Calcul du χ^2 d'indépendance

L'étude la liaison (ou plutot de la dépendance, de l'indépendance) de deux variables qualitatives se fait à partir des effectifs issus du tri croisé des variables. On calcule un indicateur de type distance nommé χ^2 *d'indépendance* que l'on compare à une valeur théorique maximale lue dans une table (la fameuse "*table du* χ^2").

1. Formules

Soient $n_{i,j}$ les valeurs du tri croisé. On note $n_{i.}$ le total de la ligne i et $n_{.j}$ le total de la colonne j. Soit $n_{..}$ le total général :

$$\begin{aligned} n_{..} &= \sum_{i,j} n_{i,j} \\ &= \sum_{i} n_{i.} \\ &= \sum_{j} n_{.j} \end{aligned}$$

Le χ^2 est alors calculé comme un χ^2 classique avec des valeurs observées et des valeurs théoriques en prenant $obs_k = n_{i,j}$ et $th_k = n_{i.}\ n_{.j}/n_{..}$ où on parcourt toutes les valeurs de i et j par exemple en prenant k de 1 à $l*c$ si l est le nombre de lignes et c le nombre de colonnes avec $k(i,j) = (c-1)*i+j$.

Le nombre de degrés de liberté est alors $(l-1).(c-1)$ et la formule de calcul est donc

$$\chi^2 = \sum_{i,j} \frac{\left(n_{i,j} - \dfrac{n_{i.}\ n_{.j}}{n_{..}}\right)^2}{\dfrac{n_{i.}\ n_{.j}}{n_{..}}}$$

Chaque terme

$$\frac{\left(n_{i,j} - \dfrac{n_{i.}\ n_{.j}}{n_{..}}\right)^2}{\dfrac{n_{i.}\ n_{.j}}{n_{..}}}$$

est nommé contribution des modalités i et j au χ^2. Plus ces contributions sont grandes et plus le χ^2 est important. Il est donc d'usage d'afficher ces contributions par ordre décroissant de valeur absolue, le signe de la différence menant à la contribution étant important puisqu'il montre si le croisement est surabondant ou sousabondant par rapport à l'hypothèse d'indépendance.

2. Un exemple fictif d'application

A partir du tri croisé (fictif) entre les deux variables qualitatives `ETUD` (niveau d'études) et `SALA` (catégorie salariale)

	PRIMAIRE	BEPC	BAC	SUP
DIRE	1	0	0	0
COMME	11	4	3	6
INGE	11	1	7	0
ADMIN	6	4	11	4
MANU	7	12	11	10
HAND	1	1	1	3

Nous pouvons calculer les marges à savoir

DIRE	1	(Direction)
COMME	24	(Commerciaux)
INGE	19	(Ingénieurs)
ADMIN	25	(Administratifs)
MANU	40	(Manutentionnaires)
HAND	6	(Handicapés)

PRIMAIRE	BEPC	BAC	SUP
37	22	33	23

Les valeurs sous hypothèse d'indépendance sont donc

	PRIMAIRE	BEPC	BAC	SUP	+	Total
DIRE	0.3	0.2	0.3	0.2	+	1.0
COMME	7.7	4.6	6.9	4.8	+	24.0
INGE	6.1	3.6	5.5	3.8	+	19.0
ADMIN	8.0	4.8	7.2	5.0	+	25.0
MANU	12.9	7.7	11.5	8.0	+	40.0
HAND	1.9	1.1	1.7	1.2	+	6.0
Total	37.0	22.0	33.0	23.0	+	115.0

Et les contributions signées sont alors

-	1.430	+	0.191	+	0.287	+	0.200
-	1.392	+	0.076	+	2.194	-	0.300
-	3.907	+	1.910	-	0.439	+	3.800
+	0.519	+	0.128	-	2.041	+	0.200
+	2.677	-	2.470	+	0.020	-	0.500
+	0.448	+	0.019	+	0.303	-	2.700

ce qui nous fournit comme χ^2 la valeur 28.151 pour (6-1)$\times$(4-1) soit 15 degrés de liberté.

La value lue dans la table étant 25, on doit rejeter l'hypothèse d'indépendance.

Pour trouver les modalités dépendantes, on utilise le début de la liste des plus fortes contributions

```
      -    3.907 ligne 3 colonne 1
      +    3.800 ligne 3 colonne 4
      -    2.700 ligne 6 colonne 4
      +    2.677 ligne 5 colonne 1
...
```

On peut ici conclure à une surabondance d'ingénieurs avec un niveau d'études primaires (!) et une sousabondance d'ingénieurs avec un niveau d'études supérieures (ouf!).

3. Un exemple réel d'application

Revoici le tri croisé `SEXE` / `ETUD` de notre fameux dossier `ELF` :

```
      Homme  Femme
NR        2      1
Prim      1      5
Bepc      7     23
Bac       8     13
Sup      17     22
```

et voici les effectifs théoriques avec les marges

```
      1.1        1.9  +       3.0
      2.1        3.9  +       6.0
     10.6       19.4  +      30.0
      7.4       13.6  +      21.0
     13.8       25.2  +      39.0
+++++++++++++++++++++++++++++++++
     35.0       64.0  +      99.0
```

ainsi que les contributions avec signe de différence

```
      -  0.832     +  0.455
      +  0.593     -  0.324
      +  1.226     -  0.671
      -  0.045     +  0.024
      -  0.748     +  0.409
```

La valeur du chi-deux est 5.32698 pour 4 degrés de liberté alors que la valeur dans la table est 9.4877 ce qui permet de conclure qu'il y a ici indépendance même si comme nous l'avions fait remarquer il y a une différence de profils entre les hommes et les femmes au niveau de la répartition entre niveaux d'études... ce qui se voit bien avec les deux plus fortes contributions.