Université de Picardie Jules Verne **2017-2018**
UFR des Sciences

Licence mention Mathématiques - Semestre 3
Statistique

Les tests de khi-deux

1. Conformité à un modèle théorique

Dans une population donnée, on étudie un caractère X pouvant prendre r modalités et on cherche à savoir si on peut considérer que ce caractère est d'un type donné. Plus précisément, désignant par p_i la probabilité d'apparition dans la population de la $i^{ème}$ modalité du caractère, 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 dernier cas, il faut choisir 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 égale à l'effectif observé de la $i^{ème}$ modalité du caractère ; l'effectif théorique étant égal à np_i.

Test de H_0 : X suit la loi théorique contre H_1 : X ne suit pas la loi théorique

Ce test s'appuie sur la distance D entre les effectifs observés et théoriques :

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

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

On sait que sous l'hypothèse H_0, D suit approximativement la loi de khi-deux à $r - 1 - k$ degrés de liberté, où k est le nombre de paramètres à estimer de la loi théorique choisie. On détermine b_α tel que $P(D \geq b_\alpha) = \alpha$ (table 4), et on décide que :

- si $d < b_\alpha$, alors on ne peut rejeter H_0 ;
- si $d \geq b_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

La qualité de l'approximation de la loi de D est satisfaisante lorsque les effectifs théoriques vérifient tous la condition $np_i \geq 5$. 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 $np_i \geq \frac{5s}{r}$, où s est égal au nombre de modalités ayant un effectif théorique $np_i < 5$.

Exemple 1 : **test de conformité à une distribution théorique**

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 à p_5 données dans le tableau ci-dessous.

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 données dans le tableau ci-dessous.

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

Population : celle qui est étudiée.
Caractère : le génotype X, à $r = 5$ modalités de probabilité théorique p_i.
Echantillon $(X_1, \ldots, X_n)$ de taille $n = 400$.
Les p_i étant donnés, il n'y a pas de paramètre à estimer : $k = 0$.

Test de H_0 : X suit la loi théorique contre H_1 : X ne suit pas la loi théorique

x_i	n_i	p_i	np_i	$n_i - np_i$	$\frac{(n_i-np_i)^2}{np_i}$
A_1	200	0,40	160	+40	10
A_2	40	0,20	80	−40	20
A_3	96	0,20	80	+16	3,2
A_4	36	0,10	40	−4	0,4
A_5	28	0,10	40	−12	3,6
	400	1	400		37,2

On calcule $d = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} = 37,2$.

On sait que sous l'hypothèse H_0, D suit approximativement la loi de khi-deux à $r - 1 - k = 4$ degrés de liberté.

On détermine b_α tel que $P(D \geq b_\alpha) = \alpha$: pour $\alpha = 0,05$, on trouve $b_\alpha = 9,49$.

Comme $d \geq b_\alpha$, on rejette l'hypothèse H_0, i.e. la conformité à la loi théorique : la répartition des génotypes dans l'échantillon n'est pas conforme à celle de la population. En prenant cette décision de rejet de H_0, on a une probabilité $\alpha = 0,05$ de se tromper.

Exemple 2 : **test de conformité à une loi de Poisson $\mathcal{P}(\lambda)$**

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 ?

Population : les coopératives. Caractère : nombre d'usagers arrivés X, à $r = 7$ modalités.

Echantillon $(X_1, \ldots, X_n)$ de taille $n = 150$ de X. On cherche à ajuster à la distribution observée une loi théorique suivie par X (i.e. les probabilités p_i des modalités de X).

Test de H_0 : X suit une loi de Poisson contre H_1 : X ne suit pas une loi de Poisson

Rappel : X suit la loi de Poisson $\mathcal{P}(\lambda)$ si X est à valeurs dans $\mathbb{N}$ et si, pour tout $k \in \mathbb{N}$, $P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$.

On a $E(X) = Var(X) = \lambda$.

