MAP 311 - Aléatoire

Leçon 9

2016-2017

Exemple : réchauffement climatique

1) La température moyenne à Paris pour les mois de juillet du XXème siècle est modélisée par une loi normale $\mathcal{N}(20, 1{,}4^2)$.
2) Les températures observées de 2001 à 2015 au mois de juillet sont :

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
22	19	21	20	18	22	21	18	20	25	21	19	23	20	22

$\rightarrow$ La moyenne empirique est $\overline{x}_{15} = 20{,}73$ et la variance empirique non-biaisée est $v_{15} = 1{,}91^2$.
$\rightarrow$ On constate que $\overline{x}_{15} > 20$, ce qui indiquerait plutôt qu'il y a un réchauffement, mais on ne peut pas en être absolument sûr...

1. 1er choix : Annoncer qu'il y a réchauffement.
   Attention, il y a un risque : s'engager dans un protocole contraignant.
2. 2ème choix : Décider qu'il n'y a pas de réchauffement.
   Attention, il y a un risque : souffrir des conséquences du réchauffement dans le futur.

Le modèle statistique

Soit $\boldsymbol{X} = (X_1, \ldots, X_n)$ un n-échantillon du modèle statistique $\boldsymbol{\mathcal{P}} = \{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$.
Soit $H_0 \subset \Theta$ donné par le contexte.
On souhaite décider, au vu des observations, si on peut accepter l'hypothèse :

$$H_0 : \quad \boldsymbol{\theta} \in H_0$$

ou au contraire refuser cette hypothèse :

$$H_1 : \quad \boldsymbol{\theta} \notin H_0$$

Exemple : Le modèle statistique est :

$$\boldsymbol{\mathcal{P}} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$$

On veut tester l'hypothèse $H_0 = \{20\} \times \mathbb{R}^+$ (ou $H_0 =]-\infty, 20] \times \mathbb{R}^+$) à partir d'un échantillon de taille $n = 15$.

Règle de décision et région critique

- On observe l'échantillon et on souhaite tester si $\boldsymbol{\theta} \in H_0$ ou pas.
- Un test est déterminé par sa région critique W qui est un sous-ensemble de l'ensemble $\mathcal{X}^n$ des valeurs possibles de $\boldsymbol{X}$.
- La règle de décision du test associé à W est la suivante.

Lorsqu'on observe $\boldsymbol{x} = (x_1, \ldots, x_n)$,
- si $\boldsymbol{x} \in W$, alors on rejette H_0 et on accepte H_1 i.e. on décide que $\boldsymbol{\theta} \in H_1$,
- si $\boldsymbol{x} \notin W$, alors on accepte H_0 et on rejette H_1 i.e. on décide que $\boldsymbol{\theta} \in H_0$.

Exemple : on peut construire un test basé sur la région critique

$$W = \left\{ \boldsymbol{x} \in \mathbb{R}^n \,,\, \left| \overline{x}_n - 20 \right| > \epsilon \right\}$$

pour un certain ϵ. Cela veut dire qu'on rejette l'hypothèse H_0 si la température moyenne de ces dernières années s'écarte "trop" de 20 deg. Comment choisir ϵ ?

Test binaire

Exemple : On considère le modèle $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \{\mu_0, \mu_1\}\}$ avec $\sigma^2 > 0$ connu et $\mu_0 > \mu_1$.
On souhaite tester $H_0 = \{\mu = \mu_0\}$ contre $H_1 = \{\mu = \mu_1\}$.

• On va accepter H_0 (resp. H_1) si la moyenne empirique $\overline{X}_n$ est grande (resp. petite), c'est-à-dire choisir la région critique

$$W(a) = \{\boldsymbol{x} \in \mathbb{R}^n, \overline{x}_n < a\}$$

pour un certain a.

• Sous H_0, $\overline{X}_n$ suit la loi $\mathcal{N}(\mu_0, \sigma^2/n)$, donc le risque de première espèce est :

