Tests non paramétriques

Introduction

Les tests précédents portaient sur des paramètres des distributions observées (moyennes, fréquences). Mais on devait faire l'hypothèse *a priori* de la normalité des distributions parentes.

Au contraire, les tests non paramétriques

– ne nécessitent pas d'hypothèse a priori sur les distributions parentes

– peuvent s'appliquer à des variables ordinales (tests sur les rangs) ou même qualitatives (khi-2)

La mise en œuvre de la plupart de ces tests se fait en deux étapes :

– on construit un protocole dérivé : signes, rangs, etc

– le test proprement dit porte sur les variables dérivées.

Il existe de nombreux tests non paramétriques. Nous n'étudierons que les plus courants.

Indépendance de deux variables nominales - Test du χ^2

Deux variables nominales X et Y observées sur un échantillon de sujets.
Nombre de modalités de X : l
Nombre de modalités de Y : c

Problème : ces deux variables sont-elles indépendantes entre elles ?

Exemple : trois groupes de musiciens : professionnels (MP), en cours de professionnalisation (MCP) et amateurs (MA).
On s'intéresse au niveau d'études des trois groupes.
Effectifs observés

	MP	MCP	MA	Total
avant bac.	7	11	4	22
bac.	12	6	5	23
post bac.	17	13	20	50
Total	36	30	29	95

Le niveau d'études et type de professionnalisation sont-ils liés ?

Hypothèses :

H_0 : Les variables X et Y sont indépendantes.
H_1 : Les variables X et Y sont dépendantes.

Statistique de test

Distance du χ^2 entre le tableau des effectifs observés et un tableau *d'effectifs théoriques* (cf. calcul infra).

Cette statistique suit une loi du χ^2 à $(l-1)(c-1)$ ddl.

Calcul de la distance du χ^2

Données observées : tableau de contingence.

Effectifs attendus (ou théoriques) si indépendance :

Dans chaque case :

$$\text{Effectif théorique} = \frac{\text{total ligne} \times \text{total colonne}}{\text{total général}}$$

Contribution de chaque case au χ^2 :

$$\text{Ctr}_i = \frac{(\text{Eff. Observé} - \text{Eff. Théorique})^2}{\text{Eff. Théorique}}$$

Distance du χ^2 : $\chi^2_{obs} = \sum \text{Ctr}_i$.

Sur l'exemple fourni :

– On choisit un seuil de 5%.
– Le nombre de ddl est : $(3-1) \times (3-1) = 4$.
– Valeur critique : $\chi^2_{crit} = 9.49$

Effectifs observés

	MP	MCP	MA	Total
avant bac.	7	11	4	22
bac.	12	6	5	23
post bac.	17	13	20	50
Total	36	30	29	95

Effectifs théoriques

	MP	MCP	MA
avant bac.	8.34	6.95	6.72
bac.	8.71	7.26	7.02
post bac.	18.95	15.79	15.26

Calcul de la "distance" du χ^2

Mod.	n_{ij}	t_{ij}	$\frac{(n_{ij}-t_{ij})^2}{t_{ij}}$
MP. < Bac	7	8.34	0.21
MP. Bac	...		1.23
MP. > Bac			0.20
MCP. < Bac			2.36
MCP. Bac			0.22
MCP. > Bac			0.49
MA. < Bac			1.09
MA. Bac			0.58
MA. > Bac			1.47
Total			7.85

On obtient : $\chi^2_{obs} \leq \chi^2_{crit}$.

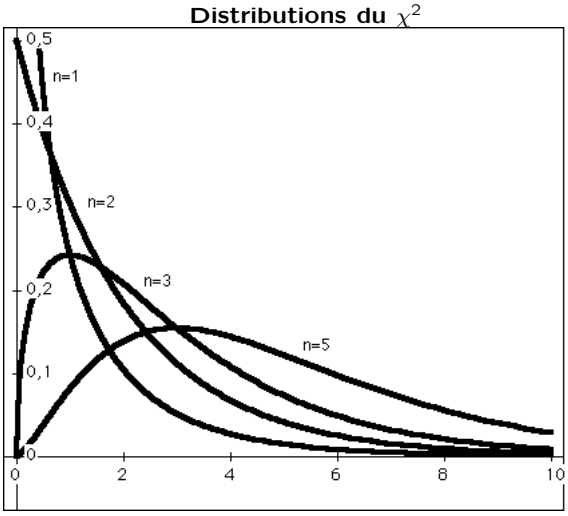
– Conclusion : On n'a pas mis en évidence de différence de niveau d'étude selon le type de professionnalisation.

