Statistique (MATH-F-315, Cours #5)

Thomas Verdebout

Université Libre de Bruxelles

Plan de la partie Statistique du cours

1. Introduction.
2. Théorie de l'estimation.
3. Tests d'hypothèses et intervalles de confiance.
4. Régression.
5. ANOVA.

Problèmes à deux échantillons : comparaison de deux moyennes

Nous considérons deux échantillons.

Echantillon 1 : $X_1, \ldots, X_{n_1}$ i.i.d. $E[X_i] = \mu_1$ et $\mathrm{Var}(X_i) = \sigma_1^2 < \infty$

$$\bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i \qquad s_X^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$$

Echantillon 2 : $Y_1, \ldots, Y_{n_2}$ i.i.d. $E[Y_i] = \mu_2$ et $\mathrm{Var}(Y_i) = \sigma_2^2 < \infty$.

$$\bar{Y} = \frac{1}{n_2} \sum_{j=1}^{n_2} Y_j \qquad s_Y^2 = \frac{1}{n_2} \sum_{j=1}^{n_2} (Y_j - \bar{Y})^2.$$

Un estimateur naturel de $\mu_2 - \mu_1$ différence $\bar{Y} - \bar{X}$ des moyennes empiriques.

Cet estimateur possède toutes les propriétés de l'estimateur $\bar{X}$: (i) non-biais, (ii) unique solution des équations de vraisemblance gaussiennes ...

Problèmes à deux échantillons : comparaison de deux moyennes

Cas 1 : Loi de $\bar{Y} - \bar{X}$, cas gaussien

Dans le cas gaussien $X_1, \ldots, X_{n_1}$ i.i.d. $\mathcal{N}(\mu_1, \sigma_1^2)$ et $Y_1, \ldots, Y_{n_2}$ i.i.d. $\mathcal{N}(\mu_2, \sigma_2^2)$:

$$\bar{Y} - \bar{X} \sim \mathcal{N}\left(\mu_2 - \mu_1, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Pas exploitable car les variances population σ_1^2 et σ_2^2 en général sont inconnues.

Hypothèse *(homogénéité des variances)* : $\sigma_1^2 = \sigma_2^2$ (on notera σ^2 leur valeur commune), alors on a

$$\bar{Y} - \bar{X} \sim \mathcal{N}\left(\mu_2 - \mu_1, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right).$$

Par ailleurs, on a

$$\frac{n_1 s_1^2}{\sigma^2} \sim \chi^2_{n_1-1} \quad \text{et} \quad \frac{n_2 s_2^2}{\sigma^2} \sim \chi^2_{n_2-1}$$

et donc (provenant de deux échantillons mutuellement indépendants, s_1^2 et s_2^2 sont indépendants)

$$\frac{n_1 s_1^2 + n_2 s_2^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}.$$

Problèmes à deux échantillons : comparaison de deux moyennes

Remarquons que

$$S^2 := \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

est un estimateur sans biais de σ^2. En effet,

$$\mathrm{E}\left[\frac{n_1 s_1^2 + n_2 s_2^2}{\sigma^2}\right] = n_1 + n_2 - 2 \quad \text{(moyenne d'une } \chi^2_{n_1+n_2-2}\text{)},$$

et donc

$$\mathrm{E}[S^2] = \mathrm{E}\left[\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\right] = \sigma^2.$$

En vertu de ce qui précède, on dispose donc d'une variable normale réduite

$$\bar{Y} - \bar{X} \sim \mathcal{N}\left(\mu_2 - \mu_1, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right) \Rightarrow \frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}(0,1)$$

et d'une variable chi-carré

$$\frac{n_1 s_1^2 + n_2 s_2^2}{\sigma^2} = \frac{(n_1 + n_2 - 2)S^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}.$$

Problèmes à deux échantillons : comparaison de deux moyennes

Indépendance entre les deux variables

- $\bar{X}$ est indépendant de s_1^2 en vertu du Lemme de Fisher
- $\bar{X}$ est indépendant de s_2^2 puisque calculé à partir de l'échantillon 2

On peut donc construire une variable de Student :

$$\frac{\dfrac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{S/\sigma} = \frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

Cas 2 : Loi de $\bar{Y} - \bar{X}$, "grands" échantillons

Lorsque n_1 et n_2 sont suffisamment grands, on a pour $\bar{X}$ et $\bar{Y}$ les lois approchées

$$\bar{X} \approx \mathcal{N}(\mu_1, \sigma_1^2/n_1) \quad \text{et} \quad \bar{Y} \approx \mathcal{N}(\mu_2, \sigma_2^2/n_2).$$

Donc,

$$\bar{Y} - \bar{X} \approx \mathcal{N}\left(\mu_2 - \mu_1, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

et, toujours pour n_1 et n_2 suffisamment grands,

$$\frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx \mathcal{N}(0, 1)$$

(approximation considérée comme raisonnable pour n_1 et $n_2 \geq 50$).

