Tests d'ajustement et d'Indépendance

I. *Test d'ajustement:*

1. *Test de χ^2*

Soit un échantillon de taille n extrait d'une population et divisé en k classes de probabilité $p_1, p_2, \dots, p_k$ d'effectifs respectifs $n_1, n_2, \dots, n_k$. On évidemment

$$\sum_{i=1}^{k} n_i = n$$

Il s'agit du test
$$\begin{cases} H_0 : F(x) = F_0(x) \\ H_1 : F(x) \neq F_0(x) \end{cases}$$

*Ou **F(x)** est la fonction de répartition de la variable échantillonnée et **$F_0(x)$** la fonction de répartition de la variable aléatoire connue.*

*En supposons que la variable étudiée suit une loi spécifique, on peut déterminer à partir de sa fonction de répartition l'effectif théorique **np_i** (sous H_0).*

On considère la quantité

$$k_n^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Qui suit une loi de χ^2 à (k-1) degrés de liberté. Le seuil de signification α étant fixé, on a $P(\chi^2_c < \chi^2_{k-1}(\alpha)) = 1-\alpha$ et si la valeur calculée k_n^2 est inferieur à $\chi^2_{k-1}(\alpha)$ lue dans la table de χ^2 , on accepte H_0.

*En estime parfois des paramètres pour déterminer **$F_0(x)$** (par exemple, m et σ^2pour une loi normale N(0,1)).*

Il y a alors réduction du nombre de degrés de liberté du χ^2_c. si on estime deux paramètres la quantité k_n^2 suit une loi à (k-1-2) degrés de liberté. On admet que $k_n^2 \sim \chi^2_{k-1}$ si np_i est supérieur à 5. Parfois on doit procéder à des regroupements.

Exemple: *soit la statistique suivante:*

Xi	0	1	2	3	4	5
n_i	7	11	8	3	2	1

Peut-on admettre que la distribution empirique peut-être ajustée par une loi de Poisson, au seuil de signification α=0.05?

On a

$$\overline{X} = 1.53 \text{ et } s^2 = 1.62$$

Si une variable X suit une loi de Poisson , on E(X)=V(X). on peut supposer que la distribution empirique soit ajustée par une loi de Poisson de paramètre λ= 1.53:

$P(X=k) = e^{-1.53} (1.53)^k /k!$ et on dresse le tableau:

x_i	n_i	$P(X=x_i)= p_i$	np_i
0	7	0.2165	6.928
1	11	0.3312	10.598
2	8	0.2534	8.1088
3	3	0.1292	4.1344
4	2	0.0494	1.5808
5	1	0.0151	0.4832

Les effectifs np_i doivent être supérieurs à 5. On regroupe les dernières classes.

x_i	n_i	np_i	$n_i - np_i$	$(n_i - np_i)^2/ np_i$
0	7	6.928	0.72	0.00075
1	11	10.598	0.4016	0.011521
2	8	8.1088	-0.1088	0.00145
3;4;5	6	6.1984	-0.1984	0.00635
Total	-	-	-	0.02376

*On a $\chi^2_c = 0.02376$ et **P** ($\chi^2 < \chi^2$ (α)) = 0.95.on trouve $\chi^2_2 = 5.991$ car 2= k-1-1 = 4-1-1 est le nombre de degrés de liberté et il y a 1 paramètre estimé **λ** et k est le nombre de classes. Comme $\chi^2_c < \chi^2$ (α) , on accepte l'hypothèse H_0 : ajustement par une loi de Poisson de paramètre **λ**= 1.53.*

2. Test de Kolmogorov

*En désignant par **F(x)** la fonction de répartition de la variable échantillonnée et **$F_n(x)$** la fonction de répartition empirique (fréquence relative cumulée) pour un échantillon de taille **n**.*

On considère la quantité $D_n = Max \left| F_n(x) - F(x) \right|$ et en compare D_n à des valeurs critiques tabulées d_n. on rejette H_0 si $D_n > d_n$

Pour α = 0.05 ; n>80 on a $D_n > 1.36 / \sqrt{n}$

α = 0.01 ; n>80 on a $D_n > 1.63 / \sqrt{n}$

II.Test d'Indépendance:

*Le test vise à rechercher s'il existe une liaison entre deux variables étudiées **X** et **Y** dans une population de **n** individus.*

on teste $\begin{cases} H_0 : X \text{ et } Y \text{ sont indépendantes} \\ H_1 : X \text{ et } Y \text{ ne sont pas indépendantes} \end{cases}$

Soit n_{ij} le nombre d'individus présentant la modalité i de X et la modalité j de Y. on présente les données sous la forme de tableau dit tableau de contingence.

