Chap 7 Tests du χ^2

I. Introduction

1) Motivation

RR

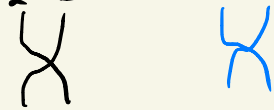
RB

BB

L'hypothèse : Dans un croisement,

père I II

mère I II



I avec $\mathbb{P}$ $\frac{1}{2}$ I : $\frac{1}{2}$

II avec $\mathbb{P}$ $\frac{1}{2}$ II : $\frac{1}{2}$

On veut tester cette hypothèse à partir de notre observation :

Génération 1 : RR BB

Génération 2 : RB

Génération 3 : RR RB BB

141 315 144 Total : 600

2) La loi multinomiale

Déf 1 : $n \geq 2$, $2 \leq k \leq n$, $p_1 + \dots + p_k = 1$. Un vecteur aléatoire $(N_1, \dots N_k)$ suit la loi multinomiale de paramètres n et $(p_1, \dots p_k)$ si

$\forall (n_1, \dots n_k) \in \mathbb{N}^k$,

$$\mathbb{P}(N_1 = n_1, \dots N_k = n_k) = \begin{cases} \frac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_n^{n_k} & n_1 + \dots + n_k = n \\ 0 & \text{sinon} \end{cases}$$

Rq : Si $k = 2$, $\mathbb{P}(N_1 = n_1, N_2 = n - n_1)$

$$= \frac{n!}{n_1!(n-n_1)!} p_1^{n_1} (1-p_1)^{n-n_1}$$

$$N_1 \sim B(n, p_1)$$

$$N_2 = n - N_1 \sim B(n, 1-p_1)$$

Rq : Si $n_1 + \dots + n_k = n$, alors

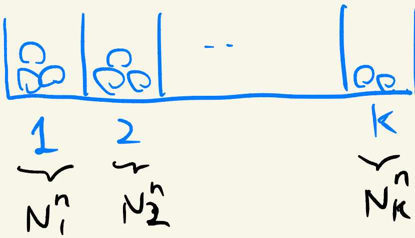
$$\frac{n!}{n_1! \dots n_k!} = \frac{n!}{n_1!(n-n_1)!} \times \frac{(n-n_1)!}{n_2!(n-n_1-n_2)!} \times \dots \times \frac{(n - \sum_{j=1}^{k-1} n_j)!}{n_k! \underbrace{(n - \sum_{j=1}^{k} n_j)!}_{0! = 1}}$$

$$= \binom{n}{n_1} \times \binom{n-n_1}{n_2} \times \dots \times \binom{n - \sum_{j=1}^{k-1} n_j}{n_k}$$

Rq : $(p_1 + \dots + p_k)^n = \sum_{\substack{n_1, \dots n_k \in \mathbb{N} \\ n_1 + \dots + n_k = n}} \frac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_n^{n_k}$

ex : Soit $(X_1, \dots X_n)$ un n-échantillon à valeurs dans $\{1, \dots k\}$ de loi $\mathbb{P}(X_1 = j) = p_j$, $1 \leq j \leq k$.

X_i



Bernoulli(p_j)

$N_j^n \sim B(n, p_j)$

On pose $N_j^n = \sum_{i=1}^n \mathbb{1}_{X_i = j}$, alors $(N_1^n, \dots N_k^n)$ suit la loi multinomiale de paramètres n et $(p_1, \dots p_k)$. En effet :

* $N_1^n + \dots + N_k^n = n$
* A une valeur $(n_1, \dots n_k)$ du k-uplet $(N_1^n, \dots N_k^n)$, il y a $\frac{n!}{n_1! \dots n_k!}$ valeurs de $(x_1, \dots x_n)$ correspondants pour le n-uplet $(X_1, \dots X_n)$. Chaque cas a une probabilité $p_1^{n_1} \dots p_k^{n_k}$.

