Les tests de χ^2

adapté du cours de S. Ducay

F. Wlazinski

Licence d'économie

1 Indépendance de 2 variables

Principe

Dans une population donnée, on étudie deux variables X et Y pouvant prendre respectivement r et s modalités. On cherche à savoir si les variables X et Y sont indépendantes.
Effectuant plusieurs échantillonnages de même taille n, on désigne par $N_{i,j}$ la variable aléatoire réelle (v.a.r.) égale à l'effectif observé du couple formé de la $i^{ème}$ modalité de la variable X et de la $j^{ème}$ modalité de la variable Y.

Remarque 1.1

Méthode

On teste H_0 contre H_1 où l'hypothèse nulle H_0 est "X et Y sont indépendantes" et l'hypothèse alternative $H_1 = \overline{H_0}$ est "X et Y ne sont pas indépendantes".
Sous l'hypothèse d'indépendance de X et Y, l'effectif théorique $\widetilde{\mathbf{n}}_{i,j}$ est égal à $\dfrac{n_{i,\star} \times n_{\star,j}}{n}$, avec $n_{i,\star} = \sum_{j=1}^{s} n_{i,j}$ et $n_{\star,j} = \sum_{i=1}^{r} n_{i,j}$

Ce test s'appuie sur la distance $D = \sum_{i=1}^{r}\sum_{j=1}^{s} \dfrac{\left(N_{i,j} - \widetilde{\mathbf{n}}_{i,j}\right)^2}{\widetilde{\mathbf{n}}_{i,j}} = \sum_{i=1}^{r}\sum_{j=1}^{s} \dfrac{N_{i,j}^2}{\widetilde{\mathbf{n}}_{i,j}} - n$ entre les effectifs observés et théoriques. Sous l'hypothèse H_0, D suit approximativement la loi de χ^2 (lire "khi deux") à $(r-1)\times(s-1)$ degrés de liberté.

En pratique, on calcule $d = \sum_{i=1}^{r}\sum_{j=1}^{s} \dfrac{\left(n_{i,j} - \widetilde{\mathbf{n}}_{i,j}\right)^2}{\widetilde{\mathbf{n}}_{i,j}} = \sum_{i=1}^{r}\sum_{j=1}^{s} \dfrac{n_{i,j}^2}{\widetilde{\mathbf{n}}_{i,j}} - n.$

Grâce à la table du χ^2 avec α et $d.d.l = (r-1)\times(s-1)$, on détermine alors la valeur plafond $b_{\max}$.
On décide que :

- si $d < b_{\max}$, alors on ne peut pas rejeter H_0. Mais on ne connait pas la probabilité que l'on se trompe en ne faisant pas de rejet.
- si $d \geq b_{\max}$, alors on rejette H_0 avec une probabilité α de se tromper.

Remarque 1.2

La qualité de l'approximation de la loi de D est satisfaisante lorsque les effectifs théoriques vérifient tous la condition $\widetilde{\mathbf{n}}_{i,j} \geq 5$.
Si ce n'est pas le cas, on peut effectuer des regroupements de lignes ou de colonnes : r et s désignent alors le nombre de modalités après le(s) regroupement(s).
Cependant, on peut ne pas faire de regroupement si les effectifs théoriques vérifient tous la condition $\widetilde{\mathbf{n}}_{i,j} \geq \dfrac{5t}{rs}$, où t est égal au nombre de couples de modalités ayant un effectif théorique $\widetilde{\mathbf{n}}_{i,j} < 5$.

Exemple 1.3

Une statistique effectuée sur 800 personnes donne la répartition suivante :

$n_{i,j}$	gros fumeurs	moyens fumeurs	petits fumeurs	non fumeurs
Avec hypertension	74	116	68	82
Sans hypertension	126	174	82	78

On veut tester, au risque 10%, l'indépendance entre l'hypertension et la consommation de tabac.

Remarque 1.4

Exemple 1.5

Dans une même catégorie sociale, un échantillon de 40 hommes a fourni 8 fumeurs et un échantillon de 60 femmes a fourni 18 fumeuses.
On se demande si la proportion de fumeurs est la même pour les deux sexes.