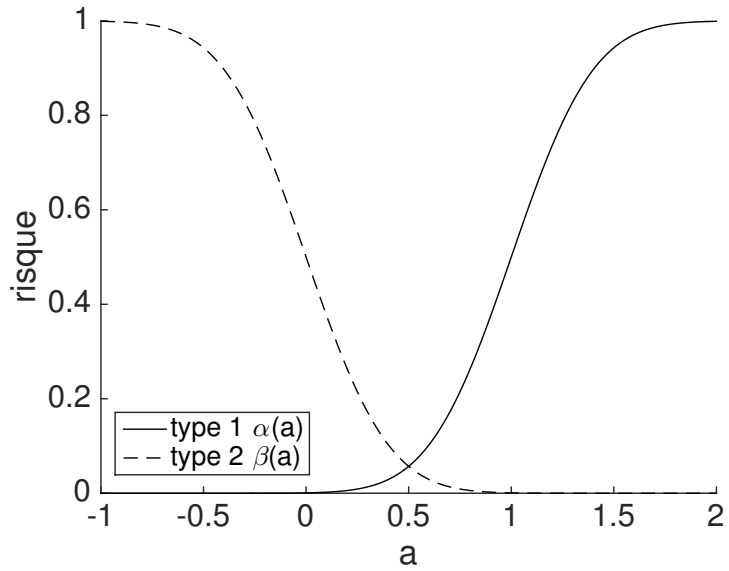
$$\alpha(a) = \mathbb{P}_{(\mu_0,\sigma^2)}(\boldsymbol{X} \in W(a)) = \mathbb{P}_{(\mu_0,\sigma^2)}(\overline{X}_n < a) = \Phi\left(\sqrt{n}\frac{a-\mu_0}{\sigma}\right),$$

avec Φ la fonction de répartition de la loi gaussienne centrée réduite.

• Le risque de seconde espèce est :

$$\beta(a) = \mathbb{P}_{(\mu_1,\sigma^2)}(\boldsymbol{X} \in W(a)^c) = \mathbb{P}_{(\mu_1,\sigma^2)}(\overline{X}_n \geq a) = \Phi\left(\sqrt{n}\frac{\mu_1-a}{\sigma}\right).$$

Il est clair que $a \to \alpha(a)$ est croissante et que $a \to \beta(a)$ est décroissante :



$\mu_0 = 1$, $\mu_1 = 0$, $\sigma = 1$, et $n = 10$.

Idéalement, on voudrait minimiser les risques de première et de deuxième espèce. Mais ceci n'est pas possible en même temps.

Par convention, on minimise en priorité le risque de première espèce.

En d'autres mots : H_0 est l'hypothèse qu'on ne veut surtout pas rejeter à tort.

Niveau d'un test

Definition

Le niveau d'un test défini par sa région critique W est le nombre

$$\alpha = \sup_{\boldsymbol{\theta} \in H_0} \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} \in W).$$

Si un test est de niveau α, alors on sait que, lorsqu'on rejette H_0, on a au plus une probabilité α de se tromper.
Parmi tous les tests de niveau inférieur à un seuil α fixé, on souhaite minimiser le risque de seconde espèce.

Test binaire

Exemple : On considère le modèle $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \{\mu_0, \mu_1\}\}$ avec $\sigma^2 > 0$ connu et $\mu_0 > \mu_1$.
On souhaite tester $H_0 = \{\mu = \mu_0\}$ contre $H_1 = \{\mu = \mu_1\}$.
On se donne $\alpha \in]0, 1[$ (risque de première espèce).

- Le test a pour région critique

$$W(a) = \{\boldsymbol{x} \in \mathbb{R}^n, \overline{x}_n < a\}$$

pour un certain a.
On calibre a de telle manière que

$$\alpha = \mathbb{P}_{(\mu_0,\sigma^2)}(\boldsymbol{X} \in W(a)) = \mathbb{P}_{(\mu_0,\sigma^2)}(\overline{X}_n < a) = \Phi\left(\sqrt{n}\frac{a-\mu_0}{\sigma}\right)$$

avec Φ la fonction de répartition de la loi gaussienne centrée réduite. Donc

$$a = \mu_0 + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(\alpha)$$

Tests pour le modèle gaussien

- On considère le modèle gaussien $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in]0, +\infty[\}$.

Soit $\mu_0 \in \mathbb{R}$. On souhaite tester $H_0 = \{\mu = \mu_0\}$ contre $H_1 = \{\mu \neq \mu_0\}$ au niveau $\alpha \in]0, 1[$.

- Rappels : 1) Si $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ et $V_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$, alors

$$\zeta_n = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sqrt{V_n}}$$

suit la loi de Student T_{n-1} sous H_0 (i.e. si $\mu = \mu_0$).
2) Sous H_1, par la loi forte des grands nombres, $\overline{X}_n - \mu_0 \overset{n\to\infty}{\longrightarrow} \mu - \mu_0 \neq 0$ et $V_n \overset{n\to\infty}{\longrightarrow} \sigma^2 > 0$ p.s.. Donc $|\zeta_n| \overset{n\to\infty}{\longrightarrow} +\infty$ p.s..

