Chapitre 4

Le test du χ^2

4.1 Les données du problème

Certains tests statistiques ont pour objet de tirer des conclusions relatives à la valeur des paramètres (moyenne, fréquence, variance) d'une ou plusieurs populations, sur la base d'informations partielles fournies par un ou plusieurs échantillons.
La même démarche peut être appliquée pour porter un "jugement" sur les caractéristiques encore plus générales de la population : la forme même de distribution du caractère étudié, la validité de sa représentation à l'aide de telle ou telle loi de probabilité particulière, les relations éventuelles entre plusieurs variables.

Concrètement, on dispose d'une distribution statistique empirique se présentant sous la forme d'une table d'effectifs ou de fréquences du caractère étudié. On désire savoir si ces effectifs ou ces fréquences sont compatibles avec une distribution théorique déterminée telle que la loi binomiale, la loi de Poisson, la loi normale ou toute autre loi de probabilité. Il s'agit en d'autres termes d'apprécier l'adéquation d'une distribution théorique particulière, en tant que représentation d'un phénomène concret observé (série empirique).

La démarche consiste donc à tester l'hypothèse selon laquelle notre échantillon serait tiré d'une population régie par une certaine loi de probabilité.
Il est évident que, même si le phénomène étudié suit effectivement la loi de probabilité dont on teste l'adéquation, les fréquences expérimentales (ou empiriques) observées sur un échantillon particulier différeront nécessairement peu ou prou des probabilités (fréquences que l'on devrait théoriquement observer selon la loi en question).

La problématique du test revient en définitive à savoir si les différences constatées entre la distribution expérimentale et la distribution théorique supposée sont explicables par l'aléa lié à la constitution de l'échantillon ou si elles sont trop importantes pour être imputables au seul hasard. En ce cas, c'est l'hypothèse de travail avancée sur la nature de la distribution qui devrait être mise en cause.

4.2 Ajustement d'une distribution observée à une distribution théorique

4.2.1 Construction du test

1. Les hypothèses du test sont les suivantes :
   - H_0 : X suit la loi théorique L,
   - H_1 : X ne suit pas L.
2. La variable observée est :
   — soit discrète et prend k valeurs x_1, x_2,. . . ,x_k

— soit continue et classée en k classes $[a_0; a_1[$, $[a_1; a_2[$,...,$[a_{k-1}; a_k[$ de centres respectifs x_1, x_2,...,x_{k-1}, x_k.

3. Les N observations de l'échantillon sont réparties sur les k valeurs de X (si X est discrète) ou sur les k classes de X (si X est continue). On a les tableaux suivants :

x_i	n_i
x_1	n_1
x_2	n_2
$\vdots$	$\vdots$
x_k	n_k

Classes	Centres x_i	Effectifs n_i
$[a_0; a_1[$	x_1	n_1
$[a_1; a_2[$	x_2	n_2
$\vdots$	$\vdots$	$\vdots$
$[a_{k-1}; a_k[$	x_k	n_k

avec $N = \sum_{i=1}^{k} n_i = n_1 + n_2 + \ldots + n_k$.

4. Sous H_0 on note p_i la probabilité dite théorique définie par
   - $p_i = p(\{X = x_i / X \rightsquigarrow L\})$ si X est discrète,
   - $p_i = p(\{X \in [a_{i-1}; a_i[/ X \rightsquigarrow L\})$ si X est continue.

   $e_i = Np_i$ est l'effectif théorique de la i-ième classe de X.

5. L'indicateur d'écart entre les distributions observées et théoriques est

$$\boxed{\sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i}} \quad (1)$$

dit χ^2 **observé ou calculé**. Cet écart suit pour N suffisamment grand une loi du χ^2_ν d'où le nom du test.

Intuitivement, on comprend que cette grandeur statistique traduise l'écart entre l'échantillon et la loi conjecturée.
Si l'ajustement était parfait, cette expression du χ^2 serait nulle, les effectifs empiriques coïncidant exactement avec les effectifs théoriques.
En revanche, plus grands sont les écarts entre les effectifs observés et les effectifs théoriques $(n_i - e_i)$ et plus forte sera la valeur du χ^2.

En outre, comme la quantité (1) ne peut pas être négative, le test d'ajustement est nécessairement un test unilatéral droit.

Le paramètre ν indiçant χ^2_ν définit le **nombre de degrés de liberté**. C'est le nom donné au nombre d'observations linéairement indépendantes qui apparaissent dans une somme de carrés. Autrement dit, c'est le nombre d'observations aléatoires indépendantes à qui l'on soustrait le nombre de contraintes imposées à ces observations.

Le nombre ν de degrés de liberté est égal à

— si les paramètres de la loi d'ajustement L sont donnés,

$$\boxed{\nu = k - 1}$$

En effet, aucun paramètre n'est à estimer puisque la loi d'ajustement est totalement spécifiée. Le χ^2 est constitué de k écarts $(n_i - e_i)$. Les écarts sont reliés par la contrainte

$$\sum(n_i - e_i) = \sum(n_i - Np_i) = \sum n_i - N\sum p_i = N - N = 0$$

En d'autres termes, lorsqu'on connaît la valeur de $k-1$ écarts, on peut en déduire la valeur du dernier qui n'est donc pas "libre" de varier de manière aléatoire,

— si la loi d'ajustement L comporte r paramètres inconnus,

$$\boxed{\nu = k - r - 1}$$

On impose de ce fait autant de contraintes supplémentaires entre les observations, diminuant d'autant le nombre de degrés de liberté.

