Détermination du nombre de degrés de liberté dans les tests d'ajustement et d'indépendance

Christophe Gonzales

20/10/2018

1 Rappels sur les tests d'ajustement et d'indépendance

Dans les tests d'ajustement et les tests d'indépendance, nous avons vu en cours que la population est supposée être répartie en k classes $C_1, \ldots, C_k$:

C_1	C_2	$\cdots$	C_{k-1}	C_k

Le principe de ces deux tests est de tirer un échantillon de taille n de la population, d'observer le nombre d'individus dans cet échantillon $n_1, \ldots, n_k$ appartenant à chacune des classes et de comparer avec ce que l'on aurait dû obtenir théoriquement si la distribution des individus dans les classes suit une certaine loi de probabilité $p_1, \ldots, p_k$:

observations :	n_1	n_2	$\cdots$	n_{k-1}	n_k
effectif théorique :	$n \times p_1$	$n \times p_2$	$\cdots$	$n \times p_{k-1}$	$n \times p_k$

Notons H_0 l'hypothèse nulle et H_1 la contre-hypothèse. On a donc :

H_0 : la population est répartie selon la distribution $p_1, \ldots, p_k$
H_1 : la population n'est pas répartie selon la distribution $p_1, \ldots, p_k$

Définissons des variables aléatoires $N_1, \ldots, N_k$ qui représentent le nombre d'individus dans l'échantillon appartenant aux classes $C_1, \ldots, C_k$ avant que celui-ci ne soit créé : avant création, comme on pourrait avoir de nombreux échantillons possibles, il s'agit donc bien de variables aléatoires ; après tirage, l'échantillon étant fixé, ce que l'on observe, ce sont les valeurs $n_1, \ldots, n_k$ des variables $N_1, \ldots, N_k$.

Un test d'ajustement consiste 1) à calculer la valeur de la statistique $D^2_{(n)}$ suivante :

$$D^2_{(n)} = \sum_{r=1}^{k} \frac{(N_r - n \times p_r)^2}{n \times p_r}, \tag{1}$$

2) à prouver que $D^2_{(n)}$ suit une loi du χ^2 à r degrés de liberté, et 3) à se servir de cette loi afin de déterminer laquelle des deux hypothèses H_0 et H_1 est la plus probable. Il est important de rappeler la définition de la loi du χ^2 :

Définition 1 La loi du χ^2 à r degrés de liberté est la loi de la somme des carrés de r variables **indépendantes** et de même loi normale centrée réduite $\mathcal{N}(0, 1)$.

Ici, le point important pour déterminer le nombre de degrés de liberté est de compter combien de variables $N_1, \ldots, N_k$ sont indépendantes. Nous ne vous montrerons pas dans MAPSI

comment démontrer que $D^2_{(n)}$ est bien constitué d'une somme de variables suivant des lois normales $\mathcal{N}(0,1)$ car cela dépasse le cadre de l'UE. Si vous êtes intéressé par la démonstration, vous la trouverez dans le Saporta, page 102 [1] :

Gilbert Saporta (2006) "Probabilités, analyse des données et statistique", Editions Technip, 2ème édition.

2 Calcul pratique du nombre de degrés de liberté

Vous l'avez compris, pour calculer le nombre de degrés de liberté, il faut compter le nombre de variables N_r indépendantes. Illustrons cela sur quelques exemples.

2.1 Test d'ajustement « classique »

Comme vu en cours, un péage d'autoroute est constitué de 10 cabines. On a comptabilisé ci-dessous le nombre de clients par cabine sur une heure :

N° cabine	1	2	3	4	5	6	7	8	9	10
Nb clients	24	14	18	20	23	13	23	24	23	18

On se demande si les clients sont distribués uniformément sur l'ensemble des cabines. Les nombres ci-dessus correspondent à $n_1, \ldots, n_{10}$. On a donc 10 variables N_r. La distribution théorique (uniforme) n'impose aucune contrainte sur les valeurs des N_r. En revanche, l'échantillon étant constitué de 200 individus, on a la contrainte :

$$\sum_{r=1}^{10} N_r = 200.$$

Par conséquent, $N_{10} = 200 - \sum_{r=1}^{9} N_r$. Donc, N_{10} n'est pas indépendant des autres variables N_r. Les variables $N_1, \ldots, N_9$, elles, sont mutuellement indépendantes : on peut vraiment choisir ces nombres N_r indépendamment les uns des autres. Comme il y a 9 variables indépendantes, il y a 9 degrés de liberté.

