Après un test du χ^2 : calcul et test de résidus

F.-G. Carpentier

05/07/2015

1 Introduction - position du problème

On reprend les notations classiques pour un tableau de contingence : soient A et B deux variables catégorielles, de modalités respectives $a_1, a_2, \dots a_p$ et $b_1, b_2, \dots b_q$ et soient $(n_{ij}), 1 \leq i \leq p, 1 \leq j \leq q$ le tableau des effectifs observés, (t_{ij}) celui des effectifs théoriques et $N = \sum_i \sum_j n_{ij}$ l'effectif total de l'échantillon.
On suppose que le test du χ^2 a montré une dépendance entre les variables A et B dans la population parente de l'échantillon étudié. Le problème que l'on se pose est alors le suivant :

Quelles sont les combinaisons (modalité-ligne, modalité-colonne) pour lesquelles l'association (attraction ou répulsion) est significative ?

2 Les résidus standardisés ou résidus de Pearson

Les résidus standardisés ([Agresti], page 38) ou résidus de Pearson [Haberman] (e_{ij}) sont définis par :

$$e_{ij} = \frac{n_{ij} - t_{ij}}{\sqrt{t_{ij}}}$$

La statistique du χ^2 est alors : $\chi^2 = \sum_i \sum_j e_{ij}^2$.
Ce résidu mesure l'attraction (résidus positifs) ou la répulsion (résidus négatifs) de la modalité a_i de A et de la modalité b_j de B. Sous hypothèse d'absence de lien, ces résidus sont des statistiques centrées. Cependant, il ne s'agit pas de variables statistiques réduites, leurs variances étant inférieures à 1. Il est possible de s'intéresser aux valeurs les plus extrêmes, mais sans que l'on puisse déterminer un seuil de significativité.

3 Les résidus ajustés ou résidus de Haberman

On introduit la fréquence p_{ij} de la combinaison de modalités (a_i, b_j) : $p_{ij} = \frac{n_{ij}}{N}$, la fréquence $p_{i\cdot}$ de la modalité a_i de A : $p_{i\cdot} = \frac{\sum_j n_{ij}}{N}$ et la fréquence $p_{\cdot j}$ de la modalité b_j de B : $p_{\cdot j} = \frac{\sum_i n_{ij}}{N}$.
On considère alors les résidus r_{ij} définis par :

$$r_{ij} = \frac{n_{ij} - t_{ij}}{\sqrt{t_{ij}(1 - p_{i\cdot})(1 - p_{\cdot j})}}$$

Ces derniers sont appelés résidus ajustés ou résidus de Haberman.
Haberman a montré que, sous hypothèse H_0 d'absence de lien, r_{ij} suit asymptotiquement une loi normale centrée réduite.

4 Calcul et test des résidus avec les logiciels usuels

4.1 Utilisation de R

La procédure `chisq.test` du package `stats` fournit comme résultats non seulement la valeur du χ^2 et son niveau de significativité, mais également les résidus de Pearson (composant `$residuals` de la liste retournée en résultat) et les résidus de Haberman (composant `$stdres` de la liste retournée en résultat)

Exemple.

```
> M <- as.table(rbind(c(212,29,11,2,3), c(318,61,6,11,13), c(160,39,9,6,12)))
> rownames(M) <- c("a1","a2","a3")
> colnames(M) <- c("b1","b2","b3","b4","b5")
> M
    b1  b2  b3  b4  b5
a1 212  29  11   2   3
a2 318  61   6  11  13
a3 160  39   9   6  12
> Xsq <- chisq.test(M)
Message d'avis :
In chisq.test(M) : l'approximation du Chi-2 est peut-être incorrecte
> Xsq

Pearson's Chi-squared test

data:  M
X-squared = 20.3583, df = 8, p-value = 0.009062

> Xsq$residuals
            b1          b2          b3          b4          b5
a1  0.93616090 -1.33963261  1.28206096 -1.48489514 -1.78406400
a2  0.09113804  0.24066234 -1.71501389  0.77521492  0.04505463
a3 -1.12090876  1.10480426  0.93998072  0.54059524  1.84188135
> Xsq$stdres
            b1          b2          b3          b4          b5
a1  2.33159333 -1.71672778  1.54215391 -1.77896194 -2.14848117
a2  0.26026470  0.35362027 -2.36537490  1.06489378  0.06221195
a3 -2.72597782  1.38245439  1.10404745  0.63240142  2.16587067
```

Sur cet exemple, le test du χ^2 donne un niveau de significativité de l'ordre de 1%. L'étude des résidus ajustés montre que pour un seuil de 1% unilatéral ($z_{crit} = 2.33$), les attractions-répulsions significatives sont : (a_1, b_1), (a_2, b_3), (a_3, b_1).
On constate également sur les valeurs numériques que les résidus ajustés sont supérieurs en valeur absolue aux résidus de Pearson (ce qui n'est pas étonnant, étant donné le résultat concernant la variance des résidus de Pearson) et que l'écart entre les deux types de résidus est d'autant plus grand que les modalités considérées ont une fréquence élevée (évident également, d'après la formule donnant les résidus ajustés).

4.2 Utilisation de Statistica

Statistica ne fournit pas de résultat *post hoc* pour le test du χ^2. Cependant, le menu Statistiques > Techniques Exploratoires Multivariées > Analyse des correspondances permet d'utiliser en entrée un tableau de contingence et de calculer des "écarts centrés-réduits" (*sic,* , en fait, ils ne sont pas réduits), qui sont en fait les résidus de Pearson.
L'exemple précédent, traité avec Statistica est disponible sur ce site.

4.3 Utilisation d'Excel ou LibreOffice Calc

Excel et LibreOffice Calc permettent le calcul complet de la statistique du χ^2 et des résidus. Un classeur contenant cet exemple, ainsi qu'un autre exemple analogue sont également disponibles sur ce site.

5 Prolongements

Dans [Beh], E. J. Beh développe une variante de l'analyse des correspondances en utilisant les résidus de Haberman au lieu des résidus de Pearson pour évaluer les écarts à l'indépendance.

Références

[Agresti] A. Agresti, *An Introduction to Categorical Data Analysis*, 2nd ed., New York, Wiley & Sons, 2002.

[Beh] Eric J. Beh, *Simple correspondence analysis using adjusted residuals*, Journal of Statistical Planning and Inference, No. 142, (2012), pp. 965-973.

[Haberman] Shelby J. Haberman, *The Analysis of Residuals in Cross-Classified Tables*, Biometrics, Vol. 29, No. 1 (Mar., 1973), pp. 205-220.