Chapitre 7 :

Les tests non paramétriques

Dans ce chapitre, nous allons examiner comment nous pouvons construire un test sur la distribution d'une certaine variable X et sur la relation de cette variable avec d'autres.

I- Le test d'ajustement :

Considérons, dans une population, une certaine variable X dont on ne connaît pas la nature de la distribution .on tire un échantillon aléatoire de cette population et l'on se demande si nous pouvons accepter ou non que l'échantillon obtenu se conforme tout à fait à une distribution particulière spécifiée.

Les hypothèses du test sont :

H_0 : La distribution observée est tirée d'une population mère caractérisée par une distribution théorique f(x, θ) où θ est un paramètre connu ou inconnu.

H_1 : La distribution observée n'est pas tirée d'une population mère caractérisée par une distribution théorique f(x, θ)

Le prélèvement de l'échantillon permettra ultérieurement d'accepter ou de rejeter H_0 avec un risque d'erreur α.

Pour effectuer ce test on utilise la loi suivante :

$$\frac{\sum_{i=1}^{n}(n_i - np_i)^2}{np_i} \rightarrow \chi^2(k-l-1)$$

k : Le nombre de classes dans la distribution

l : Nombre de paramètres inconnus de la distribution (qu'on doit estimer)

p_i : Probabilité théorique de la classe i

np_i : Effectif théorique de la classe i

n_i : Effectif observé de la classe i

règle de décision :

Soit α le niveau de risque accepté. Le seuil critique χ^2_α est donné tel que :

$P(\chi^2 > \chi^2_\alpha) = \alpha$

Soit χ^2_c la statistique calculée à partir de l'échantillon :

- Si $\chi^2_c \leq \chi^2_\alpha$: On accepte l'hypothèse H_0

- Si $\chi^2_c > \chi^2_\alpha$: On rejette l'hypothèse H_0

Remarques :

- S'il existe un ou plusieurs effectifs théoriques $np_i<5$ on effectue des regroupements

- La taille de l'échantillon doit être suffisamment grande ($n \geq 30$)

Exemple1 (ajustement à une loi uniforme) :

Le responsable des jeux d'un casino veut vérifier si les dés utilisés sont bien équilibrés. Il prend au hasard d'un dé et le jette 120 fois. Les résultats obtenus sont résumés dans le tableau suivant :

Face	1	2	3	4	5	6
Effectif	14	16	28	30	18	14

Peut-il conclure au seuil de signification α =5% que le dé est bien équilibré ?
Il s'agit de tester l'hypothèse le dé est équilibré contre l'hypothèse le dé n'est pas équilibré.

H_0 : Le dé est équilibré (la répartition des résultats est uniforme)
H_1 : Le dé n'est pas équilibré (la répartition des résultats n'est pas uniforme)

Pour décider entre H_0 et H_1 il faut calculer la statistique de χ^2_c et la comparer à celle figurant sur la table χ^2_α

face	Effectif observé (ni)	Probabilité théorique (pi)	Effectif théorique ($n.p_i$)	$(n_i-n.p_i)^2/n.p_i$
1	14	1/6	20	$6^2/20=1.8$
2	16	1/6	20	$4^2/20=0.8$
3	28	1/6	20	$8^2/20=3.2$
4	30	1/6	20	$10^2/20=5$
5	18	1/6	20	$12^2/20=0.2$
6	14	1/6	20	$6^2/20=1.8$
	n=120			χ^2_c**=12.8**

ddl= k-l-1 = 6-0-1=5

$\chi^2_{5\%}(5)$ = 11.07 (d'après la table de χ^2)

$\chi^2_{5\%}(5) < \chi^2_c$

→ On retient l'hypothèse H_1 → le dé est truqué.

Rappel

Lorsqu'on veut tester l'ajustement d'une distribution donnée à une certaines loi on est parfois amené à estimer certains paramètres de la loi théorique.

Loi	Paramètres de la loi	Estimateurs
Normale	m et σ	$\overline{X}$ et S
Poisson	λ	$\overline{X}$
Binomiale	P	$\overline{X}$ /n' n' : Taille de chaque lot

Exemple2 (ajustement à une loi de Poisson) :

Dans une usine on a effectué une étude sur le nombre de pannes mensuelles des machines de production. Les données sur les cinq dernières années sont résumées dans le tableau suivant :

Nombre de pannes mensuelles	Nombre de mois
0	9
1	15
2	18
3	11
4	6
5 et plus	1

Est-ce qu'on peut dire au seuil de signification α=1% que le nombre de pannes mensuelles suit une loi de poisson ?

