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 ?

Etude descriptive

Fréquences lignes : en l'absence de lien, les fréquences sur chaque ligne devraient être proches des fréquences de la ligne "Synthèse" :

	MP	MCP	MA	Total
avant bac.	32%	50%	18%	100%
bac.	52%	26%	22%	100%
post bac.	34%	26%	40%	100%
Synthèse	38%	32%	31%	100%

Fréquences colonnes : de même, en l'absence de lien, les fréquences dans chaque colonne devraient être proches des fréquences de la colonne "Synthèse" :

	MP	MCP	MA	Synthèse
avant bac.	19%	37%	14%	23%
bac.	33%	20%	17%	24%
post bac.	47%	43%	69%	53%
Total	100%	100%	100%	100%

Mais nos observations portent sur un échantillon. Les différences constatées peuvent-elles être expliquées par les *fluctuations d'échantillonnage* ?

Test proprement dit

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 :

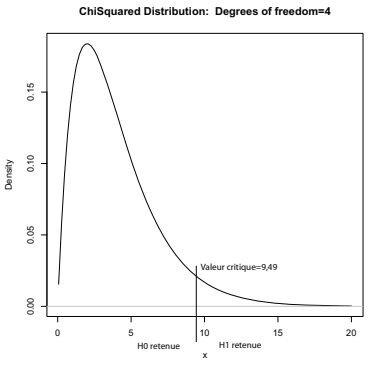
$$\text{Ctr}_i = \frac{(\text{Eff. Observé} - \text{Eff. Théorique})^2}{\text{Eff. Théorique}}$$

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 : $\chi^2_{crit} = 9.49$

Règle de décision :



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 : $\chi^2_{obs} = 7.85$, donc $\chi^2_{obs} \leq \chi^2_{crit}$.

Variante : la p-value correspondant à $\chi^2_{obs} = 7.85$ est égale à 9.7%. Elle est supérieure à 5%.

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

Remarques

1. Condition sur les effectifs théoriques minimaux :

Le test du χ^2 ne peut pas être appliqué si les effectifs sont trop faibles. On exige en général :

– que, dans le tableau des effectifs théoriques, les effectifs strictement inférieurs à 5 représentent moins de 20% des cases ;
– et que, dans le tableau des effectifs théoriques, ne figure aucun effectif inférieur à 1.

2. Dans le cas d'un tableau à deux lignes et deux colonnes, on opère souvent une correction au calcul du χ^2 : la correction de Yates. Cette correction est d'autant plus importante que les effectifs sont faibles.

3. Dans le cas d'un tableau à deux lignes et deux colonnes, si les effectifs sont trop faibles pour utiliser le test du χ^2, on peut utiliser le *test exact de Fisher*.

4. Le test du χ^2 peut notamment servir à comparer deux proportions (cf. TD avec R).

test de significativité d'un coefficient de corrélation

Deux variables numériques X et Y observées sur un échantillon de n sujets. Soit r leur coefficient de corrélation.

- Les données (x_i, y_i) constituent un échantillon
- r est une statistique
- ρ : coefficient de corrélation *inconnu* sur la population

H_0 : Indépendance de X et Y sur la population, c'est-à-dire : $\rho = 0$
H_1 : $\rho \neq 0$ (test bilatéral)

Statistique de test

– Petits échantillons : tables spécifiques. Le nombre de degrés de liberté est $ddl = n - 2$
– Grands échantillons. On calcule :

$$T = \sqrt{n-2}\ \frac{r}{\sqrt{1-r^2}}$$

T suit une loi de Student à $n-2$ degrés de liberté.

Conditions d'application

Dans la population parente, le couple (X, Y) suit une *loi normale bivariée*, ce qui implique notamment :

– la normalité des distributions marginales de X et Y ;

– la normalité de la distribution de l'une des variables lorsque l'autre variable est fixée ;

– l'égalité des variances des distributions de l'une des variables pour deux valeurs distinctes de l'autre variable.

Exemple : (Source : Article disponible en ligne à l'adresse : http ://www.cairn.info/revue-deviance-et-societe-2012-1-page-3.htm) On a mené une étude auprès de 26 collèges publics du département du Nord. 9 d'entre eux étaient classés en ZEP tandis que 17 autres ne possédaient aucune qualification particulière. On dispose, notamment, pour chaque établissement d'une mesure du niveau de violence ressenti par les élèves et par les professionnels et d'une mesure du taux d'encadrement.

Pour les 9 établissements classés en ZEP, le coefficient de corrélation entre la violence ressentie par les élèves et le taux d'encadrement est $r_1 = 0.62$, tandis que le coefficient de corrélation entre la violence ressentie par les professionnels et le taux d'encadrement est $r_2 = 0.72$. Ces coefficients sont-ils significatifs d'un lien entre les deux variables invoquées, au seuil de 5% ?

Réponse : pour $ddl = 7$ et un seuil de 5%, on lit dans la table $r_{crit} = 0.6664$. On n'a pas mis en évidence de lien entre les deux variables dans le cas des élèves (mais la taille de l'échantillon est faible). En revanche, on conclut à l'existence d'un lien entre les deux variables dans le cas des professionnels.

Etudier de même la situation des 17 établissements "ordinaires", sachant que les deux coefficients de corrélation sont alors $r_1 = 0.35$ et $r_2 = 0.68$.

Réponse : pour $ddl = 15$ et un seuil de 5%, on lit dans la table $r_{crit} = 0.4821$. Comme précédemment, on n'a pas mis en évidence de lien entre les deux variables dans le cas des élèves ; en revanche, on conclut à l'existence d'un lien entre les deux variables dans le cas des professionnels.

