Note 5 : Démonstrations « réservées aux spécialistes » (Jules 2021-2023)

Vous trouverez le détail des démonstrations dont le but est de structurer les différentes affirmations de ce roman. Elles sont destinées aux spécialistes.

1. *Démonstration N°1*

Question posée : Quel est le sens mathématique du taux attendu quand on mesure la qualité du fonctionnement d'un panel N d'institutions à l'aide d'une question ?

2. *Notations utilisées :*

On se donne une famille d'ensembles $\{G_i\}_{i\in[1,N]}$

$G_i \cap G_j = \varnothing$ *si* $i \neq j$ *pour tout i et j éléments de [1,N].*

On définit $G = \bigcup_{i\in[1,N]} G_i$.

Soit H un ensemble de valeurs discrètes.

(Par exemple : H = {1,2, 3,4,5,6}).

Si $A \subset G$ *et* $B \subset G_i$ *et on définit P(A) = card(A) / card(G) et* $P_{Gi}(B) = card(B)/card(G_i)$, *où card(E) désigne le nombre d'éléments de E.*

Pour chaque G_i *on suppose avoir une variable aléatoire :*

$$X_i : \ \omega \in G_i \ \text{----------} > X_i(\omega) \in H$$

On définit la variable X de G dans H par :

$$si \ \omega \in G_i \ , X(\omega) = X_i(\omega)$$

On note :

X_i *= a l'ensemble des* $\omega \in G_i \ tels \ que \ X_i(\omega) = a$ *et*

X *= a l'ensemble des* $\omega \in G \ tels \ que \ X(\omega) = a$

La distribution F_i de X_i est alors définie par :

$$F_i : a \in H \ \text{----------} > F_i(a) = P_{Gi}(X_i = a)$$

On définit d'une façon analogue la distribution F de G par :

$$F : a \in H \text{ ----------> } F(a) = P(X = a)$$

On remarque que les F_i *sont les distributions conditionnelles de F.*

***Par définition**, F représente le taux attendu. Définition qui est justifiée par le lemme suivant. On remarquera au passage que le taux attendu est ici une fonction !*

3. *Lemme 1:*

Alors la distribution F est la moyenne pondérée par $P(G_i)$ $i \in [1,N]$ des distributions $\{F_i\}_{i \in [1,N]}$

Il suffit de montrer que $\sum_{i \in [1,N]} P(G_i).F_i(a) = F(a)$ *pour tout* $a \in H$.

Les $\{G_i\}_{i \in [1,N]}$ *forment une partition de G et les* $\{X_i = a\}_{i \in [1,N]}$ *une partition de X=a d'où :*

$$P(G) = \sum_{i \in [1,N]} P(G_i)$$

et

$$P(\{X = a\}) = \sum_{i \in [1,N]} P(\{X_i = a\})_{i \in [1,N]}$$

Preuve :

$$\sum_{i \in [1,N]} P(G_i).F_i(a) = \sum_{i \in [1,N]} \frac{card(G_i)}{card(G)}.P_{G_i}(X_i = a) =$$

$$\sum_{i \in [1,N]} \frac{card(G_i)}{card\ (G)}.\frac{card(X_i = a)}{card(Gi)} = \sum_{i \in [1,N]} \frac{card(Xi = a)}{card\ (G)}$$

$$= \sum_{i \in [1,N]} P(X_i = a) = P(X = a) = F(a)$$

4. Démonstration N°2

Le but de cette démonstration est de regarder si la notion de taux attendu persiste quand on transforme les données en les regroupant (par exemple on rassemble les données positives en une seule valeur).