Remarques

– Correction de Yates pour les tableaux 2x2

– Condition sur les effectifs théoriques minimaux : moins de 20% des cases avec des effectifs théoriques strictement inférieurs à 5 ; pas de case avec un effectif théorique strictement inférieur à 1.

Distributions du χ^2



Tests non paramétriques sur deux groupes indépendants

Test de la médiane sur des groupes indépendants

Une variable (la variable indépendante) définit deux groupes indépendants.

Une deuxième variable ordinale ou numérique.

Hypothèses

H_0 : Les deux populations parentes ont même médiane.
H_1 : Les deux populations parentes ont des médianes différentes

Construction de la statistique de test

On détermine la médiane M de la série obtenue en réunissant les deux échantillons.

On constitue un tableau de contingence en croisant la variable indépendante et la variable dérivée "position par rapport à M"

	Gr 1	Gr 2	Ensemble
$\leq M$	N_1	N_2	$N_1 + N_2$
$> M$	N_3	N_4	$N_3 + N_4$
Total	$N_1 + N_3$	$N_2 + N_4$	N

On fait un test du χ^2 sur le tableau obtenu.

Exemple

31 basketeurs de 14 ans, répartis en deux groupes d'effectifs $n_1 = 12$ et $n_2 = 19$, selon le jugement porté par l'entraîneur (groupe G_1 : jugement négatif ; groupe G_2 : jugement positif). On a relevé la taille de chaque sujet.

G_1 : 152 163 164 173 174 176 177 177 178 178 181 184

G_2 : 167 171 172 174 175 176 176 177 179 179 180 182 183 186 188 189 189 193 195

Les deux groupes sont-ils significativement différents du point de vue de la taille ?

Détermination de la médiane

152 163 164 167 171 172 *173 174* 174 175 176 176 *176* 177 *177* $\boxed{177}$ *178 178* 179 179 180 *181* 182 183 *184* 186 188 189 189 193 195

On obtient : $Md = 177$

Tableau de contingence :

	Gr 1	Gr 2	Ensemble
$\leq Md$	8	8	16
$> Md$	4	11	15
Total	12	19	31

Effectifs théoriques et contributions au χ^2

	Gr 1	Gr 2
$\leq Md$	6.2	9.8
$> Md$	5.8	9.2

	Gr 1	Gr 2
$\leq Md$	0.52	0.33
$> Md$	0.56	0.35

Ici : $\chi^2_{obs} = 1.76$. Pour un seuil de 5%, $\chi^2_{crit} = 3.84$. On retient H_0.

Test de Wilcoxon-Mann-Whitney
Test U de Mann-Whitney

Deux groupes indépendants : deux échantillons tirés de deux populations distinctes.

Variable dépendante : ordinale ou numérique (par exemple, numérique comportant un très grand nombre de modalités).

Construction du protocole des rangs

On classe les n_1+n_2 sujets par valeurs croissantes (par exemple) de la variable. On attribue un rang à chaque sujet, avec la convention du rang moyen pour les ex æquos.

Exemple de construction du protocole des rangs

On reprend l'exemple "basket".

Deux groupes. Taille de chaque sujet.

G_1 : 152 163 164 173 174 176 177 177 178 178 181 184

G_2 : 167 171 172 174 175 176 176 177 179 179 180 182 183 186 188 189 189 193 195

Protocole des rangs :

Groupe	Taille	Rang
1	*152*	1
1	*163*	2
1	*164*	3
2	167	4
2	171	5
2	172	6
1	*173*	7
1	*174*	8.5
2	174	8.5
2	175	10
2	176	12
2	176	12
1	*176*	12
2	177	15
1	*177*	15
1	*177*	15

Groupe	Taille	Rang
1	*178*	17.5
1	*178*	17.5
2	179	19.5
2	179	19.5
2	180	21
1	*181*	22
2	182	23
2	183	24
1	*184*	25
2	186	26
2	188	27
2	189	28.5
2	189	28.5
2	193	30
2	195	31

Pour le groupe 1 :

$$W_1 = \sum R_i = 145.5$$

Pour le groupe 2 :

$$W_2 = \sum R_i = 350.5$$

Hypothèses

Soient θ_1 et θ_2 les médianes de la variable dépendante dans les populations parentes.