Intervalle de confiance

Les diverses lois ci-dessus permettent la construction d'intervalles de confiance pour $\mu_2 - \mu_1$

Cas 1 : Intervalle de confiance, échantillons gaussiens, variances égales

$$\left[\bar{Y} - \bar{X} \pm t_{n_1+n_2-2;1-\alpha/2} S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right].$$

Cas 2: Intervalle de confiance, "grands" échantillons

$$\left[\bar{Y} - \bar{X} \pm z_{1-\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right].$$

Problèmes à deux échantillons : comparaison de deux moyennes

Problème de test (unilatéral):

$$\begin{cases} H_0 : \mu_1 - \mu_2 \geq d_0 \\ H_1 : \mu_1 - \mu_2 < d_0, \end{cases}$$

Statistique de test (cas gaussien, variances égales) :

$$T = \frac{\bar{X} - \bar{Y} - d_0}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Loi sous $\mu_1 - \mu_2 = d_0$: $T \sim t_{n_1+n_2-2}$

Règle de comportement (problème unilatéral) :

$$RH_0 \quad \text{si} \quad T < t_{n_1+n_2-2;\alpha}$$

Problèmes à deux échantillons : comparaison de deux moyennes

Problème de test (bilatéral):

$$\begin{cases} H_0 : \mu_1 - \mu_2 = d_0 \\ H_1 : \mu_1 - \mu_2 \neq d_0, \end{cases}$$

Statistique de test (cas gaussien, variances égales) :

$$T = \frac{\bar{X} - \bar{Y} - d_0}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Loi sous $\mu_1 - \mu_2 = d_0$: $\quad T \sim t_{n_1+n_2-2}$

Règle de comportement (problème bilatéral) :

$$RH_0 \quad \text{si} \quad T \notin [\pm t_{n_1+n_2-2;1-\alpha/2}]$$

Echantillons appariés

L'hypothèse d'indépendance entre les deux échantillons est absolument cruciale, et il faut prendre garde à ne pas appliquer les procédures décrites ci-dessus à des échantillons n'y satisfaisant pas.

Même sous des hypothèses gaussiennes, on ne peut donc pas dire grand-chose de la loi de $\bar{Y} - \bar{X}$ quand on n'a pas l'indépendance. Il existe cependant des situations où il est raisonnable d'utiliser l'indépendance des différences
$(Y_1 - X_1), \ldots, (Y_n - X_n) =: d_1, \ldots, d_n$

Comparaison de deux proportions

Problèmes de comparaison entre deux populations.
L'expérience $\mathcal{E}$ se compose de deux schémas de Bernoulli indépendants.

$$X_1, \ldots, X_{n_1} \text{ i.i.d. } \mathrm{Bin}(1, p_1);\ p_1 \in (0,1)$$
$$Y_1, \ldots, Y_{n_2} \text{i.i.d. } \mathrm{Bin}(1, p_2);\ p_2 \in (0,1)$$

Un estimateur "naturel", sans biais, exhaustif, convergent, solution des équations de vraisemblance, etc. est donné par

$$\widehat{p_2 - p_1} := \hat{p}_2 - \hat{p}_1.$$

Quand les échantillons sont "grands",

$$\hat{p}_1 \approx \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n_1}\right) \quad \text{et} \quad \hat{p}_2 \approx \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

sont indépendants ; donc

$$\hat{p}_2 - \hat{p}_1 \approx \mathcal{N}\left(p_2 - p_1, \frac{p_2(1-p_2)}{n_2} + \frac{p_1(1-p_1)}{n_1}\right).$$

