Cours 12 – Tests d'indépendance et d'adéquation du Khi-deux.

Eya ZOUGAR *

Institut National des Sciences appliqués-INSA

Génie mathématiques GM3
Thursday 20th April, 2023

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN NORMANDIE

*Basé sur le cours de Bruno PORTIER

1. Introduction.

Nous présentons dans ce cours 2 tests dits du Khi-deux :

- le test d'indépendance du Khi-deux qui permet de vérifier l'indépendance entre 2 variables qualitatives ;
- le test d'adéquation du Khi-deux qui permet de vérifier l'adéquation à une loi de probabilité discrète donnée.

Il s'agit de tests de type non-paramétrique construits à partir d'un TLC.

2. Test de conformité construit à partir d'un TLC.
2.1. Le cadre.

On considère n données réelles $x_1, x_2, \ldots, x_n$ qui sont les mesures d'une variable quantitative.

On fait l'hypothèse fondamentale que ces données sont en fait les réalisations de n variables aléatoires $X_1, X_2, \ldots, X_n$ que l'on suppose indépendantes et de même loi F.

On s'intéresse à une caractéristique de cette loi F (espérance, variance, etc ...) ou bien à un paramètre de cette loi, en supposant que cette loi soit paramétrée (loi de Bernouilli $\mathcal{B}(p)$, loi gaussienne $\mathcal{N}(\mu, \sigma^2)$, etc ...)

On note θ cette caractéristique ou ce paramètre. On supposera θ inconnu.

2.2 Le problème et le résultat clé.

On souhaite tester, au risque α, l'hypothèse nulle

$$H_0 : \ll \theta = \theta_0 \gg$$

contre l'hypothèse alternative

$$H_1 : \ll \theta \neq \theta_0 \gg$$

où θ_0 est donné a priori (valeur de référence ou supposée). En général, $\theta_0 = 0$.

2.2 Le problème et le résultat clé.

On souhaite tester, au risque α, l'hypothèse nulle

$$H_0 : \ll \theta = \theta_0 \gg$$

contre l'hypothèse alternative

$$H_1 : \ll \theta \neq \theta_0 \gg$$

où θ_0 est donné a priori (valeur de référence ou supposée). En général, $\theta_0 = 0$.

Clé

On se place dans la situation où l'on dispose d'un estimateur $T_n = T(X_1, \ldots, X_n)$ du paramètre θ et d'un TLC pour T_n de la forme :

$$V_n(T_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{L}$$

où $\mathcal{L}$ est une loi connue et tabulée, et V_n est une variable aléatoire positive convergeant (p.s.) vers l'infini.

Comment construire un test de H_0 contre H_1 ?

Notons que cette situation est très fréquente en Statistique.
C'est par exemple le cas des tests de conformité des paramètres lorsque ces paramètres sont estimés par l'estimateur du maximum de vraisemblance.

2.3 Choix de la statistique de test.

Il nous faut donc maintenant choisir une statistique de test.
On sait que pour tout $\theta \in \mathbb{R}$,

$$V_n(T_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{L}$$

Sous H_0, le paramètre θ est égal à θ_0.

2.3 Choix de la statistique de test.

Il nous faut donc maintenant choisir une statistique de test.
On sait que pour tout $\theta \in \mathbb{R}$,

$$V_n(T_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{L}$$

Sous H_0, le paramètre θ est égal à θ_0.

- Par conséquent, sous H_0, la variable

$$Z = V_n(T_n - \theta_0)$$

converge en loi vers $\mathcal{L}$.

2.3 Choix de la statistique de test.

Il nous faut donc maintenant choisir une statistique de test.
On sait que pour tout $\theta \in \mathbb{R}$,

$$V_n(T_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{L}$$

Sous H_0, le paramètre θ est égal à θ_0.

- Par conséquent, sous H_0, la variable

$$Z = V_n(T_n - \theta_0)$$

converge en loi vers $\mathcal{L}$.

- En revanche, sous H_1, Z ne converge plus vers $\mathcal{L}$, puisque l'on a la décomposition :

$$Z = \underbrace{V_n(T_n - \theta)}_{\xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{L}} + \underbrace{V_n(\theta - \theta_0)}_{\neq 0,\ \xrightarrow[n\to\infty]{p.s.} \pm\infty}$$

Donc si, pour n assez grand, on peut approximer la loi de $V_n(T_n - \theta)$ par la loi $\mathcal{L}$, alors on dispose d'une statistique de test admissible.