- On choisit donc $W_n = \{|\zeta_n| > a\}$ pour région critique.

On calibre le test en prenant $a = t_{1-\alpha/2}(n-1)$ le quantile d'ordre $1 - \alpha/2$ de la loi de Student T_{n-1}. Ainsi, pour tout $\sigma^2 > 0$:

$$\mathbb{P}_{(\mu_0,\sigma^2)}(W_n) = \mathbb{P}_{(\mu_0,\sigma^2)}(|\zeta_n| > t_{1-\alpha/2}(n-1)) = \alpha\,,$$

ce qui montre que le niveau du test est bien α.

• On peut reprendre le raisonnement précédent pour construire un test de niveau α pour $H_0 = \{\mu \leq \mu_0\}$ contre $H_1 = \{\mu > \mu_0\}$.
$\hookrightarrow$ La région critique est alors $W_n = \{\zeta_n \geq t_{1-\alpha}(n-1)\}$.

Exemple 2 (réchauffement climatique) :
On souhaite tester $H_0 = \{\mu \leq \mu_0\}$ (absence de réchauffement) contre $H_1 = \{\mu > \mu_0\}$ (existence d'un réchauffement).
$\rightarrow$ C'est le point de vue d'un *politicien*.
La région critique est $W_n = \{\zeta_n \geq t_{1-\alpha}(n-1)\}$ avec $\zeta_n = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sqrt{V_n}}$, $\mu_0 = 20$ et $n = 15$.
On observe $\zeta_{15}^{obs} = \sqrt{15}(20{,}73 - 20)/1{,}91 = 1{,}48$.
Comme $t_{0,95}(14) = 1{,}76$, on accepte H_0 au niveau $\alpha = 5\%$.
$\hookrightarrow$ On décide qu'il n'y a pas de réchauffement climatique.

Exemple 2 (réchauffement climatique) :
On souhaite tester $H_0 = \{\mu > \mu_0\}$ (existence d'un réchauffement) contre $H_1 = \{\mu \leq \mu_0\}$ (absence de réchauffement).
$\rightarrow$ C'est le point de vue d'un *écologiste*.
La région critique est $W = \{\zeta_n \leq t_\alpha(n-1)\}$ avec $\zeta_n = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sqrt{V_n}}$,
$\mu_0 = 20$ et $n = 15$.
On observe $\zeta_{15}^{obs} = \sqrt{15}(20{,}73 - 20)/1{,}91 = 1{,}48$.
Comme $t_{0,05}(14) = -1{,}76$, on accepte H_0 au niveau $\alpha = 5\%$.
$\hookrightarrow$ On décide qu'il y a réchauffement climatique.

Exemple 2 (réchauffement climatique) :
Supposons maintenant que les données de température aient la forme :

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
22	19	21	20	20	22	24	18	20	25	21	19	24	20	23

$\rightarrow \overline{x}_{15} = 21{,}20$ et $v_{15} = 2{,}08^2$.
On souhaite tester $H_0 = \{\mu \leq 20\}$ (absence de réchauffement) contre $H_1 = \{\mu > 20\}$ (existence d'un réchauffement).

On observe $\zeta_{15}^{obs} = \sqrt{15}(21{,}20 - 20)/2{,}08 = 2{,}25$.
Comme $t_{0,95}(14) = 1{,}76$, on rejette H_0 au niveau $\alpha = 5\%$.
$\hookrightarrow$ On décide qu'il y a réchauffement climatique.

Mais si on impose le niveau $\alpha = 1\%$, alors on accepte H_0, car $t_{0,99}(14) = 2{,}62$.
$\hookrightarrow$ On décide qu'il n'y a pas de réchauffement climatique.

Comme le montre l'exemple 2 :

- Quand les données apportent peu d'information, on va toujours accepter H_0, afin d'éviter de commettre une erreur de première espèce (rejeter à tort H_0).
- Quand les données sont informatives, mais qu'on impose un niveau α très proche de 0, on va aussi toujours accepter H_0, car c'est la seule manière d'être quasi-certain de ne pas commettre une erreur de première espèce.

$\hookrightarrow$ Le choix de l'hypothèse nulle est fondamental !

Test du χ^2 (test du chi-deux) d'adéquation à une loi

Exemple : "Un dé à six faces est-il pipé ?"
On observe les fréquences d'apparition des faces lors de n lancers de dé et on les compare au vecteur $(1/6, \ldots, 1/6)$.
Si on constate qu'on s'éloigne significativement de ce vecteur, on peut rejeter l'hypothèse que le dé est équilibré et annoncer que le dé est pipé.
La question est de savoir ce qu'on entend par "significativement".