Remarque 4.2.1 *Le nombre d'observations par classes ne doit pas être faible, Np_i doit être* ***supérieur*** *à 5, $\forall i = 1, 2, \ldots, k$. Dans le cas contraire, on regroupe deux ou plusieurs classes adjacentes de façon à réaliser cette condition. On tient compte de ce regroupement pour le nombre de degrés de liberté.*

6. Pour un risque de première espèce α, la région critique est définie pour

$$\boxed{\sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i} \geq \chi^2_{\nu,1-\alpha}}$$

d'où la règle de décision :

- χ^2 observé $< \chi^2_{\nu,1-\alpha}$, on décide H_0 et $X \rightsquigarrow L$.
- χ^2 observé $\geq \chi^2_{\nu,1-\alpha}$, on décide H_1 et X ne suit pas la loi L.

Exemple 4.2.1 *Loi uniforme.*
Une statistique relative aux résultats du concours d'entrée à une grande école fait ressortir les répartitions des candidats et des admis selon la profession des parents.

Profession des candidats	Nombre de candidats	Nombre d'admis
① Fonctionnaires et assimilés	2244	180
② Commerce, industrie	988	89
③ Professions libérales	575	48
④ Propriétaires rentiers	423	37
⑤ Propriétaires agricoles	287	13
⑥ Artisans, petits commerçants	210	18
⑦ Banque, assurance	209	17
Total	4936	402

Question : Tester l'hypothèse (risque $\alpha = 0,05$) selon laquelle la profession des parents n'a pas d'influence sur l'accès à cette grande école.

Il s'agit du test d'ajustement d'une distribution théorique, on considère les hypothèses :

- H_0 : la profession des parents n'a pas d'influence sur l'accès à cette grande école, la proportion des admis est constante pour toutes les professions soit $p = \dfrac{402}{4936} \simeq 0,0814$.
- H_1 : la profession des parents influe sur l'accès à cette grande école.

Sous H_0, le nombre d'admis pour la i-ième profession est pN_i.

i	N_i	n_i effectif observé	N_ip effectif théorique	$\frac{(n_i - N_ip)^2}{N_ip}$
1	2244	180	$\frac{2244 \times 402}{4936} \simeq 182,76$	0,0416
2	988	89	$\frac{988 \times 402}{4936} \simeq 80,47$	0,9042
3	575	48	$\frac{575 \times 402}{4936} \simeq 46,83$	0,0293
4	423	37	$\frac{423 \times 402}{4936} \simeq 34,45$	0,1887
5	287	13	$\frac{287 \times 402}{4936} \simeq 23,37$	4,6050
6	210	18	$\frac{210 \times 402}{4936} \simeq 17,10$	0,0471
7	209	17	$\frac{209 \times 402}{4936} \simeq 17,02$	$\simeq 0$
Total	4936	402	402	5,8181

Le χ^2 observé vaut 5,8181. Le nombre de degrés de liberté est $7 - 1 = 6$. La table de l'annexe B fournit $\chi^2_{6;0,95} = 12,59$ donc χ^2 observé $< \chi^2_{6/0,95}$. On choisit H_0, ce qui signifie que la profession des parents n'a pas d'influence sur l'accès à cette grande école.

Exemple 4.2.2 *Loi binomiale.*
Supposons qu'on ait recueilli 300 boîtes contenant chacune trois ampoules. Dans chaque boîte, on compte le nombre d'ampoules défectueuses. On obtient les résultats suivants :

Nombre d'ampoules défectueuses x_i	Nombre de boîtes observées n_i
0	190
1	95
2	10
3	5
Total	300

Pour chaque ampoule testée, on peut observer deux états différents : l'ampoule est défectueuse ou non. Le nombre X d'ampoules défectueuses par boîte suit une loi binomiale de paramètres $n = 3$ et p.
Dans la distribution observée, le nombre d'ampoules défectueuses est de

$$0 \times 190 + 1 \times 95 + 2 \times 10 + 3 \times 5 = 130$$

soit 130 ampoules défectueuses sur un total de 900 ampoules. La proportion d'ampoules défectueuses est alors de $\frac{130}{900} \simeq 0,144$. Prenons $p = 0,15$ et réalisons alors le test suivant : soit X le nombre d'ampoules défectueuses par boîte

- $H_0 : X \rightsquigarrow \mathcal{B}(3; 0,15)$.

- H_1 : X ne suit pas cette loi binomiale.

On détermine les probabilités théoriques :

- $p_0 = p(\{X = 0/X \rightsquigarrow \mathcal{B}\}) = (0,85)^3 \simeq 0,6141$
- $p_1 = p(\{X = 1/X \rightsquigarrow \mathcal{B}\}) = C_3^1(0,15)(0,85)^2 \simeq 0,3251$
- $p_2 = p(\{X = 2/X \rightsquigarrow \mathcal{B}\}) = C_3^2(0,15)^2(0,85) \simeq 0,0574$
- $p_3 = p(\{X = 3/X \rightsquigarrow \mathcal{B}\}) = (0,15)^3 \simeq 0,0034$

On a le tableau

x_i	effectif observé n_i	p_i	effectif théorique Np_i
0	190	0,6141	184,23
1	95	0,3251	97,53
2	10	0,0574	17,22
3	5	0,0034	1,02
Total	$N = 300$	1	300

L'effectif théorique de la quatrième classe est faible : $1,02 < 5$. On effectue un regroupement de classes, les classes "2" et "3".

x_i	n_i	Np_i	$\frac{(n_i - Np_i)^2}{Np_i}$
0	190	184,23	0,18071
1	95	97,53	0,06563
2 ou 3	15	18,24	0,57553
Total	300	300	0,82187

Après le regroupement, le nombre de classes est 3, le nombre de degrés de liberté est $3 - 1 = 2$.
Au risque $\alpha = 0,01$ on a $\chi^2_{2;0,99} = 9,21$. Donc

$$\chi^2 \text{ observé} = 0,82187 < \chi^2_{2;0,99}.$$

On ne rejette pas H_0 au profit de H_1. On considère que le nombre d'ampoules défectueuses par boîte suit une loi binomiale de paramètres $n = 3$ et $p = 0,15$ au risque $\alpha = 0,01$.