5. Préambule

Avec les notations ci-dessus

Soient deux valeurs distinctes, a $\in$ H et b $\in$ H, on dit qu'on polarise a et b en c relativement à X_i (resp. X) quand on définit un ensemble T et une variable X'_i (resp. X') comme suit :

i. Pour l'ensemble T sa définition est :
   - On définit H_o comme le complémentaire de {a,b} dans H
   - Soit c un nombre, on définit T par T = $H_o \cup$ {c} (c remplace a et b)

ii. Pour la variable X'_i (resp. X') leurs définitions sont : si $\omega \in$ T avec $X_i(\omega) \neq a$ et $X_i(\omega) \neq b$ alors $X'_i(\omega) = X_i(\omega)$, *si non* $X'_i(\omega) = c$. *(respectivement pour X, où* $X'(\omega) = X(\omega)$ *sauf pour la valeur c où* $X'(\omega) = c$ *.).*

iii. De plus on suppose que l'ensemble $X'_i = c$ *est la réunion des ensembles disjoints* $X_i = a$ *et* $X_i = b$*. C'est-à-dire :* $X'_i = c = X_i = a \cup X_i = b$*. (resp. pour X).*

Comme $X_i = a$ *et* $X_i = b$*, sont disjoints :*

$$P(X'_i = c) = P(X = a) + P(X = b).$$

6. Remarque :

Pour tout i, soit K_i *la distribution de* X'_i *et K celle de X'.*

Alors : $K_i(c) = F_i(a) + F_i(b)$ *et* $K(c) = F(a) + F(b)$*).*

Preuve :
$K_i(c) = P_{G_i}(X'_i = c) = P_{G_i}(X_i = a) + P_{G_i}(X_i = b) = F_i(a) + F_i(b)$ (resp. pour X, K(c) = F(a) + F(b)).
On remarque aussi que si $x \notin \{a,b\}$, $K_i(x) = F_i(x)$ (resp. K(x) = F(x)).

7. Lemme 2 :

La polarisation des distributions du groupe G, reste la moyenne pondérée des polarisations de tous les groupes G_i

Il suffit de montrer pour la valeur c que :

$$\sum_{i \in [1,N]} P(G_i).\ K_i(c) = K(c)$$

En effet :

$$\sum_{i \in [1,N]} P(G_i).\ K_i(c) = \sum_{i \in [1,N]} P(G_i).\ (F_i(a) + F_i(b))$$
$$= \sum_{i \in [1,N]} P(G_i).\ F_i(a) + \sum_{i \in [1,N]} P(G_i).\ F_i(b) =$$
$$F(a) + F(b) = K(c)$$

8. Démonstration N°3

Le but de cette démonstration est de définir la notion de profil en croisant deux variables et de vérifier que le profil global est la moyenne pondérée de chaque profil de base. Pour cela, il faut d'abord définir la notion de « case » liée à deux variables.

9. Préambule : tableaux croisés et taux attendus

On utilise les notations ci-dessous :
Soient X et Y deux variables définies sur G et à valeur dans H et leurs restrictions $\{X_i\}_{i \in [1,N]}$ et $\{Y_i\}_{i \in [1,N]}$ sur chaque G_i .
On définit les cases de G x G liées à X et Y par :

- Si $a \in H$ et $b \in H$ alors :

$$case(a,b) = \{ (x , y) \in G \times G \ / \ X(x) = a \text{ et } Y(y) = b \}$$

et pour tout i les cases de G_i x G_i liées à X_i et Y_i par :

- Si a∈ H et b ∈ H alors :

$case_i$ (a , b) = { (x , y) ∈ G_i x G_i / X_i (x) = a et Y_i (b) = b }

10. Remarques :

Pour toute paire de couple (a,b) ≠ (c,d) de H x H on a :
case(a,b) ∩ case(c,d) = ∅ avec a, b, c, d éléments de H.
(resp. $case_i$(a,b) ∩ $case_i$ (c,d) = ∅) .
On a aussi :

$$\bigcup_{(a,b) \in HxH} case(a,b) = G \text{ x } G$$

$$\text{(resp. } \bigcup_{(a,b) \in HxH} case_i(a,b) = G_i \text{ x } G_i \text{),}$$

ce qui permet de dire que l'ensemble des cases de G x G forment une partition de G x G (resp. des G_i x G_i).
D'autre part, les $case_i$(a,b) pour i ∈ [1,N] forment une partition de la case(a,b).

