

# Statistiques décisionnelles

Charles Vin

S6 2022

## Plan du cours

1. Rappel du 1er semestre
2. Test d'ajustement :  
 $X_1, \dots, X_n$  va. iid. de loi  $\mathbb{P}_X$ 
  - (a) Est-ce que les  $X_i$  suivent la loi  $L$  ( $\mathbb{P}_X = L$ )?
  - (b) Est-ce que la loi des  $X_i$  appartient à une famille de loi? Est-ce qu'il existe  $m, \sigma^2$  tel que  $X_i \sim \mathcal{N}(m, \sigma^2)$
3. Tests de comparaison :
  - Test non paramétriques :  $(\omega, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ , On ne se restreint pas à une famille paramétrique de lois
  - Tests de comparaison :  $X_1, \dots, X_n$  jeu de données 1 et  $Y_1, \dots, Y_n$  jeu de données 2. Les  $X_i$   $Y_i$  ont-ils même loi? Les  $X_i$  et  $Y_i$  sont-ils indépendants?
4. L'ANOVA, voir cours de Mme Lavigne
5. Etudes de cas

## Table des matières

<b>1</b>	<b>Rappel sur les tests</b>	<b>3</b>
<b>2</b>	<b>Tests d'ajustement</b>	<b>4</b>
2.1	Le test d'ajustement de Kolmogorov-Smirnov	4
2.1.1	Rappels	4
2.1.2	Le test de Kolmogorov-Smirnov	5
2.1.2.1	Comment calculer en pratique $h(F_n, F)$	7
2.1.2.2	Comportement théorique de $h(F_n, F)$	8
2.1.2.3	Le test de Kolmogorov-Smirnov à 1 échantillon	9
2.1.2.4	Qu'est ce que $W_\infty$	10
2.1.2.5	Kolmogorov-Smirnov en pratique	10
2.2	Ajustement à une famille de lois	10
2.2.0.1	Adéquation à une famille d'exponentielle	11
2.2.0.2	Adéquation à une loi normale	11
2.3	Le test du $\chi^2$ d'ajustement	11
2.3.0.1	Préparatifs, introduction	12
2.3.0.2	Le test du $\chi^2$	12
2.3.0.3	Mise en place concrète :	13
2.3.0.4	Test du $\chi^2$ avec fusion des classes	13
2.4	Le test du $\chi^2$ pour une loi discrète	14
2.4.0.1	En pratique	14
2.4.0.2	Limite	15
2.5	Le test du $\chi^2$ pour une loi continue	15
2.5.0.1	Bilan de la méthode	15
2.6	Le $\chi^2$ d'ajustement à une famille paramétrique de loi	16
2.6.0.1	En pratique	16
2.7	Bilan du chapitre	17

<b>3</b>	<b>Loi de comparaison</b>	<b>17</b>
3.1	Le test d'homogénéité de Kolmogorov-Smirnov . . . . .	18
3.1.1	Test d'homogénéité de Kolmogorov-Smirnov . . . . .	18
3.1.1.1	En pratique (cas n et m grand) . . . . .	19

# 1 Rappel sur les tests

On fixe un modèle  $(\Omega, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ .  
On dit que le modèle est paramétrique s'il existe

$$d \in \mathbb{N} \text{ tel que } \Theta \in \mathbb{R}^d.$$

Sinon, on dira que le modèle est non-paramétrique.