Exemple 4.2.3 *Loi normale.*
On suppose que le rendement X (quintaux par hectares d'une parcelle de blé) suit une loi normale $\mathcal{N}(m, \sigma)$. L'observation du rendement de 1000 parcelles a donné les résultats suivants :

Rendement	[0; 10[	[10; 20[	[20; 30[	[30; 40[	[40; 50[	[50; 60[	[60; 70[	[70; 80[	[80; 90[
Nombre de parcelles	5	6	40	168	288	277	165	49	2

1. Déterminer la moyenne arithmétique et l'écart-type de la distribution observée.

- $\overline{x} = \dfrac{\sum_i n_i x_i}{N} = 49,76$

- $\sigma'^2 = \dfrac{\sum\limits_i n_i x_i^2}{N} - \overline{x}^2 = 164,5424$ donc $\sigma' \simeq 12,827$.

2. Vérifier pour un test du χ^2 avec un risque de $0,05$ si l'ajustement de la distribution observée à une loi normale $\mathcal{N}(n = 50, \sigma = 13)$ est acceptable.

   Les hypothèses du test du χ^2 sont les suivantes :
   - $H_0 : X \rightsquigarrow \mathcal{N}(50, 13)$
   - $H_1 : X$ ne suit pas $\mathcal{N}(50, 13)$

   On désigne par $[a_0; a_1[$, $[a_1; a_2[$,...,$[a_8; a_9[$ les classes et par x_1, x_2,...,x_9 les centres de ces classes. Sous H_0, $X \rightsquigarrow \mathcal{N}(50, 13)$ et $Z = \dfrac{X - 50}{13} \rightsquigarrow \mathcal{N}(0, 1)$, donc $p_i = p(\{X \in [a_{i-1}; a_i[\}) = \Pi(z_i) - \Pi(z_{i-1})$ avec $z_i = \dfrac{a_i - 50}{13}$ et $z_{i-1} = \dfrac{a_{i-1} - 50}{13}$. L'effectif théorique de la i-ème classe est $1000p_i$ et

$$\sum_i \frac{(n_i - Np_i)^2}{Np_i} \rightsquigarrow \chi^2_\nu$$

Classe $[x_{i-1}; x_i[$	n_i	z_i	$\Pi(z_i)$	p_i	Np_i	Np_i corrigé	n_i corrigé	$\frac{(n_i - Np_i)^2}{Np_i}$
$[0; 10[$	5	$-3,0769$	$0,001$	$0,0009$	$0,9$	$10,4$	11	$0,0346$
$[10; 20[$	6	$-2,3077$	$0,0105$	$0,0095$	$9,5$			
$[20; 30[$	40	$-1,5385$	$0,0620$	$0,0515$	$51,5$	$51,5$	40	$2,568$
$[30; 40[$	168	$-0,7692$	$0,2209$	$0,1589$	$158,9$	$158,9$	168	$0,5211$
$[40; 50[$	288	0	$0,5$	$0,2791$	$279,1$	$279,1$	288	$0,283$
$[50; 60[$	277	$0,7692$	$0,7791$	$0,2791$	$279,1$	$279,1$	277	$0,0158$
$[60; 70[$	165	$1,5385$	$0,9380$	$0,1589$	$158,9$	$158,9$	165	$0,234$
$[70; 80[$	49	$2,3077$	$0,9895$	$0,0515$	$51,5$	$51,5$	49	$0,1214$
$[80; 90[$	2	$3,0769$	$0,9990$	$0,0095$	$9,5$	$9,5$	2	$5,9211$
Total	1000	–	–	1	1000	1000	1000	$9,7$

   On effectue le regroupement des deux premières classes car $Np_i < 5$. Le χ^2 observé vaut $9,7$. Après le regroupement, il reste 8 classes, les deux paramètres de la loi normale sont donnés, le nombre de degrés de liberté est $\nu = (9 - 1) - 1 = 7$. À l'aide de la table, on obtient $\chi^2_{7;0,95} = 14,07$. Ainsi,

$$\chi^2 \text{ observé} < \chi^2_{7;0,95}.$$

   On choisit H_0, l'ajustement de la distribution observée à une loi normale $\mathcal{N}(50, 13)$ est acceptable.

Exemple 4.2.4 *Loi de Poisson.*
Souvent, lorsqu'on envisage une modèle pour un phénomène qu'on étudie, on ne spécifie pas complètement la loi qu'on considère. Supposons qu'on s'intéresse au nombre de voitures se présentant par minute à un poste de péage sur une autoroute. On peut se demander si cette variable aléatoire peut être modélisée par une loi de Poisson. On souhaite donc tester l'hypothèse fondamentale

$$H_0 : X \rightsquigarrow \mathcal{P}(\lambda)$$

contre l'hypothèse alternative

$$H_1 : X \text{ ne suit pas } \mathcal{P}(\lambda).$$

On ne précise pas la valeur du paramètre λ. On peut toutefois l'estimer à partir des données disponibles mais dans ce cas, $r = 1$. Le nombre de degrés sera alors $\nu = k - r - 1 = k - 2$.

