Chapitre 7 : Tests d'ajustements, d'indépendance et de corrélation

Souvent, nous cherchons à ajuster une distribution à nos données. Une fois la distribution connue, il est possible de calculer toute probabilité d'intérêt.

7.1 Test d'ajustement du Khi-deux

Soit H_0 : La population suit la distribution « x »
H_1 : la population ne suit pas la distribution « x »

L'idée est de découper le domaine de la distribution en intervalles. Dans chaque intervalle, on calcule à partir de la loi spécifiée sous H_0 la fréquence théorique attendue. On compte ensuite combien d'observations l'on retrouve dans chaque intervalle. Il suffit alors de comparer les fréquences observées aux fréquences théoriques.

Supposons que l'on divise la distribution en « k » intervalles. Soit un intervalle « i » donné. La fréquence théorique attendue pour l'intervalle « i » est $E_i=np_i$. La statistique

$$Q = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \approx \chi^2_{k-p-1}$$

où « p » représente le nombre de paramètres estimés de la loi de distribution testée sous H_0.

Note : On recommande généralement de choisir les intervalles de sorte que $E_i \geq 5 \quad \forall i$.

Note : Pour un même jeu de données, il est courant que plusieurs distributions ne puissent être rejetées par ce test.

Exemple : On a 50 données dont la répartition est la suivante :

Intervalle	[0, 0,5[	[0,5 1,0[	[1,0 1,5[	[1,5 2,0[	[2,0 2,5[	[2,5 3,0[	[3,0 , ∞ [
Nombre observé	2	23	17	4	2	0	2

Les moyenne et écart-type de l'échantillon sont : $\bar{x}$=1,168 et s=0,591
Les fréquences théoriques pour une loi normale de moyenne 1,168 et de variance $0{,}591^2$ sont :

Intervalle	<0	[0, 0,5[	[0,5 1,0[	[1,0 1,5[	[1,5 2,0[	[2,0 2,5[	[2,5 3,0[	[3,0 ∞ [
Nombre théorique (E_i)	1,20	5,25	12,94	16,23	10,38	3,38	0,559	0,05

On regroupe les classes pour avoir $E_i>5$

Intervalle	$-\infty$, 0,5[	[0,5 1,0[	[1,0 1,5[	[1,5 ∞
Nombre théorique (E_i)	6,45	12,94	16,23	14,37
Nombre observé (O_i)	2	23	17	8

On calcule : Q= 13,75 à comparer à une χ^2_{4-2-1} . Au niveau α=5%, on lit $\chi^2_{1,.05} = 3{,}84$. On rejette H_0 : la distribution suit une loi normale. (Incidemment, les données de cet exemple ont été générées suivant une loi lognormale de paramètres logarithmiques (0, 0,25)).

7.2 Test d'ajustement de Kolmogorov-Smirnov

L'idée du test est de comparer la fonction de distribution expérimentale à la fonction de répartition théorique. On mesure la différence maximale entre ces deux fonctions (en valeur absolue).

La fonction de répartition expérimentale s'obtient facilement en classant les valeurs par ordre croissant, x_1, x_2,...x_n, puis en notant : $F_e(x) = \begin{cases} 0 & x < x_1 \\ i/n & x_i \le x < x_{i+1} \\ 1 & x \ge x_n \end{cases}$.

On calcule la différence maximale par :
$D_{\max} = \max\left(\left|F_t(x) - F_e(x)\right|\right)$, le maximum se trouvant nécessairement à un des x_i dû à la forme en escalier de la fonction $F_e(x)$. F_t(x) est la distribution théorique de la distribution entièrement spécifiée sous H_0.

Les valeurs critiques de D_{max} ont été tabulées par divers auteurs[1].

n	α=0.10	α=0.05	α=0.01
5	0.51	0.56	0.67
10	0.37	0.41	0.49
15	0.30	0.34	0.40
20	0.26	0.29	0.35
25	0.24	0.26	0.32
30	0.22	0.24	0.29
40	0.19	0.21	0.25
n>40	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.63/\sqrt{n}$

Le test K-S permet de tester n'importe quelle distribution. Il est normalement plus puissant que le test du Khi-deux (i.e. il permet de rejeter plus facilement H_0) et il a l'avantage de ne pas requérir de séparer arbitrairement le domaine en intervalles.