Prop 2 : Si $(N_1, \dots N_k)$ suit la loi multinomiale de paramètres n et $(p_1, \dots p_k)$, alors

$\mathbb{E}[N_i] = np_i$, $Var(N_i) = np_i(1-p_i)$ $N_i \sim \mathcal{B}(n, p_i)$

pour $i \neq j$, $Cov(N_i, N_j) = -np_ip_j$

Dém : On pourra calculer ces quantités avec les N_j^n définies ci-dessus.

$N_j^n \sim \mathcal{B}(n, p_j)$,

donc $\mathbb{E}[N_j^n] = np_j$, $Var(N_j^n) = np_j(1-p_j)$

$i \neq j$: $Cov(N_i^n, N_j^n) = \mathbb{E}[N_i^n N_j^n] - \underbrace{\mathbb{E}[N_i^n]}_{np_i}\underbrace{\mathbb{E}[N_j^n]}_{np_j}$

$$\mathbb{E}[N_i^n N_j^n] = \sum_{\ell=1}^{n}\sum_{\ell'=1}^{n} \underbrace{\mathbb{E}[\mathbb{1}_{X_\ell = i}\mathbb{1}_{X_{\ell'} = j}]}_{= \begin{cases} 0 & \ell = \ell' \\ p_ip_j & \ell \neq \ell' \end{cases}}$$

$$= (n^2 - n)p_ip_j$$

D'où $Cov(N_i^n, N_j^n) = -np_ip_j$

3) Un résultat de convergence

$(X_1, \dots X_n)$ n-échantillon à valeurs dans $\{1, \dots k\}$. $p_j = \mathbb{P}(X_1 = j)$

$N_j^n = \sum_{i=1}^{n} \mathbb{1}_{X_i = j}$.

$\boxed{\text{Prop 3}}$ $$Z_n = {}^t\left(\frac{N_1^n - np_1}{\sqrt{np_1}}, \dots, \frac{\underbrace{N_k^n}_{\text{observé}} - \underbrace{np_k}_{\text{espéré}}}{\sqrt{np_k}} \right)$$

$$Z_n \xrightarrow[n\infty]{\mathcal{L}} \mathcal{N}_k(0, I_k - \Sigma_p)$$

où $\Sigma_p = \left(\sqrt{p_i p_j}\right)_{1 \le i,j \le k}$

Dém : On pose $Y_i = {}^t\left(\frac{\mathbb{1}_{X_i = 1} - p_1}{\sqrt{p_1}}, \dots, \frac{\mathbb{1}_{X_i = k} - p_k}{\sqrt{p_k}} \right)$

Ce sont des vecteurs aléatoires i.i.d et centrés.

Le TCL vectoriel entraîne que

$$Z_n = \underbrace{\sqrt{n}\left(\frac{Y_1 + \dots + Y_n}{n} - \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \right)}_{\frac{Y_1 + \dots Y_n}{\sqrt{n}}}$$

$$\xrightarrow[n\infty]{\mathcal{L}} \mathcal{N}_k(0, \Sigma_p)$$

avec $$(\Sigma_p)_{i,j} = \mathrm{Cov}(Y_1(i), Y_1(j))$$
$$= \mathbb{E}[Y_1(i)\, Y_1(j)]$$
$$= \frac{1}{\sqrt{p_i p_j}} \mathbb{E}\left[(\mathbb{1}_{X_1 = i} - p_i)(\mathbb{1}_{X_1 = j} - p_j) \right]$$

$$(\Sigma_p)_{i,j} = \frac{1}{\sqrt{p_i p_j}} \Big(\underbrace{\mathbb{E}[\mathbb{1}_{X_1=i, X_1=j}]}_{p_i \mathbb{1}_{i=j}} + p_i p_j - p_i \underbrace{\mathbb{E}[\mathbb{1}_{X_1=j}]}_{p_j} - p_j \underbrace{\mathbb{E}[\mathbb{1}_{X_1=i}]}_{p_i} \Big)$$