Les p_i devant être calculés à l'aide la loi de Poisson, il y a un paramètre à estimer : $k = 1$.

On calcule $\bar{x} = \frac{1}{n}\sum_{i=1}^{7} n_i x_i = 1,48$, $s^2 = \frac{1}{n}\left(\sum_{i=1}^{7} n_i x_i^2\right) - (\bar{x})^2 = 1,5696$ et $s_c^2 = \frac{n}{n-1}s^2 \simeq 1,58$.

Comme $\bar{x}$ et s_c^2 sont très proches, on pouvait effectivement penser à une loi de Poisson.

On peut alors estimer le paramètre λ à 1,48.

Sous l'hypothèse H_0, on a alors : $p_i = P(X = i) = e^{-1,48}\frac{(1,48)^i}{i!}$.

Voir le tableau en page suivante. On a, après regroupements, $r = 5$.

On calcule $d = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} \simeq 0,75$.

On sait que sous l'hypothèse H_0, D suit approximativement la loi de khi-deux à $r - 1 - k = 3$ degrés de liberté (après regroupements).

On détermine b_α tel que $P(D \geq b_\alpha) = \alpha$ (table 4) : pour $\alpha = 0,05$, on trouve $b_\alpha = 7,81$.

Comme $d < b_\alpha$, on ne peut rejeter l'hypothèse H_0, i.e. la conformité à la loi théorique de Poisson : la répartition du nombre d'usagers arrivés est conforme à une loi de Poisson. En prenant cette décision de non-rejet de H_0, on ne connait pas la probabilité de se tromper (erreur de deuxième espèce).

x_i	n_i	p_i	np_i	$n_i - np_i$	$\frac{(n_i-np_i)^2}{np_i}$
0	37	$0,2276$	$34,15$	$+2,85$	$0,24$
1	46	$0,3369$	$50,54$	$-4,54$	$0,41$
2	39	$0,2493$	$37,40$	$+1,60$	$0,07$
3	19	$0,1230$	$18,45$	$+0,55$	$0,02$
4	5	$0,0455$	$\underline{6,82}$	$-1,82$	
5	3 } 9	$0,0135$ } $0,0632$	$2,02$ } $9,46$	$+0,98$ } $-0,46$	$0,02$
6 et +	1	$0,0042$	$\underline{0,62}$	$+0,38$	
	150	1	150	0	$0,76$

On a regroupé les effectifs théoriques inférieurs à 5. On a maintenant $r = 5$.

Exemple 3 : **test de conformité à une loi Normale** $\mathcal{N}(\mu, \sigma)$

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 ?

Population : les mollusques. Caractère : taux de protéines X, à $r = 7$ modalités.

Echantillon $(X_1, \ldots, X_n)$ de taille $n = 36$ de X.

On cherche à ajuster à la distribution observée une loi théorique suivie par X (i.e. les probabilités p_i des modalités de X). Lorsque la représentation graphique (histogramme) est plutôt symétrique et "en cloche", on peut penser à une loi de Normale. A noter que ce n'est pas tout à fait le cas ici !

Test de H_0 : X **suit une loi Normale contre** H_1 : X **ne suit pas une loi Normale**

Les p_i devant être calculés à l'aide la loi Normale, il y a deux paramètres à estimer : $k = 2$.

Comme dans l'exemple 1, on calcule $\bar{x} \simeq 4,21$ et $s_c \simeq 2,86$.

On peut alors estimer les paramètres μ et σ par $4,21$ et $2,86$. Il s'agira donc de tester si X suit la loi Normale $\mathcal{N}(4,21; 2,86)$, i.e. si $U = \dfrac{X - 4,21}{2,86}$ suit la loi Normale $\mathcal{N}(0;1)$.

Voir le tableau en page suivante. On a, après regroupements, $r = 5$.

On calcule $d = \displaystyle\sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} \simeq 2,99$.

