V. Tests non paramétriques sur un échantillon

Le modèle n'est pas un modèle paramétrique

« TESTS du CHI-DEUX » :

- V.1. Test d'ajustement à une loi donnée
- V.2. Test d'indépendance de deux facteurs

101

Différentes questions peuvent se poser lorsque l'on analyse des observations :

1°) Sur une série d'observations d'une variable :

•Les tirages sont-ils i.i.d. ?

•**S'ils le sont, sont-ils issus d'une loi donnée ?**

2°) Sur une série d'observations d'un couple de variables (X, Y):

•**Les variables sont-elles indépendantes l'une de l'autre ?**

•L'une d'elles est-elle « plus grande » que l'autre ?

3°) Sur plusieurs échantillons :

•Echantillons appariés : sont-ils de même loi ?

•Echantillons indépendants : sont-ils de même loi ?

V.1. Test d 'ajustement du χ^2

- $X_1,\ldots,X_n$ i.i.d. loi P
- H_o :{ $P = P_o$ donnée} contre H_1 : { $P \neq P_o$}
- *R* est partitionné en k classes $E_1, E_2,\ldots, E_k$

$$P_o\{E_j\} = p_j$$

- Parmi les n valeurs observées, n_1 sont dans E_1, …, n_k sont dans E_k
- Effectifs théoriques : $np_1,\ldots, np_k$

☞Distance entre les n_j observés et les np_j théoriques

102

Exemple : Le nombre d'accidents mensuels à un certain carrefour est une variable aléatoire X. On observe X durant 32 mois. :

nbre accidents : 0 1 2 3 4 5

nbre de mois : 2 13 8 4 4 1

Peut-on admettre au seuil de 10% que X suit une loi de Poisson de paramètre 2? L'enjeu n'est pas la valeur du paramètre, mais le fait que la loi soit une loi de Poisson : « loi des événements rares et indépendants », cela signifie que le taux instantané d'occurrence d'un accident est constant au cours du temps, que la probabilité d'occurrence de deux accidents dans un laps de temps h est négligeable devant h : l'hypothèse de loi de Poisson reflète ainsi des hypothèses non triviales sur le mécanisme d'occurrence des accidents à ce carrefour.

Procédure de test :

- Statistique du chi-deux : $\Delta = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}$
- sous P_o, Δ tend en loi vers un χ^2 à (k-1) degrés de liberté
- on refuse $\{P = P_o\}$ si $\Delta > A$

☞ il faut que tous les np_j soient > 5

– si pour un j : $3 < npj \leq 5$ correction de Yates :

$|n_j - np_j| + 0.5$ au lieu de $n_j - np_j$

☞ si r paramètres sont estimés parmi les p_j, le degré de liberté de Δ passe à (k-1-r).

103

Que les classes soient ordonnées ou non, la statistique n'en tient pas compte : en regroupant les observations par classes, on perd éventuellement une information importante contenue dans l'échantillon de départ. Par contre, le test s'applique aussi si X est qualitative non ordonnée.

Si pour une classe, l'effectif théorique est trop petit, on regroupe cette classe avec une autre.

Exemple : test d'ajustement à $\mathcal{P}(2)$

X_j	p_j	np_j	n_j	δ_j
0	0.1353	4.331	2	1.8503*
1	0.2707	8.661	13	2.1732
2	0.2707	8.661	8	0.0505
3	0.1804	5.774	4	0.5452
4	0.0902	4.572	5	0.1883*
≥ 5	0.0527			

$k = 5$, $dl = 4$, $\alpha = 10\%$: $\Delta_\alpha = 7.78$ $\Delta_{obs.} = 4.81$

Conclusion : $\mathcal{P}(2)$ n'est pas refusée au seuil 10%

* = correction de continuité de Yates

104

On regroupe les deux derniers cas, car l'effectif théorique de la dernière case est trop petit.

On compte le nombre de cases après avoir effectué tous les regroupements nécessaires.

PROB = P(chi-deux(4) > Δ_{obs}) est égale à 0.308 : supérieure à 10%, le seuil choisi : Δ_{obs} n'est pas dans la région critique. Au seuil de 10%, on ne peut refuser l'hypothèse d'une loi de POISSON.

On a utilisé la correction de Yates pour la première et pour la dernière case. Remarquer que ce n'est pas la valeur de l'effectif observé qui importe, mais celle de l'effectif théorique calculé.

V.2. Test d'indépendance de 2 facteurs

- Sur chaque individu sont notées les réalisations de deux variables qualitatives :

  variable X à k modalités

  variable Y à r modalités

- Résultats : tableau des nombres n_{ij} d'individus ayant la modalité i de X et la modalité j de Y.
- Test de H_o = {X et Y sont indépendantes}
- Statistique employée : le chi-deux distance des n_{ij} observés avec les effectifs théoriques calculés sous H_o

105

Variable = « caractère » = « facteur »

modalité = « niveaux » = « valeurs »

Tableau des résultats = « tableau de contingence à double entrée »

Présentation de la loi théorique du couple (X,Y), kr valeurs du couple, probabilités p_{ij}, lois marginales, $p_{i.}$ et $p_{.j}$

Sous Ho : $p_{ij} = p_{i.}\, p_{.j}$

Les (k-1)+(r-1) paramètres de la loi théorique sous Ho sont estimés à partir des observations : le degré de liberté du chi-deux sera égal à kr-1-(k+r-2) = (k-1)(r-1)

Exemple :

nombre d 'accidents par an X1 = {0.1 ou 2 accidents}, X2 = {au moins 3 accidents}

âge du conducteur Y1 = {de 18 à 25}, Y2 = {de 25 à 50}, Y3 = {plus de 50}

d'où tableau à 2 lignes et 3 colonnes.

