Chapitre 8

Tests du χ^2

Sommaire

1 Introduction

Karl Pearson est un mathématicien britannique qui a établit la théorie générale de la **corrélation** et inventa la statistique du Khi-deux.
Les différents tests qui relèvent de la statistique du **Khi-deux ou Chi-deux** χ^2 ont pour objectif de déterminer dans quelle mesure les **effectifs** relatifs à un ou plusieurs **caractères qualitatifs** (ou **caractères quantitatifs** regroupés en classe) **observés** sur un ou plusieurs échantillons **sont conformes** aux **effectifs attendus** sous l'hypothèse nulle,

- soit d'égalité des distributions observées (**test d'homogénéité**),

Est-ce que la distribution des groupes sanguins présente une répartition géographique en comparant plusieurs populations géographiquement distinctes ?

- soit d'indépendance entre deux caractères qualitatifs (**test d'indépendance**)

Est-ce qu'il y a indépendance entre la couleur des yeux et la couleur des cheveux ?

- soit de conformité à une loi de probabilité connue (**test d'ajustement**) .

Est-ce que la distribution des génotypes observés pour un locus donné dans un échantillon est conforme à la distribution attendue sous le modèle de Hardy-Weinberg ?

Quelques soit le test du χ^2 réalisé, l'objectif est de déterminer si les écarts entre la distribution des **effectifs observés** et la distribution des **effectifs théoriques** est significative ou imputable uniquement aux fluctuations d'échantillonnage.

2 Principe des tests du χ^2

2.1 La statistique du χ^2

La statistique du Khi-deux χ^2 consiste à mesurer l'écart qui existe entre la distribution des effectifs théoriques t_i et la distribution des effectifs observés n_i et à tester si cet écart est suffisamment faible pour être imputable aux fluctuations d'échantillonnage.
Par exemple dans le cas d'un **test de** χ^2 **d'ajustement**, où l'on veut comparer pour un **caractère qualitatif à** k **modalités** i ou un **caractère quantitatif groupé en** k **classes** i, une distribution observée et une distribution théorique, la statistique du χ^2 est la suivante :

$$\chi^2_{obs.} = \sum_{i=1}^{k} \frac{(n_i - t_i)^2}{t_i}$$ suit une **loi de Pearson ou** χ^2

L'établissement des distributions des probabilités p_i va dépendre de la nature du test du χ^2 (hypothèse H_0) mais l'estimation des effectifs théoriques t_i sera identique à tous les tests.

si n est l'effectif total étudié, l'effectif théorique attendu, t_i pour la modalité i de la variable aléatoire X est :

$$t_i = n * p_i$$ (**loi des grands nombres en probabilité**)

Quelque soit l'hypothèse nulle testée, la stratégie est la même pour tous les tests du χ^2.

La statistique du χ^2 calculée ($\chi^2_{obs.}$) est comparée avec la valeur seuil, χ^2_{seuil} lue sur **la table du χ^2** pour *k-c ddl* (**degrés de liberté**) et pour un **risque d'erreur** α fixé.

- si $\chi^2_{obs.} \leq \chi^2_{seuil}$, l'hypothèse H_0 ne peut être rejetée : distributions des effectifs théoriques et observés ne sont pas significativement différentes
- si $\chi^2_{obs.} > \chi^2_{seuil}$, l'hypothèse H_0 est rejetée au seuil de signification α et l'hypothèse H_1 est acceptée.

2.2 Les conditions d'application

• Quelque soit le test du χ^2, la taille de la distribution des effectifs théoriques est **strictement identique** à celle des effectifs observés c'est à dire n effectif total.

• L'échantillon étudié doit être **de grande taille** $n \geq 50$

• Le test χ^2 est fondé sur l'approximation, à des lois normales, d'une **loi multinomiale**. Pour que cette approximation soit très bonne et bien que le test du χ^2 s'avère robuste, il est conseillé que les produits $t_i = n*p_i$, c'est à dire les **effectifs théoriques t_i, soient égaux ou supérieurs à 5** et de **regrouper les classes adjacentes** lorsque ce minimum est rencontré.