Remarque 1.6

Dans le cas particulier où $r = s = 2$, le test d'indépendance se confond strictement avec le test (bilatéral) d'égalité de deux proportions présenté dans un chapitre précédent. En effet, d est alors le carré de u et $b_{\max}$ le carré de z_α.

2 Homogénéité : comparaison de plusieurs échantillons

Principe

Dans une population donnée, on étudie une variable X pouvant prendre s modalités.
On dispose de r échantillons et on cherche à savoir s'ils proviennent de cette population.
Effectuant plusieurs échantillonnages, on désigne par $N_{i,j}$ la variable aléatoire réelle égale à l'effectif observé de la $j^{ème}$ modalité de la variable X dans le $i^{ème}$ échantillon.

Remarque 2.1

En pratique, pour un échantillonnage, on observe des effectifs $n_{i,j}$.

Méthode

Sous l'hypothèse d'homogénéité des échantillons, l'effectif théorique est égal à $\widetilde{\mathbf{n}}_{i,j} = \dfrac{n_{i,\star} \times n_{\star,j}}{n}$, avec $n_{i,\star} = \sum_{j=1}^{s} n_{i,j}$ et $n_{\star,j} = \sum_{i=1}^{r} n_{i,j}$.
On teste H_0 contre H_1 où H_0 est "les échantilllons sont issus de la même population" et $H_1 = \overline{H_0}$.
Ce test se déroule comme le test d'indépendance décrit dans la partie 1, même si le problème posé est de nature différente.

Exemple 2.2

On a selectionné des étudiants en économie de trois universités différentes.
Les résultats concernant les différentes filières du secondaire dont ils sont issus sont les suivants :

$n_{i,j}$	STMG	S	ES	Autres
Amiens	10	30	50	10
Lille	50	70	110	20
Reims	10	50	70	20

Les trois populations présentent-elles les mêmes proportions de filières?

3 Test de conformité de proportions et test de χ^2

Propriété 3.1

On considère une population et une variable X à deux modalités A et B, de proportions respectives p et $1-p$.
Il y a equivalence entre le test de conformité du chapitre précédent et un test de χ^2 avec $d.d.l. = 1$.

Démonstration

On a $\widetilde{\mathbf{n}}_0 = n \times p_0$ et $d = \dfrac{(nf - n \times p_0)^2}{n \times p_0} + \dfrac{(n(1-f) - n(1-p_0))^2}{n(1-p_0)} = \dfrac{(nf - n \times p_0)^2}{n \times p_0} + \dfrac{(n \times p_0 - nf)^2}{n(1-p_0)}$.

Donc $d = (nf - n \times p_0)^2 \left(\dfrac{1}{n \times p_0} + \dfrac{1}{n(1-p_0)} \right) = n^2 (f-p_0)^2 \dfrac{1}{n \times p_0 (1-p_0)} = \dfrac{(f-p_0)^2}{\dfrac{p_0(1-p_0)}{n}} = u^2$.

Par ailleurs, si U suit la loi normale $\mathcal{N}(0;1)$, alors $D = U^2$ suit la loi de χ^2 à 1 degré de liberté.
On en déduit que $z_\alpha^2 = b_{\max}$. □

Exemple 3.2

Sur un échantillon de taille $n = 100$, on a observé 90 individus de modalité A, soit une fréquence $f = \dfrac{90}{100} = 0,9$ d'individus de modalité A.
Peut-on considérer que, dans la population, la proportion d'individus de modalité A est égale à $0,8$?

Remarque 3.3

4 Conformité à un modèle théorique

Principe

Dans une population donnée, on étudie une variable X pouvant prendre r modalités et on cherche à savoir si on peut considérer que cette variable est d'un type donné. Plus précisément, désignant par p_i la probabilité d'apparition dans la population de la $i^{ème}$ modalité de la variable, on se demande si les p_i correspondent à une certaine loi de probabilité.

On choisit alors une loi théorique, par exemple, une distribution particulière (valeurs de p_i choisies arbitrairement avec $\sum_i p_i = 1$) ou une loi usuelle (loi de Poisson, loi normale, ...).

Dans ce second cas, on choisit le(s) paramètre(s) de la loi. On procède alors par estimation ponctuelle (moyenne ou variance pour le paramètre de la loi de Poisson, moyenne et écart-type pour les paramètres de la loi normale, ...).