2.4 Construction du test.

Pour tester au risque α l'hypothèse nulle

$$H_0 : \ll \theta = \theta_0 \gg$$

contre l'hypothèse alternative

$$H_1 : \ll \theta \neq \theta_0 \gg$$

On utilise la statistique de test $Z = V_n(T_n - \theta_0)$, qui suit approximativement sous H_0 la loi $\mathcal{L}$.

2.4 Construction du test.

Pour tester au risque α l'hypothèse nulle

$$H_0 : \ll \theta = \theta_0 \gg$$

contre l'hypothèse alternative

$$H_1 : \ll \theta \neq \theta_0 \gg$$

On utilise la statistique de test $Z = V_n(T_n - \theta_0)$, qui suit approximativement sous H_0 la loi $\mathcal{L}$.

- **La zone de rejet** est de la forme

$$ZR = \{Z < t_{\alpha/2}\} \bigcup \{Z > t_{1-\alpha/2}\}$$

où $t_{\alpha/2}$ et $t_{1-\alpha/2}$ sont les quantiles d'ordre $\alpha/2$ et $(1-\alpha/2)$ de la loi $\mathcal{L}$.

2.4 Construction du test.

Pour tester au risque α l'hypothèse nulle

$$H_0 : \ll \theta = \theta_0 \gg$$

contre l'hypothèse alternative

$$H_1 : \ll \theta \neq \theta_0 \gg$$

On utilise la statistique de test $Z = V_n(T_n - \theta_0)$, qui suit approximativement sous H_0 la loi $\mathcal{L}$.

- **La zone de rejet** est de la forme

$$ZR = \{Z < t_{\alpha/2}\} \bigcup \{Z > t_{1-\alpha/2}\}$$

où $t_{\alpha/2}$ et $t_{1-\alpha/2}$ sont les quantiles d'ordre $\alpha/2$ et $(1-\alpha/2)$ de la loi $\mathcal{L}$.

- On calcule alors la valeur z_{obs} de Z sur les données.

Si $t_{\alpha/2} \leq z_{obs} \leq t_{1-\alpha/2}$, alors on ne peut pas rejeter l'hypothèse nulle H_0 et on considère que le paramètre θ n'est pas significativement différent de la valeur θ_0.
Sinon, on rejette l'hypothèse nulle H_0 au risque α, et on considère que le paramètre est significativement différent de θ_0.

2.5. Remarques.

- La loi limite du TLC est généralement la loi $\mathcal{N}(0,1)$. Mais, on verra dans la suite, que l'on a parfois des lois du chi-deux.
- Attention, il s'agit d'un test asymptotique, que l'on utilise à distance finie. Toute la question sera de savoir à partir de quelle taille d'échantillon l'approximation de la loi de Z par la loi limite du TLC est de bonne qualité.

2.5.1. Exercice: Etude dans le cas de la loi Géométrique.

On dispose de n mesures $x_1, x_2, \ldots, x_n$ d'une variable quantitative discrète prenant les valeurs 1, 2, 3, ….

On fait l'hypothèse que ces mesures sont en fait les réalisations de n variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes et de même loi géométrique de paramètre inconu p.

On s'intéresse à l'estimation de p. On rappelle que si X est une variable aléatoire de loi géométrique de paramètre p, alors pour tout $x \in \mathbb{N}^*$, on a :

$$\mathbb{P}[X = x] = p(1-p)^{x-1}$$

On pourra noter $q = 1 - p$.

On rappelle que $\mathbb{E}[X] = \dfrac{1}{p}$ et $\mathbb{V}ar[X] = \dfrac{q}{p^2}$. On note P_n l'estimateur de 1/p.

Clé

2.5.1. Exercice: Etude dans le cas de la loi Géométrique.

On dispose de n mesures $x_1, x_2, \ldots, x_n$ d'une variable quantitative discrète prenant les valeurs 1, 2, 3, ….

On fait l'hypothèse que ces mesures sont en fait les réalisations de n variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes et de même loi géométrique de paramètre inconu p.

On s'intéresse à l'estimation de p. On rappelle que si X est une variable aléatoire de loi géométrique de paramètre p, alors pour tout $x \in \mathbb{N}^*$, on a :

$$\mathbb{P}[X = x] = p(1-p)^{x-1}$$

On pourra noter $q = 1 - p$.

On rappelle que $\mathbb{E}[X] = \dfrac{1}{p}$ et $\mathbb{V}ar[X] = \dfrac{q}{p^2}$. On note P_n l'estimateur de 1/p.

Clé

$$\frac{\sqrt{n}(P_n - p)}{p\sqrt{1-p}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

2.5.1. Exercice: Etude dans le cas de la loi Géométrique.

On dispose de n mesures $x_1, x_2, \ldots, x_n$ d'une variable quantitative discrète prenant les valeurs 1, 2, 3, ….