Tests de comparaison de moyennes

Comparaison de deux moyennes. Groupes indépendants

Une variable X est observée sur deux échantillons tirés au hasard dans deux populations différentes. Au vu de ce qui est observé sur les deux échantillons, les moyennes de X dans les populations parentes sont-elles égales ou différentes ? Ou encore, les moyennes observées sur ces deux échantillons sont-elles significativement différentes ?

Notations
μ_1, μ_2 : moyennes *inconnues* sur les populations parentes respectives (distributions normales de même variance)

$\overline{x_1}, \overline{x_2}$: moyennes respectives sur des échantillons de tailles n_1 et n_2

Hypothèses du test

$H_0 : \mu_1 = \mu_2$

H_1 : (test bilatéral) $\mu_1 \neq \mu_2$

Statistique de test.

$$T = \frac{\text{Différence des moyennes observées des 2 groupes}}{\text{Erreur type}}$$

L'erreur type est calculée à partir des écarts types observés dans les deux groupes et des effectifs des échantillons.

Sous H_0, T suit la loi de Student à $n_1 + n_2 - 2$ ddl. La loi de Student peut être assimilée à une loi normale centrée réduite si $n_1 > 30$ et $n_2 > 30$.

Exemple :

Un groupe de 30 adultes jeunes, et un groupe de 30 adultes âgés. On soumet les sujets des deux groupes à une épreuve de fluence orthographique. Les paramètres calculés à partir des résultats observés sont les suivants :

Fluence orthographique		
	Jeunes	Agés
n	30	30
$\overline{x}$	11.4	11.0
s_c	3.1	3.2

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$ (on fait ici un test bilatéral).

La statistique de test suit une loi de Student à $30 + 30 - 2 = 58$ ddl.

Pour un seuil de 5%, la valeur critique déduite de la table est $t_c = 2.0017$. La règle de décision est donc :

– si $-2.0017 \leq t_{obs} \leq 2.0017$, on retient H_0.
– si $t_{obs} < -2.0017$ ou $t_{obs} > 2.0017$, on rejette H_0 et on retient H_1.

Or (calcul réalisé à l'aide d'un logiciel) :

$$t_{obs} = \frac{11.4 - 11.0}{0.81} = 0.4917$$

On retient donc l'hypothèse H_0 : on n'a pas mis en évidence de différence significative de la fluence verbale.

Comparaison de deux moyennes. Groupes appariés

Une variable X est observée, dans deux conditions différentes, sur un échantillon tiré au hasard dans une population. Au vu de ce qui est observé, les moyennes de X sur la population parente dans les deux conditions sont-elles égales ou différentes ? Ou encore, les moyennes observées dans ces deux conditions sont-elles significativement différentes ?

On introduit le protocole dérivé des différences individuelles ($d_i = x_{i1} - x_{i2}$)

Notations
μ_1, μ_2 : moyennes (inconnues) de la variable X sur la population parente, dans les deux conditions.

δ : moyenne des différences individuelles sur la population ($\delta = \mu_1 - \mu_2$) (distribution normale)

n : taille de l'échantillon

$\overline{x_1}, \overline{x_2}$: moyennes respectives de la variable dans les deux conditions sur un échantillon de taille n

$\overline{d}$: moyenne des différences individuelles sur un échantillon de taille n ($\overline{d} = \overline{x_1} - \overline{x_2}$)

s_c : écart type corrigé estimant l'écart type des différences individuelles sur la population parente

Hypothèses du test

$H_0 : \mu_1 = \mu_2$, c'est-à-dire $\delta = 0$

H_1 : (test bilatéral) $\mu_1 \neq \mu_2$

Statistique de test. Cas où $n > 30$

$$t = \frac{\overline{d}}{E} \; avec \; E^2 = \frac{s_c^2}{n}$$

Sous H_0, T suit la loi de Student à $n - 1$ ddl.

Pour $n > 30$, la loi de Student peut être assimilée à une loi normale centrée réduite.

Exemple :

Temps de réaction de 10 sujets mesuré à jeun ($\overline{x_1} = 22.3$ms) et sous l'influence d'un tranquilisant ($\overline{x_2} = 31.7$ms). Ecart type corrigé de la série des différences : $s_c = 11.54$.

$H_0 : \delta = 0$
$H_1 : \delta \neq 0$ (test bilatéral, par exemple).

La statistique de test T suit une loi de Student, et, pour un seuil de 5%, la valeur critique est : $t_{crit} = 2.26$.

Or, $E^2 = \frac{11.54^2}{10}$, $E = 3.65$,
$t_{obs} = \frac{22.3 - 31.7}{3.65} = -2.58$

On conclut donc sur H_1.

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 :

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

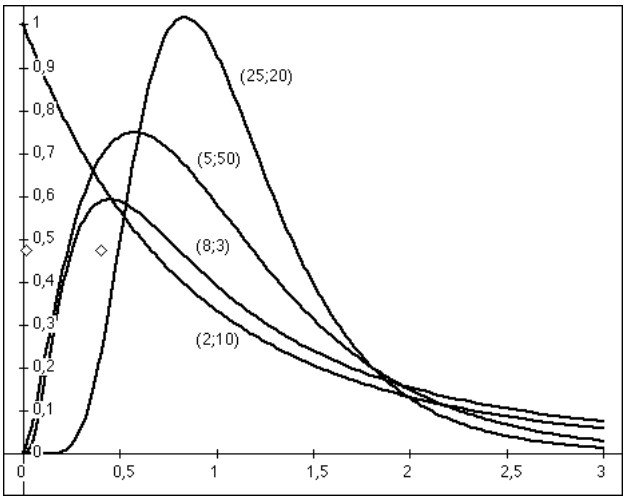
F suit une loi de Fisher avec $ddl_1 = a - 1 = 2$ et $ddl_2 = N - a = 12$.

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		