- Exemple 1.1** (de modèle paramétrique).
1.  $\Theta \subset \mathbb{R} \times \mathbb{R}, \mathbb{P}_\theta = \mathcal{N}(m, \sigma^2), \theta = (m, \sigma^2)$
  2.  $\Theta = [0, 1], \mathbb{P}_\theta = \text{Ber}(\theta), \theta \in [0, 1]$
  3.  $\Theta = \mathbb{R}_*^+, \mathbb{P}_\theta = \mathcal{E}(\theta), \theta \in \mathbb{R}_*^+$

- Exemple 1.2** (de modèle non-paramétrique).
1.  $\Theta =$  densité de probabilité sur  $\mathbb{R}, \mathbb{P}_f =$  la loi de densité  $f, f \in \Theta$
  2.  $\Theta = \{(p_i)_{i \in \mathbb{N}}, \forall i \in \mathbb{N}, p_i \in [0, 1] \sum_{i=0}^{+\infty} p_i = 1\}, \theta = (p_i)_{i \in \mathbb{N}}, \mathbb{P}_\theta =$  la loi discrète tq  $\forall k \in \mathbb{N}, \mathbb{P}(X = k) = p_k,$
  3.  $\Theta = \{\text{fonction de répartition de var.}\}, F \in \Theta, \mathbb{P}_F =$  loi de la va. dont la fonction de répartition est  $F, (\mathbb{P}_F)_{F \in \Theta}$

**Définition 1.1** (Test d'hypothèse). Soit  $\mathbb{X} = (X_1, \dots, X_n)$  un ensemble d'observations de loi  $\mathbb{P}_\theta$   
On appelle test d'hypothèse de  $H_0$  contre  $H_1$  (à  $H_0$  et  $H_1$  sont des sous-ensemble de  $\Theta$ ). toute fonction des observations à valeur dans  $\{0, 1\}$   
— à  $\phi(\mathbb{X}) = 0$  correspond à conserver  $H_0$   
— à  $\phi(\mathbb{X}) = 1$  correspond à rejeter  $H_0$  au profit de  $H_1$

$R = \phi(\{1\})$  est la zone de rejet, c'est l'ensemble des observation qui ... à un rejet de  $H_0$

*Remarque.* Si  $\phi(\mathbb{X}) = \mathbb{1}_{h(\mathbb{X}) \in R}$  on dira que  $h$  est la statistique de test et  $R$  la zone de rejet

**Exemple 1.3.**  $h(\mathbb{X}) = \sum_{i=1}^n X_i, R = [h, +\infty[.$  Test :  $\phi(\mathbb{X}) = \mathbb{1}_{\sum_{i=1}^n X_i \geq k}$

**Exemple 1.4.**  $\phi(\mathbb{X}) = 0$  le test que conserve toujours  $H_0$  est un test.

**Définition 1.2** (Erreur de première espèce & Taille du test). l'Erreur de 1ère espèce est la fonction :

$$\begin{aligned} \alpha : \Theta_0 &\rightarrow [0, 1] \\ \theta &\mapsto \mathbb{P}_\theta(\phi(\mathbb{X}) = 1) \end{aligned}$$

La taille du test  $\phi$  est

$$\alpha^* = \sup_{\theta \in \Theta_0} \alpha(\theta).$$

On dit que  $\phi$  est de niveau  $\alpha$  si

$$\alpha^* \leq \alpha.$$

Une suite de test  $(\phi_n)_{n \in \mathbb{N}}$  est de niveau asymptotique  $\alpha$  si

$$\limsup_n \alpha_n^* \leq \alpha.$$

En général on a :  $\lim_{n \rightarrow \infty} \alpha_n^* = \alpha$

*Remarque.* Pour l'erreur de 1ère espèce le meilleur test est  $\phi(\mathbb{X}) = 0$ . En effet  $\forall \theta \in \Theta_0, \mathbb{P}_\theta(\phi(\mathbb{X}) = 1) = 0$

*Remarque* (Cours de M.Thiam, def 12). Si vous préférez la formulation du 1er semestre, c'est tout aussi valable.

**Définition 1.3** (Erreur de seconde espèce et puissance). La fonction erreur de 2nd espèce d'un test  $\phi$  est

$$\begin{aligned} \underline{\beta} : \Theta_1 &\rightarrow [0, 1] \\ \theta &\mapsto \mathbb{P}_\theta(\phi(\mathbb{X}) = 0) \end{aligned}$$

C'est la probabilité de conserver à tort  $H_0$ . On appelle en général erreur de seconde espèce la quantité

$$\beta = \sup_{\theta \in \Theta_1} \underline{\beta}(\theta)$$

La fonction puissance  $\gamma$  est  $1 - \underline{\beta}$ .