Note : Lorsque les paramètres spécifiant la distribution sont estimés des mêmes données que celles utilisées dans le test, il s'ensuit un ajustement aux données que les valeurs critiques devraient refléter (ces valeurs critiques devraient être revues à la baisse). Des tables « révisées » existent

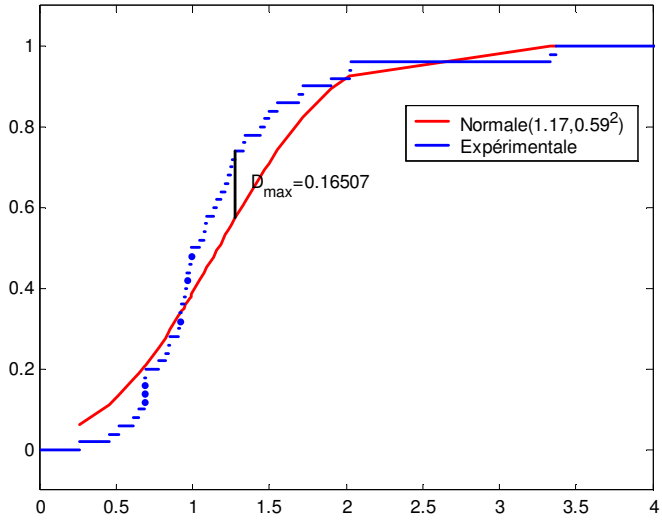
[1] Lindgren, 1962. Statistical Theory. MacMillan, New York

pour certaines distributions particulières. Dans la pratique, lorsque « n » est grand, on peut utiliser la table précédente comme test (très) approximatif (i.e. si on rejette H_0 on aurait rejeté aussi avec la bonne valeur critique; si on ne rejette pas H_0 on ne peut pas conclure).

Exemple : Mêmes données que précédemment :

x=0.27	0.68	0.78	0.92	0.96	1.05	1.16	1.26	1.47	1.91
0.45	0.68	0.82	0.92	0.96	1.08	1.18	1.28	1.49	2.02
0.52	0.69	0.84	0.93	0.98	1.09	1.22	1.33	1.56	2.03
0.61	0.69	0.85	0.94	0.99	1.10	1.23	1.34	1.69	3.33
0.65	0.69	0.91	0.96	1.00	1.14	1.25	1.44	1.72	3.37

On obtient :



Ici n=50, de la table on tire $D_{table}=1{,}36/50^{0.5}=0{,}192$. $D_{max}<D_{table}$ ici, on arrive à la conclusion contraire à celle obtenue avec le test Khi-deux, i.e. on ne peut pas rejeter l'hypothèse que la distribution soit normale[2]. Par contre, Si l'on fait le test après correction pour l'estimation des paramètres de la loi normale, on rejette H_0.

7.2.1 Test de Kolmogorov-Smirnov pour deux populations

Si l'on a deux échantillons différents et que l'on veut tester si les deux échantillons peuvent provenir de la même population, on peut utiliser le test K-S avec les mêmes valeurs critiques que précédemment. Il suffit de construire les deux fonctions de distribution expérimentales, de calculer l'écart maximal entre les deux distributions (nécessairement à une des valeurs observées) et de comparer l'écart à la valeur critique correspondante avec cette fois $n = \dfrac{n_1 n_2}{n_1 + n_2}$.

[2] Si l'on adapte les valeurs critiques pour tenir compte que les paramètres de la loi normale ont été estimés, on devrait utiliser la valeur $L_{table}=0{,}886/50^{0{,}5}=0{,}125$. Dans ce cas, on rejetterait H_0. La modification à la statistique calculée dans le cas spécifique de la loi normale a été obtenue par Lilliefors par simulation.

7.3 Test d'indépendance entre deux variables (test du Khi-deux)

Un tableau de contingence est un tableau croisant les valeurs de deux variables (qualitatives ou quantitatives, discrètes ou continues. L'on note la fréquence d'observation des différentes valeurs des deux variables. Pour une variable continue, celle-ci est découpée en intervalles. Il s'agit en quelque sorte de la généralisation à deux variables du concept d'histogramme.