2.3 Les degrés de liberté

Le nombre de **degrés de liberté** (*ddl*) est égal au **nombre de composantes indépendantes** de la statistique du χ^2.

Le nombre de composantes indépendantes d'une distribution théorique ayant k modalités (effectifs théoriques supérieurs ou égaux à 5) correspond au nombre de termes de la **statistique du χ^2.** Mais comme on impose que la taille de la distribution des effectifs théoriques soit identique à la taille de la distribution des effectifs observés n, le $k^{ème}$ effectif théorique est contraint d'où

Le **nombre de degrés de liberté maximum** est donc ***k*-1.**
avec k le nombre **de termes** du χ^2 (effectifs théoriques ≥ 5)

Toutes les relations supplémentaires imposées pour le calcul des effectifs théoriques conduisent à réduire d'une unité le nombre de degrés de liberté. Le nombre de composantes non indépendantes ou **contraintes** dépendra de la **nature du test du χ^2** (*n* étant une de ces contraintes, commune à tous les tests du χ^2) .

> **Le nombre de degrés de liberté est donc *k-c* avec**
> *k* le nombre **de termes** du χ^2 (effectifs théoriques ≥ 5) et
> *c* le nombre **de contraintes** entre les distributions comparées.

3 Test du χ^2 d'ajustement

Le **test du χ^2 d'ajustement** correspond à la comparaison d'une distribution de fréquences observées et d'une distribution de fréquences théoriques. Ce test est fréquemment utilisé en **génétique**, où l'on confronte les résultats expérimentaux de croisements pour **un caractère donné** à ceux résultant d'une transmission mendélienne de ce caractère. Le champ d'application de ces méthodes ne se limite pas à la génétique.

En effet l'utilisation des **tests d'hypothèse** tels que nous les avons définis, implique la réalisation de certaines hypothèses comme par exemple **la normalité de la variable étudiée**. Il est donc nécessaire de comparer la distribution observée des valeurs à celle attendue dans le cas d'une distribution normale de celles-ci.

3.1 Principe du test

Le principe du test du χ^2 d'ajustement est d'estimer à partir d'une loi de probabilité connue ou **inférée**, les effectifs théoriques pour les différentes modalités du caractère étudié (caractère qualitatif ou quantitatif regroupé en classe) et les comparer aux effectifs observés dans un échantillon. Deux cas peuvent se présenter :

- soit **la loi de probabilité est spécifiée *a priori*** car elle résulte par exemple d'un modèle déterministe tel que la distribution mendélienne des caractères, l'évolution de la taille d'une population, etc.
- soit **la loi de probabilité théorique n'est pas connue *a priori*** et elle est déduite des caractéristiques statistiques mesurées sur l'échantillon (distribution des fréquences, moyenne et variance)(**statistiques descriptives**).

3.2 Application et décision

L'établissement des distributions théoriques de probabilité se réfèrent aux **lois de probabilité**. A chaque modalité ou valeur de la variable aléatoire X, les probabilités associées à la loi de probabilité sont calculées ainsi que les **effectifs théoriques attendues** sous cette loi :

	Modalité du caractère A				
	A_1	A_2	... A_i	A_k	
Effectif observé n_i	n_1	n_2	.....n_i	n_k	$n = \sum_{i=1}^{k} n_i$
p_i	p_1	p_2	.....p_i	p_k	$\sum_{i=1}^{k} p_i = 1$
Effectif théorique $t_i = n * p_i$	t_1	t_2	.....t_i	t_k	$n = \sum_{i=1}^{k} t_i$

Remarque : Si le caractère A ne présente que deux modalités A = succès et $\overline{A}$ = échec, le test du χ^2 d'ajustement revient à la comparaison d'une fréquence observée et d'une fréquence théorique (**test de conformité**).

La statistique du **Khi deux** χ^2 consiste à mesurer l'écart qui existe entre la distribution théorique et la distribution observée et à tester si cet écart est suffisamment faible pour être imputable aux fluctuations d'échantillonnage.