**Exemple 1.5.** Le test  $\phi(\mathbb{X}) = 0$  (le test stupide) a une erreur de seconde espèce qui vaut 1.

$$\mathbb{P}_\theta(\phi(\mathbb{X}) = 0) = 1.$$

et sa puissance vaut 0

**Définition 1.4** (p-valeur). Si pour tout niveau  $\alpha$ , on a construit un test  $\phi_\alpha$ . Soit  $\mathbb{X}$  une observation.

$$p(\mathbb{X}) = \inf\{\alpha \in [0, 1] \text{ tel que } \phi_\alpha(\mathbb{X}) = 1\}.$$

Si on choisit un niveau  $\alpha$

$$\alpha < p(\mathbb{X}), \text{ on conserve } H_0.$$

Et si  $\alpha \geq p(\mathbb{X})$  on rejette  $H_0$

**Définition 1.5** (Test consistant). Une suite de tests  $\phi_n$  est dite consistant si pour tout  $\theta \in \Theta_1$

$$\gamma_n(\theta) \xrightarrow{n \rightarrow \infty} 1.$$

## 2 Tests d'ajustement

Le but de ce chapitre est de répondre à la question suivante :  
Étant donnée un échantillon  $X_1, \dots, X_n$  et une loi de proba sur  $\mathbb{R}$  nommée  $\mathcal{L}$

Est-ce que les  $X_i \sim \mathcal{L}$ .

- $H_0$  = les  $X_i$  ont pour loi  $\mathcal{L}$
- $H_1$  = les  $X_i$  n'ont pas pour loi  $\mathcal{L}$

Comment comprendre ce problème?

1. En général, on peut utiliser les fonction de répartition. La question devient  $F_X = F$  contre  $F_X \neq F$  (en tout point de  $\mathbb{R}$ )
2. Si les  $X_i$  sont à support dans  $\{1, \dots, K\}$ . La question devient  $\forall i \in \{1, \dots, K\}, \hat{p}_i = p_i$  contre  $\exists i \text{ tq } \hat{p}_i \neq p_i$  où  $\hat{p}_i = P(X = i)$  et  $p_i = P(L = i)$

Énorme problème : On ne connaît pas la loi des  $X_i$ , on connaît juste  $n$  réalisations.

Problème plus difficile : Ajustement à une famille de lois? Est-ce que les  $X_i$  proviennent d'une loi normale? (sans en connaître les paramètres)

*Remarque.* Cette question est fondamentale pour valider un modèle

### 2.1 Le test d'ajustement de Kolmogorov-Smirnov

#### 2.1.1 Rappels

**Définition 2.1** (Fonction de répartition). Soit  $X$  une variable aléatoire réelle, sa fonction de répartition est la fonction

$$F_X : \mathbb{R} \rightarrow [0, 1] \\ t \mapsto P(X \leq t)$$

Elle caractérise la loi de  $X$ .

Si  $X$  est à densité,  $F_X$  est continue. Les discontinuité de  $F_X$  sont les valeurs  $t_0 \in \mathbb{R}$  tel que  $P(X = t_0) > 0$ .

**Exemple 2.1.** — Si  $X \sim \text{Unif}(0, 1)$

$$F_X(t) = P(X \leq t) = \int_0^t \mathbb{1}_{[0,1]}(x) dx = \begin{cases} 0 & \text{si } t \leq 0 \\ t & \text{si } t \in [0; 1] \\ 1 & \text{si } t \geq 1 \end{cases}.$$

— Si  $X \sim \mathcal{E}(\lambda)$

$$F_X(t) = \int_0^t \lambda e^{-\lambda x} dx = \begin{cases} 0 & \text{si } t < 0 \\ 1 - e^{-\lambda t} & \text{si } t \geq 0 \end{cases}.$$