11. Lemme 3

La probabilité de la case(a,b) est la moyenne pondérée par P(G_i) des $case_i$(a,b) quand i parcourt [1,N]

Preuve :
Il faut montrer que :

$$P(case(a,b)) = \sum_{i\in[1,N]} P(G_i) \text{ x } P_{Gi}(case_i\,(a,b))$$

On a d'une part:

$$P(case(a,b)) = \frac{card(\,case(a,b))}{card\ G}$$

Et d'autre part

$$\sum_{i\in[1,N]} P(G_i) \text{ x } P_{G_i}(case_i\,(a,b)) =$$

$$\sum_{i\in[1,N]} \frac{card\ (\,G_i\,)}{card\ (G)} \text{ x } \frac{card\ (case_i(a,b))}{card\ (G_i)} =$$

$$\frac{1}{card\ G} \sum_{i\in[1,N]} card\ (case_i(a,b))$$

Comme les $case_i$(a,b) pour $i \in$ [1,N] forment une partition de la case (a,b).

$$\frac{1}{card\ G} \sum_{i\in[1,N]} card\ (\text{case}_i\ (a,b)) =$$

$$\frac{1}{card\ (G)} card\ (\bigcup_{i\in[1,N]} \text{case}_i\ (a,b)) =$$

$$\frac{1}{card\ G}\ card\ (case(a,b)) = \frac{card(\ case(a,b))}{card\ G} =$$

$$P(\text{case (a,b)})$$

Ce qui termine la démonstration

12. <u>Lemme 4 : polarisation de cases</u>

On suppose qu'on regroupe un nombre S fini de cases différentes de G x G qui forment un ensemble R avec

$$R = \bigcup_{s\in S} case(s).$$

Les ensemble $R_i = R \cap G_i$ sont un ensemble de cases disjointes de G_i x G_i tel que pour tout $i \in$ [1,N] :

$$R_i = \bigcup_{s\in S} case_i(s) \text{ et } \bigcup_{i\in[1,N]} \bigcup_{s\in S} case_i(s) = R$$

Alors on a:

$$P(R) = \sum_{i\in[1,N]} P(G_i).P_{G_i}(R_i)$$

Il suffit de décomposer, R en cases qui sont forcément disjointes ainsi que chaque R_i et d'appliquer le lemme ci-dessus et les règles des probabilités sur la réunion d'ensembles disjoints.

Voici le calcul détaillé :

$$\sum_{i\in[1,N]} P(G_i)P_{G_i}(R_i) = \sum_{i\in[1,N]} P(G_i)P_{G_i}(\bigcup_{s\in S} case_i(s)) =$$

$$\sum_{i\in[1,N]} P(G_i) \sum_{s\in S} P_{G_i}(case_i(s)) = \sum_{s\in S} \sum_{i\in[1,N]} P(G_i)P_{G_i}(case_i(s))$$

D'après le lemme ci-dessus :

$$\sum_{i\in[1,N]} P(G_i)P_{G_i}(case_i(s)) = P\big(case(s)\big)$$

D'où :

$$\sum_{s\in S} \sum_{i\in[1,N]} P(G_i)P_{G_i}(case_i(s))$$

$$= \sum_{s\in S} P\big(case(s)\big) = P(\bigcup_{s\in S} case(s)) = P(R)$$

13. Démonstration N°4

Le but de cette annexe est de montrer qu'il est dangereux de manipuler des résultats au niveau global même si au niveau local, tout semble se présenter d'une façon idéale. C'est pourquoi, un contre-exemple est présenté avant de montrer un résultat conforme à ce qu'on attend sur le plan humain.

Avec les notations ci-dessus on suppose que pour tout i , X_i et Y_i sont indépendants ce qui permet de dire qu'une décision prise sur les bases des résultats des indicateurs $\{X_i\}_{i\in N}$ n'impactera pas les résultats des indicateurs $\{Y_i\}_{i\in N}$.
Alors on se pose la question : quand est-il de l'indépendance de X et Y ?
Malheureusement, en général les indicateurs X et Y ne sont pas indépendants.