On fait l'hypothèse que ces mesures sont en fait les réalisations de n variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes et de même loi géométrique de paramètre inconu p.

On s'intéresse à l'estimation de p. On rappelle que si X est une variable aléatoire de loi géométrique de paramètre p, alors pour tout $x \in \mathbb{N}^*$, on a :

$$\mathbb{P}[X = x] = p(1-p)^{x-1}$$

On pourra noter $q = 1 - p$.

On rappelle que $\mathbb{E}[X] = \dfrac{1}{p}$ et $\mathbb{V}ar[X] = \dfrac{q}{p^2}$. On note P_n l'estimateur de 1/p.

Clé

$$\frac{\sqrt{n}(P_n - p)}{p\sqrt{1-p}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

$$\frac{\sqrt{n}(P_n - p)}{P_n\sqrt{1-P_n}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

2.6. Etude par simulations dans le cas de la loi de Weibull.

2.6.1. Pourquoi faire des simulations.

Les simulations permettent, dans le cadre des tests statistiques asymptotiques, d'étudier

- le niveau empirique du test
- ainsi que la puissance empirique du test.

Cette étude permet de voir si le test est bien calibré et puissant.

Pour étudier le niveau empirique du test, on se place sous H_0, et on compare le pourcentage de rejet à tort, au niveau théorique que l'on a fixé.

Pour étudier la puissance empirique du test, on se place sous H_1, et on compare le pourcentage de rejet (bonnes décisions) à 100%.

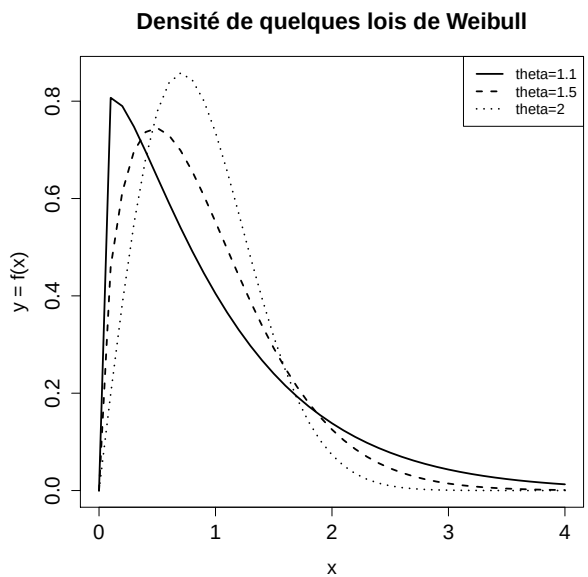
2. Loi de Weibull standard (fiabilité).
2.6.1. Densité.

On considère ici n données réelles positives $x_1, \ldots, x_n$. On suppose que ces données sont les réalisations de n variables aléatoires i.i.d de loi de Weibull standard.

La densité de la loi de Weibull standard est définie par:

$$f(x) = \theta\, x^{\theta-1} \exp\left(-x^{\theta}\right), \quad \forall x \geq 0$$

avec $\theta > 1$. Notons que $F(x) = 1 - \exp\left(-x^{\theta}\right)$ pour tout $x \geq 0$.



2.6.2. Etude dans le cas de la loi de Weibull standard.

On se place dans le cadre de la loi de Weibull standard de paramètre θ.

On souhaite pouvoir étudier par simulations le niveau et la puissance empirique d'un test de conformité du paramètre θ de la forme :

$$H_0 : "\theta = \theta_0" \quad \text{contre} \quad H_1 : "\theta \neq \theta_0"$$

où θ_0 est donné a priori.

Clé

L'estimateur du maximum de vraisemblance T_n n'a pas d'expression explicite, il est seulement défini par :

$$T_n = \arg\min_{z>0} \left(-n\log(z) - (z-1)\sum_{j=1}^{n} \log(X_j) + \sum_{j=1}^{n} X_j^z \right).$$

T_n est un estimateur asymptotiquement sans biais et convergent.
TLC:

$$\sqrt{g''(T_n)}(T_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

avec $g''(z) = \dfrac{n}{z^2} + \displaystyle\sum_{j=1}^{n} (\log(X_j))^2 X_j^z$.

2.6.2. Etude dans le cas de la loi de Weibull standard.

On considére comme statistique de test

$$Z = \sqrt{g''(T_n)}\,(T_n - \theta_0)$$

avec $g(z) = -n\log(z) - (z-1)\sum_{j=1}^{n}\log(x_j) + \sum_{j=1}^{n} x_j^z$.