— Si  $X \sim \mathcal{B}(p)$

$$F_X(t) = \begin{cases} 0 & \text{si } t < 0 \\ 1 - p & \text{si } t \in [0; 1[ \\ 1 & \text{si } t \geq 1 \end{cases}$$

**Définition 2.2** (Pseudo inverse de la fonction de répartition). Soit  $X$  une var. de fonction de répartition  $F_X$ . On pose

$$F_X^{-1} : ]0, 1[ \rightarrow \mathbb{R} \\ x \mapsto \inf\{t \in \mathbb{R}, F_X(t) \geq x\}$$

On l'appelle inverse généralisé de  $F_X$  et elle coïncide avec l'inverse si  $F_X$  est bijective. Elle vérifie la propriété fondamentale

$$\forall x \in ]0, 1[, \forall t \in \mathbb{R}, F_X^{-1} \leq t \Leftrightarrow x \leq F_X(t).$$

**Théorème 2.1.** Soit  $X$  une var. de fonction de répartition  $F_X$  et une variable uniforme  $U$  sur  $[0, 1]$  alors

$$X \text{ et } F_X^{-1}(U) \text{ ont même loi.}$$

*Preuve :* Soit  $t \in \mathbb{R}$

$$P(F_X^{-1}(U) \leq t) = P(U \leq F_X(t)) \text{ comme } \{F_X^{-1}(U) \leq t\} = \{U \leq F_X(t)\}.$$

Or  $F_X(t) \in [0, 1]$  donc

$$P(U \leq F_X(t)) = F_X(t).$$

Ainsi  $F_X^{-1}$  et  $X$  ont la même fonction de répartition et donc la même loi □

Nouveau cours du 20/01

## 2.1.2 Le test de Kolmogorov-Smirnov

**But :** Si on a  $X_1, \dots, X_n$  observation iid. Est-ce que la fonction de répartition des  $X_i$  est une certaine fonction  $F_L$  **donnée**?

$\Leftrightarrow F_X = F_L \Leftrightarrow$  La loi des  $X_i$  est la même que  $L$

**Exemple 2.2.** Se demander si les  $X_i \sim \mathcal{E}(1)$  revient à demander : Est-ce que  $\forall t \in \mathbb{R}, F_X(t) = (1 - e^{-t}) \mathbb{1}_{t \geq 0}$

**Autre reformulation :** Est-ce que mes observations sont cohérentes avec l'hypothèse  $F_{X_i} = F$ ? Il va donc falloir estimer  $F_{X_i}$  et la comparer à  $F$

**Définition 2.3** (Fonction de répartition empirique). Soit  $X_1, \dots, X_n$  un échantillon iid. On appelle **fonction de répartition empirique** de  $X_1, \dots, X_n$  la fonction

$$F_n : \mathbb{R} \rightarrow [0, 1] \\ t \mapsto \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq t}$$

Illustration graphique 1 :

**Rappels :**

1.  $\forall t \in \mathbb{R}, F_n(t) \xrightarrow[p.s]{n \rightarrow +\infty} F_{X_1}(t)$
2. De plus  $\forall t \in \mathbb{R}$  fixé

$$\frac{\sqrt{n}}{\sqrt{F_X(t)(1 - F_X(t))}} (F_n(t) - (F_X(t))) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} Z \text{ de loi } \mathcal{N}(0, 1).$$

Ce n'est rien d'autre que le TCL pour la suite de variables iid.  $(Y_i = \mathbb{1}_{X_i \leq t})_{i \in \mathbb{N}}$

