MATHÉMATIQUES ET SCIENCES HUMAINES

H. ROUANET

Note méthodologique. Mesures, proportions et probabilités : sur l'emploi des formulations ensemblistes en inférence statistique

Mathématiques et sciences humaines, tome 80 (1982), p. 83-89

<http://www.numdam.org/item?id=MSH_1982__80__83_0>

Math. Sci. hum. (20e année, n°80, *1982*, p.83-89)

Note méthodologique

MESURES, PROPORTIONS ET PROBABILITES : Sur l'emploi des formulations ensemblistes en inférence statistique

H. ROUANET

Est autem proportionalis similitudo proportionum.

1 - MESURES ET PROPORTIONS

La diversité des mesures qui interviennent en inférence statistique est manifeste. Dans tous les cas, on a au moins deux sortes de mesures : celles qui renvoient à des *données* (ce sur quoi est fondée l'inférence) ; et celles qui renvoient à un *modèle* (ce sur quoi porte l'inférence). En outre, interviennent généralement d'autres mesures, sous forme de "distributions classiques" (Gauss, etc.), ou de distributions d'échantillonnage. On notera également que certaines de ces mesures sont "normalisées" (de masse totale égale à l'unité) alors que d'autres ne le sont nullement : notamment les distributions d'effectifs, qui sont des mesures à valeurs entières.

En dépit (ou à cause ?) de cette diversité, on a souvent l'impression, en lisant les textes de statistique, d'une double tendance, d'abord à normaliser toutes les mesures, ensuite à les désigner indistinctement sous le vocable unique de "probabilité". Cette tendance ne vas pas sans conséquences fâcheuses. Tout d'abord, la normalisation est toujours un appauvrissement - sans parler des mesures qui de par leur nature ne se laissent pas normaliser (mesures infinies). Quant à l'emploi intempestif de la terminologie probabiliste, à propos de mesures qui ne formalisent en rien une incertitude, telles que les distributions d'effectifs ou de fréquences, il suffira de rappeler comment de modestes fréquences (effectifs normalisés) en arrivent à être affublées de l'appellation pompeuse de "probabilités observées" !*

(*) Nous avons évoqué par ailleurs (Rouanet et Lépine, 1976), à propos de Benzécri, qui en a fait une de ses spécialités, tous les bénéfices (rhétoriques) que peut procurer l'usage artificieux de la terminologie probabiliste.

A notre sens, bien des difficultés se trouveront écartées si l'on adopte (ou revient à) la conception de bon sens selon laquelle la probabilité est une notion dont la caractérisation formelle (mesure normalisée) n'épuise par le contenu conceptuel (formaliser l'incertitude). Cette conception amènera à réserver les formulations probabilistes aux situations où elle s'impose vraiment, c'est-à-dire celles où il s'agit de formaliser une incertitude, et à envisager dans les autres cas des formulations plus neutres, plus abstraites et générales, que nous proposons d'appeler des *formulations ensemblistes*. Nous envisagerons dans cette note la mise en oeuvre effective de telles formulations ensemblistes. Nous verrons en particulier comment on peut faire jouer au terme de "proportion" un rôle important, comparable à bien des égards à celui du terme de "probabilité" dans les formulations probabilistes.

Proportion, calcul des proportions

(1) *Dans le langage courant*, proportion désigne un "rapport de grandeur entre les parties d'une chose, entre les parties et le tout" (Littré) : proportion de la hauteur à la largeur d'une façade, proportion d'azote dans dans l'air, etc. En mathématiques, la proportion désigne traditionnellement l'égalité de deux rapports $\left(\frac{a}{b} = \frac{c}{d}\right)$. Remarquons en passant qu'une proportion peut être un nombre irrationnel, témoin le nombre d'or $\left(\frac{\sqrt{5} - 1}{2}\right)$, proportion de la largeur à la longueur d'un certain rectangle (présumé le plus harmonieux).

(2) *En théorie de la mesure*, nous retiendrons, de l'idée courante de proportion, le rapport suivant :

$$\text{proportion} = \frac{\text{mesure d'une partie}}{\text{mesure du "tout"}}$$

Soit maintenant un espace mesurable $\mathcal{Q}$ muni d'une mesure m. Prenons pour "tout" une partie (mesurable) A_o de messa finie non nulle. Si A est une partie (mesurable) de A_o, on définira la notion de proportion (conditionnelle) par le rapport $P(A|A_o) = \frac{m(A)}{m(A_o)}$.