On prendra $\theta_0 = 1,5$ et un risque de 5%.

On sait dans ce contexte que l'approximation de la loi de Z sous H_0 par une loi $\mathcal{N}(0,1)$.

2.6.3. Etude du niveau et de la puissance empirique.

Pour étudier le niveau empirique on simulera des échantillons d'une loi de Weibull de paramètres $\theta = 1,5$. Pour étudier la puissance empirique, on simulera des échantillons d'une loi de Weibull de paramètres θ =1,7, 2 et 2,3. Le tableau ci-dessous présente les résultats obtenus sur la base de 1000 échantillons de taille $n = 50, 100, 200, 500, 1000$ et 2000.

	Niveau empirique	Puissance empirique		
n	$\theta = 1,5$	$\theta = 1,7$	$\theta = 2$	$\theta = 2,3$
50	5,9	18,7	72,7	97,1
100	4,3	35	96,9	100
200	4,5	64	100	100
500	6,3	96,5	100	100
1000	5,6	100	100	100
2000	4,2	100	100	100

2.6.4. Commentaires.

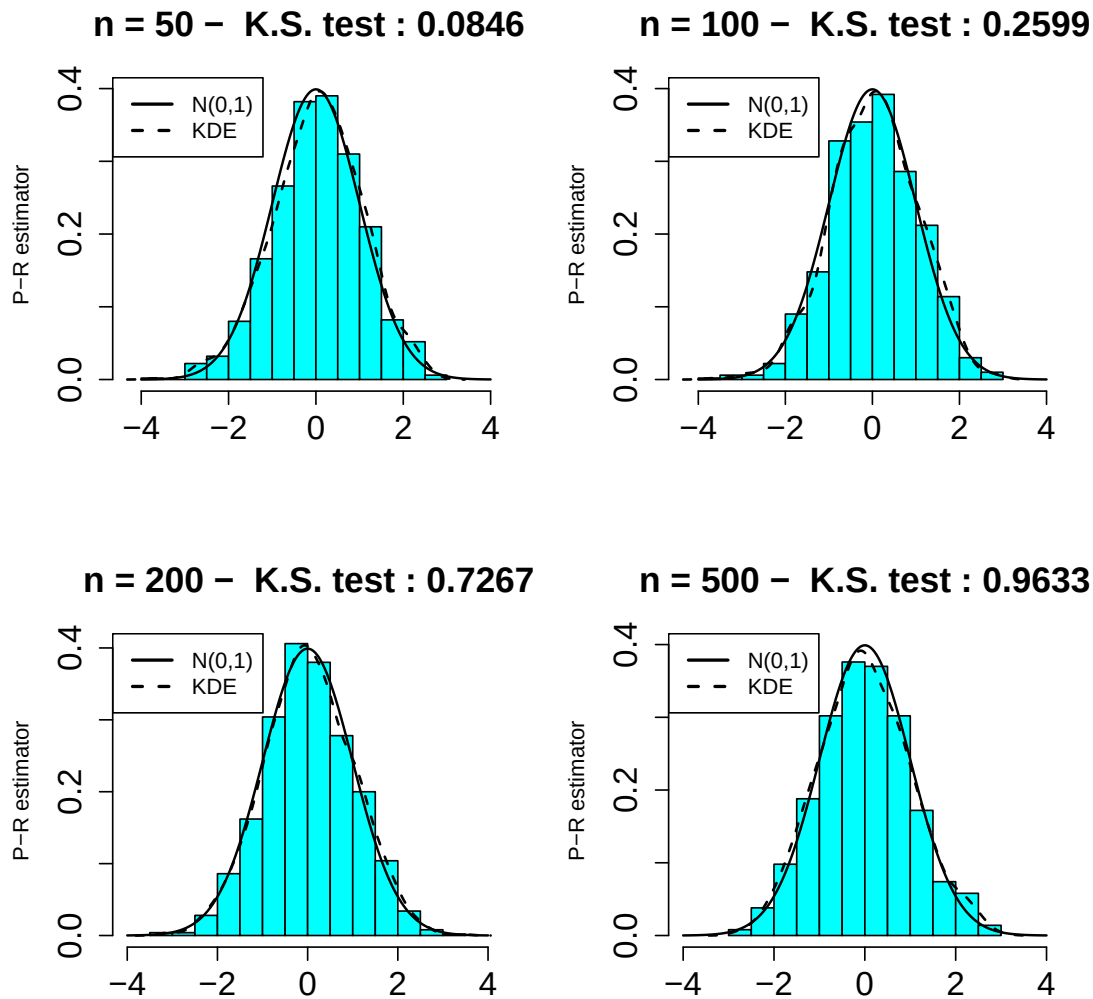
On constate que le niveau empirique est bon, proche du niveau théorique égal à 5%.