Pour prouver cela, il suffit de trouver un contre-exemple.

14. Contre-exemple

On suppose que N = 2 et que H= { a , b }, ou a et b sont des réels distincts.

On construit deux tableaux croisés T_1 cases associées à G_1 et T_2 cases associées à G_2 qui représentent des variables X_1 et Y_2 (resp. X_2 et Y_2) indépendantes. Puis un troisième tableau T, réunion des deux premiers, de façon à avoir une représentation du tableau croisé de X et Y. Le but du contre-exemple est de montrer que le tableau croisé T, ne représente pas des variables de X et Y indépendantes.

Les trois tableaux suivants représentent le nom des croisements de X et Y, sur le groupe G1, G2 et G.

	T_1	
$X_1 = 2$	$case_1$ (1 , 2)	$case_1$ (2 , 2)
$X_1 = 1$	$case_1$ (1 , 1)	$case_1$ (2 , 1)
	$Y_1 = 1$	$Y_1 = 2$

	T_2	
$X_2 = 2$	$case_2$ (1 , 2)	$case_2$ (2 , 2)
$X_2 = 1$	$case_2$ (1 , 1)	$case_2$ (2 , 1)
	$Y_2 = 1$	$Y_2 = 2$

	T	
X = 2	case (1 , 2)	case (2 , 2)
X = 1	case (1 , 1)	case (2 , 1)
	Y = 1	Y = 2

On suppose que les valeurs prises par X_1, X_2 et Y_1, Y_2 sont telles que les cases se remplissent comme ci-dessous :

	T_1	
50	0	50
0	0	0
	0	50

	T_2	
0	0	0
50	50	0
	50	0

Le tableau T se rempli alors de la façon suivante :

	T	
50	0	50
50	50	0
	50	50

On calcule les pourcentages par groupe :

	T_1	
100%	0%	100%
0%	0%	0%
	0%	100%

	T_2	
0%	0%	0%
100%	100%	0%
	100%	0%

	T	
50%	0%	50%
50%	50%	0%
	50%	50%

X_1 et Y_1 (resp. X_2 et Y_2) sont indépendantes comme on le vérifie facilement. Qu'en est-il de X et Y ?

Si c'était le cas chaque case du dernier tableau devraient être les produits des profils. Par exemple la case (1,1) devrait

prendre la valeur 50 %*50 % = 25 %, comme c'est fait dans le tableau théorique ci-dessous :

	T théorique	
50%	25%	25%
50%	25%	25%
	50%	50%

Ce qui est évidemment faux.
L'existence de contre-exemples est donc prouvée.

15. Définitions préalables :

On pose :

$$\alpha_i(a) = P_i(X_i = a) - P(X=a)$$

Et

$$\beta_i(b) = P_i(Y_i = b) - P(Y = b)$$

Les écarts entre la distribution conditionnelle et la distribution globale.

16. Lemme 5 :

Avec les notations ci-dessus on a :

$$\sum_{i\in[1,N]} P(\Omega_i)\,\alpha_i(a) = 0$$

C'est-à-dire la somme des écarts pondérés est nulle.

Preuve :

Remarquons que la famille $\{X_i = a\}_{i\in[1,N]}$ forme une partition de l'ensemble X=a d'où :

On a :

$$\mathrm{P}(\mathrm{X}=\mathrm{a}) = \sum_{\mathrm{i}\in[1,\mathrm{N}]} \mathrm{P}(\mathrm{X}_\mathrm{i}=\mathrm{a}) = \sum_{\mathrm{i}\in[1,\mathrm{N}]} \frac{card(\mathrm{X}_\mathrm{i}=\mathrm{a})}{\mathrm{card}\,\Omega} =$$
$$\sum_{\mathrm{i}\in[1,\mathrm{N}]} \frac{card(\mathrm{X}_\mathrm{i}=\mathrm{a})}{\mathrm{card}\,\Omega_\mathrm{i}} \frac{\mathrm{card}\,\Omega_\mathrm{i}}{\mathrm{card}\,\Omega} = \sum_{\mathrm{i}\in[1,\mathrm{N}]} \mathrm{P}(\Omega_\mathrm{i})\mathrm{P}_i(\mathrm{X}_\mathrm{i}=\mathrm{a})$$