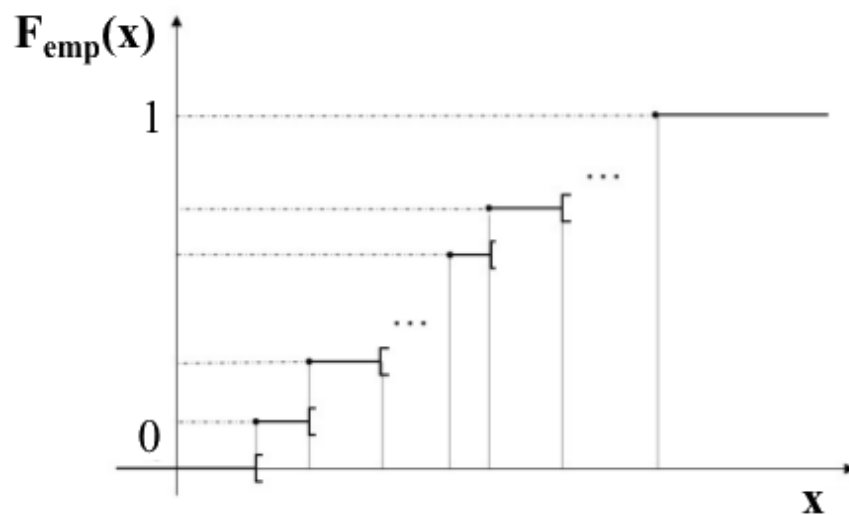


Figure 1 – Exemple de fonction de répartition empirique

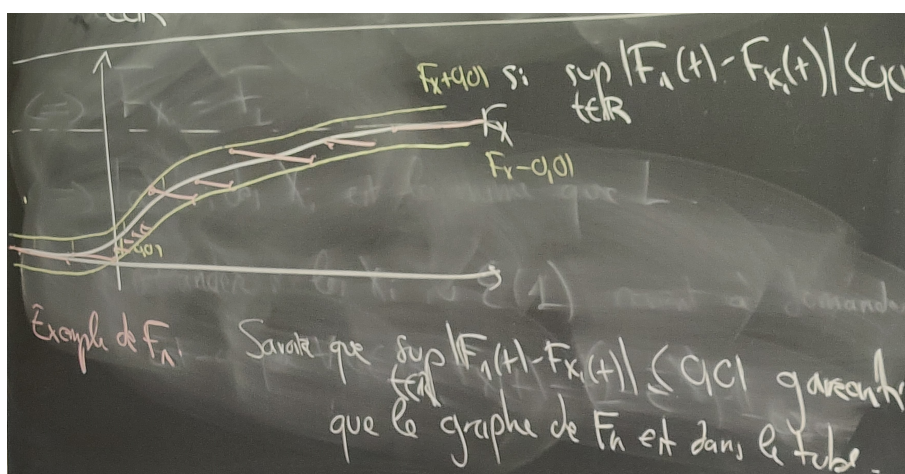


Figure 2 – Illustration graphique de Glivenko-Cantelli

**Théorème 2.2** (Glivenko-Cantelli).  $(X_i)_{i \in \mathbb{N}}$  une suite de va. iid. alors

$$\sup_{t \in \mathbb{R}} |F_n(t) - F_{X_1}(t)| \xrightarrow[n \rightarrow +\infty]{p.s.} 0.$$

*Illustration graphique 2 :*

Ce théorème montre que la bonne quantité pour savoir si  $F_X = F$  à  $F$  est une certaine fonction donnée est

$$h(F_n, F) = \sup_{t \in \mathbb{R}} |F_n(t) - F_{X_1}(t)|.$$

— Si  $F = F_X$  alors d'après le théorème de Glivenko-Cantelli :

$$h(F_n, F) \xrightarrow[n \rightarrow +\infty]{p.s.} 0.$$

— Si je me suis trompé et que  $F \neq F_X$ , alors

$$h(F_n, F) \xrightarrow[n \rightarrow +\infty]{p.s.} \sup_{t \in \mathbb{R}} |F_n(t) - F_{X_1}(t)|.$$

En effet  $F_n \rightarrow F_{X_i}$  donc

$$\begin{aligned} h(F_n, F) &= \sup_{t \in \mathbb{R}} |F_n(t) - F_{X_1}(t)| \\ &\xrightarrow[n \rightarrow +\infty]{p.s.} \sup_{t \in \mathbb{R}} |F_n(t) - F_{X_1}(t)| > 0 \end{aligned}$$

De manière informelle, on a envie de dire

- Si  $h(F_n, F)$  est petit alors  $F_X = F$
- Si  $h(F_n, F)$  n'est pas petit alors  $F_X \neq F$

### 2.1.2.1 Comment calculer en pratique $h(F_n, F)$ ?

Données :  $X_1, \dots, X_n$  des valeurs.  $F$  une fonction de répartition cible.