Comparaison de deux proportions

Un estimateur “naturel”, sans biais, exhaustif, convergent, solution des équations de vraisemblance, etc. est donné par

$$\widehat{p_2 - p_1} := \hat{p}_2 - \hat{p}_1.$$

Quand les échantillons sont “grands”,

$$\hat{p}_1 \approx \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n_1}\right) \quad \text{et} \quad \hat{p}_2 \approx \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

sont indépendants ; donc

$$\hat{p}_2 - \hat{p}_1 \approx \mathcal{N}\left(p_2 - p_1, \frac{p_2(1-p_2)}{n_2} + \frac{p_1(1-p_1)}{n_1}\right).$$

Comparaison de deux proportions

Soit la forme unilatérale du problème du test de l'égalité des paramètres p_1 et p_2 :

$$\begin{cases} H_0 & : \quad p_2 - p_1 \geq 0 \\ H_1 & : \quad p_2 - p_1 < 0. \end{cases}$$

Soit $\hat{p}$, un estimateur de la valeur commune, sous $p_2 = p_1$, de p_1 et p_2 :

$$\hat{p} := \frac{\sum_{i=1}^{n_1} X_i + \sum_{j=1}^{n_2} Y_j}{n_1 + n_2} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}.$$

Statistique de test : $\hat{p}_2 - \hat{p}_1$

Loi sous $p_2 = p_1$:

$$\hat{p}_2 - \hat{p}_1 \approx \mathcal{N}\left(0, \hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right),$$

Règle de comportement :

$$RH_0 \quad \text{si} \quad \hat{p}_2 - \hat{p}_1 < z_\alpha \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Les test chi-carré (chi-deux)

La loi multinomiale: Le schéma multinomial se rencontre lorsqu'une expérience aléatoire $\mathcal{E}$ donne lieu à I résultats possibles ($I = 2$ est le schéma de Bernoulli). Résultat 1 avec proba p_1, 2 avec proba p_2, etc. Considérons les variables aléatoires $n_1, \ldots, n_I$ comptant le nombre de fois que chacun des I résultats s'est présenté au cours de n répétitions de $\mathcal{E}$:

résultat 1	$n_1 \sim \mathrm{Bin}(n, p_1)$; donc $\mathrm{E}[n_1] = np_1$
résultat 2	$n_2 \sim \mathrm{Bin}(n, p_2)$; donc $\mathrm{E}[n_2] = np_2$
$\vdots$	$\vdots$
résultat i	$n_i \sim \mathrm{Bin}(n, p_i)$; donc $\mathrm{E}[n_i] = np_i$
$\vdots$	$\vdots$
résultat I	$n_I \sim \mathrm{Bin}(n, p_i)$; donc $\mathrm{E}[n_I] = np_I$
Total	n

Le vecteur aléatoire $\mathbf{n} := (n_1, \ldots, n_I)'$ est dit *vecteur multinomial* de paramètres $p_1, \ldots, p_I$ et d'exposant n, ce que l'on note

$$\mathbf{n} := (n_1, \ldots, n_I)' \sim \mathrm{Mult}(n; p_1, \ldots, p_I).$$

La vraisemblance multinomiale se calcule facilement :

Les test chi-carré (chi-deux)

Un résultat asymptotique

On peut montrer que, si $\mathbf{n} \sim \mathrm{Mult}(n; p_1, \ldots, p_l)$,

$$Q^{(n)} := \sum_{i=1}^{l} \frac{(n_i - np_i)^2}{np_i} \approx \chi^2_{l-1}.$$

En réalité, il s'agit d'un résultat asymptotique (convergence en loi). Mais l'approximation est considérée satisfaisante pourvu que $np_i \geq 5$ pour au moins 80% des valeurs de i, et $np_i \geq 1$ pour toutes.

Si $p_i = p_i(\boldsymbol{\theta})$, où $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$ est un paramètre de dimension k, que l'intérieur de Θ n'est pas vide, et que $\hat{\boldsymbol{\theta}} = \mathrm{Arg}\max_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{l} n_i \log p_i(\boldsymbol{\theta})$ (donc $\hat{\boldsymbol{\theta}}$ est l'estimateur maximum de vraisemblance de $\boldsymbol{\theta}$), alors on peut montrer que, si $n \sim \mathrm{Mult}(n; p_1(\boldsymbol{\theta}), \ldots, p_l(\boldsymbol{\theta}))$,

$$\hat{Q}^{(n)} := \sum_{i=1}^{l} \frac{(n_i - np_i(\hat{\boldsymbol{\theta}}))^2}{np_i(\hat{\boldsymbol{\theta}})} \approx \chi^2_{l-1-k} \ .$$