L'hypothèse testée est la suivante :

H_0 : la distribution observée est conforme à la distribution théorique.

H_1 : la distribution observée ne s'ajuste pas à la distribution théorique.

$$\chi^2_{obs.} = \sum_{i=1}^{k} \frac{(n_i - t_i)^2}{t_i}$$

k modalités du caractère étudié

avec n_i l'effectif observé et t_i l'effectif théorique attendu sous H_0

$\chi^2_{obs.}$ est comparée avec la valeur seuil, χ^2_{seuil} lue sur **la table du χ^2** pour *k-c ddl* (**degrés de liberté**) et pour un **risque d'erreur** α fixé.

Remarque : Il est impératif que les conditions d'application soient vérifiées : taille de l'échantillon $n \geq 50$ et les $np_i \geq 5$.

Exemple :

Soit le locus biallélique codant pour la glucose 6 phosphate déhydrogénase (G6PDH), enzyme participant au métabolisme énergétique (dégradation des sucres), l'analyse électrophorétique des génotypes chez l'anophèle, vecteur de la malaria, donne la répartition suivante

FF = 44, FS = 121, SS = 105.

La répartition des génotypes est-elle conforme au modèle de Hardy-Weinberg ?

3.3 Ajustements à différentes lois de probabilité connues

3.3.1 Ajustement à une loi binomiale

Application

Est-ce que la distribution du nombre de filles observées dans 320 fratries de 5 enfants suit une loi binomiale de paramètre $\mathcal{B}(5, 0{,}5)$?

X : Nbre de filles (i)	0	1	2	3	4	5
Nbre de fratries observées (n_i)	18	56	110	88	40	8

La distribution théorique suit **une loi binomiale** $\mathcal{B}(n, p)$ avec n :nbre d'épreuves

p : probabilité du succès

$$p_k = P(X = k) = C_n^k p^k q^{n-k}$$

k : nbre de valeurs prises X

Le nombre de degrés de liberté est :

nombre de termes du χ^2 ($\leq k$) **moins** le nombre de contraintes c

- $c = 1$ (n) si p est connue
- $c = 2$ (n et $\hat{p}$) si p est inconnue avec $\hat{p} = \dfrac{\sum_{i=1}^{k} n_i x_i}{n \sum_{i=1}^{k} n_i} = \dfrac{nombre\ de\ succès}{nombre\ d'observation}$

Exemple :

Refaire le test du χ^2 d'ajustement en utilisant pour p, la fréquence du nombre de filles dans les fratries de 5 enfants, son estimation faite à partir des données de l'échantillon.

3.3.2 Ajustement à une loi de poisson

Application

Est-ce que le nombre de cas graves traités chaque jour par un vétérinaire sur une période de 200 jours suit une loi de poisson ?

X : Nbre de cas graves (i)	0	1	2	3	4	5 et plus
Nbre de jours (n_i)	50	74	50	21	4	1

La distribution théorique suit une **loi de poisson** $\mathcal{P}(\lambda)$

$$p_k = P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$ k : nombre de valeurs prises X

Le nombre de degrés de liberté est :

nombre de terme du χ^2 ($\leq k$) **moins** le nombre de contraintes c

- $c = 1$ (n) si λ est connu
- $c = 2$ (n et $\hat{\lambda}$) si λ est inconnu avec $\hat{\lambda} = \overline{x}$

Remarque : La distribution de poisson n'étant pas bornée lorsque $X \to +\infty$, il est nécessaire de borner la distribution en estimant la probabilité de la dernière classe par différence avec la somme des probabilités qui est de 1.

Exemple : En reprenant les données relatives à **la cécidomyie du hêtre**, peut-on affirmer que la répartition du nombre de galles par feuille suit une loi de poisson ?

3.3.3 *Ajustement à une loi normale*

Application

Le caractère « taille » mesuré sur 1000 individus peut-il être considéré comme suivant une loi normale ?

