REVUE DE STATISTIQUE APPLIQUÉE

MICHÈLE NEUILLY

Comparaison d'histogrammes expérimentaux

Revue de statistique appliquée, tome 41, n° 4 (1993), p. 73-96

<http://www.numdam.org/item?id=RSA_1993__41_4_73_0>

Rev. Statistique Appliquée, 1993, **XLI** (4), 73-95

COMPARAISON D'HISTOGRAMMES EXPÉRIMENTAUX (1)

Michèle Neuilly

2A place Général de Gaulle, 13100 Aix-en-Provence

RÉSUMÉ

Les exemples qui viennent les premiers à l'esprit pour la comparaison d'histogrammes expérimentaux sont souvent d'ordre médical (p.ex. distribution des différentes catégories de fumeurs dans une population de cancéreux et dans une population saine). Mais des applications importantes existent également dans le domaine des mesures, par exemple comparaison d'un spectre gamma à celui d'une substance de référence ou des spectres gamma de deux échantillons entre eux. C'est pour ce type d'applications que J. Dorlet a proposé en 1975 un test utilisant une fonction qu'il a appelée la distance entre deux histogrammes.

Si n_{i1} et n_{i2} sont les nombres d'observations dans la $i$ème classe, respectivement pour la première et la seconde population, Dorlet définit l'angle θ par la relation :

$$\cos\theta = \frac{\Sigma n_{i1}\, n_{i2}}{\sqrt{\Sigma n_{i1}^2 \times \Sigma n_{i2}^2}}$$

et la distance d par :

$$d = 2\sin\frac{\theta}{2} = \sqrt{2(1-\cos\theta)}$$

L'étude de la distribution de cette fonction n'a pas été faite par Dorlet. Dans le présent article, la moyenne, la variance et les coefficients de Pearson β_1 et β_2 de la variable d^2 sont donnés sous la forme du terme principal d'un développement en série de $1/N$, N étant le nombre total d'observations. Deux cas possibles ont été envisagés :

- l'un des histogrammes correspond à une population théorique (probabilités p_i connues dans chaque classe),
- les deux histogrammes sont expérimentaux et établis à partir de nombres N_1 et N_2 d'observations du même ordre de grandeur, mais qui peuvent être différents (on verra que le cas où N_1 est faible par rapport à N_2 peut se ramener au cas précédent).

Les caractéristiques de la distribution de d^2 sont des fonctions des probabilités p_i, ce qui ne pose pas de problème dans le premier cas considéré. Mais, si l'on compare deux histogrammes expérimentaux, les probabilités p_i doivent être estimées en regroupant les

(1) Ce travail a été réalisé en partie au C.E.A. (Commissariat à l'Energie Atomique), Département de Sécurité des Matières Radioactives, Centre d'Etudes de Fontenay-aux-Roses, B.P. 6, 92265 Fontenay-aux-Roses.

observations dans un histogramme unique (histogramme total). On admet que l'erreur introduite par cette estimation est du même ordre de grandeur que l'approximation initiale (développement en $1/N$ limité au terme principal).

Les caractéristiques de la distribution de d^2 permettent de l'assimiler à une distribution du type VI de Pearson (fonction linéaire d'une variable F de Snedecor), donc de calculer le seuil de décision d'un test de comparaison à l'aide des tables usuelles.

Le test de la distance peut également être utilisé pour comparer entre eux plus de deux histogrammes expérimentaux. Si q est le nombre d'histogrammes, d_j la distance de l'un d'entre eux à l'histogramme total, μ_j l'espérance mathématique de d_j^2 et σ_j son écart-type, la fonction discriminante du test est :

$$R = \frac{1}{q}\Sigma\left[\frac{d_j^2 - \mu_j}{\sigma_j}\right]^2$$

Les caractéristiques de la distribution de R ont été calculées en fonction de celles des carrés des distances d_j, ce qui permet d'assimiler R à une variable de type VI de Pearson et de déterminer le seuil de décision du test.

Des applications numériques tirées de la littérature sont données pour chaque test proposé.

Mots-clés : *Loi Multinomiale, Comparaisons.*

SUMMARY

Examples of the comparison of experimental histograms which come to mind first of all are often of a medical kind (such as the distribution of different categories of smoker in a cancerous population and in a healthy one). However, important applications can also be found in the measurement field, for example the comparison of a gamma spectrum with that of a reference substance, or between the gamma spectra of two samples.

It was for this type of application that J. Dorlet proposed, in 1975, a test using a function which he called the distance between two histograms.

If n_{i1} and n_{i2} are the number of observations in the 1st class, for the first and second populations respectively, Dorlet defines the angle θ by the relationship :

$$\cos\theta = \frac{\Sigma n_{i1}\, n_{i2}}{\sqrt{\Sigma\, n_{i1}^2 \times \Sigma n_{i2}^2}}$$

and the distance by :

$$d = 2\sin\frac{\theta}{2} = \sqrt{2(1-\cos\theta)}$$

The study of the distribution of this function was not performed by Dorlet. In the present article, the mean, the variance and the Pearson coefficients β_1 and β_2 for the variable d^2 are given in the form of the main term of a series development of $1/N$, where N is the total number of observations. Two possible cases have been considered :

- one of the histograms corresponds to a theoretical population (probabilities p_i known in each class),
- the two histograms are experimental and are determined from numbers of observations N_1 and N_2, of the same order of magnitude, but which may be different (it will be seen that, if N_1 is small compared to N_2, the case falls within the range of the previous one).

The characteristics of the distribution of d^2 are functions of the probabilities p_i. It is not a problem in the first case considered. However, if two experimental histograms are compared, the probabilities p_i must be estimated by collecting all the observations into a single histogram (total histogram). The error introduced by this estimate is supposed of the same order of magnitude as the initial approximation (development in $1/N$ restricted to the main term).

The characteristics of the distribution of d^2 allow it to be assimilated to a type VI Pearson distribution (linear function of a Snedecor F variable), and thus the decision threshold of a comparison test can be calculated from the usual tables.