La puissance empirique augmente avec la taille de l'échantillon pour atteindre 100%.

Sans surprise, le test est moins puissant lorsque la valeur de θ est proche de $\theta_0 = 1,5$, moins puissant c'est à dire qu'on a plus de mal à rejeter H_0 à raison..

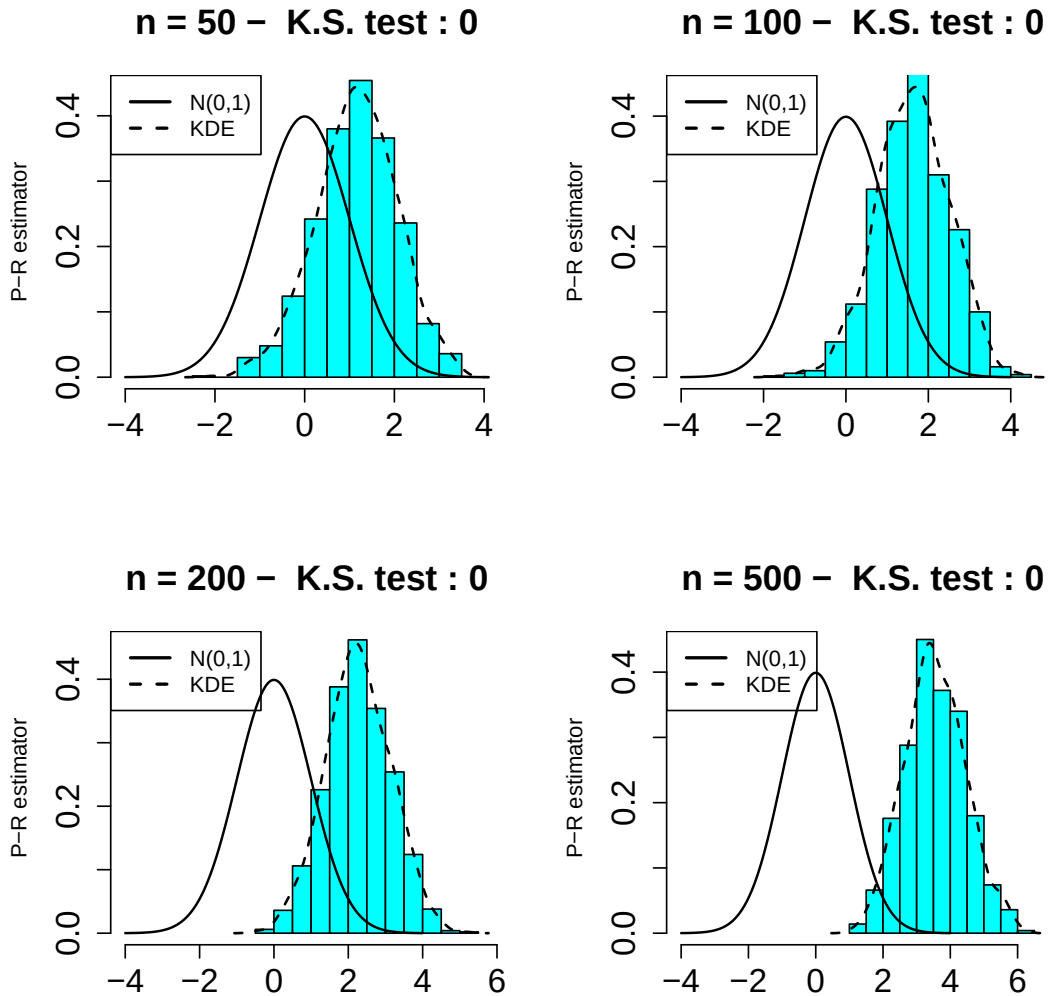
2.6.5. Visualisation du niveau empirique.

Le graphique ci-dessous permet de valider la loi de la statistique de test sous H_0.



2.6.6. Visualisation de la puissance empirique du test.

Le graphique ci-dessous permet d'illustrer le pouvoir séparateur du test sous H_1 ($\theta = 1,7$).



3. Test d'indépendance du Khi-deux.
3.1. Le problème.

On dispose de n données $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ qui sont les mesures d'un couple de variables à valeurs dans $\{a_1, a_2, \ldots, a_p\} \times \{b_1, b_2, \ldots, b_q\}$.