Or :

$$\mathrm{P}_i(\mathrm{X}_\mathrm{i}=\mathrm{a}) = \mathrm{P}\,(\mathrm{X}=\mathrm{a}) + \alpha_\mathrm{i}(\mathrm{a})$$

Donc :

$$\mathrm{P}(\mathrm{X}=\mathrm{a}) =$$
$$\sum_{\mathrm{i}\in[1,\mathrm{N}]} \mathrm{P}(\Omega_\mathrm{i})\mathrm{P}(\mathrm{X}=\mathrm{a}) + \sum_{\mathrm{i}\in[1,\mathrm{N}]} \mathrm{P}(\,\Omega_\mathrm{i})\alpha_\mathrm{i}(\mathrm{a}) =$$
$$\mathrm{P}(\mathrm{X}=\mathrm{a}) \sum_{\mathrm{i}\in[1,\mathrm{N}]} \mathrm{P}(\Omega_\mathrm{i}) + \sum_{\mathrm{i}\in[1,\mathrm{N}]} \alpha_\mathrm{i}(\mathrm{a})\ \mathrm{P}(\,\Omega_\mathrm{i}) =$$
$$\mathrm{P}(\mathrm{X}=\mathrm{a}).1 + \sum_{\mathrm{i}\in[1,\mathrm{N}]} \alpha_\mathrm{i}(\mathrm{a})\ \mathrm{P}(\,\Omega_\mathrm{i}) =$$
$$\mathrm{P}(\mathrm{X}=\mathrm{a}) + \sum_{\mathrm{i}\in[1,\mathrm{N}]} \alpha_\mathrm{i}(\mathrm{a})\ \mathrm{P}(\,\Omega_\mathrm{i})$$

D'où :

$$\sum_{\mathrm{i}\in[1,\mathrm{N}]} \alpha_\mathrm{i}(\mathrm{a})\ \mathrm{P}(\,\Omega_\mathrm{i}) = 0$$

17. <u>Théorème de conservation « approximative » de l'indépendance</u>

Avec les mêmes notations que ci-dessus, on suppose que pour tout i les variables Xi et Yi sont indépendantes. Alors on a :

$$\mathrm{P}(\,\mathrm{X}=\mathrm{a} \cap \mathrm{Y}=\mathrm{b}\,) =$$
$$\mathrm{P}(\mathrm{X}=\mathrm{a})\ \mathrm{P}(\mathrm{Y}=\mathrm{b}) + \sum_{\mathrm{i}\in[1,\mathrm{N}]} \alpha_\mathrm{i}(\mathrm{a})\ \beta_\mathrm{i}(\mathrm{b})\ \mathrm{P}(\Omega_\mathrm{i})$$

Démonstration :

Pour tout i on a :

$$P_i(X_i=a \cap Y_i=b) = P_i(X_i=a)\, P_i(Y_i=b)$$

On remarque que :

$P_i(X_i=a) = P(X=a) + \alpha_i(a)$ et que $P_i(Y_i=b) = P(Y=b) + \beta_i(b)$

D'où en développant

(i) $P_i(X_i=a \cap Y_i=b)$

On obtient les quatre termes suivants :

(ii) $P_i(X_i=a)\, P_i(Y_i=b) = [P(X=a) + \alpha_i(a)]\, [P(Y=b) + \beta_i(b)]$

$= P(X=a) \cdot P(Y=b)] + P(Y=b).\alpha_i(a) + P(X=a) \cdot \beta_i(b) + \alpha_i(a) \cdot \beta_i(b)$

En multipliant les termes de la ligne ci-dessus (ii) par $P(\Omega_i)$ et en faisant la somme sur i on obtient les termes suivants :

Terme 1 :

$$\sum_{i\in[1,N]} P(\Omega_i) \cdot P(X = a) \cdot P(Y = b). = P(X=a) \cdot P(Y=b)$$

