omega (A_\beta) .$$

(2) est satisfaite si

$$P_\Omega (A_\alpha) (1 - P_\Omega (A_\alpha)) P_\Omega (A_\beta) (1 - P_\Omega (A_\beta)) = 0 ,$$

En dehors de ce cas trivial, il est possible qu'elle ne soit satisfaite pour aucun couple (A_α, A_β).

Ainsi si card Ω est premier, car (2) équivaut à

$$\text{card} (\Omega_\alpha \wedge \Omega_\beta) \text{ card } \Omega = \text{card } \Omega_\alpha \text{ card } \Omega_\beta ,$$

si card $(\Omega_\alpha \wedge \Omega_\beta) \neq 0$, card Ω doit donc diviser card Ω_α (mais alors $P_\Omega (A_\alpha) = 1$) ou card Ω_β .

V.3 — *Indépendance de deux questions et de deux sous-anneaux*

Deux alternatives A_α, A_β sont indépendantes si tout élément du partage $(A_\alpha, \overline{A_\alpha})$ est indépendant de tout élément du partage $(A_\beta, \overline{A_\beta})$.

De même, nous dirons que les deux questions Q_λ, $Q_{\lambda'}$ (ou les deux classes $\widehat{Q}_\lambda$, $\widehat{Q}'_{\lambda'}$) sont indépendantes si $\forall$ i, j A_λ^i et $A_{\lambda'}^j$ sont indépendantes, c'est-à-dire que

$$P_\Omega (A_\lambda^i \wedge A_{\lambda'}^j) = P_\Omega (A_\lambda^i) P_\Omega (A_{\lambda'}^j)$$

où les A_λ^i (resp. les $A_{\lambda'}^j$) sont les alternatives associées aux partages (ou partitions) qui définissent Q_λ (resp. $Q_{\lambda'}$) ou $\widehat{Q}_\lambda$ (resp. $\widehat{Q}_{\lambda'}$).

Tout élément de l'anneau $\mathcal{A}_\lambda$ (réunion de A_λ^i incompatibles) est alors indépendant de tout élément de l'anneau $\mathcal{A}_{\lambda'}$ (réunion de $A_{\lambda'}^j$ incompatibles) et quand il en est ainsi, on dit que les deux *sous anneaux* $\mathcal{A}_\lambda$ et $\mathcal{A}_{\lambda'}$ sont indépendants.

V.4 — *Indépendance de deux applications mesurables*

Soient $X_1 = Y_1 \circ R$ et $X_2 = Y_2 \circ R$ deux applications R-mesurables sur Ω. Elles sont indépendantes si pour tout $B \subset \mathfrak{X}$, les questions définies par $Y_1^{-1}(B)$ et $Y_2^{-1}(B)$ sont indépendantes ou encore si les deux sous-anneaux $Y_1^{-1}(B)$ et $Y_2^{-1}(B)$ sont indépendants ou encore si les deux sous-anneaux $Y_1^{-1}(\mathfrak{X})$ et $Y_2^{-1}(\mathfrak{X})$ sont indépendants.

V.5 — *Indépendance de n questions*

On définit comme pour deux l'indépendance de n questions et de n sous-anneaux, donc de n applications mesurables. On dit alors que n alternatives A_{α_i} sont indépendantes si les n sous-anneaux $\mathcal{A}_\alpha$ qu'elles engendrent sont indépendants. (On remarquera que si $n \geqslant 3$, il ne suffit pas que

$$P\left(\bigwedge_{i=1}^{n} A_{\alpha_i}\right) = \prod_{i=1}^{n} P(A_{\alpha_i})$$

mais qu'il faut écrire toutes les relations obtenues en remplaçant dans celles-ci certains A_{α_i} par leur contraire $\overline{A_{\alpha_i}}$.

V.6 — A la notion d'indépendance se rattache celle *d'échantillon représentatif*, base de la théorie des sondages.

On dit qu'une partie Ω_α de Ω est un *échantillon représentatif* pour un sous-anneau $\mathcal{A}_\lambda$ de $\mathcal{A}$ (ou pour la question Q_λ) si l'alternative A_α est indépendante de tout élément de $\mathcal{A}_\lambda$ ou encore, en notant $\mathcal{A}_\alpha$ le sous-anneau unitaire engendré par A_α, si les sous-anneaux $\mathcal{A}_\alpha$ et $\mathcal{A}_\lambda$ sont indépendants.

On a donc $P_\Omega(A) = P_{\Omega_\alpha}(A) \qquad \forall A \in A_\lambda \qquad (1)$

et la détermination de P_Ω sur $\mathcal{A}_\lambda$ nécessite seulement de connaître P_{Ω_α} c'est-à-dire d'interroger les éléments de Ω_α.

Bien entendu, pour vérifier cette condition exactement, il faut interroger toute la population ; la notion d'échantillon représentatif prend tout son intérêt si l'on fait de fréquents recensements et qu'on se contente dans (1) d'une approximation.

V.7 — Bien souvent les conditions d'indépendance seront seulement satisfaites avec une certaine approximation et on cherchera à déterminer une valeur approchée $\widehat{P}_\Omega$ de P_Ω qui, elle, satisfasse *exactement* les conditions d'indépendance.

Ainsi si un questionnaire L comporte n questions Q_i supposées *indépendantes pour* $\widehat{P}_\Omega$, $\widehat{P}_\Omega$ sera entièrement déterminée sur $\mathcal{A}$ par sa restriction aux alternatives primaires A_i^j.

En effet l'hypothèse d'indépendance permet de calculer

$$\widehat{P}_\Omega \left(\bigwedge_{i=1}^{n} A_i^{j_i} \right) = \prod_{i=1}^{n} \widehat{P}_\Omega (A_i^{j_i})$$

et comme les $\bigwedge_{j=1}^{n} A_i^{j_i}$ sont les atomes de $\mathcal{A}$, ceci détermine complètement $\widehat{P}_\Omega$ sur $\mathcal{A}$. On peut donc, à *condition d'expliciter cette hypothèse* d'indépendance, restreindre *dans ce cas particulier* la publication du recensement à la diffusion[7] des $\widehat{P}_\Omega$ $(A_i^{j_i})$.