Exemple :

		Variable 2			
		Valeur (ou intervalle) 1	Valeur (ou intervalle) 2	Valeur (ou intervalle) 3	
Variable 1	Valeur (ou intervalle) 1	n_{11}	n_{12}	n_{13}	$n_{1.}$
	Valeur (ou intervalle) 2	n_{21}	n_{22}	n_{23}	$n_{2.}$
		$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{..}$

Sous hypothèse d'indépendance, la distribution conjointe est simplement le produit des distributions marginales, i.e. $f_{ij} = f_i f_j$. Si l'on estime f_{ij} par $n_{ij}/n_{..}$ et f_i par $n_{i.}/n$, on devrait donc avoir $n_{ij} \approx \frac{n_{i.} n_{.j}}{n..}$.
L'idée est de calculer l'écart entre les deux termes, le n_{ij} observé (noté O_{ij}) et le n_{ij} prédit ou théorique (noté E_{ij}), si cet écart devient trop important, on devra rejeter l'hypothèse que les variables sont indépendantes. On calcule :

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(n_{ij} - n_{i.}n_{.j}/n_{..}\right)^2}{n_{i.}n_{.j}/n_{..}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

La statistique Q est distribuée approximativement suivant une $\chi^2_{(r-1)(c-1)}$ où « r » et « c » désignent le nombre de valeurs ou intervalles des deux variables.

Note : Comme pour le test d'ajustement, il faut que la fréquence théorique $E_{ij} \geq 5 \quad \forall i, j$ pour que le test soit valide.

Exemple : Une flotte d'autobus est équipée de 4 types de pneus (A, B, C ,D). On mesure le kilométrage parcouru avant usure du pneu. On construit 3 classes de kilométrage (en milliers) <20, [20,30], >30 . On a obtenu les résultats suivants :

Observé	A	B	C	D	Total
<20	26	23	15	32	96
[20,30]	118	93	116	121	448
>30	56	84	69	47	256
Total	200	200	200	200	800

Les deux variables sont-elles indépendantes?

On calcule le tableau des fréquences théoriques :

Théorique	A	B	C	D	Total
<20	24	24	24	24	96
[20,30]	112	112	112	112	448
>30	64	64	64	64	256
Total	200	200	200	200	800

et $Q=(26-24)^2/24+(23-24)^2/24+....+(69-64)^2/64+(47-64)^2/64=22,82$.
On compare à une $\chi^2_{(4-1)(3-1),.05} = \chi^2_{6,.05} = 12,59$. Q>12,59, donc on rejette l'hypothèse que le kilométrage obtenu soit indépendant de la marque de pneus.

7.4 Test sur le coefficient de corrélation simple entre deux variables quantitatives suivant une distribution binormale

Soit l'hypothèse H_0 : $\rho = 0$. Sous cette hypothèse, on a :

$$\frac{r^2(n-2)}{1-r^2} \sim F_{1,n-2}$$

De façon équivalente, on a aussi :

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

Cette dernière statistique permettant un test unilatéral ou bilatéral.

Le test découle directement.

Note : Si H_0 est plutôt du type $\rho = \rho_0$ alors il faut recourir à un autre test. On utilise alors le fait que : $1/2 \ln\left(\frac{1+r}{1-r}\right) \approx N\left(1/2 \ln\left(\frac{1+\rho_0}{1-\rho_0}\right), \frac{1}{n-3}\right)$ pour construire le test.

Exemple : Dix échantillons de sols ont été prélevés pour lesquels on a mesuré la porosité (n) et la conductivité hydraulique (K) en laboratoire. On a obtenu une corrélation de 0,6 entre log(K) et n. Cette corrélation est-elle significative au niveau $\alpha = 0,05$?

Ici, on pouvait prévoir que la corrélation serait positive, il semble donc plus indiqué d'effectuer un test unilatéral. on calcule $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,6\sqrt{8}}{\sqrt{1-0,6^2}} = 2,12$. Comme $t_{8,.05}$=1,86, on rejette H_0, i.e. la corrélation observée est significative.

7.5 Test sur le coefficient de corrélation de rang (Spearman) entre deux variables quantitatives

Le coefficient de corrélation de rang n'est rien d'autre que le coefficient de corrélation usuel calculé sur les rangs plutôt que les données brutes. L'avantage est que ce coefficient n'exige pas une relation linéaire entre les deux variables (il faut néanmoins que les deux variables soient reliées de façon monotone). Les tests précédents s'appliquent à ce coefficient pour fournir un test approximatif. Les cas d'égalité sont traités de différentes façons dans la littérature, une de celles-ci consistant à octroyer le rang moyen aux valeurs égales.