Il faut d'abord estimer le paramètre λ de la loi de poisson à partir de la moyenne des observations :

Nombre de pannes mensuelles x_i	Nombre de mois ni	$X_i\ n_i$
0	9	0
1	15	15
2	18	36
3	11	33
4	6	24
5 et plus	1	5
	$\sum n_i = 60$	$\sum x_i n_i = 113$

$\hat{\lambda} = \overline{X} = \sum x_i n_i / \sum n_i = 113/60 = 1.88$

Il s'agit de tester l'hypothèse que le variable aléatoire nombre de pannes mensuelles qu'on peut noter X suit une loi de poisson de paramètre

λ = 1.88 contre l'hypothèse que cette variable ne suit pas une loi de poisson.

H_0 : X ↝ P(1.88)

H1 : X ↜ P(1.88)

Nombre de pannes x_i	Effectif observé n_i	Prob théorique p_i $P(X=x_i) = e^{-1.88}1.88^x/\ x!$	Effectif théorique $e_i = n \times p_i$	Calcul de X^2 $= \sum(n_i - e_i)^2/e_i$
0	9	0.152	9.12	0.0016
1	15	0.286	17.16	0.272
2	18	0.269	16.14	0.214
3	11	0.168	10.08	0.084
4	6 } 7	0.079	4.74 } 6.528	0.0341
5 et plus	1 }	0.0298	1.788 }	
Total	**n=60**			χ^2_c **= 0.6057**

Après regroupement des effectifs théoriques inférieures à 5 on dispose de 5 classes d'où ddl= k-l-1 = 5-1-1 = 3

$\chi^2_{1\%}(3)$ = 11.345 (d'après la table de χ^2)

$\chi^2_{1\%}(3)$ > χ^2_c

→ On retient l'hypothèse H_0 → La variable aléatoire nombre de panne suit une loi de poisson de paramètre $\lambda = 1.88$

Exemple 3 (ajustement à une loi binomiale) :

Dans une entreprise fabriquant des tubes de verre, on effectue un contrôle visuel sur un échantillon de 320 lots formé chacun de 5 tubes.

Tubes défectueux	Nombre de lots
0	18
1	56
2	110
3	88
4	40
5	8
Total	320

Est-ce qu'on peut dire au seuil de signification α=5% que le nombre de tubes défectueux suit une loi binomiale?

Le nombre de tubes défectueux varie de 0 à 5, le premier paramètre de la loi n, est donc connue (n=5), quant à la proportion p des tubes défectueux p, elle peut être estimée à partir de la moyenne des observations :

Nombre de tubes défectueux x_i	Nombre de lot ni	$x_i\, n_i$
0	18	0
1	56	56
2	110	220
3	88	264
4	40	160
5	8	40
	$\sum n_i = 320$	$\sum x_i n_i = 740$

$\widehat{P} = \bar{X} = \sum x_i n_i / n \sum n_i = 740 / (5x320) = 0.46$

Il s'agit de tester l'hypothèse que la variable aléatoire nombre de tubes défectueux qu'on peut noter X suit une loi de binomiale de paramètre n=5, p=0.46 contre l'hypothèse que cette variable ne suit pas une loi binomiale.

H_0 : X ↛ B(5 ; 0.46)

H1 : X ↚ B(5 ; 0.46)

Nombre de pannes x_i	Effectif observé n_i	Prob théorique p_i $P(X=x_i) = C_5^x\ 0.46^x \times 0.54^{5-x}$	Effectif théorique $e_i = n \times p_i$	Calcul de X^2 $= \sum(n_i - e_i)^2/e_i$
0	18	0.045	14.4	0.9
1	56	0.193	61.8	0.54
2	110	0.332	106.3	0.13
3	88	0.286	91.3	0.12
4	40	0.123	39.4	0.01
5	8	0.021	6.8	0.21
Total	**n=320**	$\sum$ **=1**	$\sum$ = 320	χ_c^2 **= 1.91**

ddl= k-l-1 = 6-1-1 = 4

$\chi^2_{5\%}(4)$ = 9.49 (d'après la table de χ^2)

$\chi^2_{5\%}(4)$ > χ^2_c

→ On retient l'hypothèse H_0 → La variable aléatoire nombre de tubes défectueux suit une loi binomiale de paramètre n=5 et p=0.46

II- Le test d'indépendance :

On considère deux variable X et Y dans une même population. La question et de savoir si ces deux variable sont indépendantes ou liées. Le test de Khi-deux permet de répondre à cette question.