Test du χ^2 (test du chi-deux) d'adéquation à une loi

On observe un n-échantillon $\boldsymbol{X} = (X_1, \ldots, X_n)$ de variables aléatoires i.i.d. à valeurs dans un espace fini $\mathcal{X} = \{a_1, \ldots, a_k\}$.
La loi est paramétrée par $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ avec $\mathbb{P}_{\boldsymbol{\theta}}(X_1 = a_j) = \theta_j$ pour $j \in \{1, \ldots, k\}$. Le paramètre $\boldsymbol{\theta}$ vit dans l'ensemble

$$\Theta = \left\{ \boldsymbol{\theta} \in [0,1]^k, \sum_{i=1}^{k} \theta_i = 1 \right\}$$

Pour $\boldsymbol{\theta}^{(0)} \in \Theta$ fixé, on souhaite tester $H_0 = \{\boldsymbol{\theta}^{(0)}\}$ contre $H_1 = \Theta \backslash \{\boldsymbol{\theta}^{(0)}\}$.

Exemple : Dans le cas du dé à six faces, $\mathcal{X} = \{1, \ldots, 6\}$ et $\boldsymbol{\theta}^{(0)} = (1/6, \ldots, 1/6)$.

Proposition

1) L'EMV de $\boldsymbol{\theta}$ est donné par le vecteur des fréquences empiriques d'apparition des différentes valeurs possibles :

$$\hat{\boldsymbol{\theta}}_n = \Big(\frac{N_1(\boldsymbol{X})}{n}, \dots, \frac{N_k(\boldsymbol{X})}{n}\Big), \qquad N_i(\boldsymbol{x}) = \sum_{j=1}^{n} \mathbf{1}_{a_i}(x_j).$$

2) L'EMV est non-biaisé, convergent, asymptotiquement normal : Sous $\mathbb{P}_{\boldsymbol{\theta}}$,

$$\sqrt{n}\big(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\big) \overset{n\to+\infty}{\longrightarrow} \mathcal{N}(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta})),$$

en loi, avec

$$C(\boldsymbol{\theta})_{jj'} = \begin{cases} \theta_j - \theta_j^2 & \text{si } j = j', \\ -\theta_j \theta_{j'} & \text{si } j \neq j'. \end{cases}$$

Preuve : 1) vraisemblance $p_n(\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{j=1}^{n} \prod_{i=1}^{k} \theta_i^{\mathbf{1}_{a_i}(x_j)} = \prod_{i=1}^{k} \theta_i^{N_i(\boldsymbol{x})}$.
EMV obtenu par maximisation de $\log p_n(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{k} N_i(\boldsymbol{x}) \ln(\theta_i)$.
2) TCL vectoriel.

Idée à la base du test : le vecteur $\hat{\boldsymbol{\theta}}_n$ est censé être plus proche de $\boldsymbol{\theta}^{(0)}$ sous H_0 que sous H_1.
Afin de quantifier la "proximité", on utilise la pseudo-distance du χ^2 :

$$\zeta_n = n \sum_{j=1}^{k} \frac{(\hat{\theta}_{n,j} - \theta_j^{(0)})^2}{\theta_j^{(0)}} .$$

Proposition

1) Sous H_0, ζ_n converge en loi quand $n \to +\infty$ vers une variable aléatoire Z de loi χ^2 à $k-1$ degrés de liberté.
2) Sous H_1, ζ_n tend presque sûrement vers $+\infty$.

Note : $k-1 =$ "dimension" de Θ.
En pratique, on considère que l'approximation en loi sous H_0 par χ^2_{k-1} est valide si $n \min_{j=1,\dots,k} \theta_j^{(0)} \geq 5$.
$\hookrightarrow$ Test de région critique $W_n = \{\zeta_n > x_{1-\alpha}(k-1)\}$ où $x_{1-\alpha}(k-1)$ est le quantile d'ordre $1-\alpha$ de la loi χ^2_{k-1}.

Exemple : On considère un dé à 6 faces. On souhaite tester au niveau 5% si le dé n'est pas pipé. Ici $\mathcal{X} = \{1, \dots, 6\}$ et $H_0 = \{(1/6, \dots, 1/6)\}$.