The test of the distance may also be used to compare more than two experimental histograms with one another. If q is the number of histograms, d_j the distance from one of them to the total histogram, μ_j the mathematical expectation of d_j^2 and σ_j its standard error, the discriminating function of the test is :

$$R = \frac{1}{q}\Sigma\left[\frac{d_j^2 - \mu_j}{\sigma_j}\right]^2$$

The characteristics of the distribution of R were calculated with respect to those of the squares of the distances d_j, thus allowing it to be assimilated to a type VI Pearson variable and the decision threshold of the test to be determinated.

Numerical applications, taken from the references, are given for each test proposed.

Key-words : *Multinomial distribution, Comparisons.*

1. Introduction

Les comparaisons d'histogrammes expérimentaux sont généralement exécutées à l'aide du test de «Chi-2», en utilisant des «tableaux de contingence» plus ou moins compliqués [2] [3] [4]. Ces techniques, mises au point pour le dépouillement de données biologiques ou médicales pourraient s'appliquer aux problèmes d'exploitation de résultats de mesure, par exemple la comparaison d'un spectre gamma à celui d'une substance de référence ou des spectres gamma de deux échantillons entre eux.

Toutefois, pour ce dernier type d'application, J. Dorlet a proposé en 1975 un test différent utilisant une fonction d qu'il a appelée la distance entre deux histogrammes [1]. L'utilisation de ce test est restée très limitée car la distribution de d n'avait pas été calculée jusqu'à présent. C'est l'objet du travail présenté ici.

Dans un premier temps, le calcul a porté sur la distance entre un histogramme expérimental et une population théorique (probabilités connues). Le test a ensuite été étendu à la comparaison d'histogrammes expérimentaux.

2. Comparaison d'un histogramme à une population théorique

2.1 Définition

Si n_i est le nombre d'observations situées dans la $i$ème classe de l'histogramme et p_i la probabilité correspondante, Dorlet définit l'angle θ par relation :

$$f = \cos\theta = \frac{\Sigma n_i p_i}{\sqrt{\Sigma n_i^2 \cdot \Sigma p_i^2}} \tag{1}$$

et la distance d par :

$$d = 2\sin\frac{\theta}{2} = \sqrt{2(1-f)} \tag{2}$$

Si, N étant le nombre total d'observations, chaque nombre n_i était égal à Np_i, la valeur de f serait égale à l'unité et celle de d à zéro.

En réalité, les variables n_i obéissent à une loi multinomiale, chaque variable ayant pour espérance la valeur Np_i. Les moments centrés de la distribution sont donnés dans la référence [2] jusqu'à l'ordre 4. Pour le calcul à faire ici, il fallait disposer des moments suivants jusqu'à l'ordre 8. Ce calcul a été fait en suivant la technique proposée par la référence [2]. Si le lecteur désire en avoir les résultats, l'auteur se tient à sa disposition pour les lui communiquer.

2.2 Caractéristiques de la distribution de d^2

Le calcul étant déjà assez compliqué, il a été limité à la détermination du terme principal du développement en $1/N$ représentant chaque moment de la fonction d^2. En effet, les études d'histogrammes ne sont intéressantes qui si N est assez grand, ce qui permet de limiter le developpement.

La variable n_i a été remplacée par la variable centrée :

$$x_i = \frac{n_i - Np_i}{N}$$

Les moments d'ordre 2 de ces variables sont proportionnels à $1/N$, les moments d'ordre 3 et 4 proportionnels à $1/N^2$, ceux d'ordre 5 et 6 proportionnels à $1/N^3$, ceux d'ordre 7 et 8 proportionnels à $1/N^4$. On peut donc utiliser des développements limités en fonction des x_i.

En utilisant la relation (2), on obtient, pour le terme principal du carré de la distance :

$$d^2 \sim \frac{\Sigma x^2}{\Sigma p^2} - \frac{(\Sigma px)^2}{(\Sigma p^2)^2} \tag{3}$$

L'espérance mathématique de cette quantité est de la forme a_1/N :

Pour avoir a_1, il faut calculer, à partir des moments centrés de n_i, les espérances mathématiques de $[\Sigma x^2]$ et $(\Sigma px)^2$:

$$E\left[\Sigma x^2\right] = \frac{1}{N}\,\Sigma p_i(1-p_i) = \frac{1}{N}\left[1-\Sigma p^2\right]$$
$$E\left[(\Sigma px)^2\right] = E\left[\Sigma p_i^2 x_i^2 + 2\,\Sigma\Sigma p_i p_j x_i x_j\right]$$
$$= \frac{1}{N}\left[\Sigma p_i^3(1-p_i) - 2\,\Sigma\Sigma p_i^2 p_j^2\right] = \frac{\Sigma p^3 - (\Sigma p^2)^2}{N}$$

On obtient :

$$a_1 = \frac{1}{(\Sigma p^2)^2}\left[\Sigma p^2 - \Sigma p^3\right] \tag{4}$$

La variance de d^2 est calculée par la formule :

$$\sigma^2 = E(d^4) - \mu^2, \text{ avec } \mu = E(d^2)$$

Le terme principal de d^4 est, d'après (3), de la forme :

$$d^4 \sim \frac{(\Sigma x^2)^2}{(\Sigma p^2)^2} - \frac{2(\Sigma x^2)(\Sigma px)^2}{(\Sigma p^2)^3} + \frac{(\Sigma px)^4}{(\Sigma p^2)^4}$$

Il faut donc calculer les espérances mathématiques de $(\Sigma x^2)^2$, $(\Sigma x^2)(\Sigma px)^2$ et $(\Sigma px)^4$. Pour la deuxième de ces quantités, par exemple, on écrit :

$$(\Sigma x^2)(\Sigma px)^2 = \Sigma p_i^2 x_i^4 + \Sigma\Sigma p_i^2\, x_i^2 x_j^2 + 2\Sigma\Sigma p_i p_j x_i^3 x_j$$
$$+ 2\,\Sigma\Sigma\Sigma p_j p_k x_i^2 x_j x_k$$