La mesure du TOUT ("le grand tout") pourra être infinie (mesure des longueurs sur la droite, des surfaces dans la plan, etc.). Si elle est finie, la proportion $P(A|\mathcal{Q})$ pourra être notée également P(A), et appelée *proportion de la mesure* m *correspondant à la partie* A. Si la mesure se trouve déjà normalisée ($m(\mathcal{Q}) = 1$), cette proportion n'est autre que la mesure de cette partie A.

Dans le contexte proprement probabiliste, on retrouve naturellement la notion de probabilité conditionnelle posée comme notion première : Cf.Renyi, Suppes, etc.

Le *calcul des proportions*, sur le plan formel, coïncidera avec le calcul des probabilités. On définira de la même façon la notion de *proportion composée* P(B∧A), celle de *proportion conditionnelle* P(A/B), on écrira de même le "théorème des proportions totales", etc. Avantage terre à terre mais appréciable : on pourra *maintenir les écritures familières*, du fait que proportion a la même initiale que probabilité.

2. QUELQUES APPLICATIONS EN INFERENCE STATISTIQUE

Espace d'observation, distributions et variables

Comme notion fondamentale de la statistique on prendra (cf. Rouanet, 1981) celle d'espace d'observation, notion primitive correspondant à l'idée d'ensemble (ou plus généralement espace muni d'une tribu de parties mesurables) "dans lequel on observe". Une distribution (au sens statistique) sera une mesure positive (de masse totale finie) sur un "espace d'observation". Mais il en ira du terme de distribution comme de celui de probabilité : on n'appellera pas "distribution" n'importe quelle mesure sur un espace d'observation (nous en donnerons un exemple plus loin).

(Etant donnée une distribution sur un espace d'observation $\mathcal{U}$ une valeur observable u (élément de $\mathcal{U}$) sera appelée valeur de la distribution, et une classe U (partie mesurable de $\mathcal{U}$) sera appelée classe de la distribution).

Souvent une distribution s'introduit comme mesure-image, à partir d'une application y (à valeurs dans un espace d'observation $\mathcal{U}$) dont l'ensemble de départ est muni d'une pondération P; une telle application y sera appelée une variable (au sens statistique).

Dans la suite du texte :

1) Nous illustrerons les formulations ensemblistes à propos d'une part des distributions d'effectifs, d'autre part des distributions de probabilité.

2) Nous présenterons les formulations ensemblistes que nous proposons comme formulations de rechange à propos des distributions classiques et des distributions d'échantillonnage.

1 - Distribution d'effectifs (et de fréquences)

Une famille d'observations, ou protocole peut être représentée par une application x : I → $\mathfrak{U}$; $\mathfrak{U}$ est un espace d'observation, I ensemble fini sera appelé support du protocole. Si U est une classe de l'espace d'observation, l'effectif de cette classe est le cardinal $|x^{-1}(U)|$ de l'image réciproque de la classe U par l'application x et la distribution des effectifs est la mesure-image, par cette application de la mesure de dénombrement sur le support. La fréquence de la classe U : $|x^{-1}(u)|/|x^{-1}(\mathfrak{U})|$ n'est autre que la proportion des observations dont la valeur appartient à U, on dira encore la proportion des valeurs de la variable-protocole qui appartiennent à la classe U.

N.B.: Le terme de "proportion", employé à propos d'une variable, renverra toujours à la mesure fondamentale définie sur le domaine de la variable (ensemble de départ de l'application), donc ici à la pondération uniforme sur le support du protocole, et jamais à une mesure qui serait définie directement sur l'espace d'observation, indépendamment des données. Ainsi, pour un espace d'observation $\mathfrak{U}$ fini, on ne parlera pas de "proportion des observations" à propos de $|U|/|\mathfrak{U}|$, fraction des valeurs observables qui appartiennent à la classe U.

2 - Distribution de probabilité

Une distribution de probabilité sur un espace d'observation $\mathfrak{U}$ est souvent définie à partir d'un espace fondamental Ω (espace mesurable) muni d'une mesure normalisée P dite de probabilité, et d'une variable dite aléatoire (application mesurable) : y : $\Omega \to \mathfrak{U}$; la probabilité de la classe U n'est autre que la proportion des valeurs de la variable aléatoire qui appartiennent à U.

