Le test du χ^2

Les résultats tirés d'échantillons ont une variabilité qui ne correspond pas toujours aux résultats théoriques que l'on pourrait attendre du simple jeu des probabilités. Afin de vérifier si cette variabilité est justifiée ou non, on propose de comparer les distributions théoriques et observées au travers de la statistique du χ^2.

Définition du χ^2

La statistique du χ^2 est une mesure des écarts entre les fréquences observées et théoriques. Elle s'écrit

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \ ... \ + \frac{(o_k - e_k)^2}{e_k} = \sum_{j=1}^{k} \frac{(o_j - e_j)^2}{e_j}$$

où, si le nombre d'événements est N, on a

$$\sum_{j=1}^{k} o_j = \sum_{j=1}^{k} e_j = N.$$

Si $\chi^2 = 0$, les fréquences observées et théoriques sont similaires, plus χ^2 est grand plus la différence entre les données et le modèle est importante.
La distribution d'échantillonnage du χ^2 est très proche de la distribution du même nom

$$Y = Y_0 \chi^{\nu-2} \exp\left(\frac{\chi^2}{2}\right)$$

si les fréquences théoriques sont au moins égales à 5. L'approximation s'améliore avec des valeurs plus élevées.
ν est le nombre de degré de liberté et

$$\nu = k - 1.$$

Test de signification

On teste l'hypothèse H_0 pour laquelle il n'existe pas de différence entre les distributions théoriques et observées. Si la valeur calculée du χ^2 est supérieure à une valeur critique χ_α au seuil α, on conclut que les données diffèrent significativement du modèle et l'on rejette H_0 au seuil correspondant.
Le même test peut être mené si les données sont trop proches du modèle. Il faudra alors vérifier si la valeur du χ^2 est inférieure à une valeur critique $\chi_{(1-\alpha)}$. Dans ce cas, le test est trop bon, il faudra alors peut être remettre en cause le caractère aléatoire de la sélection des données.

Tableaux de contingences

Le test du χ^2 peut s'appliquer à des systèmes dépendant de plusieurs variables. On parle alors de tableaux de contingences.
Dans le cas le plus simple à deux variables, si les variables comprennent h et k catégories, nous avons

$$\chi^2 = \sum_{j=1}^{hk} \frac{(o_j - e_j)^2}{e_j}$$

avec

$$\nu = (h-1)(k-1).$$

Correction de continuité de Yates

Afin de mieux analyser des données discrètes, on peut faire des corrections de continuité. La statistique du χ^2 devient alors

$$\chi^2 = \sum_{j=1}^{k} \frac{(|o_j - e_j| - 0.5)^2}{e_j}$$

avec

$$\nu = k - 1.$$

Pour information, il s'agit des cas pour lesquels on connaît les paramètres de la population (i.e. on ne les estime pas à partir des paramètres de l'échantillon).