Terme 2

$$\sum_{i\in[1,N]} [\, P(Y = b) \cdot P(\Omega_i)\, \alpha_i(a)] =$$
$$P(Y = b) \sum_{i\in[1,N]} [P(\Omega_i). \alpha_i(a)] = 0$$

d'après le lemme précédent

Terme 3

$$\sum_{i\in[1,N]} [\, P(X = a). P(\Omega_i). \beta_i\,] = 0$$

Enfin Terme 4 :

$$\sum_{i\in[1,N]} \alpha_i(a). \beta_i(b). P(\Omega_i)$$

D'où le résultat :

$$\sum_{i\in[1,N]} P(\Omega_i)\,.\,P_i(X_i = a \cap Y_i = b) \;=$$

$$P(X=a)\,.\,P(Y=b) + \sum_{i\in[1,N]} \alpha_i(a).\beta_i\,(b)\;\; P(\Omega_i)$$

D'autre part en multipliant le terme (i) par $P(\Omega_i)$ et en faisant la somme sur i on obtient

$$\sum_{i\in[1,N]} P_i(X_i = a \cap Y_i = b).P(\Omega_i) =$$

$$\sum_{i\in[1,N]} \frac{card(Xi=a\cap Yi=b)}{card(\Omega i)}.\frac{card(\Omega i)}{card(\Omega)} =$$

$$\frac{1}{card(\Omega)} \sum_{i\in[1,N]}\; card(X_i = a \cap Y_i = b) =$$

$$\frac{card(X = a \cap Y = b)}{card(\Omega)} = P(\,X = a \cap Y = b\,)$$

D'où le résultat cherché :

$$P(\,X = a \cap Y = b\,) \;=$$

$$P(X=a)\;P(Y = b) + \sum_{i\in[1,N]} \alpha_i(a).\beta_i\,(b)\;\; P(\Omega_i)$$

18. <u>Remarque</u>

Dans la pratique on sélectionne les questions (variables) dont les distributions conditionnelles $F_{i\in[1,N]}$ sont proches de la distribution globale F. De ce fait, les termes $|\alpha i(a)|$et $|\beta i(b)|$ sont généralement inférieurs à 10%, ce qui entraine que le terme $\left|\sum_{i\in[1,N]} \alpha i(a)\;\beta i(b)\;\; P(\Omega i)\right|$est inférieur à 1% car :

$$\left|\sum_{i\in[1,N]} \alpha i(a)\;\beta i(b)\;\; P(\Omega i)\right| \leq \sum_{i\in[1,N]} |\;\alpha i(a)\;\beta i(b)|\;P(\Omega i) \leq$$

$$\sum_{i\in[1,N]} \frac{1}{100} P(\Omega i) \leq \;\frac{1}{100}$$

Les apports de ce papier justifient au moins partiellement beaucoup de pratiques de l'administration. Cependant, il faut rester prudent comme le montrent le contre-exemple ci-dessus.

D'autres questions peuvent se poser qui demanderaient plus de développement. La plus pertinente est la question sur les corrélations entre variables. Question similaire à la question sur la conservation de l'indépendance. Je laisse à ceux qui désireraient continuer ce travail le soin de le développer.

Je remercie

Je désire remercier Gilbert Saporta, professeur émérite, ex-titulaire de la chaire de statistiques appliquées au Conservatoire National des Arts et Métiers. Il m'a amené à formaliser le cadre théorique des données que j'avais récoltées dans différents établissements.

Merci aussi à :

Xavier Bry	Maître de conférences en statistiques (UM2 Montpellier)
Ali Gannoun	Professeur des Universités (UM2 Montpellier)
Enzo mon petit fils	Étudiant en Master 2 mécanique des solides et des structures : Modélisation et Simulation

qui ont validé la partie mathématiques .

Gabriel Pitiot-Cohen
Auteur
12, rue Léopold Morice - 30900 Nîmes
Gabriel.pitiot@evallib.fr

Date de parution : Janvier 2024