Les modalités des variables ne sont pas nécessairement ordonnées : situation géographique, secteur d'activité, etc…

La statistique utilisée ne tient pas compte de l'ordre dans lequel sont écrites les modalités.

Si on fait des regroupements de classes, ce sont des regroupements de lignes ou des regroupements de colonnes

X= nbre d'accidents, Y= âge conducteur

x \ y	]18, 25]	]25, 50]	]50,..	total
x = 0,1,2	23	54	16	93
x >= 3	22	21	14	57
Total	45	75	30	150

x \ y	]18, 25]	]25, 50]	]50,..	total
x = 0,1,2	27.9	46.5	18.6	93
x >= 3	17.1	28.5	11.4	57
Total	45	75	30	150

$D_{obs.}$ = 6.405, dl = (3-1)(2-1) = 2, $D_{5\%}$ = 5.99

$$\hat{p}_{ij} = \frac{n_{i\bullet}}{N}\frac{n_{\bullet j}}{N} \text{ et l'effectif théorique est } N\hat{p}_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{N}$$

On calcule ainsi les effectifs théoriques du second tableau :

par exemple, 27.9 = 93*45/150

Tous les effectifs théoriques sont supérieurs à 5, on peut donc calculer le chi-deux :

$$D_{observé} = \frac{(23-27.9)^2}{27.9} + \frac{(54-46.5)^2}{46.5} + \frac{(16-18.6)^2}{18.6}$$
$$+ \frac{(22-17.1)^2}{17.1} + \frac{(21-28.5)^2}{28.5} + \frac{(14-11.4)^2}{11.4} = 6.405$$

6.405 > 5.99 : au seuil de 5%, on refuse l'hypothèse d'indépendance entre l'age du conducteur et le nombre d'accidents

Exemple tableau 2*2

Test de comparaison de fréquences

n1=	200
nA1=	128
n2=	150
nA2=	109
F1=	0,640
F2=	0,727
F=	0,6771
Z^2=	2,9449

Z =1,7161 **PROB = 0,0431**

unilatéral 1%	2,3263
unilatéral 2%	2,0537
unilatéral 4%	1,6449
unilatéral 10%	1,2816

Tableau de contingence 2x2

	A	nonA	total
Lozère	128	72	200
Aveyron	109	41	150
total	237	113	350

effectifs théoriques

	A	nonA	total
population1	135,43	64,57	200
population2	101,57	48,43	150
total	237	113	350

Δ = 2,9449 **PROB= 0,0862**

seuil 2%	5,4119
seuil 4%	4,2179
seuil 10%	2,7055
seuil 20%	1,6424

107

Lorsque les deux variables sont à 2 modalités (tableau de contingence 2x2), la question posée est :

La loi de la variable X sachant Y= y est-elle la même dans les deux cas y = 1 ou y = 2 ?

Cela est équivalent à la question suivante :

La probabilité que X=1 est-elle la même sachant Y=1 que sachant Y=2,

ou encore : p_1 est-elle égale à p_2?

On est ramenés à un problème de comparaison de fréquences entre deux populations : on avait déjà un test pour tester $\{p_1 = p_2\}$contre $\{p_1 < p_2\}$ (ou unilatéral dans l'autre sens, ou bilatéral).

Ici, il s'agit de faire le test bilatéral de $\{p_1 = p_2\}$contre $\{p_1 \neq p_2\}$.

Comparaison des deux tests : $|Z| > u$ lu dans N(0;1)

ou $\Delta > d$ lu dans chi-deux à 1 dl.

$|Z| > u$ est équivalent à $Z^2 > u^2$, lu dans chi-deux 1 dl (donc $u^2 = d$ pour un même seuil)

On a donc, au seuil α donné, deux régions critiques : $Z^2 > d$, ou $\Delta > d$

Pas d'affolement : les deux statistiques Z^2 et Δ sont identiques, les deux tests sont donc en fait identiques.

V.3. Test des séquences

- H_o = {X_1,…, X_n sont i.i.d.} (test d'homogénéité)
- remplacer chacun des x_i par son signe
- N_1=nbre de « + », N_2=nbre de « - »
- R = nombre de séquences
- refus de Ho si $\{R \le C_1(N_1,N_2)\}$ ou $\{R \ge C_2(N_1,N_2)\}$
- si $N_1 < 20$ et $N_2 < 20$: tables calculées à consulter

$$\frac{R-M}{S} \xrightarrow{Loi} N(0;1),$$

$$M = 1 + \frac{2N_1N_2}{n} \text{ et } S^2 = \frac{2N_1N_2(2N_1N_2 - n)}{n^2(n-1)}$$

108

Signe, ou position par rapport au précédent (suite croissante, ou décroissante)

On pourra l'appliquer aux résidus d'un ajustement MCO : il permettra de déceler éventuellement une relation non linéaire : résidus systématiquement positifs au début et à la fin, négatifs au milieu :

R = 3, N1 = 40, N2 = 35 sur n=75 observations

M = 1+ 2*35*40/75=38.33

S2 = 18.33

U = - 8.25

refus de Ho au seuil de 5% si |U| > 2

Ici, 8.25 > 2 : on refuse Ho

Remarque : hors programme du partiel en 2003-2004, mais au programme de l'examen de septembre.