On fait l'hypothèse fondamentale que les données $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ sont les réalisations d'un couple (X, Y) de variables aléatoires discrètes, la variable X prenant ses valeurs dans $\{a_1, a_2, \ldots, a_p\}$, et la variable Y dans $\{b_1, b_2, \ldots, b_q\}$.

On ne connait pas la loi du couple (X, Y).

On souhaite tester le fait que ces deux variables sont indépendantes c'est à dire tester au risque $\alpha \in]0, 1[$ l'hypothèse nulle

$$H_0 : \ll X \text{ et } Y \text{ sont indépendantes} \gg$$

contre l'hypothèse alternative

$$H_1 : \ll X \text{ et } Y \text{ ne sont pas indépendantes} \gg$$

3.2. L'idée du test.

La construction du test repose sur la définition suivante:

Les variables X et Y sont indépendantes si et seulement si pour tout couple $(i,j) \in \{1,\dots,p\} \times \{1,\dots,q\}$,

$$\mathbb{P}\left[X = a_i \text{ et } Y = b_j\right] = \mathbb{P}\left[X = a_i\right] \times \mathbb{P}\left[Y = b_j\right].$$

Ainsi, pour savoir si les variables aléatoires X et Y sont indépendantes, il suffit de comparer des estimateurs des probabilités jointes au produit d'estimateurs des probabilités marginales.

3.3. Notations.

On associe à chaque réalisation (x_k, y_k) le couple de variables aléatoires (X_k, Y_k).

Les vecteurs $(X_1, Y_1), \ldots, (X_n, Y_n)$ sont indépendants et de même loi que (X, Y).

Pour tout entier $i = 1, \ldots, p$ et $j = 1, \ldots, q$, on note

- $O_{i,j} = \sum_{k=1}^{n} 1_{\{X_k = a_i, Y_k = b_j\}}$
- $O_{i,\bullet} = \sum_{j=1}^{q} O_{i,j}$ qui désigne le nombre de données pour lesquelles $X = a_i$;
- $O_{\bullet,j} = \sum_{i=1}^{p} O_{i,j}$, qui désigne le nombre de données pour lesquelles $Y = b_j$.

Enfin, on note

- $E_{i,j} = \dfrac{O_{i,\bullet} \times O_{\bullet,j}}{n}$.

3.4. Le résultat théorique.

Pour tout couple (i, j),

- $O_{i,j}/n$ est un estimateur de $\mathbb{P}[X = a_i, Y = b_j]$
- alors que $E_{i,j}/n$ est un estimateur de $\mathbb{P}[X = a_i] \times \mathbb{P}[Y = b_j]$.

Sous l'hypothèse d'indépendance, ces deux quantités doivent être proches.

Pour savoir si les variables sont indépendantes, on va évaluer la distance entre les valeurs observées $O_{i,j}$ et les valeurs attendues s'il y avait indépendance $E_{i,j}$, et en fonction de la valeur de cette distance, on décidera de rejeter ou non l'hypothèse d'indépendance.

Statistique de Test

On introduit la variable

$$Z = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

elle vérifie que sous H_0,

$$Z \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_{(p-1)(q-1)}$$

ce résultat étant faux sous H_1.

3.5. Construction du test.

Pour tester H_0 contre H_1 au risque α, on utilise la statistique de test

$$Z = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

qui, sous H_0, suit approximativement une loi du χ^2 à $(p-1)(q-1)$ degrés de liberté.

La zone de rejet du test est de la forme

$$ZR_{H_0} = \left\{ Z > k_{1-\alpha} \right\}$$

où $k_{1-\alpha}$ est le quantile d'ordre $(1-\alpha)$ d'une loi du χ^2 à $(p-1)(q-1)$ ddl.

On calcule ensuite la valeur z_{obs} de la statistique de test Z sur les données et on compare z_{obs} à $k_{1-\alpha}$:

- si $z_{obs} \leq k_{1-\alpha}$ (ou p-value $\geq \alpha$) alors on ne peut pas rejeter l'hypothèse H_0 selon laquelle les variables sont indépendantes ;
- sinon, on rejette H_0 au risque α.

3.6. Exemple : couleur des cheveux.
3.6.1. Le problème et les données.

On étudie l'influence du sexe sur la couleur des cheveux d'élèves d'un district écossais.

Nous souhaitons savoir si la couleur des cheveux est indépendante du sexe.

Pour cela, on dispose d'un échantillon de 3883 élèves et du tableau de données suivant :

Sexe	Blond	Roux	Châtain	Brun	Noir de jais	Total
Garçon	592	119	849	504	36	2100
Fille	544	97	677	451	14	1783
Total	1136	216	1526	955	50	3883

On désignera par X la variable "Couleur des Cheveux"
et par Y la variable "Sexe". On supposera que les données, notées $(x_k, y_k)_{1 \leq k \leq 3883}$, sont des réalisations indépendantes du couple (X, Y).