X : taille en cm (x_i)	< 155	[155-165]	[165-175]	[175-185]	>185
Nbre d'individus (n_i)	1	70	500	379	50

La distribution théorique suit une **loi normale** $\mathcal{N}(\mu,\sigma)$

$P(a \le X \le b) = P(z_a \le Z \le z_b) = \pi(b) - \pi(a)$ (voir **probabilités**)

avec la **variable centrée réduite** $Z = \dfrac{X-\mu}{\sigma}$ et k : nombre de classes de la variable X

Le nombre de degrés de liberté est :

nombre de terme du χ^2 ($\le k$) **moins** le nombre de contraintes c

- $c = 1$ (n) si μ et σ connues
- $c = 2$ (n , $\hat{\mu}$) si μ inconnue avec $\hat{\mu} = \bar{x}$ (même chose si σ inconnue)
- $c = 3$ (n , $\hat{\mu}$, $\hat{\sigma}$) si μ et σ inconnues avec $\hat{\mu} = \bar{x}$ et $\hat{\sigma}^2 = \dfrac{n}{n-1} s^2$

Remarque : • La **loi normale n'étant pas bornée aux deux extrémités de la distribution**, lorsque $X \to \pm\infty$, il est nécessaire de borner la distribution en estimant la probabilité des deux classes extrèmes par différence avec 0 et 1.

• Si n < 50, le test **non paramétrique de Lilliefors** permet de tester la normalité d'une variable dans le cas de faibles effectifs.

Exemple : En reprenant les données relatives à **la longueur de la rectrice de la gélinotte huppée**, peut-on affirmer que cette mesure suit une loi normale ?

4 Test du χ^2 d'égalité de distributions

Comme pour le test du χ^2 d'ajustement, on considère **un caractère** (quantitatif groupé en classe ou qualitatif) présentant plusieurs modalités (*p* **modalités**) mais définis sur **plusieurs échantillons indépendants** (*q* **échantillons**) .L'hypothèse H_0 testée est « **l'égalité des *q* distributions observées** du caractère étudié «. Ce test s'apparente aux **tests d'homogénéité**.

4.1 Principe du test

La statistique du **Khi deux** χ^2 va permettre de mesurer l'écart qui existe entre les *q* **distributions des effectifs observés** pour la variable qualitative *X* **sous l'hypothèse d'égalité des distributions** dans les *q* populations comparées. On teste si cet écart est suffisamment faible pour être imputable aux fluctuations d'échantillonnage.

• Les données sont structurées sous forme d'un **tableau des effectifs observés** ou **table de contingence**.

Caractère A

	modalité 1		*modalité i*		*modalité p*	Total
Echantillon 1	n_{11}		n_{i1}		n_{p1}	$n_{.1}$
Echantillon j	n_{1j}		n_{ij}		n_{pj}	$n_{.j}$
Echantillon q	n_{1q}		n_{iq}		n_{pq}	$n_{.q}$
Total	$n_{1.}$		$n_{i.}$		$n_{p.}$	$n_{..}$= N

La nomenclature commune aux tables de contingence est basée sur deux indices i et j :

l'effectif n_{ij} est celui de la **colonne i** et de la **ligne j** avec $1 \leq \mathbf{i} \leq p$ et $1 \leq \mathbf{j} \leq q$

l'effectif $n_{i.}$ est la somme des effectifs de la **colonne i**

l'effectif $n_{.j}$ est la somme des effectifs de la **ligne j**

l'effectif $n_{..}$ est l'effectif total de la table de contingence

• Le **tableau des effectifs attendus** sous l'hypothèse H_0 : les q échantillons proviennent de q populations où la distribution en fréquence du caractère étudié est identique :