$$= \frac{1}{\sqrt{p_i p_j}} \left(p_i \mathbb{1}_{i=j} - p_i p_j \right)$$

$$= \begin{cases} -\sqrt{p_i p_j} & i \neq j \\ 1 - p_i & i = j \end{cases}$$

II. Test du χ^2 d'adjustement

On a un vecteur aléatoire de loi multinomiale de paramètres n et $\underbrace{p}_{\text{inconnu}} = (p_1, \dots p_k)$. On veut tester

$$H_0 : \text{“} p = p^0 \text{”} \quad \text{contre} \quad H_1 : \text{“} p \neq p^0 \text{”}.$$

ex : Fleurs $\qquad p^0 = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$

$\mathbb{P}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
	RR (1)	RB (2)	BB (3)

$n = 600$

$$\hat{p}_1 = \frac{141}{600}, \quad \hat{p}_2 = \frac{315}{600}, \quad \hat{p}_3 = \frac{144}{600}$$

$\boxed{\text{Th4}}$ $(X_1, \dots X_n)$ un n-échantillon à valeurs dans $\{1, \dots k\}$, $p_j = \mathbb{P}(X_1 = j)$. $N_j^n = \sum_{i=1}^{n} \mathbb{1}_{X_i = j}$.

On introduit

$$D_n^2 = \sum_{j=1}^{k} \frac{(N_j^n - n p_j^\circ)^2}{n p_j^\circ}$$

Alors $*$ si $p = p^\circ$, $D_n^2 \xrightarrow[n \infty]{\mathcal{L}} \chi^2(k-1)$

$*$ si $p \neq p^\circ$, $D_n^2 \xrightarrow[n \infty]{p.s.} +\infty$

Dem : $*$ $p = p^\circ$. $D_n^2 = \|Z_n\|^2$

avec $Z_n = {}^t\left(\frac{N_1^n - n p_1^\circ}{\sqrt{n p_1^\circ}}, \dots \frac{N_k^n - n p_k^\circ}{\sqrt{n p_k^\circ}} \right)$

$$Z_n \xrightarrow[n \infty]{\mathcal{L}} Y \sim \mathcal{N}_k(0, I_k - \Gamma_{p^\circ})$$

$$(\Gamma_{p^\circ})_{ij} = \sqrt{p_i^\circ p_j^\circ}$$

$\|\cdot\|^2$ est une fonction $\mathcal{C}^0$,

donc $\underbrace{\|Z_n\|^2}_{D_n^2 =} \xrightarrow[n \infty]{\mathcal{L}} \|Y\|^2$

$$\Gamma_{p^\circ} = \underbrace{\begin{pmatrix} \sqrt{p_1^\circ} \\ \vdots \\ \sqrt{p_k^\circ} \end{pmatrix}}_{\sqrt{p^\circ}} (\sqrt{p_1^\circ}, \dots \sqrt{p_k^\circ})$$

$V = \text{vect}(v_1, \dots v_p)$

$X = \text{mat}_{can}(v_1, \dots v_p)$.

$\Pi_V = X({}^tX X)^{-1}\, {}^tX$

Pourquoi $\Pi_{\text{vect}(\sqrt{p^\circ})} = \Sigma_{p^\circ}$? $\Pi_{\text{vect}(\sqrt{p^\circ})} = \sqrt{p^\circ}\,(\underbrace{{}^t\sqrt{p^\circ}\sqrt{p^\circ}}_{\|\sqrt{p^\circ}\|^2 = \sum_{i=1}^{n} p_i^\circ = 1})^{-1}\,{}^t\sqrt{p^\circ}$