- Lors de $n = 100$ lancers du dé on observe les résultats suivants :

$$N_1 = 20,\ N_2 = 13,\ N_3 = 17,\ N_4 = 12,\ N_5 = 23,\ N_6 = 15.$$

On évalue

$$\zeta_{100}^{obs} = n \sum_{j=1}^{k} \frac{(\hat{\theta}_{n,j} - \theta_j^{(0)})^2}{\theta_j^{(0)}} = 100 \sum_{j=1}^{6} \frac{(N_j/100 - 1/6)^2}{1/6} \simeq 5{,}36$$

Sous l'hypothèse H_0 ("le dé n'est pas pipé"), ζ_{100} suit une loi de χ^2_5. On a $x_{0,95}(5) = 11{,}07$. Comme $\zeta_{100}^{obs} < 11{,}07$, on accepte, au niveau 5%, l'hypothèse H_0 que le dé n'est pas pipé.

Exemple : On considère un dé à 6 faces. On souhaite tester au niveau 5% si le dé n'est pas pipé. Ici $\mathcal{X} = \{1, \dots, 6\}$ et $H_0 = \{(1/6, \dots, 1/6)\}$.

• Lors de $n = 1000$ lancers du dé on observe les résultats suivants :

$$N_1 = 200,\ N_2 = 130,\ N_3 = 170,\ N_4 = 120,\ N_5 = 230,\ N_6 = 150.$$

On évalue

$$\zeta_{1000}^{obs} = n \sum_{j=1}^{k} \frac{(\hat{\theta}_{n,j} - \theta_j^{(0)})^2}{\theta_j^{(0)}} = 1000 \sum_{j=1}^{6} \frac{(N_j/1000 - 1/6)^2}{1/6} \simeq 53{,}6$$

Comme $\zeta_{1000}^{obs} > 11{,}07$, on rejette, au niveau 5%, l'hypothèse H_0.

• Dans le premier cas où $n = 100$, les proportions observées étaient les mêmes que dans le second cas où $n = 1000$, mais on a cependant accepté l'hypothèse que le dé n'était pas pipé car on n'avait pas assez de données pour rejeter avec suffisamment d'assurance cette hypothèse.

Exemple : Dans sa célèbre expérience, Mendel croise des pois jaunes lisses (AB) et des pois verts ridés (ab) et observe la répartition des 4 phénotypes à la deuxième génération (AB, Ab, aB, ab). Selon la théorie de Mendel, la distribution des phénotypes suit la loi $\boldsymbol{\theta}_0 = (9/16, 3/16, 3/16, 1/16)$.

• L'expérience de Mendel porte sur 556 observations et la répartition qu'il obtient est $(315, 101, 108, 32)$. On teste $H_0 = \{\boldsymbol{\theta}_0\}$ contre $H_1 = \{\boldsymbol{\theta}_0\}^c$. On calcule

$$\zeta_n^{obs} = n \sum_{j=1}^{4} \frac{(\hat{\theta}_{n,j} - \theta_{0,j})^2}{\theta_{0,j}}$$

$$= 556 \left(\frac{(\frac{315}{556} - \frac{9}{16})^2}{\frac{9}{16}} + \frac{(\frac{101}{556} - \frac{3}{16})^2}{\frac{3}{16}} + \frac{(\frac{108}{556} - \frac{3}{16})^2}{\frac{3}{16}} + \frac{(\frac{32}{556} - \frac{1}{16})^2}{\frac{1}{16}} \right) = 0{,}47$$

Au niveau $\alpha = 5\%$, une région critique est $W_n = \{\zeta_n \in [0{,}22, 9{,}35]^c\}$ où $[0{,}22, 9{,}35]$ est un intervalle de confiance d'une loi χ^2_3 au niveau 0,95. Comme $\zeta_n^{obs} \in [0{,}22, 9{,}35]$, on accepte H_0 : on décide que la distribution des phénotypes est bien $\boldsymbol{\theta}_0$.

Test d'adéquation à une famille de lois

• Peut-on répondre à une question du type : les observations sont-elles géométriques, gaussiennes, etc ?
Il ne s'agit plus de tester l'adéquation d'observations à une loi donnée, mais à une famille de lois.