Caractère A

	modalité 1		*modalité i*		*modalité p*	Total
Echantillon 1	$\frac{n_{1.} \times n_{.1}}{N}$		$\frac{n_{i.} \times n_{.1}}{N}$		$\frac{n_{p.} \times n_{.1}}{N}$	$n_{.1}$
Echantillon j	$\frac{n_{1.} \times n_{.j}}{N}$		$\frac{n_{i.} \times n_{.j}}{N}$		$\frac{n_{p.} \times n_{.j}}{N}$	$n_{.j}$
Echantillon q	$\frac{n_{1.} \times n_{.q}}{N}$		$\frac{n_{i.} \times n_{.q}}{N}$		$\frac{n_{p.} \times n_{.q}}{N}$	$n_{.q}$
Total	$n_{1.}$		$n_{i.}$		$n_{p.}$	$n_{..}$= N

Sous H_0, l'effectif attendu t_{ij} correspondant à la **modalité i** du caractère A (A_i) pour **l'échantillon j** peut être obtenu de la façon suivante :

$P(A_i \cap$ échantillon **j**$) = P(A_i) \times P($échantillon **j**$)$ (**deux évènements indépendants**)

d'où $P_{ij} = \frac{n_{i.}}{N} \times \frac{n_{.j}}{N} = \frac{t_{ij}}{N}$ avec t_{ij} effectif attendu

d'où $t_{ij} = N * P_{ij}$ ainsi $t_{ij} = \frac{n_{i.} \times n_{.j}}{N}$

Tous les effectifs attendus sont obtenus par **le rapport du produit des distributions marginales sur l'effectif total de la table de contingence.**

$$t_{ij} = \frac{n_{i.} \times n_{.j}}{N}$$

Ainsi, le **nombre de degrés de liberté** correspondant au nombre d'effectifs estimés indépendants est $(p - 1)(q - 1)$. Les effectifs associés à la colonne p peuvent être obtenus par différence avec la distribution marginale des lignes (p-1) et inversement pour les effectifs associés à la ligne q (q-1) (cases indépendantes grisées dans la **table de contingence**).

4.2 Application et décision

L'hypothèse testée est la suivante :

H_0 : la distribution de fréquence du caractère étudié est identique pour les différentes populations comparées.

H_1 : la distribution de fréquence du caractère étudié diffère entre les différentes populations comparées.

$$\chi^2_{obs.} = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(n_{ij} - t_{ij}\right)^2}{t_{ij}}$$

p : nombre de colonnes, q : nombre de lignes

avec n_{ij} l'effectif observé et t_{ij} l'effectif théorique attendu sous H_0

$\chi^2_{obs.}$ est comparée avec la valeur seuil, χ^2_{seuil} lue sur **la table du χ^2** pour (p-1)(q-1)*ddl* (**degrés de liberté**) et pour un **risque d'erreur** α fixé.

- si $\chi^2_{obs.} > \chi^2_{seuil}$ l'hypothèse H_0 est rejetée au risque d'erreur α : les différents échantillons sont extraits de populations ayant des distributions différentes du caractère étudié.
- si $\chi^2_{obs.} \leq \chi^2_{seuil}$ l'hypothèse H_0 est acceptée: les différents échantillons sont extraits de populations ayant la même distribution du caractère étudié.

Remarque : La statistique du Khi-deux χ^2 ne peut être calculée que si les effectifs théoriques t_{ij} sont **supérieurs à 5**. Dans ce cas, il faut regrouper à la fois toute la ligne et toute la colonne correspond à la case possédant une valeur t_{ij} **inférieur à 5**.

Exemple :

Les groupes sanguins A,B,AB et O ont été déterminés dans trois échantillons (E_1 : France, E_2 : Roumanie, E_3 : Proche-Orient) d'hommes adultes. La répartition des groupes sanguins dépend-elle d'un facteur géographique ?

	A	B	AB	O
E1	54	14	6	51
E2	45	14	8	31
E3	33	34	12	33

4.3 Cas particulier de la comparaison de deux fréquences

Le test χ^2de **comparaison de deux fréquences** est un cas particulier du test de comparaison de plusieurs distributions. Dans ce cas le caractère étudié présente **deux modalités** (A = succès, $\bar{A}$ = échec) et est étudié sur **deux échantillons indépendants** extraits de deux populations. On fait l'hypothèse que les deux échantillons proviennent de 2 populations dont les **probabilités de succès sont identiques** : $H_0 : p_1 = p_2$.