H_0 : $\theta_1 = \theta_2$
H_1 : Choix à faire entre :
$\theta_1 \neq \theta_2$ (hypothèse bilatérale),
$\theta_1 < \theta_2$ (hypothèse unilatérale à gauche)
$\theta_1 > \theta_2$ (hypothèse unilatérale à droite)

Construction de la statistique de test

• *n_1 et n_2 petits : utilisation de tables*

On calcule la somme des rangs du plus petit des deux échantillons : W

On compare W aux valeurs critiques W_s ou W'_s fournies par la table.

Sur l'exemple, test unilatéral à gauche au seuil de 5% :

Somme des rangs du groupe 1 : $W_1 = 145.5$

Valeur critique lue dans la table, pour un seuil de 5% : $W_s = 150$

$W_1 < W_s$: on conclut sur H_1 : les deux populations diffèrent du point de vue de la taille, les sujets de la première population ont une taille moins grande.

• *Lorsque $n_1 \geq 10$ et $n_2 \geq 10$: approximation par une loi normale*

$\overline{R_1}$: moyenne des rangs observés sur le premier échantillon

$\overline{R_2}$: moyenne des rangs observés sur le deuxième échantillon

$$Z = \frac{\overline{R_1} - \overline{R_2}}{E} \; avec \; E^2 = \frac{(n_1 + n_2 + 1)(n_1 + n_2)^2}{12 n_1 n_2}$$

Sous H_0, Z suit une loi normale centrée réduite.

Sur l'exemple :

$\overline{R_1} = 12.13$; $\overline{R_2} = 18.45$; $E^2 = 11.23$; $Z = -1.88$

Comme précédemment, on conclut sur H_1.

Remarque. Statistica calcule la statistique U de Mann-Whitney, liée aux sommes de rangs W_1 et W_2 par :

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - W_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - W_2$$

$$U = \min(U_1, U_2)$$

Tests non paramétriques sur deux groupes appariés

Test du signe

Un échantillon de sujets, placés dans deux conditions expérimentales différentes : groupes appariés.

Variable dépendante : ordinale ou numérique.

– protocole du signe des différences individuelles
– on élimine les différences nulles

D_+ : nombre de différences positives
D_- : nombre de différences négatives
$N = D_+ + D_-$: nombre total d'observations après élimination des différences nulles.

Hypothèses du test :

H_0 : les différences sont dues au hasard : dans la population parente, la fréquence des différences positives est 50%.

H_1 : Cette fréquence n'est pas 50% (test bilatéral)
ou (tests unilatéraux)
Cette fréquence est inférieure à 50%
Cette fréquence est supérieure à 50%

• *Cas des petits échantillons ($N \leq 30$)*

Sous H_0, la variable statistique "nombre de sujets présentant une différence positive sur un échantillon de taille N" suit une *loi binomiale de paramètres N et 0.5*.

On raisonne en termes de "niveau de significativité".

Par exemple, dans le cas d'un test unilatéral tel que
H_1 : fréquence inférieure à 50%
on calcule la fréquence cumulée $P(X \leq D_+)$ de D_+ pour la loi binomiale $B(N, 0.5)$.
Pour un seuil α donné :
Si $P(X \leq D_+) < \alpha$ on retient H_1
Si $P(X \leq D_+) \geq \alpha$ on retient H_0

Exemple.

14 sujets observés dans deux conditions. 2 différences positives, 10 différences négatives, 2 différences nulles.

La statistique de test D_+ suit une loi binomiale de paramètres $N = 12$ et $p = 0.5$.

Calcul du niveau de significativité de $D_{+,obs}$:

$P(D_+ = 0) = C_{12}^0\ 0.5^{12} = 0.0002441$

$P(D_+ = 1) = C_{12}^1\ 0.5^{12} = 0.0029297$

$P(D_+ = 2) = C_{12}^2\ 0.5^{12} = 0.0161133$

D'où : $P(D_+ \leq 2) = 0.019 = 1.9\%$

Au seuil de 5% unilatéral, on retient donc H_1.