On sait que sous l'hypothèse H_0, D suit approximativement la loi de khi-deux à $r - 1 - k = 2$ degrés de liberté.

On détermine b_α tel que $P(D \geq b_\alpha) = \alpha$ (table 4) : pour $\alpha = 0,05$, on trouve $b_\alpha = 5,99$.

Comme $d < b_\alpha$, on ne peut rejeter l'hypothèse H_0, i.e. la conformité à la loi théorique Normale : la répartition du taux de protéines est conforme à une loi Normale. En prenant cette décision de non-rejet de H_0, on ne connait pas la probabilité de se tromper (erreur de deuxième espèce).

Classes de X	n_i	Classes de U : $]u_i;u_{i+1}[$	$\phi(u_i)$	$p_i = \phi(u_{i+1}) - \phi(u_i)$	np_i	$n_i - np_i$	$\frac{(n_i-np_i)^2}{np_i}$
		$-\infty$	0				
$]-\infty;1,5]$	8	$]-\infty;-0,95]$		0,1711	6,16	+1,84	0,55
		$-0,95$	0,1711				
$]1,5;3]$	7	$]-0,95;-0,42]$		0,1661	5,98	+1,02	0,17
		$-0,42$	0,3372				
$]3;4,5]$	4	$]-0,42;0,10]$		0,2026	7,29	−3,29	1,49
		0,10	0,5398				
$]4,5;6]$	9	$]0,10;0,63]$		0,1959	7,05	+1,95	0,54
		0,63	0,7357				
$]6;7,5]$	2	$]0,63;1,15]$		0,1392	5,01	−3,01	
		1,15	0,8749				
$]7,5;9]$	3	$]1,15;1,67]$		0,0776	2,79	+0,21	0,24
		1,67	0,9525				
$]9;+\infty[$	3	$]1,67;+\infty[$		0,0475	1,71	+1,29	
		$+\infty$	1				
	36			1	36		2,99

On a regroupé les effectifs inférieurs à 5, c'est-à-dire les trois dernières classes, ce qui donne :

$]6;+\infty[$	8				9,51	− 1,51	0,24

2. Indépendance de 2 caractères

Dans une population donnée, on étudie deux caractères X et Y pouvant prendre respectivement r et s modalités. Effectuant plusieurs échantillonnages de même taille n, on désigne par $N_{i,j}$ la variable aléatoire égale à l'effectif observé du couple formé de la $i^{ème}$ modalité du caractère X et de la $j^{ème}$ modalité du caractère Y. En pratique, pour un échantillon, on observe des effectifs $n_{i,j}$.

Sous l'hypothèse d'indépendance de X et Y, l'effectif théorique est égal à $np_{i,j} = \frac{n_{i,\bullet}n_{\bullet,j}}{n}$, avec $n_{i,\bullet} = \sum_{j=1}^{s} n_{i,j}$ et $n_{\bullet,j} = \sum_{i=1}^{r} n_{i,j}$.

Test de H_0 : X et Y sont indépendantes contre H_1 : X et Y ne sont pas indépendantes

Ce test s'appuie sur la distance D entre les effectifs observés et théoriques :

$$D = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(N_{i,j} - np_{i,j})^2}{np_{i,j}} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{N_{i,j}^2}{np_{i,j}} - n \ .$$

On calcule $d = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{i,j} - np_{i,j})^2}{np_{i,j}} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{i,j}^2}{np_{i,j}} - n.$

On sait que sous l'hypothèse H_0, D suit approximativement la loi de khi deux à $(r-1)(s-1)$ degrés de liberté. On détermine le réel b_α tel que $P(D \geq b_\alpha) = \alpha$ (table 4), et on décide que :

- si $d < b_\alpha$, alors on ne peut rejeter H_0 ;
- si $d \geq b_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

La qualité de l'approximation de la loi de D est satisfaisante lorsque les effectifs théoriques vérifient tous la condition $np_{i,j} \geq 5$. 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 $np_{i,j} \geq \frac{5t}{rs}$, où t est égal au nombre de couples de modalités ayant un effectif théorique $np_{i,j} < 5$.