Une écriture susceptible de plusieurs "lectures" : "langage des probabilités" et"langage des proportions". - On connaît, en calcul des probabilités, la convention d'écriture classique qui consiste à substituer, à $P(y^{-1}(U))$, l'écriture $P(y \in U)$, avec la lecture : "la probabilité que la variable y appartienne à U". Une telle écriture pourra être adoptée dans les formulations ensemblistes, avec la nouvelle "lecture" en termes de proportion : "la proportion des valeurs de la variable y qui appartiennent à U" ou encore : "la proportion des termes de la variable y dont la valeur appartient à U".

3 - Distributions statistiques classiques

Les propriétés des distributions statistiques classiques (Gauss, etc), font traditionnellement l'objet de formulations probabilistes; exemple : "pour une distribution normale réduite, la probabilité d'une valeur supérieure à 1,96 est égale à 0,025", etc. La formulation ensembliste correspondante sera la

la suivante : "Pour une distribution normale réduite, la proportion des valeurs supérieures à 1,96 est égale à 0,025"(*).

Souvent les propriétés de la distribution sont exprimées à l'aide de la variable "application identique de l'espace d'observation sur lui-même"; on pourra alors reprendre les écritures familières telles que la suivante : $P(z>1,96) = 0,025$ (z désignant une variable normale réduite), avec la nouvelle lecture : "la proportion des valeurs de z supérieures à 1,96 est égale à 0,025.

Les formulations ensemblistes apparaîtront préférables chaque fois que les distributions classiques interviennent en tant que "distributions abstraites", par exemple lors de leur définition. Dans M.-P. et B. Lecoutre (1979) et Rouanet (1981), on a systématiquement fait usage des formulations ensemblistes, en termes de proportions, pour la présentation des distributions statistiques classiques.

4 - Distribution d'échantillonnage

Traditionnellement, les notions relatives à l'échantillonnage font également l'objet de formulations probabilistes, souvent sous le couvert du fameux "schéma d'urne". Par exemple, pour définir la distribution binomiale, on considère une population de N individus, dont A possèdent un certain caractère; on introduit l'idée "d'échantillonage au hasard avec remise", et on en déduit la distribution d'échantillonnage de la statistique Fréquence F du caractère dans un échantillon de taille n : $P(F=k/n) = \binom{n}{k} (\frac{A}{N})^k (\frac{N-A}{N})^{n-k}$ en lisant l'écriture P(F=k/n) : "la probabilité que la Fréquence F soit égale à $\frac{k''}{n}$.

Pour redéfinir la distribution binomiale selon des formulations ensemblistes, il suffira de définir un échantillon comme une application d'un ensemble à n éléments (le support d'un protocole) dans un ensemble à N éléments (la population). Le nombre des échantillons (i.e. applications pour lesquelles la Fréquence est égale à $\frac{k}{n}$) est égal à $\binom{n}{k} A^k (N-A)^{n-k}$; d'où, puisque le nombre total d'échantillons est N^n, la proportion $P(F=\frac{k}{n}) = \binom{n}{k} (\frac{A}{N})^k (\frac{N-A}{N})^{n-k}$ de ceux pour lesquels F est égal à $\frac{k}{n}$.

1) Pour la distribution binomiale, ainsi que pour l'hypergéométrique(**)

(*) En général, ces proportions ne seront plus, comme les fréquences, des nombres rationnels, mais, on l'a vu plus haut à propos du nombre d'or, il n'y a, à la réflexion, aucune raison d'exiger d'une proportion qu'elle soit un nombre rationnel.

(**) Pour l'hypergéométrique, les échantillons pourront être définis encore plus simplement, comme les parties de même taille n de la population.

etc. l'espace des échantillons est fini et la proportion P est le rapport d'un nombre d'échantillons au nombre total des échantillons. Dans le cas des modèles d'échantillonnage plus généraux, chers à la "statistique mathématique", l'espace des échantillons sera quelconque et le nombre d'échantillons infini, mais rien ne s'oppose à l'extension du langage ensembliste. Notamment, la notion classique d'échantillon de taille n extrait "au hasard" d'une distribution de probabilité π sur un espace d'observation $\mathcal{U}$ se réduit formellement au concept de mesure (normalisée) "équiproduit", construite à l'aide de π sur l'espace d'observation produit $\mathcal{U}^n$; on peut donc définir la notion purement ensembliste d'*échantillon-équiproduit*, qu'on pourra, sauf ambiguïté, appeler en bref, échantillon de la distribution π. Ainsi on pourra parler d'un échantillon $(X_1, X_2, \ldots, X_n)$ d'une distribution gaussienne de moyenne μ et de variance σ^2, etc. Si M désigne la moyenne (statistique) d'un tel échantillon, on pourra, en désignant par P la mesure équiproduit correspondante, maintenir les écritures familières, avec la lecture en termes de proportion. Exemple :