L'espérance mathématique de cette quantité est calculée à l'aide des moments de x (cf. annexe) :

$$E\left[(\Sigma x^2)(\Sigma px)^2\right] \sim \frac{1}{N^2}\left[3\,\Sigma p^4 - 6\,\Sigma p^5 + 3\,\Sigma p^6 + \Sigma\Sigma p_i^3 p_j - \Sigma\Sigma p_i^4 p_j\right.$$
$$\left.-7\,\Sigma\Sigma p_i^3 p_j^2 + 9\,\Sigma\Sigma p_i^4 p_j^2 - 2\,\Sigma\Sigma\Sigma p_i p_j^2 p_k^2 + 18\,\Sigma\Sigma\Sigma p_i^2 p_j^2 p_k^2\right].$$

Pour se ramener aux sommes des puissances de p, les sommes triples sont transformées à l'aide des identités suivantes (valables avec Σp égal à 1) :

$$2\,\Sigma\Sigma\Sigma p_i p_j^2 p_k^2 \equiv (\Sigma p^2)^2 - \Sigma p^5 - \Sigma\Sigma p_i^4 p_j - 2\,\Sigma\Sigma p_i^3 p_j^2$$
$$6\,\Sigma\Sigma\Sigma p_i^2 p_j^2 p_k^2 \equiv (\Sigma p^2)^3 - \Sigma p^6 - 3\,\Sigma\Sigma p_i^4 p_j^2$$

D'autres identités sont ensuite utilisées pour transformer les sommes doubles.

La variance de d^2 est de la forme $\sigma^2 \sim \frac{a_2}{N^2}$, avec :

$$a_2 = \frac{2}{(\Sigma p^2)^4}\left[(\Sigma p^3)^2 - 2\,\Sigma p^4 \cdot \Sigma p^2 + (\Sigma p^2)^3\right] \tag{5}$$

Des calculs analogues ont été exécutés pour obtenir $E(d^6)$ et $E(d^8)$, d'où les moments centrés de d^2 :

$$\mu_3 = E(d^6) - 3\mu\sigma^2 - \mu^3$$
$$\mu_4 = E(d^8) - 4\mu\mu_3 - 6\mu^2\sigma^2 - \mu^4$$

Le moment d'ordre 3 de d^2 est de la forme a_3/N^3 avec :

$$a_3 = \frac{8}{(\Sigma p^2)^6}\left[3\Sigma p^4 \cdot \Sigma p^3 \cdot \Sigma p^2 - 3\Sigma p^5(\Sigma p^2)^2 + \Sigma p^3 \cdot (\Sigma p^2)^3 - (\Sigma p^3)^3\right] \tag{6}$$

Le moment d'ordre 4 de d^2 est de la forme :

$$\mu_4 \sim 3\,\sigma^4 + \frac{a_4}{N^4} = \frac{3\,a_2^2 + a_4}{N^4} \tag{7}$$

avec :

$$a_4 = \frac{48}{(\Sigma p^2)^8}\left[(\Sigma p^3)^4 - 4\,\Sigma p^4(\Sigma p^3)^2\Sigma p^2 - 4\,\Sigma p^6(\Sigma p^2)^3\right.$$
$$\left.+\Sigma p^4(\Sigma p^2)^4 + 2(\Sigma p^4)^2(\Sigma p^2)^2 + 4\,\Sigma p^5\,\Sigma p^3(\Sigma p^2)^2\right] \tag{8}$$

Les quantités a_1, a_2, a_3 et a_4 peuvent être calculées si l'on connaît par hypothèse les probabilités (cf. exemple 5.1. de contrôle d'une loi équiprobable). Souvent ces probabilités doivent être estimées. Les classes de l'histogramme doivent alors être correctement définies de façon que les estimations p_i ne soient liées que par une seule relation (somme égale à l'unité). On verra un exemple au paragraphe 5.3.

Les coefficients de Pearson sont indépendants du nombre N d'observations :

$$\beta_1 = \frac{a_3^2}{a_2^3} \text{ et } \beta_2 = 3 + \frac{a_4}{a_2^2}$$

2.3 Forme de la distribution de d^2

La distribution de d^2 a été assimilée à une distribution de Pearson de façon à pouvoir utiliser les tables courantes.

L'assimilation est faite en identifiant les quatre premiers moments des deux lois. Le type de loi de Pearson est choisi suivant la valeur de :

$$k = \frac{\beta_1(\beta_2+3)^2}{4(4\beta_2-3\beta_1)(2\beta_2-3\beta_1-6)}$$

Si ce coefficient est supérieur à l'unité, la loi de Pearson est de type VI, c'est-à-dire de la forme :

$$d^2 = \frac{1}{N}\left[a_0 + a_1 + \frac{w\nu_1}{\nu_2}F(\nu_1,\nu_2)\right] \tag{9}$$

où F a la même forme que la variable de Fisher-Snedecor correspondant à ν_1 et ν_2 degrés de liberté, mais ν_1 et ν_2 ne sont pas forcément des nombres entiers. Le calcul des paramètres de la formule (9) est classique [2] [3]. Il faut calculer :

$$r = \frac{6(\beta_2-\beta_1-1)}{3\beta_1-2\beta_2+6}$$

Puisque β_1 est positif par définition, on voit que k et la quantité $r+1 = \frac{4\beta_2-3\beta_1}{3\beta_1-2\beta_2+6}$ sont toujours de signes opposés.

Les paramètres de (9) sont donnés par :

$$\begin{aligned}
\nu_1 &= r\left[1-\sqrt{\frac{k}{k-1}}\right]\\
\nu_2 &= 2(1-r)\\
w &= 2\sqrt{a_2(r+1)(1-k)}\\
&\text{où } (r+1) \text{ et } (1-k) \text{ sont tous deux négatifs}\\
a_0 &= \frac{\nu_1 w}{2r}
\end{aligned}$$

2.4 Cas particuliers

Deux cas limites ont été envisagés, celui de la loi équiprobable parce qu'il conduit à des résultats très simples, et celui où l'une des probabilités p_0 est voisine de l'unité parce qu'il permet de vérifier les calculs des paragraphes précédents.