• *Cas des grands échantillons : approximation par une loi normale ($N > 30$)*

$D = \max(D_+, D_-)$

$$Z = \frac{2D - 1 - N}{\sqrt{N}}$$

Z suit une loi normale centrée réduite.

Remarque. Dans le cas d'un test unilatéral, la zone de rejet est toujours située "à droite".

Exemple.

40 sujets observés dans deux conditions. 10 différences positives, 30 différences négatives, 0 différence nulle.

On a ici : $D = 30$ et $Z = \dfrac{60 - 1 - 40}{\sqrt{40}} = 3.00$

Au seuil de 1% unilatéral, on retient H_1 : les différences négatives sont significativement plus nombreuses que les différences positives.

Test de Wilcoxon sur des groupes appariés
Test T, ou test des rangs signés

Un échantillon de sujets, placés dans deux conditions expérimentales différentes : groupes appariés.

Variable dépendante : numérique.

Soit θ la médiane des différences individuelles dans la population parente.

On construit :

– le protocole des effets individuels d_i
– le protocole des valeurs absolues de ces effets $|d_i|$
– le protocole des rangs appliqués aux valeurs absolues, en éliminant les valeurs nulles.

T_+ : somme des rangs des observations tq $d_i > 0$
T_- : somme des rangs des observations tq $d_i < 0$
N = nombre de différences non nulles
$T_m = \min(T_+, T_-)$;
$T_M = \max(T_+, T_-)$

Hypothèses

H_0 : $\theta = 0$
H_1 : choix à faire entre :
$\theta \neq 0$ (hypothèse bilatérale),
$\theta < 0$ (hypothèse unilatérale à gauche)
$\theta > 0$ (hypothèse unilatérale à droite)

Statistique de test

• *Cas des petits échantillons*
$N \leq 15$ *: utilisation de tables spécialisées*

On compare T_m aux valeurs critiques indiquées par la table.

Exemple

On a testé huit sujets dans deux conditions A_1 et A_2. On obtient le protocole suivant :

Suj.	A_1	A_2	d_i	$\lvert d_i \rvert$	r_{i+}	r_{i-}
s1	100	105	5	5	1	
s2	70	63	-7	7		2
s3	40	50	10	10	3	
s4	123	98	-25	25		4
s5	92	60	-32	32		5
s6	120	78	-42	42		6
s7	172	119	-53	53		7
s8	173	101	-72	72		8
T					4	32

On trouve $T_+ = 4$, $T_- = 32$ et donc $T_m = 4$.

Au seuil de 5% unilatéral, on lit dans la table : $T_{crit} = 5$.

Comme $T_m < T_{crit}$, on conclut à une différence significative entre les conditions A_1 et A_2 au seuil de 5% unilatéral.

• *Cas des grands échantillons*
$N > 15$ *: approximation par une loi normale*

$$Z = \frac{T_M - 0.5 - \frac{N(N+1)}{4}}{E}$$

avec

$$E^2 = \frac{N(N+1)(2N+1)}{24}$$

Sous H_0, Z suit une loi normale centrée réduite.

Analyse de Variance à un facteur

Exemple introductif : Test commun à trois groupes d'élèves. Moyennes observées dans les trois groupes : $\overline{x}_1 = 8$, $\overline{x}_2 = 10$, $\overline{x}_3 = 12$.

Question : s'agit-il d'élèves "tirés au hasard" ou de groupes de niveau ?

Première situation :

	Gr1	Gr2	Gr3
	6	8	10.5
	6.5	8.5	10.5
	6.5	8.5	11
	7	9	11
	7.5	9.5	11
	8	10	12
	8	11	13
	10	11	13
	10	12	14
	10.5	12.5	14
$\overline{x}_i$	8	10	12

Deuxième situation :

	Gr1	Gr2	Gr3
	5	4	6
	5.5	5.5	7
	6	7.5	9
	6	9	10
	6.5	9.5	11
	7	10	12
	7.5	11	13
	10	13	14
	12	15	18
	14.5	15.5	20
$\overline{x}_i$	8	10	12

Démarche utilisée : nous comparons la dispersion des moyennes (8, 10, 12) à la dispersion à l'intérieur de chaque groupe.