On effectue 200 comptages au péage.

x_i	0	1	2	3	4	5	6	7	8	≥ 9	Total
n_i	6	15	40	42	37	30	10	12	8	0	200
$n_i x_i$	0	15	80	126	148	150	60	84	64	0	727

où x_i est le nombre de voitures par minute lors de la i-ième l'observation et n_i est l'effectif correspondant. Par exemple, $x_1 = 0$ et $n_1 = 6$ c'est-à-dire que lors de 6 observations, il y a 0 voiture.
La moyenne arithmétique de cette distribution observée est

$$\overline{x} = \frac{\sum\limits_i n_i x_i}{\sum\limits_i n_i} = \frac{727}{200} = 3,635 \simeq 3,5$$

On peut tester l'hypothèse

$$H_0 : X \rightsquigarrow \mathcal{P}(\lambda = 3,5).$$

x_i	n_i	p_i	Np_i	Np_i corrigé	n_i corrigé	$\frac{(n_i - Np_i)^2}{Np_i}$
0	6	0,0302	6,04	6,04	6	0,00026
1	15	0,1057	21,14	21,14	15	1,78333
2	40	0,1850	37	37	40	0,24324
3	42	0,2158	43,16	43,16	42	0,03118
4	37	0,1888	37,76	37,76	37	0,01530
5	30	0,1322	26,44	26,44	30	0,47933
6	10	0.0771	15,42	15,42	10	1,90508
7	12	0,0385	7,7	7,7	12	2,40130
8	8	0,0169	3,38	5,34	8	1,32502
≥ 9	0	0,0098	1,96			
Total	200	1	200	200	200	8,18404

On a $p_i = p(\{X = x_i / X \rightsquigarrow \mathcal{P}(3,5)\})$ donc

- $p_0 = p(\{X = 0 / X \rightsquigarrow \mathcal{P}(3,5)\}) = e^{-3,5} \simeq 0,0302$ et
- $p_1 = p(\{X = 1 / X \rightsquigarrow \mathcal{P}(3,5)\}) = e^{-3,5} 3,5 \simeq 0,1057$.

On a effectué le regroupement des deux dernières classes car l'effectif théorique y est inférieur à 5. Après ce regroupement, le nombre de classes est de 9. Le nombre de degrés de liberté est $9 - 1 - 1 = 7$. Au risque $\alpha = 0,01$, $\chi^2_{7;0,99} = 18,48$ donc

$$\chi^2 \text{ observé} = 8,18404 < \chi^2_{7;0,99}.$$

On ne rejette pas l'hypothèse H_0 et $X \rightsquigarrow \mathcal{P}(\lambda = 3,5)$ au risque $\alpha = 0,01$.

4.3 Comparaison de distributions observées. Test d'indépendance. Test d'homogénéité

4.3.1 Présentation du test

Le test du χ^2 est également utilisé pour tester l'indépendance de deux variables aléatoires. Considérons l'exemple suivant :

Exemple 4.3.1 On a posé à des parents la question suivante : « Dans cette liste, quelle est la qualité que vous souhaitez transmettre prioritairement à votre (vos) fille(s) ? »

	Femmes		Total	Hommes	Total
	Actives	Au foyer			
l'honnêteté	47	50	97	108	205
le sens du devoir	16	21	37	46	83
la patience	7	7	14	8	22
la coquetterie	4	3	7	6	13
l'esprit de famille	18	24	42	38	80
l'indépendance	21	12	33	30	63
le sens créatif	9	4	13	16	29
le dévouement	7	8	15	6	21
les qualités ménagères	11	20	31	24	55
la volonté	20	15	35	36	71
les bonnes manières	10	8	18	26	44
la réussite des études	16	12	28	34	62
	186	184	370	378	748

On peut par exemple se demander si les pères et les mères attachent la même importance aux qualités proposées. Autrement dit, on peut regarder si la qualité retenue est indépendante du sexe du parent. On veut tester les hypothèses

. H_0 : la qualité préférée est indépendante du sexe du parent interrogé,

. H_1 : cette qualité dépend du sexe.

On peut aussi se demander si la qualité retenue par les mères est indépendante du fait qu'elles travaillent ou non.

. H_0 : la qualité préférée par la mère est indépendante de son activité,

. H_1 : la qualité préférée par la mère n'est pas indépendante de son activité.

4.3.2 Construction du test

1. On considère deux variables aléatoires X et Y (le plus souvent qualitatives), X prenant les modalités x_1, x_2,..., x_k et Y les modalités y_1, y_2,...,y_p. On considère la loi de probabilité du couple (X, Y),

$$p_{ij} = p(\{X = x_i\} \cap \{Y = y_j\})$$

et les lois marginales de X et de Y

$$p_{i.} = p\{X = x_i\} = \sum_{j=1}^{p} p_{ij} \text{ et } p_{.j} = p\{Y = y_j\} = \sum_{i=1}^{k} p_{ij}$$

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_p	$p_{i.}$
x_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	p_{1p}	$p_{1.}$
x_2	p_{21}	p_{22}	$\cdots$				$p_{2.}$
$\vdots$				$\vdots$			$\vdots$
x_i			$\cdots$	p_{ij}	$\cdots$		$p_{i.}$
$\vdots$				$\vdots$			$\vdots$
x_k							$p_{k.}$
$p_{.j}$	$p_{.1}$	$p_{.2}$	$\cdots$	$p_{.j}$	$\cdots$	$p_{.p}$	1

2. On veut tester l'indépendance des variables X et Y.
   . H_0 : X et Y sont indépendantes,
   . H_1 : X et Y ne sont pas indépendantes.

3. Sous H_0, $p_{ij} = p_{i.} \times p_{.j}$ donc
$$p(\{X = x_i\} \cap \{Y = y_j\}) = p\{X = x_i\}p\{Y = y_j\}$$
$$\forall i = 1, \ldots, k,\ \forall j = 1, \ldots, p$$
   On peut reformuler les hypothèses H_0 et H_1
   . $H_0 : p_{ij} = p_{i.} \times p_{.j},\ \forall i,\ \forall j$
   . $H_1 : \exists i,\ \exists j,\ p_{ij} \neq p_{i.} \times p_{.j}$

4. Comme les valeurs p_{ij}, $p_{i.}$ et $p_{.j}$ sont inconnues, on les estime à l'aide des données relatives aux variables X et Y. Ces N données sont le plus souvent présentées sous la forme d'un tableau de contingence.