• On observe un n-échantillon $\boldsymbol{X} = (X_1, \ldots, X_n)$ de variables aléatoires i.i.d. à valeurs dans un espace fini $\mathcal{X} = \{a_1, \ldots, a_k\}$.
La loi est paramétrée par $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ avec $\mathbb{P}_{\boldsymbol{\theta}}(X_1 = a_j) = \theta_j$ pour $j \in \{1, \ldots, k\}$. Le paramètre $\boldsymbol{\theta}$ vit dans l'ensemble

$$\Theta = \left\{ \boldsymbol{\theta} \in [0,1]^k, \sum_{i=1}^{k} \theta_i = 1 \right\}$$

• On souhaite tester l'hypothèse nulle H_0 contre $H_1 = \Theta \backslash H_0$ pour un certain sous-ensemble $H_0 \subset \Theta$.

Exemple : On considère $n = 200$ rouleaux d'un jeu de grattage contenant 100 tickets chacun. Pour $k = 1, \dots, n$, on observe X_k le nombre de tickets gagnants dans le $k$ème rouleau.
On veut tester l'hypothèse $H_0 =$"la loi du nombre de tickets gagnants par rouleau est une binomiale" (en d'autres mots, on veut vérifier si les tickets gagnants sont bien uniformément répartis dans les rouleaux, sans présumer de la proportion de tickets gagnants).
Ici

$$\mathcal{X} = \{0, \dots, 100\}$$

$$\Theta = \left\{ \boldsymbol{\theta} \in [0,1]^{101}, \sum_{j=0}^{100} \theta_j = 1 \right\}$$

et

$$H_0 = \left\{ \boldsymbol{\theta} \in \Theta, \exists p \in [0,1], \theta_j = \binom{100}{j} p^j (1-p)^{100-j} \, \forall j = 0, \dots, 100 \right\}$$

• On paramétrise $H_0 = \{\boldsymbol{\theta}^{\boldsymbol{\pi}}, \boldsymbol{\pi} \in \boldsymbol{\Pi}\}$, où
- $\boldsymbol{\Pi}$ est une partie d'intérieur non vide de $\mathbb{R}^h$ avec $h < k-1$,
- $\boldsymbol{\pi} \mapsto \boldsymbol{\theta}^{\boldsymbol{\pi}}$ est une application de $\boldsymbol{\Pi}$ dans Θ.
• Idée : utiliser un estimateur $\hat{\boldsymbol{\pi}}_n$ de $\boldsymbol{\pi}$ à valeurs dans $\boldsymbol{\Pi}$ (très souvent, ce sera l'EMV de $\boldsymbol{\pi}$) et comparer les vecteurs $\hat{\boldsymbol{\theta}}_n$ et $\boldsymbol{\theta}^{\hat{\boldsymbol{\pi}}_n}$.
Si ces vecteurs sont suffisamment proches, on pourra accepter H_0.
La question est de savoir ce que veut dire "suffisamment proches".

Proposition

Soit

$$\zeta_n = n \sum_{j=1}^{k} \frac{(\hat{\theta}_{n,j} - \theta_j^{\hat{\boldsymbol{\pi}}_n})^2}{\theta_j^{\hat{\boldsymbol{\pi}}_n}}.$$

Sous des hypothèses de régularité non-précisées (vérifiées en général lorsque $\hat{\boldsymbol{\theta}}_n$ est l'EMV de $\boldsymbol{\theta}$ et $\hat{\boldsymbol{\pi}}_n$ est l'EMV de $\boldsymbol{\pi}$) :
- *Sous H_0, ζ_n converge en loi vers $Z \sim \chi^2_{k-h-1}$.*
- *Sous H_1, ζ_n tend presque sûrement vers $+\infty$.*

Note : $k - h - 1 =$ "dimension" de $\Theta -$ "dimension" de H_0.

Exemple : On note $N_j = \mathrm{Card}(k = 1\ldots, 200\,,\, X_k = j)$ et on observe :

j	0	1	2	3	4	5	6	7	8	9	10	11	12	≥ 13
N_j	1	7	14	29	36	41	26	17	17	8	1	1	2	0

On regroupe par paquets de tailles supérieures à 5 :

j	≤ 1	2	3	4	5	6	7	8	≥ 9
N_j	8	14	29	36	41	26	17	17	12

L'EMV de p est :

$$\hat{p}_{200} = \frac{\sum_{k=1}^{200} X_k}{100 \times 200} = \frac{\sum_j jN_j}{100 \times 200} \simeq 0{,}05.$$

On compare $\hat{\theta}_{200,j} = N_j/200$ et $\theta_j^{\hat{p}_{200}}$:

j	≤ 1	2	3	4	5	6	7	8	≥ 9
$\hat{\theta}_{200,j}$	0,040	0,070	0,145	0,180	0,205	0,130	0,085	0,085	0,060
$\theta_j^{\hat{p}_{200}}$	0,037	0,081	0,140	0,178	0,180	0,150	0,106	0,065	0,063