Cas particulier : $r = s = 2$.

Dans ce cas, le test d'indépendance se confond strictement avec le test (bilatéral) d'égalité de deux proportions présenté dans le chapitre précédent. En effet, d est alors le carré de u et b_α le carré de u_α.

Exemple 4 : **test d'indépendance**

Une statistique effectuée sur 800 personnes donne la répartition suivante :

$n_{i,j}$	gros fumeurs	moyen fum.	petits fum.	non fum.	$n_{i\bullet}$
hypertension	74	116	68	82	340
pas d'hypert.	126	174	82	78	460
$n_{\bullet j}$	200	290	150	160	800

Tester au risque 10% l'indépendance entre l'hypertension et la consommation de tabac.

Les deux caractères sont X : hypertension et Y : consommation de tabac.
On a $r = 2$ et $s = 4$.

Sous l'hypothèse d'indépendance de X et Y, les effectifs théoriques sont $np_{i,j} = \dfrac{n_{i\bullet}\, n_{\bullet j}}{n}$.

$np_{i,j}$	gros fumeurs	moyen fum.	petits fum.	non fum.	
hypertension	85	123,25	63,75	68	340
pas d'hypert.	115	166,75	86,25	92	460
	200	290	150	160	800

Par exemple, $np_{1,2} = \dfrac{n_{1\bullet}\, n_{\bullet 2}}{n} = \dfrac{340 \times 290}{800} = 123,25$.

$\frac{(n_{ij}-np_{i,j})^2}{np_{i,j}}$	gros fumeurs	moyen fum.	petits fum.	non fum.
hypertension	1,424	0,426	0,283	2,882
pas d'hypert.	1,052	0,315	0,209	2,130

Par exemple, $\dfrac{(n_{12} - np_{1,2})^2}{np_{1,2}} = \dfrac{(116 - 123,25)^2}{123,25} = 0,426$.

On obtient : $d = \displaystyle\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{i,j} - np_{i,j})^2}{np_{i,j}} = 8,721$.

On sait que sous l'hypothèse H_0, D suit approximativement la loi de khi deux à $(r-1)(s-1) = 3$ degrés de liberté.

On détermine le réel b_α tel que $P(D \geq b_\alpha) = \alpha$ (table 4) : pour $\alpha = 0,10$, on trouve $b_\alpha = 6,25$.

Comme $d \geq b_\alpha$, on rejette l'hypothèse H_0 avec une probabilité α de se tromper : on rejette donc l'indépendance des deux caractères.

Remarque.

Si on teste au risque $\alpha = 0,05$, on a $b_\alpha = 7,81$, et donc $d \geq b_\alpha$: même décision qu'avec $\alpha = 0,10$, et on a diminué la probabilité de se tromper.

Si on teste au risque $\alpha = 0,025$, on a $b_\alpha = 9,35$, et donc $d < b_\alpha$: on ne rejette pas H_0 mais on ne connait pas la probabilité de se tromper (erreur de deuxième espèce).

Exemple 5 : **test d'indépendance de deux caractères à $r = 2$ et s = 2 modalités**

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.

On a déjà traité cette question dans un précédent chapitre par un **test d'homogénéité** (comparaison de deux proportions).

Population 1 : hommes. Variable X_1 de loi de Bernoulli $\mathcal{B}(p_1)$, où p_1 est la proportion d'hommes fumeurs. Echantillon de taille $n_1 = 40$ de X_1. Estimation de p_1 : $f_1 = \frac{8}{40} = 0,2$.

Population 2 : femmes. Variable X_2 de loi de Bernoulli $\mathcal{B}(p_2)$, où p_2 est la proportion de femmes fumeuses. Echantillon de taille $n_2 = 60$ de X_2. Estimation de p_2 : $f_2 = \frac{18}{60} = 0,3$.