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_p	effectif de X : $n_{i.}$
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1p}	$n_{1.}$
x_2	n_{21}	n_{22}					$n_{2.}$
$\vdots$				$\vdots$			$\vdots$
x_i			$\cdots$	n_{ij}	$\cdots$		$n_{i.}$
$\vdots$				$\vdots$			$\vdots$
x_k							$n_{k.}$
effectif de Y : $n_{.j}$	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.p}$	N

   Soient
   . n_{ij} l'effectif de $\{X = x_i\} \cap \{Y = y_j\}$,
   . $n_{i.} = \sum_{j=1}^{p} n_{ij}$ l'effectif de $\{X = x_i\}$,
   . $n_{.j} = \sum_{i=1}^{k} n_{ij}$ l'effectif de $\{Y = y_j\}$.

Parfois, on dispose du tableau des fréquences observées f_{ij} (fréquence de l'événement $\{X = x_i\} \cap \{Y = y_j\}$) où

$$f_{ij} = \frac{n_{ij}}{N},\ f_{i.} = \frac{n_{i.}}{N} \text{ et } f_{.j} = \frac{n_{.j}}{N}.$$

On estime les probabilités p_{ij}, $p_{i.}$ et $p_{.j}$ respectivement par

. $\hat{p}_{ij} = f_{ij} = \dfrac{n_{ij}}{N}$

. $\hat{p}_{i.} = f_{i.} = \dfrac{n_{i.}}{N}$

. $\hat{p}_{.j} = f_{.j} = \dfrac{n_{.j}}{N}$

5. Sous H_0, $p_{ij} = p_{i.} \times p_{.j}$ avec les estimations $\dfrac{n_{ij}}{N} \simeq \dfrac{n_{i.}}{N} \times \dfrac{n_{.j}}{N}$ soit $n_{ij} \simeq \dfrac{n_{i.}n_{.j}}{N}$, l'observation n_{ij} doit être proche d'une quantité notée $v_{ij} = \dfrac{n_{i.}n_{.j}}{N}$ appelée **effectif théorique**.

   Cet effectif théorique représente l'effectif qu'on doit approximativement observer si l'hypothèse d'indépendance est vraie.

   On construit une mesure de l'écart entre l'effectif observé et l'effectif théorique

$$\chi^2 = \sum_{i=1}^{k} \left[\sum_{j=1}^{p} \frac{(n_{ij} - v_{ij})^2}{v_{ij}} \right]$$

   Si l'hypothèse d'indépendance est vraie, cette quantité doit être petite.

   On peut montrer que si les variables X et Y sont indépendantes alors pour N grand, $X^2 \rightsquigarrow \chi^2_\nu$ avec $\nu = (k-1)(p-1)$ et on utilise la règle de décision usuelle.

   - χ^2 observé $\geq \chi^2_{\nu,1-\alpha}$, H_0 est rejetée au profit de H_1, les variables X et Y ne sont pas indépendantes,
   - χ^2 observé $< \chi^2_{\nu,1-\alpha}$, l'hypothèse H_0 est acceptée, c'est-à-dire que les variables X et Y sont indépendantes.

4.3.3 Exemples

1. Reprenons l'exemple 4.3.1 :

   On a les hypothèses

   . H_0 : la qualité préférée est indépendante du sexe du parent interrogé.

   . H_1 : la qualité préférée n'est pas indépendante du sexe du parent interrogé.

Sexe \ Qualité	1	2	3	4	5	6	7	8	9	10	11	12	$n_{i.}$
F	97	37	14	7	42	33	13	15	31	35	18	28	370
H	108	46	8	6	38	30	16	6	24	36	26	34	378
$n_{.j}$	205	83	22	13	80	63	29	21	55	71	44	62	$748 = N$

   On constitue le tableau des effectifs théoriques $v_{ij} = \dfrac{n_{i.}n_{j.}}{N}$, par exemple $v_{11} = \dfrac{205 \times 370}{748} \simeq 101.40$, $v_{21} = \dfrac{205 \times 378}{748} \simeq 103.60$

n_{ij}	v_{ij}	$\frac{(n_{ij}-v_{ij})^2}{v_{ij}}$	n_{ij}	v_{ij}	$\frac{(n_{ij}-v_{ij})^2}{v_{ij}}$
97	101.40	0.19093	16	14.66	0.12248
108	103.6	0.18687	15	10.39	2.04544
37	41.06	0.40145	6	10.61	2.00303
46	41.94	0.39303	31	27.21	0.52790
14	10.88	0.89471	24	27.79	0.51688
8	11.12	0.87540	35	35.12	0.00041
7	6.43	0.05053	36	35.88	0.00040
6	6.57	0.04945	18	21.76	0.64971
42	39.57	0.14923	26	22.24	0.63568
38	40.43	0.14605	28	30.67	0.28244
33	31.16	0.10865	34	31.33	0.22754
30	31.84	0.10633	Total = 748	748	10.63976
13	14.34	0.12522			

Le χ^2 observé est 10.63976, le nombre de degrés de liberté est

$$(2-1)(12-1) = 11.$$

La table fournit

$$\chi^2_{11,0.95} = 19.68$$

Comme χ^2 observé $< \chi^2_{11,0.95}$, on ne rejette pas H_0 au risque de 5%. On admet l'indépendance entre la qualité préférée pour la fille et le sexe du parent interrogé et ceci au risque $\alpha = 0.05$.

2. Avec les mêmes données on peut aussi tester les hypothèses suivantes :
   . H_0 : la qualité préférée par la mère est indépendante de son activité,
   . H_1 : la qualité préférée par la mère n'est pas indépendante de son activité.