On calcule

$$\zeta_{200}^{obs} = 200 \sum_j \frac{(\hat{\theta}_{200,j} - \theta_j^{\hat{p}_{200}})^2}{\theta_j^{\hat{p}_{200}}} \simeq 3{,}74\,.$$

Ici on a $9 - 1 - 1 = 7$ degrés de liberté, le seuil pour le test $\{\zeta_{200} \geq x_{0,95}(7)\}$ au niveau 0,05 est $x_{0,95}(7) = 14{,}07$.
Comme on a $\zeta_{200}^{obs} < x_{0,95}(7)$, on accepte l'hypothèse nulle "la loi du nombre de tickets gagnants par rouleau est une binomiale".

Test d'indépendance

• Ici on suppose que les observations sont des paires (Y_i, Z_i), et on se demande si ces deux quantités sont indépendantes.
• On observe un n-échantillon $((Y_1, Z_1), \ldots, (Y_n, Z_n))$ de vecteurs aléatoires i.i.d., avec Y_i à valeurs dans $\{b_1, \ldots, b_d\}$ et Z_i à valeurs dans $\{c_1, \ldots, c_m\}$.
• (Y_i, Z_i) est à valeurs dans l'espace fini $\mathcal{X} = \{b_1, \ldots, b_d\} \times \{c_1, \ldots, c_m\}$ de cardinal dm.

Test d'indépendance

On pose

$$\hat{\theta}_{jl} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(b_j,c_l)}(Y_i, Z_i), \qquad \hat{\pi}_j^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{b_j}(Y_i), \quad \hat{\pi}_l^{(2)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{c_l}(Z_i).$$

On note

$$\zeta_n = n\sum_{j=1}^{d}\sum_{l=1}^{m} \frac{(\hat{\theta}_{jl} - \hat{\pi}_j^{(1)}\hat{\pi}_l^{(2)})^2}{\hat{\pi}_j^{(1)}\hat{\pi}_l^{(2)}}$$

ζ_n mesure la "distance" entre la matrice $\hat{\boldsymbol{\theta}}$ des fréquences des couples (b_j, c_l) et la matrice $\hat{\boldsymbol{\pi}}$ des produits des fréquences marginales.
Si les deux coordonnées Y_i et Z_i sont indépendantes, $\hat{\boldsymbol{\theta}}$ et $\hat{\boldsymbol{\pi}}$ sont proches.

Proposition

Sous l'hypothèse d'indépendance, la suite $(\zeta_n)_n$ converge en loi vers une variable $\chi^2_{(d-1)(m-1)}$.
Sous l'hypothèse alternative $(\zeta_n)_n$ converge p.s. vers $+\infty$.

Remarque : $(d-1)(m-1) = dm - 1 - h$ avec $h = d - 1 + m - 1$.

Exemple : On considère les résultats au concours de l'X de deux lycées :

	Admis	Recalés	Présentés
Henri IV	81	17	98
Le Parc	136	17	153
Total	217	34	251

On désire tester l'hypothèse selon laquelle les élèves des deux lycées ont le même taux de réussite à l'X.

• Chaque observation est de la forme (Y_i, Z_i) où Y_i est à valeurs dans $\{H, L\}$ et $Y_i = H$, resp. L, signifie que l'étudiant vient du lycée H, resp. lycée L, et Z_i est à valeurs dans $\{A, R\}$ et $Z_i = A$, resp. R, signifie que l'étudiant a été admis, resp. recalé.

$$\zeta_{251}^{obs} = 251 \sum_{j\in\{H,L\}, l\in\{A,R\}} \frac{(\hat{\theta}_{jl} - \hat{q}_j \hat{p}_l)^2}{\hat{q}_j \hat{p}_l},$$

avec

$$\hat{q}_H = \frac{98}{251}, \quad \hat{q}_L = \frac{153}{251}, \quad \hat{p}_A = \frac{217}{251}, \quad \hat{p}_R = \frac{34}{251},$$

$$\hat{\theta}_{HA} = \frac{81}{251}, \quad \hat{\theta}_{HR} = \frac{17}{251}, \quad \hat{\theta}_{LA} = \frac{136}{251}, \quad \hat{\theta}_{LR} = \frac{17}{251}.$$

Exemple : On considère les résultats au concours de l'X de deux lycées :

	Admis	Recalés	Présentés
Henri IV	81	17	98
Le Parc	136	17	153
Total	217	34	251

On désire tester l'hypothèse selon laquelle les élèves des deux lycées ont le même taux de réussite à l'X.