Si l'histogramme comporte $(K+1)$ classes correspondant toutes à la même probabilité, celle-ci a pour valeur :

$$p = (K+1)^{-1}$$

Les formules des paragraphes 2.1. et 2.2. se simplifient :

$$f = \frac{p\Sigma n_i}{\sqrt{\Sigma n_i^2 \cdot (K+1)p^2}} = \frac{N}{\sqrt{(K+1)\,\Sigma n_i^2}}$$

Les variables x_i sont égales à :

$$x_i = \frac{n_i}{N} - p$$

Donc :

$$\Sigma x_i^2 = \frac{\Sigma n_i^2}{N^2} + (K+1)p^2 - 2p = \frac{\Sigma n_i^2}{N^2} - \frac{1}{K+1}$$

et $f = \dfrac{1}{\sqrt{(K+1)\Sigma x_i^2 + 1}}$

Donc :

$$d^2 = 2(1-f) \sim (K+1)\Sigma x_i^2 = (K+1)\Sigma \frac{(n_i - Np)^2}{N^2} = \frac{1}{N}\Sigma \frac{(n_i - Np)^2}{Np}$$

On retrouve, au facteur $1/N$ près, la statistique du test de «Chi-2» :

$$d^2 = \frac{1}{N}\chi^2(K)$$

On peut vérifier que les formules (4) à (8) donnent bien les caractéristiques de cette distribution.

Dans le second cas envisagé, la probabilité p_0 est voisine de l'unité, et les autres probabilités p_i (i positif) faibles devant p_0. On peut faire le calcul dans un cas analogue à celui de la loi de Poisson, c'est-à-dire en supposant que p_i tend vers zéro et N vers l'infini de façon que le produit Np_i reste fini. Dans ce cas, on voit, en désignant par K le nombre de valeurs de i supérieures à zéro [$(K+1)$ classes], que la relation (2) devient :

$$d^2 = \frac{1}{N^2}\sum_1^K (n_i - Np_i)^2$$

D'autre part, la forme de la densité de probabilité de la loi multinomiale montre que ce cas correspond à la distribution de K variables de Poisson indépendantes. On peut donc recalculer les caractéristiques de d^2 en utilisant cette loi de distribution.

Les résultats sont les suivants :

$$a_1 = \sum_1^K p_i$$

$$a_2 = 2\sum_1^K p_i^2$$

$$a_3 = 8\sum_1^K p_i^3$$

$$a_4 = 48\sum_1^K p_i^4$$

On peut retrouver ces résultats à partir des formules (4) à (8) en remarquant que :

$$\sum_0^K p_i^m = p_0^m + \sum_1^K p_i^m = \left(1 - \sum_1^K p_i\right)^m + \sum_1^K p_i^m$$

$$= \sum_1^K p_i^m + 1 - m\sum_1^K p_i + \frac{m(m-1)}{2}\left(\sum_1^K p_i\right)^2 + \ldots$$

où le développement peut être limité puisque les probabilités p_i sont petites devant l'unité. Ce calcul confirme la validité des résultats dans le cas général.

3. Comparaison de deux histogrammes expérimentaux

3.1 Définition de la distance

Le cas étudié est celui de deux histogrammes correspondant respectivement aux nombres N_1 et N_2 d'observations. Ces nombres doivent être du même ordre de grandeur, c'est-à-dire que $1/N_1^2$ doit être négligeable devant $1/N_2$ et $1/N_2^2$ négligeable devant $1/N_1$.

L'angle θ est défini par :

$$f = \cos\theta = \frac{\Sigma n_{i1} n_{i2}}{\sqrt{\Sigma n_{i1}^2 \times \Sigma n_{i2}^2}} \qquad (10)$$

où n_{i1} et n_{i2} sont les nombres d'observations dans la $i$ème classe de chaque histogramme. La distance d est encore donnée par la relation (2).

3.2 Caractéristiques de la distribution de d^2

Les variables n_{i1} et n_{i2} sont remplacées par les variables centrées :

$$x_i = \frac{n_{i1} - N_1 p_i}{N_1} \qquad y_i = \frac{n_{i2} - N_2 p_i}{N_2}$$

où p_i est la probabilité correspondant à la $i$ème classe, supposée commune aux deux populations.

Un calcul analogue à celui du paragraphe 2.2. donne :

$$d^2 \sim \frac{\Sigma(x_i - y_i)^2}{\Sigma p_i^2} - \left[\frac{\Sigma p_i(x_i - y_i)}{\Sigma p_i^2}\right]^2 \tag{11}$$

Si d_x représente la distance du premier histogramme à la population théorique, d_x et d sont données par des relations de même forme, la première en fonction des x_i et la seconde en fonction des quantités :

$$z_i = x_i - y_i$$

Le calcul des moments centrés d'ordre 2, 4, 6 et 8 de la variable z_i a été fait à l'aide des moments des variables x_i et y_i. On constate qu'ils sont obtenus à partir de ceux de la variable x_i en remplaçant $1/N$ par :

$$\frac{1}{N_1} + \frac{1}{N_2}$$

Les calculs faits pour d_x sont donc utilisables pour la variable d.

L'espérance mathématique de d^2 est :

$$\mu \sim a_1 \left[\frac{1}{N_1} + \frac{1}{N_2}\right] \tag{12}$$

où a_1 est donné par la relation (4).

La variance de d^2 est :

$$\sigma^2 \sim a_2 \left[\frac{1}{N_1} + \frac{1}{N_2}\right]^2 \tag{13}$$

où a_2 est donné par la relation (5).

En particulier, si N_1 et N_2 ont la même valeur N, on voit que :

$$\mu \sim \frac{2a_1}{N} = 2\mu_x$$

$$\sigma \sim \sqrt{\frac{4a_2}{N^2}} = 2\sigma_x$$

Le moment d'ordre 3 de d^2 est :

$$\mu_3 \sim a_3 \left[\frac{1}{N_1} + \frac{1}{N_2}\right]^3 \tag{14}$$

où a_3 est donné par la relation (6).

Le moment d'ordre 4 de d^2 est de la forme :

$$\mu_4 \sim 3\sigma^4 + a_4 \left[\frac{1}{N_1} + \frac{1}{N_2}\right]^4 = (3a_2^2 + a_4)\left[\frac{1}{N_1} + \frac{1}{N_2}\right]^4 \tag{15}$$

où a_4 est donné par la relation (8).