Z \ Y	1	2	3	4	5	6	7	8	9	10	11	12	$n_{i.}$
Femmes actives	47	16	7	4	18	21	9	7	11	20	10	16	186
Femmes au foyer	50	21	7	3	24	12	4	8	20	15	8	12	184
$n_{.j}$	97	37	14	7	42	33	13	15	31	35	18	28	$370 = N$

On détermine les effectifs théoriques v_{ij}, par exemple $v_{11} = \dfrac{97 \times 186}{370}$, puis l'indicateur d'écart, χ^2 observé et on conclut.

3. On peut aussi tester les hypothèses
   . H_0 : la qualité préférée est indépendante du sexe du parent actif interrogé,
   . H_1 : la qualité préférée n'est pas indépendante du sexe du parent actif interrogé.

T \ Y	1	2	3	4	5	6	7	8	9	10	11	12	$n_{.j}$
Femmes actives	47	16	7	4	18	21	9	7	11	20	10	16	186
Hommes	108	46	8	6	38	30	16	6	24	36	26	34	378
$n_{i.}$	155	62	15	10	56	51	25	13	35	56	36	50	$564 = N$

On détermine les effectifs théoriques v_{ij}, par exemple $v_{11} = \dfrac{155 \times 186}{564}$, puis l'indicateur d'écart χ^2 observé et on conclut.

Remarque 4.3.1 Il est nécessaire que les v_{ij} soient supérieurs à 5, si ce n'est pas le cas, on regroupe des lignes ou des colonnes adjacentes.

4.4 Comparaison de proportions

Les précédents tests du χ^2 peuvent également être utilisés pour
. comparer une proportion observée à une proportion théorique,
. comparer deux proportions.

4.4.1 Comparaison d'une proportion observée à une proportion théorique

On souhaite comparer la proportion p_0 d'individus possédant le caractère C, observée sur un échantillon de taille N, à la proportion théorique p.

La variable qualitative est à deux modalités
. les effectifs observés sont Np_0, $N(1-p_0)$
. les effectifs théoriques sont Np, $N(1-p)$

L'indicateur d'écart est alors :

$$\chi^2_{\text{observé}} = \frac{(Np_0 - Np)^2}{Np} + \frac{[N(1-p_0) - N(1-p)]^2}{N(1-p)}$$
$$\Leftrightarrow \chi^2_{\text{observé}} = \frac{N^2(p_0-p)^2}{Np} + \frac{N^2(p_0-p)^2}{N(1-p)}$$
$$\Leftrightarrow \chi^2_{\text{observé}} = N(p_0-p)^2\left[\frac{1}{p} + \frac{1}{1-p}\right] = \frac{N(p_0-p)^2}{p(1-p)}$$
$$\Leftrightarrow \chi^2_{\text{observé}} = \frac{N(p_0-p)^2}{p(1-p)}$$

On considère les hypothèses
. $H_0 : p = p_0$, le nombre de degrés de liberté est 1,
. $H_1 : p \neq p_0$.

Si $\chi^2_{\text{observé}} \geq \chi^2_{1,1-\alpha}$, H_0 est rejetée au profit de H_1.

Exemple 4.4.1 Pour l'année 1967, le pourcentage des candidats reçus au baccalauréat Mathématiques Élémentaires a été de $p = 64\%$. M$^{\text{r}}$ X a présenté 40 candidats, 31 furent déclarés reçus. Peut-on dire que M$^{\text{r}}$ X prépare mieux les candidats à l'examen ?

La proportion de reçus dans la classe de M$^{\text{r}}$ X est de $p_0 = \dfrac{31}{40} = 0.775$
Soient

. $H_0 : p = p_0$, les deux proportions sont identiques, M$^{\text{r}}$ X ne prépare pas mieux ses candidats à l'examen,

. $H_1 : p \neq p_0$.

L'indicateur d'écart est $\chi^2 = \dfrac{N(p_0 - p)^2}{p(1-p)}$. On a par conséquent

$$\chi^2 = 40\frac{(0.64 - 0.775)^2}{0.64 \times 0.36} = 3.164062$$

Le nombre de degrés de liberté est 1. La table nous donne $\chi^2_{1,0.95} = 3.841$ donc

$$\chi^2 \text{ observé} < \chi^2_{1,0.95},$$

on ne rejette pas H_0.

Par conséquent, la différence observée entre les proportions n'est pas significative et M$^{\text{r}}$ X ne prépare pas mieux ses candidats à l'examen au risque $\alpha = 5\%$. On a également $\chi^2_{1,0.90} = 2.71$ donc

$$\chi^2 \text{ observé} \geq \chi^2_{1,0.90},$$

on rejette H_0 au risque $\alpha = 10\%$.

4.4.2 Comparaison de deux proportions

S'il s'agit de comparer deux proportions p_1 et p_2 observées sur deux échantillons de taille N_1 et N_2 prélevés dans les populations P_1 et P_2, on construit le tableau de contingence suivant :

variable \ population	P_1	P_2	Total
C	N_1p_1	N_2p_2	$N_1p_1 + N_2p_2$
$\overline{C}$	$N_1(1-p_1)$	$N_2(1-p_2)$	$N_1(1-p_1) + N_2(1-p_2)$
Total	N_1	N_2	$N_1 + N_2$

On procède ensuite comme dans les cas précédents.

Exemple 4.4.2 Une enquête sur la population active de la région parisienne révèle que sur 2400 hommes et 1600 femmes interrogées, les cadres moyens sont au nombre de 314 et 182 respectivement. Peut-on dire que l'accession à cette catégorie est égalitaire pour les deux sexes ?

. H_0 : homogénéité ou accession égalitaire,

. H_1 : accession égalitaire.

La fréquence des cadres chez les hommes est $f_H = \dfrac{314}{2400}$,

celle des femmes est $f_F = \dfrac{182}{1600}$.