Les test chi-carré (chi-deux)

Le test chi-carré d'ajustement

Le test chi-carré d'ajustement est un test sur la valeur du paramètre $p_1, \ldots, p_l$ d'une multinomiale. Soit donc $\mathbf{n} \sim \mathrm{Mult}(n; p_1, \ldots, p_l)$ un vecteur observé, de loi multinomiale. Le problème de test est

$$\begin{cases} H_0 \ : \ p_1 = p_1^0, p_2 = p_2^0, \ldots, p_l = p_l^0 \\ H_1 \ : \ \text{il existe au moins un } i \text{ tel que } p_i \neq p_i^0. \end{cases}$$

Statistique de test :

$$Q^{(n)} := \sum_{i=1}^{l} \frac{(n_i - np_i^0)^2}{np_i^0}.$$

Loi sous H_0 : $Q^{(n)} \approx \chi^2_{l-1}$.

Règle de comportement :

$$RH_0 \text{ si } \quad Q^{(n)} > \chi^2_{l-1;1-\alpha}.$$

Ce test est une généralisation du test de l'hypothèse bilatérale (concernant une seul paramètre $p \in (0,1)$) $H_0 : p = p_0$ vu plus haut.

Le test chi-carré d'homogénéité

Généralisons au cas multinomial et à $J \geq 2$ échantillons le problème de comparaison de deux probabilités, et supposons avoir observé J multinomiales indépendantes

$$\begin{aligned}
\mathbf{n}_1 &\sim \mathrm{Mult}(n_1; p_{11}, \ldots, p_{I1}) \\
&\vdots \\
\mathbf{n}_j &\sim \mathrm{Mult}(n_j; p_{1j}, \ldots, p_{Ij}) \\
&\vdots \\
\mathbf{n}_J &\sim \mathrm{Mult}(n_J; p_{1J}, \ldots, p_{IJ}).
\end{aligned}$$

Les n_j, $j = 1, \ldots, J$ sont des constantes *fixées* par les conditions expérimentales. Au total, on dispose donc de $n = n_1 + \cdots + n_j + \cdots + n_J = \sum_{j=1}^{J} n_j$ réalisations d'expériences multinomiales.

Les test chi-carré (chi-deux)

Les observations se présentent sous forme d'un tableau $I \times J$ de fréquences observées, appelé *table de contingence* :

1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
I	n_{I1}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I\bullet}$
	n_1	$\cdots$	n_j	$\cdots$	n_J	n

Les test chi-carré (chi-deux)

A ce tableau de fréquences observées correspond le tableau des paramètres

1	p_{11}	$\cdots$	p_{1j}	$\cdots$	p_{1J}	
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
i	p_{i1}	$\cdots$	p_{ij}	$\cdots$	p_{iJ}	
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
I	p_{I1}	$\cdots$	p_{Ij}	$\cdots$	p_{IJ}	
	1	$\cdots$	1	$\cdots$	1	

des J multinomiales observées. L'hypothèse d'homogénéité est l'hypothèse sous laquelle les paramètres de ces J multinomiales coïncident :

$$\begin{cases} H_0 \ : p_{ij_1} = p_{ij_2} & \forall i = 1, \ldots, I \ ; \ j_1, j_2 = 1, \ldots, J \\ H_1 \ : \ \exists i, j_1, j_2 \ : \ p_{ij_1} \neq p_{ij_2}. \end{cases}$$

Les test chi-carré (chi-deux)

Procédure de test: Notons p_i la valeur commune (inconnue) de $p_{i1} \dots p_{iJ}$ sous H_0.
On a, en vertu du théorème d'addition des variables chi-carré,

$$Q^{(n)} := \sum_{j=1}^{J} \underbrace{\sum_{i=1}^{I} \frac{(n_{ij} - n_j p_i)^2}{n_j p_i}}_{\approx \chi^2_{I-1}, \text{ indépendantes}} \approx \chi^2_{J(I-1)}.$$

Les valeurs des p_i étant inconnues, il faut les remplacer par des estimateurs. Notons $\hat{p}_i := \frac{n_{i\bullet}}{n}$ leurs estimateurs maximum de vraisemblance. Les quantités $n_j p_i$ sont alors estimées par $n_j \hat{p}_i = \frac{n_{i\bullet} n_j}{n}$.