Soit $v \in \mathbb{R}^k$,

$$\Pi_{\text{vect}(\sqrt{p^\circ})} v = \sqrt{p^\circ}\langle \sqrt{p^\circ}, v\rangle$$

$$= \underbrace{\sqrt{p^\circ}\,{}^t\sqrt{p^\circ}}_{\text{matrice de projection orthogonale.}}\, v$$

$$= \sqrt{p^\circ}\,{}^t\sqrt{p^\circ}$$

$$= \Sigma_{p^\circ}$$

$$I_k - \Sigma_p = \Pi_{\text{vect}(\sqrt{p^\circ})^\perp}$$

Par le théorème de Cochran, si $W \sim N_k(0, I_k)$,

alors $\Pi_{\text{vect}(\sqrt{p^\circ})^\perp} W \sim N_k(0, I_k - \Sigma_{p^\circ})$

et de plus $\|\Pi_{\text{vect}(\sqrt{p^\circ})^\perp} W\|^2 \sim \chi^2(\underbrace{k-1}_{\dim(\text{Vect}(\sqrt{p^\circ})^\perp)})$

Par conséquent, $\|Y\|^2 \sim \chi^2(k-1)$ et on a montré

$$D_n^2 \xrightarrow[n\infty]{\mathcal{L}} \chi^2(k-1)$$

* Si $p \neq p^\circ$, alors il existe j tel que $p_j \neq p_j^\circ$.

$$D_n^2 \geq \frac{(N_j^n - n p_j^\circ)^2}{n p_j^\circ} = n\,\frac{(N_j^n/n - p_j^\circ)^2}{p_j^\circ}$$

Par la loi des grands nombres,

$$\frac{N_j^n}{n} = \frac{\sum_{i=1}^{n} \mathbb{1}_{X_i=j}}{n} \xrightarrow[n\infty]{p.s.} \mathbb{E}[\mathbb{1}_{X_1=j}] = p_j \neq p_j^0$$

Donc $$\frac{(N_j^n/n - p_j^0)^2}{p_j^0} \xrightarrow[n\infty]{p.s.} \frac{(p_j - p_j^0)^2}{p_j^0} > 0$$

Donc $$D_n^2 \xrightarrow[n\infty]{p.s.} +\infty.$$

Cor 5: Soit $\alpha \in]0,1[$. Le test du χ^2 de H_0: "$p = p^0$" contre H_1: "$p \neq p^0$" de niveau α est donné par

$$\mathbb{1}_{D_n^2 > c_{1-\alpha}^{(K-1)}}$$

quantile d'ordre $1-\alpha$ de $\chi^2(K-1)$

ex : Fleurs

$$D_n^2 = \frac{(141 - 600\times\frac{1}{4})^2}{600\times\frac{1}{4}} + \frac{(315-600\times\frac{1}{2})^2}{600\times\frac{1}{2}} + \frac{(144-600\times\frac{1}{4})^2}{600\times\frac{1}{4}}$$

$$\simeq 1,53$$

$\alpha = 5\%$, $c_{1-\alpha}^{(2)} \simeq 5,99$, $\mathbb{1}_{D_n^2 > c_{1-\alpha}^{(2)}} = 0$,

donc on accepte H_0 : "$p = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$".

Rq : Puissance asymptotique (puissance $\pi : \begin{cases} \Theta_1 \to [0,1] \\ \theta \mapsto \mathbb{P}_\theta(\phi(X)=1) \end{cases}$)

Pour $p \neq p^0$, on a vu que

$$\liminf_{n\infty} \frac{D_n^2}{n} \geq c > 0$$

Donc pour tout $t>0$, $\underbrace{\mathbb{P}_p(D_n^2 > t)}_{\pi(p)} \xrightarrow[n\infty]{} 1$

III. Variantes

1) Test du χ^2 d'adjustement à une famille paramétrisée de loi

Dans certains cas pratiques, il se peut que on soit obligé d'estimer certains paramètres avant d'appliquer un test du χ^2. Dans ce cas on utilise le théorème suivant :

$\boxed{\text{Th 6}}$: $\Theta \subseteq \mathbb{R}^d$, $k > d+1$ et

$p : \theta \in \Theta \mapsto (p_1(\theta), \dots p_k(\theta)) \in \mathbb{R}^k$.

On suppose que p est injective et de classe $\mathcal{C}^2$, et les $p_j(\theta)$ ne s'annulent jamais sur Θ.