Supposons maintenant que la loi théorique soit une loi normale $\mathcal{N}(\mu, \sigma^2)$, avec μ et σ connus d'avance. Là encore, cette distribution n'impose aucune contrainte sur les N_r, donc il y a aussi 9 degrés de liberté.

On voit donc que le nombre de paramètres définissant la loi théorique n'a aucune importance sur le nombre de degrés de liberté dès lors que ces paramètres n'imposent pas de contraintes sur les valeurs des variables N_r.

2.2 Test d'indépendance

Dans un test d'indépendance, on crée un tableau de contingence dans lequel on indique le nombre d'observations des couples $(X = A_i, Y = B_j)$, comme indiqué dans le tableau ci-dessous :

1. Il n'est pas précisé dans le Saporta comment obtenir l'inverse de la matrice Σ, mais ce résultat est immédiat si vous appliquez la formule de Sherman-Morrison sur l'inverse d'une somme de deux matrices.

$X \backslash Y$	B_1	B_2	$\cdots$	B_j	$\cdots$	B_J	$total$
A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1.}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I.}$
$total$	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.J}$	n

Nous avons vu en cours que les fréquences d'apparition n_{ij}/n tendent vers les probabilités $P(X = A_i, Y = B_j)$. On calcule les marginales $n_{i.} = \sum_j n_{ij}$ et $n_{.j} = \sum_i n_{ij}$, de telle sorte que $n_{i.}/n$ et $n_{.j}/n$ tendent respectivement vers les probabilités $P(X = A_i)$ et $P(Y = B_j)$. Par conséquent, s'il y a indépendance, on devrait avoir :

$$P(X = A_i, Y = B_j) \approx n_{ij}/n = P(X = A_i) \times P(Y = B_j) \approx n_{i.}/n \times n_{.j}/n.$$

$P(X) \times P(Y)$ est la loi théorique, qui vaut exactement $n_{i.}/n \times n_{.j}/n$ et qui est supposée fixée à cette valeur pour notre échantillon. Cela crée des contraintes sur les variables N_{ij}. En effet, on a :

$$\forall\ i \in \{1, \ldots, I\}, \quad \sum_{j=1}^{J} N_{i,j} = n_{i.}$$
$$\forall\ j \in \{1, \ldots, J\}, \quad \sum_{i=1}^{I} N_{i,j} = n_{.j}$$

On en déduit donc que :

$$\forall\ i \in \{1, \ldots, I\}, \quad N_{iJ} = n_{i.} - \sum_{j=1}^{J-1} N_{i,j}$$
$$\forall\ j \in \{1, \ldots, J\}, \quad N_{Ij} = n_{.j} - \sum_{i=1}^{I-1} N_{i,j}$$

Notons qu'il y a I contraintes sur la première ligne, J contraintes sur la 2ème ligne, mais N_{IJ} est exprimé sur les 2 lignes. Donc, globalement, il y a $I + J - 1$ contraintes. Comme le tableau contient $I \times J$ variables, on en déduit qu'il y a $I \times J - (I + J - 1) = (I - 1) \times (J - 1)$ variables indépendantes, donc $(I - 1) \times (J - 1)$ degrés de liberté.

2.3 Test d'ajustement avec contraintes

Reprenons l'exemple du péage :

N° cabine	1	2	3	4	5	6	7	8	9	10
Nb clients	24	14	18	20	23	13	23	24	23	18

Supposons maintenant que la distribution théorique est une distribution symétrique P que l'on estime de la manière suivante :

$$\forall r \in \{1, \ldots, 10\},\ P(X = r) = P(X = 11 - r) = \frac{n_r + n_{11-r}}{2n},$$

où $n = 200$ est la taille de l'échantillon. Dans ce cas, on a les contraintes suivantes :

$$\forall r \in \{1, \ldots, 5\},\ N_r + N_{11-r} = n_r + n_{11-r}. \tag{2}$$

On a également $\sum_{r=1}^{10} N_r = n$, mais cette contrainte est déjà incluse dans les 5 contraintes de l'équation (2) puisqu'elles induisent :

$$\sum_{r=1}^{10} N_r = \sum_{r=1}^{5} (N_r + N_{11-r}) = \sum_{r=1}^{5} n_r + n_{11-r} = n.$$

Ici, on a donc 5 contraintes et 10 variables N_r. Donc il y a $10 - 5 = 5$ variables indépendantes et donc 5 degrés de liberté.