• Table de contingence des **effectifs observés** (voir nomenclature **A** et **B**)

Table A

	Succès	Echecs	Effectifs
Echantillon 1	k_1	$n_1 - k_1$	n_1
Echantillon 2	k_2	$n_2 - k_2$	n_2
Total	k_1+k_2	$(n_1+n_2)-(k_1+k_2)$	n_1+n_2

ou

Table B

Succès	Echecs	Total
n_{11}	n_{21}	$n_{.1}$
n_{12}	n_{22}	$n_{.2}$
$n_{1.}$	$n_{2.}$	$n_{..}$= N

• Table de contingence des **effectifs attendus** sous $H_0 : p_1 = p_2$

	Succès	Echecs	Total
Echantillon 1	$\frac{n_{1.} \times n_{.1}}{N}$	$\frac{n_{2.} \times n_{.1}}{N}$	$n_{.1}$
Echantillon 2	$\frac{n_{1.} \times n_{.2}}{N}$	$\frac{n_{2.} \times n_{.2}}{N}$	$n_{.2}$
Total	$n_{1.}$	$n_{2.}$	$n_{..}$= N

L'hypothèse testée est la suivante :

$H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$

$$\chi^2_{obs.} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(n_{ij} - t_{ij}\right)^2}{t_{ij}}$$ suit une **loi du Khi-deux** χ^2

avec n_{ij} l'effectif observé et t_{ij} l'effectif théorique attendu sous H_0

$\chi^2_{obs.}$ est comparée avec la valeur seuil, χ^2_{seuil} lue sur **la table du χ^2** pour 1 *ddl* (**degrés de liberté**) et pour un **risque d'erreur** α fixé.

- si $\chi^2_{obs.} > \chi^2_{seuil}$ l'hypothèse H_0 est rejetée au risque d'erreur α : les deux échantillons sont extraits de deux populations ayant des probabilités de succès respectivement p_1 et p_2.
- si $\chi^2_{obs.} \leq \chi^2_{seuil}$ l'hypothèse H_0 est acceptée: les deux échantillons sont extraits de deux populations ayant même probabilité de succès p.

Remarque :

- La statistique du Khi-deux χ^2 ne peut être calculée que si les effectifs théoriques t_{ij} sont **supérieurs à 5**. Dans ce cas, il faut regrouper à la fois toute la ligne et toute la colonne correspond à la case possédant une valeur t_{ij} **inférieur à 5**.
- La statistique du Khi-deux χ^2 d'une **table de contingence** 2 x 2 avec **1 *ddl*** correspond au carré d'une **variable normale centrée réduite** ε^2 (démonstration).

Exemple :

Reprendre l'exemple de **l'impact des travaux dirigés dans la réussite** à l'examen de statistique avec le test du Khi-deuxχ^2.

5 Test du χ^2 d'indépendance

5.1 Principe du test

Le test du χ^2 d'indépendance constitue une autre formulation du **test de comparaison de plusieurs distributions**. Dans ce cas ce sont les distributions relatives à **deux caractères** (quantitatifs groupés en classe ou qualitatifs) présentant **plusieurs modalités** et définis sur une même population qui sont comparées. On fait l'hypothèse qu'il y a **indépendance entre les deux caractères** dans la population :

H_0 : les deux caractères sont indépendants.

H_1 : les deux caractères ne sont pas indépendants

• Les données sont structurées sous forme d'un **tableau des effectifs observés** pour les deux caractères comparés ou **table de contingence**.