Les coefficients de Pearson de d^2 sont donc les mêmes que ceux de d_x^2 :

$$\beta_1 = \frac{a_3^2}{a_2^3} \qquad \beta_2 = 3 + \frac{a_4}{a_2^2}$$

Pour utiliser les relations (12) à (15), il faudrait connaître les probabilités p_i. Celles-ci seront estimées par :

$$p_i = \frac{n_{i1} + n_{i2}}{N_1 + N_2}$$

Puisque la somme des n_{i1} est égale à N_1 et celle des n_{i2} égale à N_2, la somme des p_i est égale à l'unité. S'il y a $(K + 1)$ classes, les estimations p_i forment un ensemble à K degrés de liberté. On verra dans l'exemple numérique 5.2. que la définition des classes à comparer doit tenir compte de cette contrainte.

Nous avons admis que l'erreur introduite par l'estimation des p_i est du même ordre de grandeur que celle faite en limitant les développements en $1/N$ à leur terme principal.

Puisque les coefficients de Pearson sont les mêmes que ceux de d_x^2, la forme de la distribution est du même type et les formules du paragraphe 2.3 sont applicables.

4. Comparaison de q distributions

Si l'on dispose de q histogrammes (repérés par l'indice j), on peut vouloir vérifier s'ils correspondent tous à la même loi de probabilité.

Les probabilités, supposées communes aux q histogrammes, sont calculées par la relation :

$$p_i = \frac{\sum_j n_{ij}}{\sum_j N_j}$$

où N_j est le nombre d'observations dans le $j$ème histogramme. On suppose tous les N_j du même ordre du grandeur (c'est-à-dire $1/N_k^2$ négligeable devant $1/N_j$ quels que soient j et k).

Les valeurs des p_i permettent de calculer les coefficients a_1, a_2, a_3 et a_4 communs aux q populations.

On en déduit :

– l'espérance de d_j^2 : $\mu_j = \frac{a_1}{N_j}$,

– la variance de d_j^2 : $\sigma_j^2 = \frac{a_2}{N_j^2}$,

– les valeurs de β_1 et β_2 communes à tous les histogrammes.

La statistique du test proposé pour la comparaison est :

$$R = \frac{1}{q}\Sigma\left[\frac{d_j^2 - \mu_j}{\sigma_j}\right]^2 \tag{16}$$

Son espérance mathématique est 1 si tous les histogrammes correspondent aux mêmes probabilités.

Les caractéristiques de R sont calculées dans ce cas :

$$\text{Var}(R) = \frac{1}{q}(\beta_2 - 1) \tag{17}$$

$$\beta_1(R) = \frac{(\beta_4 - 3\,\beta_2 + 2)^2}{q(\beta_2 - 1)^3} \tag{18}$$

$$\beta_2(R) = 3 + \frac{\beta_6 - 4\,\beta_4 + 3\,\beta_2^2 + 12\,\beta_2 - 6}{q(\beta_2 - 1)^2} \tag{19}$$

avec :

$$\beta_4 = \frac{\mu_6}{\sigma^6} \qquad \beta_6 = \frac{\mu_8}{\sigma^8}$$

(ces coefficients sont les mêmes pour tous les histogrammes).

Les valeurs de β_4 et β_6 peuvent être calculées dès que la loi de d_j^2 a été mise sous la forme (9) :

$$d_j^2 = \frac{1}{N_j}\left[a_0 + a_1 + \frac{w\,\nu_1}{\nu_2}F(\nu_1, \nu_2)\right]$$

En effet, la référence [2] donne les moments non centrés de la variable F :

$$E(F^k) = \left(\frac{\nu_2}{\nu_1}\right)^k f_k \tag{20}$$

avec :

$$f_k = \frac{[\nu_1 + 2(k-1)][\nu_1 + 2(k-2)] \ldots\ldots \nu_1}{(\nu_2 - 2)(\nu_2 - 4) \ldots\ldots (\nu_2 - 2k)}$$

Si l'on pose :

$$z = \frac{\nu_1}{\nu_2} F(\nu_1, \nu_2)$$

l'espérance de z_k est égale à f_k et on a donc :

$$\mu_6(z) = f_6 - 6f_1f_5 + 15f_1^2f_4 - 20f_1^3f_3 + 15f_1^4f_2 - 5f_1^6$$
$$\mu_8(z) = f_8 - 8f_1f_7 + 28f_1^2f_6 - 56f_1^3f_5 + 70f_1^4f_4 - 56f_1^5f_3 + 28f_1^6f_2 - 7f_1^8$$

Pratiquement les valeurs de f_k sont calculées par récurrence avec :

$$\left| \begin{array}{l} f_1 = \dfrac{\nu_1}{\nu_2 - 2} \\ f_{k+1} = f_k \cdot \dfrac{\nu_1 + 2k}{\nu_2 - 2k - 2} \end{array} \right. \tag{21}$$

On en déduit les paramètres β_4 et β_6 de la distribution de d^2 :

$$\beta_4 = \frac{w^6 \ \mu_6(z)}{a_2^3} \tag{22}$$

$$\beta_6 = \frac{w^8 \mu_8(z)}{a_2^4} \tag{23}$$

Ce qui permet d'obtenir les coefficients de Pearson de R, donc le seuil du test de comparaison.

5. Exemples numériques

5.1 Loi équiprobable

A l'aide d'une calculette, $N = 300$ nombres ont été tirés au hasard dans le domaine de 0 à 10. Leur répartition est donnée dans le tableau 1.

Les résultats sont répartis en $(K + 1) = 10$ classes correspondant chacune à la probabilité $p = 0,10$.

Au niveau de probabilité 95 %, la valeur expérimentale de d^2 doit, si la distribution est bien équiprobable, être inférieure à un seuil égal, d'après le paragraphe 2.4, à :

$$d^2_{0,95} = \frac{1}{300} \chi^2_{0,95}(9) = 0,0563$$

Tableau 1

Classe i	n_i
$0-1$	40
$1-2$	28
$2-3$	25
$3-4$	22
$4-5$	33
$5-6$	27
$6-7$	32
$7-8$	31
$8-9$	26
$9-10$	36

La valeur de d^2 est calculée par les formules du paragraphe 2.4 :

$$f = \frac{30}{\sqrt{0,1\ \Sigma n^2}} = 0,9854$$
$$d^2 = 2(1-f) = 0,0291$$

La valeur de d^2 est bien inférieure au seuil; on ne peut pas refuser l'hypothèse d'une loi équiprobable.

