REVUE DE STATISTIQUE APPLIQUÉE

F. HATTON
F. FACY
F. LAURENT

Une méthode simple de comparaisons partielles

Revue de statistique appliquée, tome 24, n° 4 (1976), p. 75-78
<http://www.numdam.org/item?id=RSA_1976__24_4_75_0>

© Société française de statistique, 1976, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

UNE MÉTHODE SIMPLE DE COMPARAISONS PARTIELLES

F. HATTON *, F. FACY *, F. LAURENT *

INTRODUCTION

Au cours d'une analyse statistique est souvent abordée l'étude de la répartition d'un caractère qualitatif dichotomique dans divers groupes de sujets, répartition matérialisée par les pourcentages observés dans chaque groupe (soit k le nombre de ces groupes).

Le test global du X^2 de répartitions observées met en évidence une éventuelle différence significative, traduisant l'existence d'une différence globale entre ces k pourcentages. Mais on désire, bien sûr, affiner ces résultats et voir où se situent exactement les différences ; en somme, on veut classer et même ordonner les k pourcentages $p_1, p_2 \ldots p_i \ldots p_k$.

Ceci peut être réalisé à l'aide d'une des méthodes de comparaisons partielles entre les pourcentages pris deux à deux, quand k est relativement petit. Mais il y a de nombreux problèmes où k est "grand" : il en est ainsi par exemple, pour la comparaison des départements français où k est supérieur à 90. La méthode des comparaisons partielles paraît alors difficilement applicable :

– en effet, le nombre de tests nécessaires s'accroît très rapidement puisqu'il est égal à $\frac{k(k-1)}{2}$,

– mais surtout, les résultats ainsi obtenus sont difficilement interprétables du fait même de la multiplicité des tests réalisés.

Aussi, propose-t-on ici, une méthode de classification plus simple, basée sur la comparaison de chaque pourcentage $p_1, p_2 \ldots p_i \ldots p_k$ au pourcentage moyen P observé pour l'ensemble de la population.

Cette méthode, si elle ne répond pas exactement à l'objectif précédemment décrit, permet toutefois de différencier 3 sortes de groupes :

– groupes pour lesquels le pourcentage observé p_i ne diffère pas significativement du pourcentage moyen P,

(*) Section "Statistique, Epidémiologie et Informatique" I.N.S.E.R.M. – D.R.M.S. 44, Chemin de Ronde – F – 78110 Le Vesinet.

Mots-clés : **Test de comparaison.**

– groupes pour lesquels p_i diffère significativement de P, et tels que $p_i > P$

– groupes pour lesquels p_i diffère significativement de P, et tels que $p_i < P$

METHODE PROPOSEE

– Soit une population de N sujets, séparée en k groupes de n_i sujets : $n_1, n_2 \ldots n_i \ldots n_k$;

– soit un caractère qualitatif dichotomique, par exemple C absent ou présent, les indices 0 et 1 correspondant respectivement à C absent et C présent ;

on dispose donc des données suivantes :

Groupe	Nombre total de sujets	Nombre de sujets présentant C	Pourcentages de sujets présentant C
1	n_1	n_{11}	p_1
2	n_2	n_{21}	p_2
i	n_i	n_{i1}	p_i
k	n_k	n_{k1}	p_k
TOTAL	N	N_1	P

On désire savoir pour chaque groupe i si le pourcentage observé p_i diffère significativement du pourcentage P observé pour l'ensemble de la population. On teste donc l'hypothèse nulle $P = P_i$ contre l'hypothèse alternative $P \neq P_i$.

Pour cela, on calculera l'écart réduit ϵ entre P et p_i :

$$\epsilon = \frac{P - p_i}{\sqrt{\text{variance } (P - p_i)}}$$

On sait que P et p_i ne sont pas indépendants.

$$P - p_i = \frac{N_1}{N} - \frac{n_{i1}}{n_i}$$

On calcule facilement que

$$P - p_i = \frac{N - n_i}{N}\left[\frac{N_1 - n_{i1}}{N - n_i} - \frac{n_{i1}}{n_i}\right],$$

$\frac{N_1 - n_{i1}}{N - n_i}$ correspondant au pourcentage de sujets présentant le caractère C parmi les sujets des $(k - 1)$ groupes, autres que le groupe i (ensemble de la population, à l'exclusion des sujets du groupe i).