On compare les deux proportions f_H et f_F observées sur deux échantillons des populations masculine et féminine. On construit le tableau de contingence

sexe \ activité	C	$\overline{C}$	Total
H	314	2086	2400
F	182	1418	1600
Total	496	3504	4000

On a

$$v_{11} = \frac{2400 \times 496}{4000} = 297.6, \quad v_{12} = \frac{2400 \times 3504}{4000} = 2102.4$$
$$v_{21} = \frac{1600 \times 496}{4000} = 198.4, \quad v_{12} = \frac{2400 \times 3504}{4000} = 1401.6$$

L'indicateur d'écart est $\chi^2 = \sum \frac{(n_{ij} - v_{ij})^2}{v_{ij}}$ et χ^2 observé $= 2.5792$. Le nombre de degrés de liberté est $(2-1) \times (2-1) = 1$. Pour un risque α de 5%, $\chi^2_{1,0.95} = 3.84$ et

$$\chi^2 \text{ observé} < \chi^2_{1,0.95}.$$

On décide H_0, l'accession à la catégorie cadre est égalitaire pour les deux sexes.

4.5 Exercices

Exercice 27 En lançant successivement 60 fois un dé, un joueur obtient les résultats suivants :

Faces x_i	1	2	3	4	5	6
Effectifs n_i	15	7	4	11	6	17

Le dé est-il truqué ?

Exercice 28 On a enregistré le nombre X de clients entrant dans un magasin en 1 minute. On a obtenu le tableau suivant :

Nombre de clients x_i	Nombre de minutes où $X = x_i$ (où il est entré x_i clients)
0	23
1	75
2	68
3	51
4	30
5	10
plus de 5	7

Peut-on admettre que les arrivées sont régies par une loi de Poisson de paramètre $m = 2$ (au seuil $\alpha = 0,05$) ?

Exercice 29 Une enquête sur les chiffres d'affaires mensuels de 103 magasins de détail a donné les résultats suivants (en milliers d'euros) :

Classes de chiffres d'affaires	Centres de classes	Nombre de magasins
$5,5$ à moins de $6,5$	6	2
$6,5$ à moins de $7,5$	7	3
$7,5$ à moins de $8,5$	8	12
$8,5$ à moins de $9,5$	9	27
$9,5$ à moins de $10,5$	10	23
$10,5$ à moins de $11,5$	11	15
$11,5$ à moins de $12,5$	12	12
$12,5$ à moins de $13,5$	13	5
$13,5$ à moins de $14,5$	14	2
$14,5$ à moins de $15,5$	15	2

Peut-on considérer que l'échantillon est tiré d'une loi normale ?

Exercice 30 On a étudié le nombre de garçons dans 1883 familles de 7 enfants. Les résultats sont présentés en fonction du nombre x_i de garçons, rangés de 0 à 7 :

Nombre de garçons x_i	Effectif des familles n_i
0	27
1	111
2	287
3	480
4	529
5	304
6	126
7	19
Total	1883

Peut-on admettre au seuil de 5% que le nombre x_i de garçons par famille obéisse à une loi binomiale ? Laquelle ?

Exercice 31 Une entreprise achète une machine dont le fabricant assure que 95% des pièces qu'elle permet d'usiner satisfont aux normes exigées. Sur un échantillon de 100 pièces, 9 ne satisfont pas aux normes.
Peut-on admettre, au seuil de 5% que les caractéristiques réelles de la machine ne correspondent pas aux garanties du fournisseur ? (Résoudre ce test de comparaison d'une fréquence à une valeur standard à l'aide du test du χ^2).

Exercice 32 On fait passer une épreuve à deux groupes, l'un de 50 personnes et l'autre de 30 personnes.
Dans le premier groupe, il y a 42 succès à l'épreuve.
Dans le second groupe, il y a 18 succès à l'épreuve.
L'hypothèse à tester est que les résultats sont les mêmes pour les deux populations d'où l'on a extrait les deux échantillons.

n_i	Succès	Échecs	Total
1^{er} groupe	42	8	$N_1 = 50$
$2^{\text{ème}}$ groupe	18	12	$N_2 = 30$
Total	60	20	$N = 80$

Exercice 33 Il s'agit de rechercher chez 160 adolescents asthmatiques de la région Aquitaine un lien entre l'intensité de l'asthme et la présence ou l'absence d'éczéma durant l'année d'étude ou antérieurement. On a établi le tableau suivant :

Asthme \ Eczéma	Présent	Passé	Jamais	Total
Fort	22	28	28	78
Faible	12	33	37	82
Total	34	61	65	160

On souhaite examiner l'hypothèse

H_0 : « l'intensité de l'asthme et la présence d'eczéma sont indépendantes »

1. Déterminer dans un tableau les effectifs théoriques à $0,1$ au plus près.
2. Préciser le nombre de degrés ν de liberté.
3. Calculer le χ^2.
4. Quelle conclusion peut-on déduire au seuil de signification $0,95$ (avec pour risque d'erreur 5%) ?

Exercice 34 On cherche à savoir si les résultats électoraux d'un village du Sud-Est de la France correspondent aux résultats nationaux pour une élection où les cinq candidats A,B,C,D et E se présentent :

	A	B	C	D	E
Nombre de votants du village	475	364	1968	1633	560
Résultats nationaux en %	$10,3$	$7,1$	$40,8$	$32,1$	$9,7$

1. Faire un tableau comportant les résultats des 5000 votants du village et les résultats théoriques si l'hypothèse de ressemblance est vérifiée.
2. Calculer le χ^2 associé.
3. Interpréter en utilisant la table du χ^2 au risque de 1%.

Exercice 35 Au cours d'une élection, il y a 5 candidats en présence.
Une enquête d'opinion est faite sur un échantillon de 875 sujets, 410 hommes et 465 femmes.