Le test habituel aurait consisté dans le calcul de :

$$\chi^2 = \frac{\Sigma(n_i - Np_i)^2}{Np_i} = \frac{\Sigma(n_i - 30)^2}{30}$$
$$= 8,93$$

Cette valeur doit être comparée au seuil :

$$\chi^2_{0,95}(9) = 16,9$$

Les deux tests sont équivalents.

5.2 Influence du tabac sur le cancer

Cet exemple est tiré de la référence [4]. Le tableau 2 donne la répartition des fumeurs et des non fumeurs dans une population de cancéreux et une population saine, en fonction des occupations professionnelles et de l'âge. Les effectifs étant trop faibles pour que le test de χ^2 soit valable, les auteurs proposaient une modification de ce test pour analyser les données.

Un premier test consiste dans la comparaison des histogrammes totaux, cancéreux et non cancéreux, pour vérifier la validité de la population témoin en ce qui concerne la répartition des âges et des occupations.

Tableau 2

Catégorie	Cancéreux			Non cancéreux			p_i %
	Fumeurs f_{i1}	Non fum.	Somme n_{i1}	Fum. f_{i2}	Non fum.	Somme n_{i2}	
Femmes à la maison							
< 45 ans	0	2	2	0	7	7	2,88
45-54	2	5	7	1	24	25	10,22
55-64	3	6	9	0	49	49	18,53
$\geqslant$ 65	0	11	11	0	42	42	16,93
«Cols Blancs»							
< 45 ans	3	0	3	2	6	8	3,51
45-54	2	2	4	2	18	20	7,67
55-64	2	4	6	2	23	25	9,90
$\geqslant$ 65	0	6	6	1	11	12	5,75
Autres occupations							
< 45 ans	1	0	1	3	10	13	4,47
45-54	4	1	5	1	12	13	5,75
55-64	0	6	6	1	19	20	8,31
$\geqslant$ 65	1	3	4	0	15	15	6,07
Somme		46	64		236	249	100

Les effectifs à comparer sont désignés par n_{i1} et n_{i2} sur le tableau (colonnes 4 et 7).

Dans chaque classe, la probabilité p_i, supposée commune aux deux populations, est calculée par :

$$p_i = \frac{n_{i1} + n_{i2}}{N_1 + N_2} = \frac{n_{i1} + n_{i2}}{313}$$

Elle est donnée dans la dernière colonne du tableau 2.

La distance entre les histogrammes est calculée par les formules (2) et (10). On obtient :

$$d^2 = 0,07275$$

Les probabilités p_i permettent d'utiliser la formule (4) qui donne :

$$a_1 = 7,8230$$

On en déduit l'espérance mathématique de d^2 :

$$\mu = a_1 \left(\frac{1}{N_1} + \frac{1}{N_2} \right) = 0,15365$$

Puisque la valeur expérimentale de d^2 est inférieure à cette valeur, il n'est pas utile de continuer le test pour conclure à la validité de la population témoin.

Pour étudier l'influence du tabac, il faut répondre à la question : y a-t-il plus de fumeurs chez les cancéreux que chez les non cancéreux ?

Dans chaque population, chaque catégorie est partagée en fumeurs et non fumeurs. Si f_{i1} est le nombre de fumeurs dans la $i$ème catégorie de la population 1 (cancéreux), le nombre de non fumeurs dans la même catégorie est $(n_{i1} - f_{i1})$. On définit de même les effectifs f_{i2} et $(n_{i2} - f_{i2})$ dans la population 2 (non cancéreux).

Si les probabilités d'être fumeur sont les mêmes dans les deux populations, elles sont estimées par :

$$p_{fi} = \frac{f_{i1} + f_{i2}}{N_1 + N_2} = \frac{f_{i1} + f_{i2}}{313}$$

La probabilité de ne pas fumer dans la $i$ème catégorie s'en déduit immédiatement : elle est égale à $(p_i - p_{fi})$. Les effectifs des non fumeurs ne doivent donc pas être pris en compte puisque les formules du paragraphe 3.1 supposent que les estimations de probabilités sont liées par une seule relation (somme égale à l'unité).

Chaque population est partagée en 13 classes : 12 classes $(i = 1, 2 \ldots 12)$ correspondent aux fumeurs de chaque catégorie, la classe résiduelle, notée zéro, correspondant aux non fumeurs. Les effectifs sont donnés dans le tableau 3 ainsi que les probabilités correspondantes.

La valeur de d^2 est :

$$d^2 = 0,01856$$

Son espérance est :

$$\mu = 0,002425 \qquad (a_1 = 0,12346)$$

et son écart type :

$$\sigma = 0,00121 \qquad (a_2 = 0,0037887)$$

Donc :

$$d^2 = \mu + 13,4\sigma$$

Tableau 3

i (Classe)	n_{i1}	n_{i2}	p_i %
Non fumeurs $i = 0$	46	236	90,10
Fumeurs			
1	0	0	0
2	2	1	0,96
3	3	0	0,96
4	0	0	0
5	3	2	1,60
6	2	2	1,28
7	2	2	1,28
8	0	1	0,32
9	1	3	1,28
10	4	1	1,60
11	0	1	0,32
12	1	0	0,32
Somme	64	249	100

Pour savoir si cette valeur est significative, il faut calculer les caractéristiques de la distribution de d^2. On obtient :

$$\mu_3 = 1,928 \cdot 10^{-9}$$
$$\mu_4 = 3\sigma^4 + 3,987 \cdot 10^{-12}$$

Donc :

$$\beta_1 = 1,191 \text{ et } \beta_2 = 4,866$$

D'où $d^2 \leftrightarrows \left(\frac{1}{64} + \frac{1}{249}\right)[0,00526 + 0,1170F(\nu_1, \nu_2)]$

$$= 0,00010 + 0,00230F(\nu_1, \nu_2)$$

avec $\nu_1 = 7,74$ et $\nu_2 = 202$

On peut donc calculer, à l'aide de la table de F, la limite supérieure de d^2 au niveau de confiance 99,5% soit :

$$d^2_{0,995} = 0,0065$$

La valeur expérimentale de d^2 est très supérieure à ce seuil. Il faut donc conclure à l'influence du tabac sur l'apparition de cancers.