Les échantillons sont indépendants.

Test (bilatéral) de $H_0 : p_1 = p_2 = p$ contre $H_1 : p_1 \neq p_2$.

On a $n_1 f_1 = 8 \geq 5$, $n_1(1-f_1) = 32 \geq 5$, $n_2 f_2 = 18 \geq 5$, $n_2(1-f_2) = 42 \geq 5$.

Sous l'hypothèse H_0, $U = \frac{F_1 - F_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})p(1-p)}}$ suit approximativement la loi normale $\mathcal{N}(0;1)$, et en regroupant les deux échantillons, on peut estimer p par $f_{1,2} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2} = \frac{8+18}{40+60} = 0,26$. En remplaçant p par $f_{1,2}$, on ne modifie pas la loi approchée de U.

On calcule $u = \frac{f_1 - f_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})f_{1,2}(1-f_{1,2})}} = \frac{0,2-0,3}{\sqrt{\left(\frac{1}{40} + \frac{1}{60}\right)0,26(1-0,26)}} \simeq -1,12$.

On détermine u_α tel que $P(-u_\alpha < U < u_\alpha) = 1-\alpha$, i.e. $u_\alpha = \phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ (table 2) : pour $\alpha = 0,05$, on trouve $u_\alpha = 1,96$.

Comme $u \in]-u_\alpha, u_\alpha[$, on ne peut rejeter H_0 : la proportion de fumeurs ne diffère pas significativement entre les deux sexes. Pour cette décision de non-rejet, on ne connait pas la probabilité de se tromper (erreur de deuxième espèce).

On peut également traiter cette question par un **test d'indépendance** des deux caractères X : sexe, à $r = 2$ modalités (hommes, femmes), et Y : être fumeur, à $s = 2$ modalités (fumeur, non fumeur).

Test de H_0 : X et Y sont indépendantes contre H_1 : X et Y ne sont pas indépendantes

Ce test s'appuie sur la distance D entre les effectifs observés et théoriques : $D = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(N_{i,j} - np_{i,j})^2}{np_{i,j}}$

Sous l'hypothèse H_0 d'indépendance de X et Y, les effectifs théoriques sont $np_{i,j} = \frac{n_{i\bullet}\, n_{\bullet j}}{n}$.

$n_{i,j}$	fumeurs	non fum	$n_{i\bullet}$
hommes	8	32	40
femmes	18	42	60
$n_{\bullet j}$	26	74	100

$np_{i,j}$	fumeurs	non fum	
hommes	10,4	29,6	40
femmes	15,6	44,4	60
	26	74	100

$\frac{(n_{ij}-np_{i,j})^2}{np_{i,j}}$	fumeurs	non fum	
hommes	0,55	0,19	
femmes	0,37	0,13	
			1,24

On obtient : $d = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - np_{i,j})^2}{np_{i,j}} = 1,24$.

On sait que sous l'hypothèse H_0, D suit approximativement la loi de khi deux à $(r-1)(s-1) = 1$ degré de liberté.

On détermine le réel b_α tel que $P(D \geq b_\alpha) = \alpha$ (table 4) : pour $\alpha = 0,05$, on trouve $b_\alpha = 3,84$.

Comme $d < b_\alpha$, on ne rejette pas l'hypothèse H_0 d'indépendance de X et Y : on peut donc considérer que les caractères "sexe" et "être fumeur" sont indépendants, ce qui signifie que les proportions de fumeurs chez les hommes et chez les femmes ne diffèrent pas significativement. Cela correspond aux résultats du test d'homogénéité précédent. En prenant cette décision de non-rejet de H_0, on ne connait pas la probabilité de se tromper (erreur de deuxième espèce).

Remarque.