Comparer a moyennes sur des groupes indépendants

Plan d'expérience : $\mathcal{S} < \mathcal{A}_a >$

Une variable $\mathcal{A}$, de modalités $A_1, A_2, \ldots, A_a$ définit a groupes indépendants.

Variable dépendante X mesurée sur chaque sujet.
x_{ij} : valeur observée sur le i-ème sujet du groupe j.

Problème : La variable X a-t-elle la même moyenne dans chacune des sous-populations dont les groupes sont issus ?

Conditions d'application :
– distribution normale de X dans chacun des groupes
– Egalité des variances dans les populations.

Hypothèses du test :

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_a$
H_1 : Les moyennes ne sont pas toutes égales.

Exemple :

15 sujets évaluent 3 couvertures de magazine. Sont-elles équivalentes ?

	C1	C2	C3	
	13	17	14	
	5	15	16	
	11	9	14	
	9	9	14	
	7	15	12	
$\overline{x}_i$	9	13	14	12

Variation (ou somme des carrés) totale :

$SC_T = (13-12)^2 + (5-12)^2 + \ldots + (12-12)^2 = 174$

Décomposition de la variation totale :

Score d'un sujet = Moyenne de son groupe + Ecart

C1	C2	C3
9	13	14
9	13	14
9	13	14
9	13	14
9	13	14

C1	C2	C3
4	4	0
-4	2	2
2	-4	0
0	-4	0
-2	2	-2

Variation (ou somme des carrés) inter-groupes :

$SC_{inter} = (9-12)^2 + (9-12)^2 + \ldots + (14-12)^2 = 70$

Variation (ou somme des carrés) intra-groupes :

$$SC_{intra} = 4^2 + (-4)^2 + \ldots + (-2)^2 = 104$$

Calcul des carrés moyens :

$$CM_{inter} = \frac{SC_{inter}}{a-1} = 35 \ ; \ CM_{intra} = \frac{SC_{intra}}{N-a} = 8.67$$

Statistique de test :

$$F_{obs} = \frac{CM_{inter}}{CM_{intra}} = 4.04$$

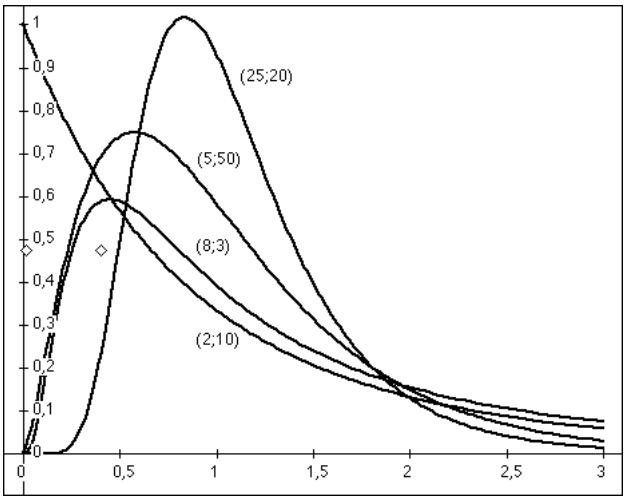
F suit une loi de Fisher avec $ddl_1 = a - 1 = 2$ et $ddl_2 = N - a = 12$.

Résultats

Source	Somme carrés	ddl	Carré Moyen	F
$\mathcal{C}$	70	2	35	4.04
Résid.	104	12	8.67	
Total	174	14		

Pour α=5%, $F_{crit} = 3.88$: H_1 est acceptée

Distributions du F de Fisher



Remarque

Si 2 groupes, équivaut à un T de Student. $F = T^2$

Pour les deux situations proposées en introduction :

Situation 1

Analysis of Variance Table

Response : x1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2	80.000	40.000	17.008	1.659e-05 ***
Residuals	27	63.500	2.352		

Situation 2

Analysis of Variance Table

Response : x2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2	80.00	40.00	2.7136	0.08436 .
Residuals	27	398.00	14.74		