5.3 Cas où N_1 est petit devant N_2

L'exemple est présenté dans la référence [3], p. 237. Le tableau 4 donne, pour une population de $N = 20073$ habitants, la répartition des cas de maladie par tranches d'âge. Le nombre total $N_1 = 350$ des malades est très petit devant celui des non-malades ($N_2 = 19723$).

Tableau 4

Age	Malades	Non malades	Somme
0-20	153	6 670	6 823
20-40	139	8 279	8 418
40-60	52	4 160	4 212
> 60	6	614	620
Somme	350	19 723	20 073

On ne peut donc pas appliquer le test de la distance entre deux populations (malades et non malades). Il faut considérer l'ensemble des individus comme appartenant à la même population et vérifier si les effectifs observés correspondent à l'hypothèse : l'apparition de la maladie est indépendante de l'âge.

Si cette hypothèse est réalisée, la probabilité d'être malade est estimée par :

$$p_m = \frac{350}{20\,073} = 1,744\%$$

Les probabilités P_i d'appartenir à la $i$ème tranche d'âge sont obtenues à partir de la dernière colonne du tableau 4. Si l'âge n'influe pas sur l'apparition de la maladie, la probabilité d'être malade, pour un individu de la $i$ème tranche d'âge est :

$$p_i = p_m \times P_i$$

Celle de ne pas être malade dans la même tranche d'âge s'en déduit immédiatement : elle est égale à $P_i - p_i$. Il ne faut donc pas tenir compte des effectifs correspondants puisque les probabilités estimées doivent être liées par une seule relation.

La population est partagée en 5 classes : les nombres $n_i (i = 1, 2, 3, 4)$ sont ceux des malades dans chaque tranche d'âge, le nombre n_o résiduel étant celui des non malades (cf. tableau 5).

La distance entre les effectifs n_i et les probabilités p_i est donnée par :

$$d^2 = 4,374 \times 10^{-6}$$

Tableau 5

i	Classe	n_i	P_i %	p_i %
o	Non mal.	19 723		98,26
	Malades			
1	0-20	153	33,99	0,593
2	20-40	139	41,94	0,731
3	40-60	52	20,98	0,366
4	> 60	6	3,09	0,054
Sommes		20 073		100

La valeur de a_1 est :

$$a_1 = 0,018166$$

d'où l'espérance de d^2 :

$$\mu = 0,905 \cdot 10^{-6}.$$

L'écart-type de d^2 est calculé à partir de :

$$a_2 = 0,0002222 : \quad \sigma = 0,743 \cdot 10^{-6}$$

Donc :

$$d^2 - \mu = 4,67\sigma$$

Pour vérifier si cet écart est significatif, il faut calculer les coefficients de Pearson, à l'aide des formules (6), (7) et (8). On obtient :

$$a_3 = 5,879 \cdot 10^{-6} \quad \text{et} \quad a_4 = 0,2425 \cdot 10^{-6}$$

d'où :

$$\begin{aligned} \mu_3 &= 7,269 \cdot 10^{-19} \quad \text{et} \quad \mu_4 - 3\sigma^4 = 1,494 \cdot 10^{-24} \\ \beta_1 &= 3,148 \quad \text{et} \quad \beta_2 = 7,910 \end{aligned}$$

En appliquant les formules du paragraphe 2.3, on trouve que d^2 doit être de la forme :

$$\begin{aligned} d^2 &= \frac{1}{N}[0,00068 + 0,0172F(\nu_1, \nu_2)] \\ &= [0,034 + 0,857F(\nu_1, \nu_2)] \times 10^{-6} \end{aligned}$$

avec :

$$\nu_1 = 2,865 \text{ et } \nu_2 = 123$$

La table de F donne, au niveau de confiance 99 %, la valeur :

$$F_{0.99}(2,9;120) \# 4$$

ce qui correspond à :

$$d^2_{0.99} \# 3,4 \cdot 10^{-6}$$

La valeur expérimentale de d^2 dépasse largement ce seuil. Il faut donc admettre que la fréquence de la maladie dépend de l'âge (plus importante chez les enfants). Le test de «Chi-2» conduisait à la même conclusion [3].

5.4 Comparaison de plusieurs histogrammes

Cet exemple est tiré de la référence [5].

Le tableau 6 donne, pour 3 520 cas de décès par cancer de travailleurs, leur répartition suivant la dose cumulée de radiations (en centirads) et l'âge au moment de la mort. On compare donc entre eux les histogrammes correspondant $q = 5$ niveaux de dose.

Tableau 6

Dose / Age (ans)	0	1-19	20-99	100-499	> 500	p_i %
< 40	108	55	58	24	9	7,22
40-49	185	82	137	74	17	14,06
50-59	331	137	200	155	58	25,03
60-69	360	162	248	184	53	28,61
> 70	352	189	251	74	17	25,09
N_j	1336	625	894	511	154	
d^2_j	0,00221	0,02019	0,00810	0,07670	0,14369	
μ_j	0,00242	0,00516	0,00361	0,00632	0,02096	
σ_j	0,00185	0,00396	0,00277	0,00485	0,01608	
$\frac{d^2_j - \mu_j}{\sigma_j}$	− 0,111	3,792	1,620	14,521	7,631	

Les probabilités p_i sont calculées par la relation :

$$p_i = \frac{\sum_j n_{ij}}{3\,520}$$

Elles sont données dans la dernière colonne du tableau 6. On en déduit les quantités :

$$a_1 = 3,2285 \qquad a_2 = 6,1349$$

et les valeurs de μ_j et σ_j reportées dans le tableau 6.