Candidats	A	B	C	D	E	Totaux
Hommes	25	75	105	130	75	410
Femmes	147	116	17	80	105	465
Total	172	191	122	210	180	875

1. Déterminer à l'aide d'un tableau les effectifs théoriques à $0,01$ au plus près.
2. Préciser le nombre de degrés de liberté ν.
3. Calculer le χ^2.

4. Quelle conclusion peut-on déduire au risque de 1% ?

Exercice 36 On veut savoir si le rythme cardiaque est différent entre la population adulte urbaine de l'agglomération bordelaise et la population rurale de l'Aquitaine.
Pour cela, on choisit un échantillon aléatoire de 150 personnes adultes de l'agglomération bordelaise et un échantillon aléatoire de 400 personnes adultes vivant dans des communes rurales de moins de 5000 habitants de la région Aquitaine. Ces observations conduisent au tableau suivant :

Population \ Rythme R	$R < 65$	$65 \leq R < 70$	$70 \leq R < 75$	$75 \leq R < 80$	$80 \leq R$
Urbaine	6	21	53	60	10
Rurale	32	80	163	112	13

On considère l'hypothèse :

H_0 : « Les deux caractères *rythme cardiaque* et *vie urbaine ou rurale* sont indépendants »

1. Déterminer dans un tableau les effectifs théoriques à $0,01$ au plus près.
2. Préciser le nombre de degrés ν de liberté.
3. Calculer le χ^2.
4. Quelle conclusion peut-on déduire au seuil de signification $0,95$ (avec pour risque d'erreur 5%) ?

Exercice 37 Trois groupes d'étudiants de même niveau, constitués indépendamment les uns des autres, sont soumis à des méthodes pédagogiques différentes :
- Le premier groupe reçoit un enseignement traditionnel.
- Le second groupe bénéficie d'un renforcement pédagogique dans le cadre des méthodes traditionnelles.
- Le troisième groupe expérimente une nouvelle méthode pédagogique.

Les étudiants des trois groupes passent en fin d'année un même examen. On enregistre les résultats suivants :
- Sur le premier groupe d'effectif égal à $N_1 = 115$, $K_1 = 82$ étudiants sont admis.
- Sur le deuxième groupe d'effectif égal à $N_2 = 90$, $K_1 = 71$ étudiants sont admis.
- Sur le troisième groupe d'effectif égal à $N_3 = 35$, $K_3 = 26$ étudiants sont admis.

Peut-on admettre que l'une des trois méthodes est plus efficace que les autres (au seuil $\alpha = 0,05$) ?

Exercice 38 Sur un échantillon de 200 ménages choisis au hasard, on a étudié la propension moyenne à épargner (variable Y) en fonction du revenu disponible (variable X).
Pour la variable X, on a distingué 3 classes :
- x_1 : faibles revenus,
- x_2 : revenus intermédiaires,
- x_3 : revenus élevés.

De même, les taux d'épargne ont été classés en 3 niveaux :
- y_1 : faibles taux,
- y_2 : taux intermédiaires,
- y_3 : taux élevés.

Les résultats sont présentés dans la table de contingence :

Revenus \ Épargne	y_1	y_2	y_3	Total $N_{i.}$
x_1	53	14	6	73
x_2	15	58	8	81
x_3	7	10	29	46
Total $N_{.j}$	75	82	43	200

Existe t-il une relation entre le taux d'épargne et le niveau de revenu disponible ?

Exercice 39 Une étude est menée dans une petite université sur l'absentéisme des étudiants. On aimerait savoir si certaines plages horaires sont plus propices à une absence aux cours qu'une autre. Pour cela, on a relevé, au cours d'un mois, le nombre d'absences d'étudiants aux cours d'une petite composante à différents moments de la journée.

Heures de la journée	Nombre d'étudiants absents
8-10	25
10-12	15
13-15	18
15-17	32

En considérant cet échantillon tiré au hasard, peut-on dire, au seuil de 5%, que les absences des étudiants aux cours se répartissent uniformément tout au long de la journée ?

Exercice 40 Une entreprise fabriquant des produits alimentaires sucrés veut élargir sa gamme de barres de céréales en en lançant une nouvelle sur le marché. Le directeur du marketing décide de faire une enquête de goût en faisant tester ce nouveau produit à 200 personnes. Le test a lieu en aveugle, et les personnes ont donc à se prononcer sur leur préférence concernant la nouvelle barre et quatre autres barres de céréales concurrentes. Les produits étant appelés A (la nouvelle barre), B, C, D et E, les résultats du test dont les suivants :

Barres	A	B	C	D	E
Nombre de préférences	40	35	55	40	30

Au seuil 5%, peut-on dire au vu des résultats d'échantillon que la nouvelle barre a meilleur goût que les autres ?

Exercice 41 Un responsable qualité d'une entreprise fabriquant de l'appareillage électronique a mesuré la durée de vie de 60 dispositifs électroniques d'un même type. Il a obtenu les résultats suivants :

Durée de vie (en heures)	Nombre de dispositifs
$[250; 270[$	3
$[270; 290[$	5
$[290; 310[$	15
$[310; 330[$	22
$[330; 350[$	13
$[350; 370[$	2

Les donnée permettent-elles, au seuil de 5%, de penser que la durée de vie d'un dispositif électronique de ce type est distribué selon une loi normale ?

Exercice 42 On a observé le nombre de défauts de pièces de tissu traitées par un teinturier. Les résultats de 50 observations sont reproduits dans le tableau ci-dessous (par exemple, 8 des 50 pièces présentaient 3 défauts). Des études antérieures avaient permis de faire l'hypothèse que le nombre de défauts par pièce pouvait être considéré comme une variable aléatoire X obéissant à une loi de Poisson. Les observations permettent-elles de confirmer cette hypothèse, au seuil de 5% ?

x_i	0	1	2	3	4	5
n_i	6	14	16	8	4	2