Y / X	y_1	Y_2		Y_j	...	Y_q	Total
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	$n_{2.}$
			...		...		
x_i	n_{i1}	n_{11}	...	n_{ij}	...	n_{iq}	$n_{i.}$
. . .	. . .	. . .	...	. . .	...	. . .	. . .
x_p	n_{p1}	n_{p2}	...	n_{pj}	...	n_{pq}	$n_{p.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.q}$	n

L'effectif théorique au rang (i,j) est

$$c_{ij} = \frac{n_{i.}n_{.j}}{n}$$

Car sous l'hypothèse $\mathbf{H_0}$ *, on a* $\mathbf{p_{ij} = p_{i.}\, p_{.j}}$ *. l'effectif théorique au rang (i,j) est* $\mathbf{n\, p_{i.}\, p_{.j}}$ *. la quantité*

$$k_n^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\left(n_{ij} - np_{i.}p_{.j}\right)^2}{np_{i.}p_{.j}}$$

est une relation $\boldsymbol{\chi^2}$ *à pq-1 degrés de liberté.*

On estime les valeurs $\mathbf{p_{i.}}$ *et* $\mathbf{q_{.j}}$ *par les quantités*

$\frac{n_{i.}}{n}$ et $\frac{n_{.j}}{n}$ *respectivement et on cherche alors la quantité*

$$k_n^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

qui est une réalisation du $\boldsymbol{\chi^2}$ *à (p-1)(q-1) degrés de liberté car on a procédé à (p-1)(q-1) estimations des* $\mathbf{p_{i.}}$ *et* $\mathbf{p_{.j}}$

Pour un seuil fixé α , on a $\mathbf{P}\,(\boldsymbol{\chi^2} < \boldsymbol{\chi^2_{(p-1)(q-1)}}\,(\boldsymbol{\alpha})\,) = \mathbf{1} - \boldsymbol{\alpha}$ *et on accepte* $\mathbf{H_0}$ *si* $\boldsymbol{\chi^2_c} < \boldsymbol{\chi^2_{(p-1)(q-1)}}\,(\boldsymbol{\alpha})$ *ou* $\boldsymbol{\chi^2_{(p-1)(q-1)}}\,(\boldsymbol{\alpha})$ *est une valeur lue dans la table du* $\boldsymbol{\chi^2}$*.*

<u>Exemple:</u>

On traite trois échantillons de malades atteints d'une maladie M, selon leur classe d'âges et leur sexe. On veut savoir si l'âge (A) et le sexe (S) peuvent être considérés comme des facteurs indépendants à un seuil de signification de 5 %.

Soit le tableau résumant le nombre de malades suivant le sexe dans les classes d'âge.

A / S	A < 30 ans	30 ans≤ A < 60 ans	60 ans ≤ A	Total
Hommes	**58**	**90**	**49**	**197**
Femmes	**14**	**17**	**14**	**45**
Total	**72**	**107**	**63**	**242**

On forme un tableau d'effectifs théorique. Il t y a (197 / 242)=81.14 % d'hommes (sur le total des malades). On cherche les 81.14% de 72 ; 107 et 63 (nombre total des malades selon la tranche d'âge) et on complète le tableau puisque les sommes marginales restent les même.

*On trouve alors (197 / 242) 72 = 58.624% , (197 / 242) 107 = 87.14% **,** (197 / 242) 632 = 51.28%.*

le même tableau aurait pu trouvé en calculant les (45 / 242) = 18.6% de femmes (sur le total des malades)puis en complète le tableau de la même façon que précédemment.

A / S	A < 30 ans	30 ans≤ A < 60 ans	60 ans ≤ A	Total
Hommes	**58.62**	**87.14**	**51.28**	**197**
Femmes	**13.38**	**19.9**	**11.72**	**45**
Total	**72**	**107**	**63**	**242**

On calcule ensuite

$$k_n^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

$$\chi_c^2 = \frac{(58-58.62)^2}{58.62} + \frac{(14-13.28)^2}{13.28} + \cdots + \frac{(14-11.72)^2}{11.72} = 1.099$$

Le nombre de degré de liberté est : (p-1)(q-1)=(2-1(3-1)=2 ou p est le nombre de lignes et q le nombre de colonnes du tableau.

La lecture de la table du $\chi^2{}_2$ donne
$\chi_2^2\ (\alpha) = 5.991$ *ou α=0.05*

Comme 1.099 = $\chi^2{}_c < \chi^2{}_2(0.05)=5.991$, on accepte l'hypothèse selon laquelle la répartition des malades par classes d'âge n'est pas affectée par le sexe.