Les valeurs :

$$a_3 = 25,3968 \qquad a_4 = 164,6023$$

permettent de calculer les coefficients β_1 et β_2 communs aux cinq histogrammes :

$$\beta_1 = 2,7934 \qquad \beta_2 = 7,3734$$

et, par suite, la variance de R, à l'aide de la relation (17) :

$$\text{Var}(R) = \frac{1}{5}(\beta_2 - 1) = 1,2747$$

La valeur expérimentale de R est :

$$R = 57,219 = 1 + 49,79\sigma_R$$

Pour avoir la limite supérieure de R, il faut étudier la distribution des carrés des distances.

Les valeurs de β_1 et β_2 correspondent à la loi :

$$d_j^2 = \frac{1}{N_j}\left[a_0 + a_1 + \frac{w\,\nu_1}{\nu_2} F(\nu_1, \nu_2)\right]$$

avec :

$$\nu_1 = 3,27 \qquad \nu_2 = 119 \qquad w = 110,96$$

On en déduit immédiatement les coefficients f_k par les formules (21), et donc $\mu_6(z)$ et $\mu_8(z)$:

$$\mu_6(z) = 2,2699 \cdot 10^{-8} \qquad \text{d'où } \beta_4 = 183,471$$
$$\mu_8(z) = 5,6468 \cdot 10^{-10} \qquad \text{d'où } \beta_6 = 9\,160$$

Les formules (18) et (19) donnent alors les coefficients de Pearson de la variable R :

$$\beta_1(R) = 20,613 \qquad \beta_2(R) = 45,695$$

Ces coefficients correspondent à la loi de distribution :

$$R = 0,3415 + 0,5662F(0,87; 14, 3)$$

Les tables de F sont établies pour les valeurs entières de ν_1 et ν_2 donc pour ν_1 au moins égal à 1. Elles ne permettent donc pas d'avoir un seuil pour F.

On peut cependant remarquer que, pour $\nu_1 = 1$ et $\nu_2 = 14$, le seuil au niveau de probabilité 99,5 % est :

$$F_{0,995}(1, 14) = 11,06$$

ce qui correspondrait, pour R, à une limite voisine de 7. La valeur expérimentale $R = 57$ lui est très supérieure. On peut donc conclure à l'influence de la dose de radiation sur l'âge de la mort par cancer.

6. Conclusions

Les tests fondés sur l'utilisation de la distance se prêtent à de nombreuses applications. Le présent article a été limité au calcul de seuils de décision, moyennant quelques approximations qu'il faudra sans doute justifier ultérieurement, par exemple par des simulations. De même, l'efficacité des tests proposés devrait faire l'objet d'études ultérieures.

La détermination des seuils de décision fait appel à des formules qui ne sont pas toujours très simples mais qui peuvent être programmées et utilisées avec des calculateurs de faible capacité (calculettes de poche avec une trentaine de mémoires). Ce travail ne paraît pas excessif devant celui qui a été nécessaire pour la collecte des données.

En principe, la comparaison de deux histogrammes expérimentaux nécessite des effectifs totaux du même ordre de grandeur. Mais on a vu, grâce à un exemple numérique, que cette contrainte pouvait être contournée sans difficulté.

La statistique R utilisée pour la comparaison de plusieurs histogrammes a une distribution extrêmement dissymétrique. On peut alors se heurter à des difficultés pratiques lorsque l'assimilation à une loi de type VI de Pearson conduit à un nombre ν_1 inférieur à l'unité. Une table a été établie au CEA pour des valeurs de ν_1 non entières, comprises entre 0,3 et 2, et ν_2 entier, de 1 à 10. Elle peut être communiquée par l'auteur aux lecteurs qui le désireraient.

Références

(1) DORLET J., Rapport EUR 5667 e/f (1977).

(2) KENDALL M. et STUART A., The advanced theory of statistics – 1 – Distribution theory – Ch. Griffin a. Co, 4° édition 1960.

(3) CETAMA, Statistique appliquée à l'exploitation des mesures-Masson 2° édition 1986.

(4) MANTEL N. et HAENSZEL W., Statistical aspects from retrospective studies of disease – J. of the National Cancer Institute – vol. 22 - n° 4 p. 719/748 – 1959.

(5) MANCUSO T.F., STEWART A. et KNEALE G., Radiation exposures of HANDFORD workers dying from cancer and the other causes – Health Physics – vol. 33 - p. 369/385-1977.

Annexe

Moments centrés de la distribution multinomiale

On désigne par $(abc\ldots)$ le moment centré :

$$\mu_{abc\ldots} = E(n_1 - Np_1)^a \cdot (n_2 - Np_2)^b \cdot (n_3 - Np_3)^c \ldots\ldots$$

Moments d'ordre 2

$$(2) \quad = E(n_1 - Np_1)^2 = Np_1(1 - p_1)$$
$$(11) \quad = E(n_1 - Np_1)(n_2 - Np_2) = -Np_1p_2$$

Moments d'ordre 3

$$(3) \quad = Np_1(1 - p_1)(1 - 2p_1)$$
$$(21) \quad = -Np_1p_2(1 - 2p_1)$$
$$(111) \quad = 2Np_1p_2p_3$$

Moments d'ordre 4

$$(4) \quad = 3N^2p_1^2(1 - p_1)^2 + Np_1(1 - p_1)(1 - 6p_1 + 6p_1^2)$$
$$(31) \quad = -3N^2p_1^2p_2(1 - p_1) - Np_1p_2(1 - 6p_1 + 6p_1^2)$$
$$(22) \quad = N^2p_1p_2(1 - p_1 - p_2 + 3p_1p_2) - Np_1p_2(1 - 2p_1 - 2p_2 + 6p_1p_2)$$
$$(211) \quad = (-N^2 + 2N)p_1p_2p_3(1 - 3p_1)$$
$$(1111) \quad = (3N^2 - 6N)p_1p_2p_3p_4$$

Pour ne pas rallonger trop l'article, nous ne donnons pas l'expression des moments d'ordre 5 à 8 (qui sont utilisés au paragraphe 2.2, pour le calcul des quatres premiers moments de d^2), mais ils peuvent être communiqués par l'auteur aux lecteurs qui en feraient la demande.