Dans cette dernière présentation, on a supposé travailler sur un échantillon de $n = 100$ personnes, issues de la population générale hommes/femmes, et sur lesquelles on a observé les deux variables "sexe" et "être fumeur". On aurait pu présenter ce travail en termes d'homogénéité des deux populations "hommes" et "femmes", et donc de comparaison de deux échantillons : voir le paragraphe suivant.

3. Homogénéité : comparaison de plusieurs échantillons

Dans une population donnée, on étudie un caractère X pouvant prendre s modalités.

On dispose de r échantillons pouvant provenir de cette population.

Effectuant plusieurs échantillonnages, on désigne par $N_{i,j}$ la variable aléatoire égale à l'effectif observé de la $j^{ème}$ modalité du caractère X dans le $i^{ème}$ échantillon. En pratique, pour un échantillonnage, on observe des effectifs $n_{i,j}$.

Sous l'hypothèse d'homogénéité des échantillons, l'effectif théorique est égal à $np_{i,j} = \frac{n_{i,\bullet} n_{\bullet,j}}{n}$, avec $n_{i,\bullet} = \sum_{j=1}^{s} n_{i,j}$ et $n_{\bullet,j} = \sum_{i=1}^{r} n_{i,j}$.

Test de H_0 : les échantilllons sont issus de la même population contre $H_1 = \overline{H_0}$

Ce test se déroule comme le test d'indépendance décrit au paragraphe 2, même si le problème posé est de nature différente.

Exemple 6

Dans deux échantillons de populations d'une même espèce, d'effectifs respectifs 100 et 400, on dénombre 4 phénotypes. Les résultats sont les suivants :

n_{ij}	A_1	A_2	A_3	A_4
e_1	10	30	50	10
e_2	60	120	180	40

Les deux populations présentent-elles les mêmes proportions de phénotypes ?

Autrement dit, les deux populations sont-elle identiques en termes de répartition des phénotypes ?

Ce qui nous amène à tester si les deux échantillons proviennent de la même population.

On a $r = 2$ échantillons et $s = 4$ modalités pour le caractère phénotype.

$n_{i,j}$	A_1	A_2	A_3	A_4	$n_{i\bullet}$
e_1	10	30	50	10	100
e_2	60	120	180	40	400
$n_{\bullet j}$	70	150	230	50	500

$np_{i,j}$	A_1	A_2	A_3	A_4	$n_{i\bullet}$
e_1	14	30	46	10	100
e_2	56	120	184	40	400
$n_{\bullet j}$	70	150	230	50	500

$\frac{(n_{i,j}-np_{i,j})^2}{np_{i,j}}$	A_1	A_2	A_3	A_4
e_1	1,14	0	0,35	0
e_2	0,29	0	0,09	0

On obtient : $d = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{i,j} - np_{i,j})^2}{np_{i,j}} = 1,87.$

On sait que sous l'hypothèse H_0, D suit approximativement la loi de khi deux à $(r-1)(s-1) = 3$ degrés de liberté. On détermine le réel b_α tel que $P(D \geq b_\alpha) = \alpha$ (table 4) : pour $\alpha = 0,05$, on trouve $b_\alpha = 7,81$.

Comme $d < b_\alpha$, on ne peut rejeter l'hypothèse H_0 : les deux échantillons proviennent de la même population.

4. Exercices

Sauf mention explicite, les tests seront réalisés au risque 5%.

Exercice 1.

On a effectué le croisement de balsamines blanches avec des balsamines pourpres. En première génération les fleurs sont toutes pourpres. En deuxième génération, on obtient quatre catégories avec les effectifs suivants :

couleur	pourpre	rose	blanc lavande	blanc
effectif	1790	547	548	213

Peut-on accepter l'hypothèse de répartition mendélienne $\left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right)$?

Exercice 2.

Dans une usine de production d'un laboratoire pharmaceutique, on a dénombré pendant deux mois, soit 50 jours d'activité, le nombre de pannes quotidiennes. On a consigné les résultats dans le tableau suivant :

x_i	0	1	2	3	4 et +
n_i	21	18	7	3	1

où n_i est le nombre de jours où l'on a observé x_i pannes.