Caractère A

Caractère B	*modalité 1*		*Modalité i*		*modalité p*	total
modalité 1	n_{11}		n_{i1}		n_{p1}	$n_{.1}$
modalité j	n_{1j}		n_{ij}		n_{pj}	$n_{.j}$
modalité q	n_{1q}		n_{iq}		n_{pq}	$n_{.q}$
Total	$n_{1.}$		$n_{i.}$		$n_{p.}$	$n_{..}$= N

avec l'effectif n_{ij} correspond au nombre d'individus ayant la **modalité i du caractère A** et la **modalité j du caractère B** avec $1 \leq i \leq p$ et $1 \leq j \leq q$

l'effectif $n_{i.}$ est la somme des effectifs de la **colonne i**

l'effectif $n_{.j}$ est la somme des effectifs de la **ligne j**

l'effectif $n_{..}$ est l'effectif total de la table de contingence

• Le **tableau des effectifs attendus** sous l'hypothèse H_0 : indépendance entre le caractère A et le caractère B.

Caractère A

Caractère B	*modalité 1*		*modalité i*		*modalité p*	Total
modalité 1	$\frac{n_{1.} \times n_{.1}}{N}$		$\frac{n_{i.} \times n_{.1}}{N}$		$\frac{n_{p.} \times n_{.1}}{N}$	$n_{.1}$
modalité j	$\frac{n_{1.} \times n_{.j}}{N}$		$\frac{n_{i.} \times n_{.j}}{N}$		$\frac{n_{p.} \times n_{.j}}{N}$	$n_{.j}$
modalité q	$\frac{n_{1.} \times n_{.q}}{N}$		$\frac{n_{i.} \times n_{.q}}{N}$		$\frac{n_{p.} \times n_{.q}}{N}$	$n_{.q}$
Total	$n_{1.}$		$n_{i.}$		$n_{p.}$	$n_{..}$= N

Sous H_0, l'effectif attendu t_{ij} correspondant à la **modalité i** du caractère A (A_i) et à la modalité **j** du caractère B (B_j) peut être obtenu de la façon suivante :

$P(A_i \cap B_j) = P_{ij} = P(A_i) \times P(B_j)$ sous H_0 : **indépendance** entre les deux caractères

d'où $P_{ij} = \frac{n_{i.}}{N} \times \frac{n_{.j}}{N} = \frac{t_{ij}}{N}$ avec t_{ij} effectif attendu

d'où $t_{ij} = N * P_{ij}$ ainsi $t_{ij} = \frac{n_{i.} \times n_{.j}}{N}$

5.2 Application et décision

L'hypothèse testée est la suivante :

H_0 : Indépendance entre le caractère A et le caractère B

H_1 : Non indépendance entre le caractère A et le caractère B

$$\chi^2_{obs.} = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n_{ij} - t_{ij})^2}{t_{ij}}$$

p : nombre de colonnes, q : nombre de lignes

avec n_{ij} l'effectif observé et t_{ij} l'effectif théorique attendu sous H_0

$\chi^2_{obs.}$ est comparée avec la valeur seuil, χ^2_{seuil} lue sur **la table du χ^2** pour $(p-1)(q-1)ddl$ (**degrés de liberté**) et pour un **risque d'erreur** α fixé.

- si $\chi^2_{obs.} > \chi^2_{seuil}$ l'hypothèse H_0 est rejetée au risque d'erreur α : il n'y a pas indépendance statistique entre les deux caractères étudiés dans la population.
- si $\chi^2_{obs.} \leq \chi^2_{seuil}$ l'hypothèse H_0 est acceptée: les deux caractères étudiés dans la population sont statistiquement indépendants.

Remarque : La statistique du Khi-deux χ^2 ne peut être calculée que si les effectifs théoriques t_{ij} sont **supérieurs ou égaux à 5**. Dans ce cas, il faut regrouper à la fois toute la ligne et toute la colonne correspond à la case possédant une valeur t_{ij} **inférieur à 5**.

Exemple :

Sur un échantillon de la population française, on a noté pour chaque personne, la couleur des yeux et celle des cheveux (naturelle). Peut-on conclure à l'indépendance de ces deux caractères qualitatifs ?

Cheveux / Yeux	Noirs	Bruns	Blonds	Roux
Marrons	152	247	83	11
Vert-gris	73	114	37	8
Bleus	36	102	127	10