Soit π ce dernier pourcentage, on a donc :

$$P - p_i = \frac{N - n_i}{N} [\pi - p_i]$$

$$\text{variance } (P - p_i) = \left(\frac{N - n_i}{N}\right)^2 \text{ variance } (\pi - p_i)$$

où π et p_i sont indépendants.

En cas d'hypothèse nulle, on aurait : $\pi = p_i = P$

Donc,

$$\text{variance } (\pi) = \frac{P(1 - P)}{N - n_i - 1} \# \frac{P(1 - P)}{N - n_i} \text{ compte tenu de la taille de N.}$$

De même, $$\text{variance } (p_i) \# \frac{P(1 - P)}{n_i}$$

En définitive,

$$\text{variance } (P - p_i) = \left(\frac{N - n_i}{N}\right)^2 \left[\frac{P(1 - P)}{N - n_i} + \frac{P(1 - P)}{n_i}\right]$$

soit

$$\text{variance } (P - p_i) = \frac{N - n_i}{N} \left[\frac{P(1 - P)}{n_i}\right]$$

En définitive, on a donc la formule de ϵ

$$\epsilon = \frac{|P - p_i|}{\sqrt{\frac{N - n_i}{N} \times \frac{P(1 - P)}{n_i}}}$$

La différence entre P et p_i sera significative au risque α donné si :

$$\frac{|P - p_i|}{\sqrt{\frac{N - n_i}{N} \times \frac{P(1 - P)}{n_i}}} \geqslant \epsilon_\alpha ,$$

ϵ_α étant la valeur lue sur la table de la loi normale réduite pour le risque α choisi.

Soit si $$|P - p_i| \sqrt{\frac{N n_i}{N - n_i}} \geqslant \epsilon_\alpha \sqrt{P(1 - P)} \qquad (1)$$

En résumé :

– $|P - p_i| \sqrt{\frac{N n_i}{N - n_i}} \leqslant \epsilon_\alpha \sqrt{P(1 - P)}$: pas de différences significative.

– $|P - p_i| \sqrt{\frac{N n_i}{N - n_i}} \geqslant \epsilon_\alpha \sqrt{P(1 - P)}$: différence significative.

et dans ce cas :

si $P - p_i > 0, \quad p_i < P$

si $P - p_i < 0, \quad p_i > P$

DISCUSSION

– L'intérêt du test proposé est évidemment limité et ne permet pas d'ordonner tous les k pourcentages les uns par rapport aux autres.

Cependant, grâce à son utilisation, il est possible de différencier, parmi ces k pourcentages, trois sous-ensembles différents par rapport au pourcentage moyen global observé pour l'ensemble de la population et ceci en fonction d'un risque α donné :

– groupes où les pourcentages sont inférieurs au pourcentage global,

– groupes où les pourcentages ne diffèrent pas du pourcentage global,

– groupes où les pourcentages sont supérieurs au pourcentage global.

Cette classification constitue une première approche du problème qui nous intéresse ici et paraît tout spécialement intéressante pour les comparaisons de pourcentages ou de taux observés dans les départements français aussi bien en ce qui concerne la mortalité que la morbidité.

Il serait en outre possible de l'affiner en constituant des sous-ensembles plus nombreux en fonction du degré de signification, c'est-à-dire en recherchant la signification des tests pour diverses valeurs de α.

– Toutefois, à cet égard il faut noter que le nombre de tests effectués étant assez grand, puisqu'il est égal à (k – 1), il est difficile d'attribuer une valeur formelle à ce risque d'erreur α.

On devra donc agir avec prudence, et ne tenir compte que des différences significatives pour un risque α inférieur à la valeur habituellement retenue.

– Pour terminer, on peut remarquer que la formule trouvée se rapproche du test de comparaison d'un pourcentage observé à un pourcentage théorique que l'on aurait pu effectuer en considérant le pourcentage global P comme un pourcentage théorique ; on aurait alors adopté le test suivant

$$|P - p_i| \geqslant \epsilon_\alpha \sqrt{\frac{P(1-P)}{n_i}} \quad \text{soit} \quad |P - p_i| \sqrt{n_i} \geqslant \epsilon_\alpha \sqrt{P(1-p_i)} \quad (2)$$

En comparant les formules (1) et (2) on voit que le test qui est proposé ici, est plus puissant puisqu'intervient un facteur correctif supérieur à l'unité, $\frac{N}{N - n_i}$.

Remerciements :

Nous adressons nos plus vifs remerciements à Monsieur LELLOUCH pour les conseils qu'il a bien voulu nous donner.