Effectuant plusieurs échantillonnages de même taille n, on désigne par N_i la variable aléatoire réelle égale à l'effectif observé de la $i^{ème}$ modalité de la variable ; l'effectif théorique étant égal à $n \times p_i$.

Méthode

On teste H_0 contre H_1 où H_0 est "X suit la loi théorique" et $H_1 = \overline{H_0}$ est "X ne suit pas la loi théorique".
Ce test s'appuie sur la distance D entre les effectifs observés et théoriques :

$$D = \sum_{i=1}^{r} \frac{(N_i - n \times p_i)^2}{n \times p_i} = \sum_{i=1}^{r} \frac{N_i^2}{n \times p_i} - n \ ,$$

On sait que sous l'hypothèse H_0, D suit approximativement la loi de χ^2 à $r-k-1$ degrés de liberté, où k est le nombre de paramètres à estimer de la loi théorique choisie.

En pratique, pour un échantillon, on observe un effectif n_i pour la $i^{ème}$ modalité de la variable et on calcule $d = \displaystyle\sum_{i=1}^{r} \frac{(n_i - n \times p_i)^2}{n \times p_i} = \sum_{i=1}^{r} \frac{n_i^2}{n \times p_i} - n$.

Grâce à la table du χ^2 avec α et $d.d.l = r - k - 1$, on détermine la valeur plafond $b_{\max}$.

On décide que :

- si $d \leq b_{\max}$, alors on ne peut pas rejeter H_0.
- si $d > b_{\max}$, alors on rejette H_0 avec une probabilité α de se tromper.

Remarque 4.1

La qualité de l'approximation de la loi de D est satisfaisante lorsque les effectifs théoriques vérifient tous la condition $n \times p_i \geq 5$. Si ce n'est pas le cas, on peut regrouper certains effectifs de modalités voisines, r désignant alors le nombre de modalités après le(s) regroupement(s). Cependant, on peut ne pas faire de regroupement si les effectifs théoriques vérifient tous la condition $n \times p_i \geq \dfrac{5s}{r}$, où s est égal au nombre de modalités ayant un effectif théorique $n \times p_i < 5$.

Exemple 4.2

Dans une population vivante, on enregistre la présence de 5 génotypes, notés A_1 à A_5, et auxquels une théorie attribue les probabilités $p_1 = 0,4$, $p_2 = 0,2$, $p_3 = 0,2$, $p_4 = 0,1$ et $p_5 = 0,1$.

Sur un échantillon de $n = 400$ individus choisis au hasard dans la population, on désigne par n_i le nombre d'individus de génotype A_i. Les n_i sont donnés dans le tableau ci-dessous.

x_i	n_i
A_1	200
A_2	40
A_3	96
A_4	36
A_5	28

Peut-on dire, au risque $\alpha = 0,05$, que la répartition des génotypes dans l'échantillon est conforme à celle de la population?

Exemple 4.3

Une enquête effectuée auprès du comptoir de 150 coopératives agricoles a permis d'étudier l'arrivée dans le temps des usagers de ces coopératives. Pendant l'unité de temps, soit une heure, on a obtenu les résultats suivants :

Nombre d'usagers arrivés	0	1	2	3	4	5	6
Nombre de coopératives	37	46	39	19	5	3	1

Peut-on admettre que le nombre d'usagers arrivés dans cette population suit une loi de Poisson?

Exemple 4.4

Lors d'une étude sur un caractère X d'une population P, on obtient les résultats suivants :

Valeur de X	$[0;4[$	$[4;6[$	$[6;8[$	$[8;14]$
Effectif	6	10	16	8

Peut-on admettre que le caractère X dans cette population suit une loi normale?

Exemple 4.5

Lors d'une étude biologique portant sur une certaine espèce de mollusques, on a mesuré le taux de protéines de 36 individus appartenant à cette espèce. On a obtenu les résultats suivants :

taux de protéine (en mg)	$]0;1,5]$	$]1,5;3]$	$]3;4,5]$	$]4,5;6]$	$]6;7,5]$	$]7,5;9]$	$]9;10,5]$
nombre d'individus	8	7	4	9	2	3	3

Peut-on admettre que le taux de protéines dans cette population suit une loi normale?