Pour savoir si la couleur des cheveux est indépendante du sexe, on testera l'indépendance des variables X et Y.

3.6.2. Mise en oeuvre du test d'indépendance.

Pour savoir si la couleur des cheveux est indépendante du sexe, on teste l'hypothèse nulle H_0 : $\ll X$ et Y sont indépendantes $\gg$ contre l'alternative H_1 : $\ll X$ et Y ne sont pas indépendantes $\gg$.

La statistique de test Z suit une loi du khi-deux à $(5-1)\times(2-1) = 4$ddl.

Le quantile d'ordre 0,95 d'une loi du khi-deux à 4 ddl est égale à : 9,49 et celui d'ordre 0,99 est égal à 13,28.
La zone de rejet du test à 5% est donc $\{Z > 9,49\}$, et à 1%, elle est $\{Z > 13,28\}$.

La valeur z_{obs} de la statistique de test Z sur les données vaut : $z_{obs} = 10,47$ (p-value = 0,033).

La p-value vaut ici :

$$\text{p-value} = \mathbb{P}_{H_0}\left[Z > 10,47\right] = \mathbb{P}\left[\chi^2_4 > 10,47\right] = 0,033$$

Le test est donc significatif à 5%, mais pas à 1%. On peut donc considérer que la couleur des cheveux et le sexe ne sont pas indépendants à 5%, mais le sont à 1%.

3.6.3. Illustration avec le logiciel R.

Avec le logiciel R, on utilise les instructions suivantes :

```
# construction de la table des données
Effectifs=matrix(c(592, 544, 119, 97, 849, 677,
    504, 451, 36, 14), ncol=5)
rownames(Effectifs)=c("Garçon","Fille")
colnames(Effectifs)=c("Blond", "Roux",
  "Chatain",  "Brun", "Noir de jais")
# Test d'indépendance du khi-deux
chisq.test(Effectifs)

^^IPearson's Chi-squared test

data:  Effectifs
X-squared = 10.467, df = 4, p-value = 0.03325

> qchisq(0.95,3)
[1] 7.814728
```

4. Test d'ajustement ou d'adéquation à une loi donnée.
4.1. Le problème.

On dispose de n données $x_1, x_2, \ldots, x_n$ qui sont les mesures d'une variable qualitatives à valeurs dans $\{a_1, a_2, \ldots, a_K\}$.

On fait l'hypothèse fondamentale que les données $x_1, x_2, \ldots, x_n$ sont en fait les réalisations d'une variable aléatoire discrète X à valeurs dans $\{a_1, a_2, \ldots, a_K\}$ de loi de probabilité P inconnue.

On postule cependant que la variable X est distribuée selon la loi P_0, loi donnée a priori ou loi de référence.

On souhaite vérifier ce postulat à l'aide d'un test statistique, au risque α c'est à dire tester l'hypothèse selon laquelle la loi de X est P_0.

Nous allons donc tester au risque $\alpha \in]0, 1[$ l'hypothèse nulle

$$H_0 : \ll \text{La loi de } X \text{ est } P_0 \gg \iff H_0 : \ll P = P_0 \gg$$

contre l'hypothèse alternative

$$H_1 : \ll \text{La loi de } X \text{ n'est pas } P_0 \gg \iff H_1 : \ll P \neq P_0 \gg$$

4.2. Notations.

Pour tout entier $j = 1, \ldots, n$, on associe à chaque réalisation x_j la variable aléatoire X_j.

Les variables aléatoires $X_1, X_2, \ldots, X_n$ sont indépendantes et de même loi P inconnue.

On note $p_{0,1}, p_{0,2}, \ldots, p_{0,K}$ les probabilités définissant la loi P_0, c'est à dire que si Y est une variable aléatoire de loi P_0, alors Y prend la valeur a_k avec probabilité $p_{0,k}$ pour tout entier $k = 1, \ldots, K$.

Enfin, pour tout entier $k = 1, 2, \ldots, K$, on note

- O_k l'effectif observé de la modalité a_k, c'est à dire que

$$O_k = \sum_{j=1}^{n} 1_{\{X_j = a_k\}}$$

- E_k l'effectif théorique attendu de la modalité a_k sous la loi P_0, c'est à dire que

$$E_k = n \times p_{0,k}$$

4.3. L'idée de la statistique de test du Khi-deux.

Pour tout entier $k = 1, 2, \ldots, K$, la variable O_k/n est un estimateur de $p_k = \mathbb{P}[X = a_k]$, probabilité de la modalité a_k.
Sous l'hypothèse nulle H_0 les quantités O_k/n et $p_{0,k}$ doivent coincider.