On suppose de plus que pour tout $\theta \in \Theta$,

les vecteurs $\begin{pmatrix} \frac{\partial}{\partial \theta_1} p_1(\theta) \\ \vdots \\ \frac{\partial}{\partial \theta_1} p_k(\theta) \end{pmatrix} \cdots \begin{pmatrix} \frac{\partial}{\partial \theta_d} p_1(\theta) \\ \vdots \\ \frac{\partial}{\partial \theta_d} p_k(\theta) \end{pmatrix}$

sont libres.

Soit $(X_1, \dots X_n)$ un n-échantillon de loi $p(\theta)$ avec $\theta \in \Theta$. On note $\hat{\theta}_n$ l'EMV de θ et $N_j^n = \sum_{i=1}^n \mathbb{1}_{X_i = j}$ pour $j \in \{1, \dots k\}$. Alors

$$D_n^2(\hat{\theta}_n) = \sum_{j=1}^k \frac{(N_j^n - n p_j(\hat{\theta}_n))^2}{n p_j(\hat{\theta}_n)} \xrightarrow[n\infty]{\mathcal{L}} \chi^2(k-1-d).$$

dans le Th4 $d=0$

<u>ex</u>: On a une VA Y à valeurs dans $\mathbb{N}$.

On veut savoir si Y suit une loi de Poisson.

H_0: " $Y \sim \mathcal{P}(\theta)$ pour un certain $\theta > 0$ " $\neq \left(\begin{matrix} \theta = \theta_0 \\ p = q^* \end{matrix} \text{ etc.} \right)$

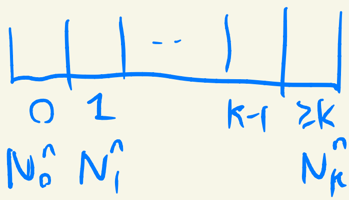
H_1: " $Y \nsim \mathcal{P}(\theta)$ pour tout $\theta > 0$ "

Comment passer à une loi multinomiale?

On construit $X_i = \begin{cases} Y_i & \text{si } 0 \le Y_i \le k-1 \\ k & \text{si } Y_i \ge k \end{cases}$.

On a des observations $X_1, \dots X_n$ à valeurs dans $\{0, 1, \dots k\}$, avec $p_j = \mathbb{P}(X_1 = j)$, $0 \le j \le k$.

$N_j^n = \sum_{i=1}^{n} \mathbb{1}_{X_i = j}$. Alors $(N_0^n, \dots N_k^n)$ suit une loi multinomiale de paramètres n et $(p_0, \dots p_k)$.



On considère $p(\theta) = (p_0(\theta), \dots p_k(\theta))$

$$p_j(\theta) = \begin{cases} \mathbb{P}(Y_1 = j) & 0 \le j \le k-1 \\ \mathbb{P}(Y_1 \ge k) & j = k \end{cases}$$

Par le théorème 6, on sait que $Y_1 \sim \mathcal{P}(\theta)$

$$\mathbb{P}_{H_0}\left(\mathbb{1}_{D_n^2(\hat{\theta}_n) > c_{1-\alpha}^{(k-1)}} = 1\right) \le \alpha$$

$k+1-1-\underbrace{d}_{1}$

Donc $\mathbb{1}_{D_n^2(\hat{\theta}_n) > c_{1-\alpha}^{(k-1)}}$ est un test de H_0 contre H_1 de niveau α.

Idée : on compte le nombre de contraintes indépendantes.