• On trouve : $\zeta_{251}^{obs} = 1{,}98$. Or ici on a 1 degré de liberté, le seuil pour le test $\{\zeta_{251} \geq x_{0,95}(1)\}$ au niveau 0,05 est $x_{0,95}(1) = 3{,}84$.
Comme on a $\zeta_{251}^{obs} < x_{0,95}(1)$, on accepte l'hypothèse nulle que les deux lycées ont le même taux de réussite.

Matrices aléatoires et simulation de fractales

Une marche aléatoire matricielle :

$$\boldsymbol{X}_{n+1} = \mathbf{A}(Y_{n+1})\boldsymbol{X}_n + \boldsymbol{B}(Y_{n+1}) \in \mathbb{R}^d, \qquad \boldsymbol{X}_0 \in \mathbb{R}^2 \text{ donné}$$

où

- $\{\mathbf{A}(1), \ldots, \mathbf{A}(k)\}$ matrices $d \times d$,
- $\{\boldsymbol{B}(1), \ldots, \boldsymbol{B}(k)\}$ vecteurs d-dimensionnels,
- $(Y_n)_n$ i.i.d. à valeurs dans $\{1, \ldots, k\}$.

Les matrices $\mathbf{A}(i)$ sont des contractions : $\|\mathbf{A}(i)\| < 1$.

Pour $k = 1$, $\boldsymbol{X}_{n+1} = \mathbf{A}\boldsymbol{X}_n + \boldsymbol{B}$, on a $\boldsymbol{X}_n \overset{n\to\infty}{\longrightarrow} \boldsymbol{X}_\infty := (\mathbf{I} - \mathbf{A})^{-1}\boldsymbol{B}$.

Pour $k > 1$, que se passe-t-il ? Une drôle de convergence...

Une spirale

$d = 2$, $k = 2$:

$$\boldsymbol{X}_{n+1} = \mathbf{A}(Y_{n+1})\boldsymbol{X}_n + \boldsymbol{B}(Y_{n+1}) \in \mathbb{R}^2$$

- $\mathbb{P}(Y_n = 1) = 0{,}9$, $\mathbb{P}(Y_n = 2) = 0{,}1$,
- $\mathbf{A}(1) = \begin{pmatrix} 0{,}839 & -0{,}303 \\ 0{,}383 & 0{,}924 \end{pmatrix}$, $\mathbf{A}(2) = \begin{pmatrix} -0{,}161 & -0{,}136 \\ 0{,}138 & -0{,}182 \end{pmatrix}$,
- $\mathbf{B}(1) = \begin{pmatrix} 0{,}232 \\ -0{,}08 \end{pmatrix}$, $\mathbf{B}(2) = \begin{pmatrix} 0{,}921 \\ 0{,}178 \end{pmatrix}$.

Une feuille

$d = 2$, $k = 4$:

$$\boldsymbol{X}_{n+1} = \mathbf{A}(Y_{n+1})\boldsymbol{X}_n + \boldsymbol{B}(Y_{n+1}) \in \mathbb{R}^2$$

- $\mathbb{P}(Y_n = 1) = 0{,}01$, $\mathbb{P}(Y_n = 2) = 0{,}07$,
  $\mathbb{P}(Y_n = 3) = 0{,}07$, $\mathbb{P}(Y_n = 4) = 0{,}85$.
- $\mathbf{A}(1) = \begin{pmatrix} 0 & 0 \\ 0 & 0{,}16 \end{pmatrix}$, $\mathbf{A}(2) = \begin{pmatrix} 0{,}2 & -0.26 \\ 0{,}23 & 0{,}22 \end{pmatrix}$,
  $\mathbf{A}(3) = \begin{pmatrix} -0{,}15 & 0{,}28 \\ 0{,}26 & 0{,}24 \end{pmatrix}$, $\mathbf{A}(4) = \begin{pmatrix} 0{,}85 & 0{,}04 \\ -0{,}04 & 0{,}85 \end{pmatrix}$.
- $\mathbf{B}(1) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\mathbf{B}(2) = \begin{pmatrix} 0 \\ 1{,}6 \end{pmatrix}$,
  $\mathbf{B}(3) = \begin{pmatrix} 0 \\ 0{,}44 \end{pmatrix}$, $\mathbf{B}(4) = \begin{pmatrix} 0 \\ 1{,}6 \end{pmatrix}$.

Bon été