Par conséquent, il en est de même pour les quantités O_k et E_k.

Pour savoir si la loi inconnue de X est la loi théorique P_0, on va évaluer la distance globale entre les effectifs observés (O_k) et les effectifs théoriques attendus (E_k) si la loi de X était P_0,
et en fonction de la valeur de cette distance, on décidera de rejeter ou non l'hypothèse nulle.

La distance utilisée est celle du Khi-deux définie par :

$$Q^2 = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k}$$

On démontre que sous H_0, la statistique Q^2 suit approximativement une loi du Khi-deux à $(K-1)$ ddl.

4.4. Construction du test.

Pour tester H_0 contre H_1 au risque α, on utilise la statistique de test

$$Q^2 = \sum_{j=1}^{K} \frac{(O_j - E_j)^2}{E_j}$$

qui, sous H_0, suit approximativement une loi du Khi-deux à $(K-1)$ degrés de liberté.

La zone de rejet du test est de la forme $ZR_{H_0} = \left\{ Q^2 > k_{1-\alpha} \right\}$, où $k_{1-\alpha}$ est le quantile d'ordre $(1-\alpha)$ d'une loi du Khi-deux à $(K-1)$ ddl.

On calcule ensuite la valeur z_{obs} de la statistique de test Q^2 sur les données et on compare z_{obs} à $k_{1-\alpha}$:

- si $z_{obs} \leq k_{1-\alpha}$ (ou p-value $\geq \alpha$) alors on ne peut pas rejeter l'hypothèse H_0 selon laquelle la loi de X est P_0 ;
- sinon, on rejette H_0 au risque α et on considère que la loi de X est significativement différente de P_0.

4.5. Remarques.

Comme il s'agit d'un test asymptotique, il y a quelques conditions d'utilisation à respecter.

Des études par simulations, ont montré que :

- la taille de l'échantillon n doit être supérieure à 30 ;
- les effectifs théoriques doivent être tous supérieurs à 5.

4.6. Illustration pour valider une simulation.
4.6.1. Le problème et les données.

Pour illustrer le principe de simulation d'une loi discrète à nombre fini de valeurs (Section 2.3 du cours 2), nous avons simulé un échantillon de 200 réalisations d'une variable aléatoire discrète X de loi de probabilité

x	1	2	3	4	5
$\mathbb{P}[X = x]$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{20}$

Table: Loi de probabilité de la variable X.

Nous avons obtenu les résultats suivants :

x	1	2	3	4	5	Total
Effectifs	63	52	42	32	11	n=200

Table: Effectifs observés.

Comment valider cette simulation ?

Pour valider cette simulation, on peut mettre en oeuvre un test d'adéquation à une loi du khi-deux.

4.6.2. Mise en oeuvre du test d'adéquation.

On peut commencer par construire le tableau suivant :

k	1	2	3	4	5	Total
O_k	63	52	42	32	11	200
E_k	$200 \times \frac{1}{3} = 66{,}67$	$n \times \frac{1}{4}$=50	40	33,33	10	200
Dist. χ^2	0.202	0.080	0.100	0.053	0.100	z_{obs} =0,535

La statistique de test Q^2 aura pour loi sous H_0, une loi du khi-deux à 4 ddl.

La zone de rejet du test à 5% sera de la forme $\left\{Q^2 > 9,488\right\}$.

Puisque $z_{obs} = 0,535 < 9,488$, on ne peut pas rejeter l'hypothèse nulle H_0 au risque 5%, et on considère que l'échantillon simulé est bien distribué selon la loi de X.

4.6.3. Avec le logiciel R.

Avec le logiciel R, on utilise la fonction `chisq.test`.

Sur les données de l'exemple, on a :

```
rm(list=ls())
RNGkind(sample.kind = "Rounding")
set.seed(111)
prob = c(1/3, 1/4, 1/5, 1/6, 1/20)
P = cumsum(prob)
N = 200
# Simulation de l'échantillon
x=rep(NA, N)
for(i in 1:N) x[i] = (1:length(P))[runif(1)<P][1]
# Test d'adéquation du Chi-deux
qchisq(0.95,4) # Quantile chi-deux
[1] 9.488
chisq.test(table(x), p = prob)
    Chi-squared test for given probabilities
data:  table(x)
X-squared = 0.535, df = 4, p-value = 0.97
```