⚝ Si $a_1, \dots a_r$ sont des vecteurs de $\mathbb{R}^k$

linéairement indépendants et $Z \sim \mathcal{N}_k(0, I_k)$

$(\star)$ $\forall 1 \leq i \leq r, \langle a_i, Z \rangle = 0$

Conditionné à $(\star)$, $\|Z\|^2 \sim \chi^2(k-r)$

(conditionné à $(\star)$, $Z \overset{\mathcal{L}}{=} \Pi_{V^\perp} Z$

avec $V = \text{vect}(a_1, \dots a_r)$

on applique le th de Cochran)

ex : $D_n^2 = \sum_{j=1}^{n} \underbrace{\frac{(N_j^n - np_j^0)^2}{np_j^0}}_{Z_n(j)^2} = \|Z_n\|^2 \sim \chi^2(d-1)$

Contrainte : $\sum_{j=1}^{n} \underbrace{\sqrt{np_j^0}\, Z_n(j)}_{N_j^n - np_j^0} = 0$

Dans notre cas :

$D_n^2(\hat{\theta}_n) = \|Z_n\|^2$

avec $Z_n(j) = \dfrac{N_j^n - n p_j(\hat{\theta}_n)}{\sqrt{n p_j(\hat{\theta}_n)}}$

Quelles sont les contraintes ?

$\#1 \to \sum_{j=1}^{n} \sqrt{n p_j(\hat{\theta}_n)}\, Z_n(j) = 0$

$\#d \to$ EMV : vraisemblance $V_x(\theta) = \prod_{j=1}^{k} p_j(\theta)^{N_j^n}$

$$\ell_x(\theta) = \sum_{j=1}^{n} N_j^n \ln p_j(\theta)$$

$$\begin{cases} \partial_{\theta_1} \ell_x(\hat{\theta}_n) = 0 \\ \vdots \\ \partial_{\theta_d} \ell_x(\hat{\theta}_n) = 0 \end{cases} \to \sum_{j=1}^{n} N_j^n \frac{\partial_{\theta_1} p_j(\hat{\theta}_n)}{p_j(\hat{\theta}_n)} = 0$$

N_j^n ↓ relié à $Z_n(j)$

Nb de contraintes $= d+1$,

$$\|Z_n\|^2 \underset{n\infty}{\sim} \chi^2(k-1-d)$$

2) Test du χ^2 d'indépendance

$\boxed{\text{Th 7}}$ On observe un n-échantillon $((X_1, Y_1) .. (X_n, Y_n))$ avec X_i à valeurs dans $\{1, .. k\}$ et Y_i à valeurs dans $\{1, .. \ell\}$. Pour tout

$i \in \{1, .. k\}$ et $j \in \{1, .. \ell\}$.

On pose

$$N_{i,j} = \sum_{r=1}^{n} \mathbb{1}_{(X_r, Y_r) = (i,j)}$$

$$N_{i,\cdot} = \sum_{r=1}^{n} \mathbb{1}_{X_r = i}$$

$$N_{\cdot,j} = \sum_{r=1}^{n} \mathbb{1}_{Y_r = j}.$$

$i \backslash j$	1	..	ℓ	
1				
⋮		$N_{i,j}$		$N_{i,\cdot}$
k				
		$N_{\cdot,j}$		

Si les $(X_i)_{1 \leq i \leq n}$ sont indépendantes des $(Y_i)_{1 \leq i \leq n}$,

alors

$$D_n^2 = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{(N_{i,j} - N_{i,\cdot} N_{\cdot,j}/n)^2}{N_{i,\cdot} N_{\cdot,j}/n} \xrightarrow[n\infty]{\mathcal{L}} \chi^2((k-1)(\ell-1))$$

Csq : H_0 : " Les X_i sont indépendantes des Y_j "

contre H_1 : " les X_i ne sont pas indépendantes des Y_j "

$\mathbb{1}_{D_n^2 > c_{1-\alpha}^{((k-1)(\ell-1))}}$ est un test de H_0 contre H_1 de niveau α.