1) Calculer la moyenne et la variance de cette distribution.

2) Tester l'ajustement à cette distribution d'une loi de Poisson.

Exercice 3. *D'après partiel de novembre 2013*

Un chercheur s'est intéressé aux facteurs qui déterminent le choix des UE optionnelles par les étudiants. Il a posé la question suivante à un échantillon de 50 étudiants : « Parmi les quatre facteurs proposés, lequel est le plus important lorsque vous sélectionnez une UE optionnelle ? ». Les étudiants devaient choisir un des quatre facteurs suivants : l'intérêt pour le contenu du cours, le degré de complexité de l'examen, le professeur et l'horaire auquel ont lieu les enseignements. Voici les résultats que le chercheur a obtenu :

Facteur	Cours	Examen	Professeur	Horaire
Nombre d'étudiants	18	16	7	9

Effectuer un test statistique au risque 5% pour répondre à la question suivante : le chercheur peut-il conclure que les quatre facteurs ont la même influence dans le choix des étudiants ? Le résultat est-il le même au risque 10 % ? En cas de décisions contradictoires avec les deux risques 5 % et 10 %, préciser et justifier la décision à retenir.

Exercice 4. *D'après examen de janvier 2014*

Dans une agence de location de voitures, le directeur veut savoir quelles sont les voitures qui n'ont roulé qu'en ville pour les revendre immédiatement.

Pour cela, il y a dans chaque voiture une boîte noire qui enregistre le nombre d'heures pendant lesquelles la voiture est restée au point mort, au premier rapport, au deuxième rapport, ..., au cinquième rapport.

On sait qu'une voiture qui ne roule qu'en ville passe en moyenne 10% de son temps au point mort, 5% en première, 30% en seconde, 30% en troisième, 20% en quatrième et 5% en cinquième.

1) Sur une voiture, on a observé sur 2000 heures de conduite la répartition des rapports suivante :

Rapport	PM	1	2	3	4	5
Nombre d'heures	210	94	564	630	390	112

Effectuer un test statistique au risque 5% pour répondre à la question suivante : la voiture n'a-t-elle roulé qu'en ville ? Présenter le détail des calculs permettant d'effectuer ce test.

2) En utilisant le logiciel R, on a obtenu les résultats suivants :

```
> observed1 = c(210,94,564,630,390,112)
> chisq.test(observed1,p=c(0.10,0.05,0.30,0.30,0.20,0.05))

        Chi-squared test for given probabilities

data:  observed1
X-squared = 6.21, df = 5, p-value = 0.2863
```

Expliquer ce que réalisent les deux instructions saisies, indiquer ce que représentent les trois valeurs calculées et interpréter les résultats obtenus. Comparer avec les résultats du 1)b).

Exercice 5.

A la suite du même traitement, on a observé 40 bons résultats chez 70 malades jeunes et 50 bons résultats chez 100 malades agés.

Peut-on dire qu'il y a indépendance entre l'âge du malade et l'effet du traitement ?

Exercice 6. *D'après examen de janvier 2005*

En novembre 2004, beaucoup d'étudiants ont déclaré être stressé par les changements consécutifs à la mise en place du LMD. C'est pourquoi la Faculté leur a proposé de suivre un stage de relaxation proposant plusieurs méthodes différentes : méthode "be cool", méthode "be aware", méthode "be zen". A l'issue du stage, on leur a demandé comment ils se sentaient : moins, autant ou plus stressé qu'avant le stage. On a obtenu la répartition suivante des étudiants :

	moins	autant	plus
be cool	30	15	15
be aware	10	5	15
be zen	15	10	35

Effectuer un test statistique adéquat pour répondre à la question suivante : peut-on considérer que la méthode de relaxation choisie a une influence sur le niveau de stress après le stage ?