$P(\frac{|M-\mu|}{\sigma/\sqrt{n}} > z_\alpha = \alpha$: "la proportion des échantillons pour lesquels la variable $\frac{|M-\mu|}{\sigma/\sqrt{n}}$ est supérieure à z_α (valeur critique bilatérale de la distribution normale réduite au seuil α) est égale à α ", etc.

Quelques commentaires en guise de conclusion

Dans la situation d'échantillonnage dans une population finie, les formulations ensemblistes "coulent de source" et c'est sans surprise que nous les trouvons dans les textes centrés sur la statistique descriptive et qui mettent en avant les préoccupations algébriques : v. notamment l'exercice intitulé "Echantillonnage" dans le chapitre Dénombrements de Jullien et Leclerc, du Tome 2 de Barbut et al. (1974).

En revanche, dans les manuels traditionnels d'inférence statistique, nous n'avons pu, à part l'exception notable de Faverge (1956), discerner les formulations ensemblistes qu'à l'état de traces (*), ce qui tout de même est assez curieux.

2) Les questions abordées dans cette note se réduisent-elles à des "questions de langage" ? Il nous paraît risqué de le soutenir, dès lors que les formulations portent sur des situations tant soit peu complexes. Reprenons le cas exemplaire de la distribution binomiale. L'énoncé ensembliste : "la proportion des échantillons pour lesquels la fréquence est égale à k/n, etc." est valide en général, tandis que l'énoncé probabiliste correspondant,

(*) Toute notre reconnaissance s'adresse d'avance à ceux qui nous signaleraient d'autres références.

lequel revient à "convertir" les proportions en probabilités, exige l'hypothèse supplémentaire d'une probabilisation uniforme des échantillons. On voit donc mal comment le passage du langage probabiliste au langage ensembliste pourrait être considéré comme une simple "traduction", puisqu'il s'accompagne d'un gain de généralité.

3) Dans l'utilisation concrète de l'inférence statistique, l'avantage des formulations ensemblistes apparaîtra décisif chaque fois que les données examinées n'ont aucune raison, objective ou subjective, d'être regardées comme résultant d'un "échantillonnage au hasard". En pareil cas, les formulations probabilistes à propos des procédures inférentielles (probabilité de "se tromper" en rejetant l'hypothèse nulle, etc.)ne sont alors guère moins fictives qu'à propos des distributions de la statistique descriptive. Cette remarque ne saurait disqualifier, à notre sens, l'usage de ces procédures, pourvu qu'il soit bien entendu que leur véritable finalité est alors de *situer les observations* par rapport à des données potentielles engendrées par un certain modèle, sans qu'il soit vraiment question de formaliser une incertitude; en l'occurrence, l'emploi des formulations ensemblistes permettra de marquer clairement le statut véritable de ces procédures.

(Nous nous proposons de revenir sur ce sujet dans un texte à venir).

BIBLIOGRAPHIE

BARBUT, M., D'ADHEMAR, C., LECLERC, B., JULLIEN, P., *"Mathématiques élémentaires, applications à la statistique et aux sciences sociales"*, Paris, Presses Universitaires de France, 1974.

FAVERGE, J.-M., *Méthodes statistiques en psychologie appliquée*, Paris, Presses Universitaires de France, 1956.

LECOUTRE, M.-P., LECOUTRE B., *Enseignement programmé sur l'utilisation d'une table de distribution normale*, Paris, C.D.U.-S.E.D.E.S., 1979.

ROUANET, H., *Documents pour l'enseignement de statistique au Certificat C1 de Psychologie : procédures statistiques et fondamentales*, texte ronéoté, UER de Mathématiques, Université René Descartes, 1981.

ROUANET, H., LEPINE D., *A propos de l' "Analyse des Données" selon Benzécri*, (suivi d'une "lettre de Commentaires" de J.P. Benzécri), *Année Psychologique*, 76, 133-144.