Dém: $p_i = \mathbb{P}(X_1 = i)$, $q_j = \mathbb{P}(Y_1 = j)$

$$\theta = \underbrace{(p_1, \dots p_{k-1}, q_1, \dots q_{\ell-1})}_{\dim\, = k+\ell-2} \qquad \begin{array}{l} p_k = 1 - p_1 - \dots - p_{k-1} \\ q_\ell = 1 - q_1 - \dots - q_{\ell-1} \end{array}$$

on veut tester

$$H_0 : \text{“} \forall i, j, \ \mathbb{P}(X_1 = i, Y_1 = j) = \underbrace{p_i q_j}_{p_{i,j}(\theta)} \text{”}$$

contre H_1 : “ —— ”

Il suffit d'appliquer le Th 6 :

$$\hat{\theta}_n = \left(\frac{N_{1,\cdot}}{n}, \dots, \frac{N_{k-1,\cdot}}{n}, \frac{N_{\cdot,1}}{n}, \dots \frac{N_{\cdot,\ell-1}}{n} \right)$$

$$p_{i,j}(\hat{\theta}_n) = \underbrace{\frac{N_{i,\cdot}}{n}}_{\hat{p}_i} \times \underbrace{\frac{N_{\cdot,j}}{n}}_{\hat{q}_j} \qquad \frac{N_{i,\cdot} \times N_{\cdot,j}}{n}$$

$$D_n^2(\hat{\theta}_n) = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{(N_{i,j} - n\, p_{i,j}(\hat{\theta}_n))^2}{n\, p_{i,j}(\hat{\theta}_n)}$$

$$\xrightarrow[n\infty]{\mathcal{L}} \chi^2(kl - 1 - (k+l-2))$$

$$= \chi^2((k-1)(l-1))$$

3) Test du χ^2 d'homogénéité

Cas particulier du χ^2 indépendance.

On observe un n-échantillon $(X_1, \dots X_n)$ et un m-échantillon $(Y_1, \dots Y_m)$. On suppose que les deux échantillons sont indépendants et à valeurs dans $\{1, \dots k\}$. On pose

$$N_i = \sum_{r=1}^{n} \mathbb{1}_{X_r = i}, \quad M_i = \sum_{r=1}^{m} \mathbb{1}_{Y_r = i}$$

$$\hat{p}_i = \frac{N_i + M_i}{n+m}.$$

Si les deux échantillons ont la même loi, alors

$$D_{n,m}^2 = \sum_{i=1}^{k} \left(\frac{(N_i - n\hat{p}_i)^2}{n\hat{p}_i} + \frac{(M_i - m\hat{p}_i)^2}{m\hat{p}_i} \right)$$

$$\underset{n,m\to\infty}{\simeq} \chi^2(k-1)$$

Csq : H_0 : " $\mathbb{P}_{X_1} = \mathbb{P}_{Y_1}$ " contre H_1 : " $\mathbb{P}_{X_1} \neq \mathbb{P}_{Y_1}$ ".

$\mathbb{1}_{D_{n,m}^2 > c_{1-\alpha}^{(k-1)}}$ est un test de H_0 contre H_1 de niveau α.

Idée : On considère le vecteur aléatoire (U,V) avec V à valeurs dans $\{0,1\}$.

Conditionné à $V=0$, $U \sim X$; conditionné à $V=1$, $U \sim Y$.

On prend un n_0 - échantillon $((U_1,V_1), \cdots (U_n,V_n))$

Supposons que $\mathbb{P}_{X_1} = \mathbb{P}_{Y_1}$, alors les U_i sont indépendantes des V_i.

$\longrightarrow$ Test d'indépendance.

$U \backslash V$	0	1	
1	N_1	M_1	$N_1 + M_1$
$\vdots$	$\vdots$	$\vdots$	
k	N_k	M_k	$N_k + M_k$
	n	m	

D^2_{n+m} (indépendance)

$$D^2_{n,m} = \sum_{i=1}^{k} \frac{\left(N_i - \frac{n(N_i+M_i)}{n+m}\right)^2}{\frac{n(N_i+M_i)}{n+m}} + \frac{\left(M_i - \frac{m(N_i+M_i)}{n+m}\right)^2}{\frac{m(N_i+M_i)}{n+m}}$$

$$\simeq \chi^2((k-1)(2-1)) = \chi^2(k-1)$$