Les hypothèses à considérer se notent :

H_0 : Les deux variables X et Y étudiées dans cette population sont **indépendantes**

H_1 : Les deux variables X et Y étudiées dans cette population sont **dépendantes**

Pour effectuer ce test, on tire un échantillon de taille n de cette population. On compare les effectifs observés n_{ij} (associés aux différentes couples de modalités de X et Y) avec les effectifs théoriques n'_{ij}.

Les observations au niveau de l'échantillon sont regroupées dans un tableau de contingence :

	y_1	….	y_j	…..	y_r	n_j
X_1	n_{11}	….	n_{1j}	….	n_{1r}	n_1
….	…….	…	….	…….	…….	…….
X_j	n_{i1}		n_{ij}	….	n_{ir}	n_j
….	…	….	…..	…..	…..	……
X_k	n_{k1}	….	n_{kj}	…..	n_{kr}	n_k
n_j	n_1	…	$n_{,j}$	….	n_r	N

(k modalités pour la variable X)

(r modalités pour la variable Y)

Les effectifs théoriques en cas d'indépendance s'obtiennent en utilisant la propriété suivante :

$$n'_{ij} = \frac{n_{i.} * n_{.j}}{n}$$

On est donc en présence de deux distributions : L'une empirique et l'autre théorique dont on cherche l'adéquation. On utilise la statistique :

$$A = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}}$$

La statistique $A \rightarrow \chi^2[(k-1)(r-1)]$ sous les hypothèses que :

- Les k*r observations sont indépendantes
- La taille n de l'échantillon est suffisamment grande (n≥30)
- Les effectifs théoriques n'_{ij} sont suffisamment grands ($n'_{ij} \geq 5$)

Règle de décision :

Soit α le niveau de risque accepté. Le seuil critique X^2 α est donné tel que :

$P(X^2 > \chi^2_\alpha) = \alpha$

Soit χ^2_c la statistique calculée à partir de l'échantillon :

- Si $\chi^2_c \leq \chi^2_\alpha$: On accepte l'hypothèse H_0

- Si $\chi^2_c > \chi^2_\alpha$: On rejette l'hypothèse H_0

Exemple :

Une entreprise de produits cosmétiques fabrique trois types de produit. Afin d'ajuster sa stratégie marketing, elle veut connaître s'il y a ou non un lien de dépendance entre l'âge du consommateur et le type de produit préféré. Une enquête sur un échantillon de 200 personnes a donné les résultats suivants :

	Produit 1	Produit 2	Produit 3
Moins de 20 ans	13	19	25
Entre 20 et 40 ans	28	29	28
Plus que 40 ans	24	18	16

Effectuer le test adéquat pouvant aider cette entreprise (α=5%)

Si on désigne par X l'âge du consommateur et par Y le type de produit préféré, les hypothèses du test s'expriment ainsi :

H_0 : X et Y sont indépendants

H_1 : X et Y sont dépendants (le type de produit préféré dépend de l'âge du consommateur).

Tableau des effectifs observés :

	Produit 1	Produit 2	Produit 3	Totaux
Moins de 20 ans	13	19	25	57
Entre 20 et 40 ans	28	29	28	85
Plus que 40 ans	24	18	16	58
Totaux	65	66	69	200

Tableau des effectifs théoriques :

	Produit 1	Produit 2	Produit 3
Moins de 20 ans	18.525	18.81	19.665
Entre 20 et 40 ans	27.625	28.05	29.325
Plus que 40 ans	18.85	19.14	20.01

Le calcul de χ_c^2 :

χ_c^2 = (18.525-13)²/18.525 + (18.81- 19)²/18.81 + (19.665 -25)²/19.665
+ (27.625-28)²/27.625 + (28.05-29)²/28.05 + (29.325-28)²/29.325
+ (18.85 – 24)²/18.85 + (19.14- 18)²/19.14 + (20.01- 16)²/20.01
= 1.647 + 0.0019 + 1.447 + 0.005 + 0.0321 + 0.059 + 1.407 + 0.0678 + 0.836
= 5.5

ddl= (3-1) x (3-1) = 2x2 = 4

$\chi^2_{5\%}(4)$ = 9.488 (d'après la table de χ^2)

$\chi_c^2 < \chi^2_{5\%}(4)$

→ On retient l'hypothèse H_0 (le type de produit préféré est indépendant de l'âge du consommateur)