La statistique de test devient: $\hat{Q}^{(n)} := \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i\bullet} n_j}{n}\right)^2}{\frac{n_{i\bullet} n_j}{n}}$.

Sa loi (approchée) sous H_0 est: $Q^{(n)} \approx \chi^2_{(I-1)(J-1)}$.

En effet ($(I-1)$ est le nombre de paramètres estimés sous H_0),

$$J(I-1) - (I-1) = (I-1)(J-1).$$

La règle de comportement est donc : RH_0 si

$$\hat{Q}^{(n)} > \chi^2_{(I-1)(J-1);1-\alpha}.$$

Les test chi-carré (chi-deux)

Le test chi-carré d'indépendance

Le contexte expérimental pour ce test est entièrement différent. Les observations sont un échantillon bivarié discret ou discrétisé, de la forme

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \text{ i.i.d.,}$$

où la loi des (X_ν, Y_ν) est décrite par (loi bivariée discrète)

les valeurs possibles des X_ν : $x_1, \ldots, x_I$

les valeurs possibles des Y_ν : $y_1, \ldots, y_J$

les probabilités $p_{ij} := P[X_\nu = x_i \text{ et } Y_\nu = y_j]$.

La présentation de ces données se fait dans un tableau à double entrée : une *table de contingence* $I \times J$, où

$n_{ij} =$ fréquence observée de la valeur (x_i, y_j) au sein de l'échantillon $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} \cdots \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$.

Les test chi-carré (chi-deux)

A cette table de fréquences observées correspond le tableau des probabilités : ici, contrairement au tableau rencontré dans le problème d'homogénéité, la somme de *toutes* les probabilités p_{ij} vaut 1.

	y_1	$\cdots$	y_j	$\cdots$	y_J	
x_1	p_{11}	$\cdots$	p_{1j}	$\cdots$	p_{1J}	$p_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	p_{i1}	$\cdots$	p_{ij}	$\cdots$	p_{iJ}	$p_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_I	p_{I1}	$\cdots$	p_{Ij}	$\cdots$	p_{IJ}	$p_{I\bullet}$
	$p_{\bullet 1}$	$\cdots$	$p_{\bullet j}$	$\cdots$	$p_{\bullet J}$	1

Les test chi-carré (chi-deux)

Hypothèse d'indépendance

L'hypothèse nulle à laquelle nous allons nous intéresser est l'hypothèse d'indépendance entre les variables X et Y

$$\begin{cases} H_0 & : \text{ indépendance entre } X \text{ et } Y, \text{ i.e. } p_{ij} = p_{i\bullet} p_{\bullet j} \ \forall i, j \\ H_1 & : \exists\ i, j : p_{ij} \neq p_{i\bullet} p_{\bullet j} \end{cases}$$

Sous H_0,

$$Q^{(n)} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n p_{i\bullet} p_{\bullet j})^2}{n p_{i\bullet} p_{\bullet j}} \approx \chi^2_{IJ-1}$$

$((n_{11}, n_{21}, \ldots, n_{IJ})$ est de loi $\mathrm{Mult}(n; p_{11}, p_{21}, \ldots, p_{IJ})$ et, sous H_0, $p_{ij} = p_{i\bullet} p_{\bullet j})$

Mais les $p_{i\bullet}$ et $p_{\bullet j}$ sont inconnus. Estimer les $p_{i\bullet}$ et les $p_{\bullet j}$ à partir de leurs estimateurs maximum de vraisemblance $\hat{p}_{i\bullet} = \dfrac{n_{i\bullet}}{n}$ et $\hat{p}_{\bullet j} = \dfrac{n_{\bullet j}}{n}$.

Les test chi-carré (chi-deux)

La statistique de test est

$$\hat{Q}^{(n)} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}}.$$

La loi sous H_0 de cette statistique est

$$\hat{Q}^{(n)} \approx \chi^2_{(I-1)(J-1)};$$

car $(I-1)+(J-1)$ est le nombre de paramètres estimés sous H_0:

$$IJ - 1 - [(I-1)+(J-1)] = (I-1)(J-1).$$

Règle de comportement : on rejette l'hypothèse d'indépendance si

$$\hat{Q}^{(n)} > \chi^2_{(I-1)(J-1);1-\alpha}.$$