**But :** Calculer  $h(F_n, F)$  de manière pratique. à  $h(F_n, F) = \sup_{t \in \mathbb{R}} |F_n(t) - F_{X_1}(t)|$  (Voir Figure. 3)

*Note* (du dessin). Le but de cette explication est de montrer graphiquement et instinctivement pourquoi on ne regarde pas pour tout  $t \in \mathbb{R}$  mais uniquement à chaque saut.

1. étape : avant  $X_{(1)}$

$$\sup_{t \leq X_{(1)}} |F_n(t) - F_{X_1}(t)| = \max\left\{\left|\frac{1}{n} - F(X_{(1)})\right|, |F(X_{(1)}) - 0|\right\}.$$

On recommence pour les différentes valeurs de  $X_{(i)}$  et on voit que la plus grande distance entre les deux courbes est forcément atteinte à un des points de saut

**Remarque (attention).** Pour chaque saut, il faut regarder 2 valeurs AVANT et APRES le saut.

Formule de calcul de  $h(F_n, F)$

$$h(F_n, F) = \max_{1 \leq i \leq n} \left( \max\left(\left|\frac{i}{n} - F(X_{(i)})\right|, \left|\frac{i-1}{n} - F(X_{(i)})\right|\right) \right).$$

*Note.* On fait le max pour tous les sauts du maximum entre la distance APRES (au moment du saut) et AVANT (juste avant le saut (i-1)).

**Exemple 2.3** (Cas concret).  $X_1 = 0.06, X_2 = 0.8, X_3 = 0.27, X_4 = 0.67, X_5 = 0.38$

$$F(t) = F_U(t) = \begin{cases} 0 & \text{si } t \leq 0 \\ t & \text{si } t \in [0, 1] \\ 1 & \text{si } t \geq 1 \end{cases}.$$

Etape 1 : On ordonne les valeurs Ici  $h(F_n, F_U) = 0.22$

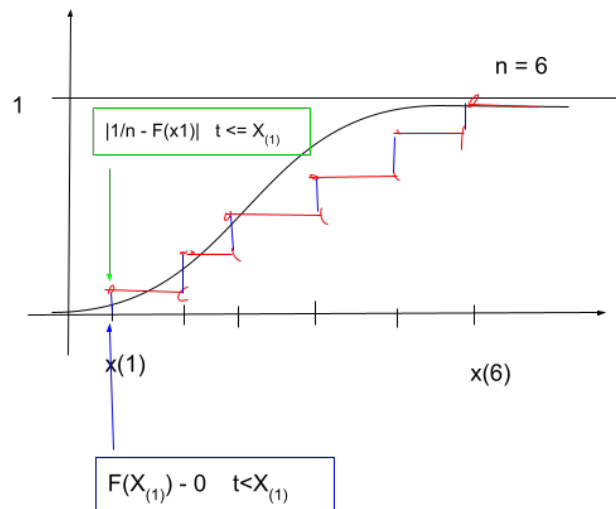


Figure 3 – Figure pour trouver la fonction  $h(F_n, F)$

$X_{(i)}$	0.06	0.27	0.38	0.67	0.8
$F_n$	0.2	0.4	0.6	0.8	1
$F$	0.06	0.27	0.38	0.67	0.8
Après le saut : $\frac{i}{n} - F(X_{(i)})$	0.14	0.13	<b>0.22</b>	0.13	0.2
Avant le saut : $\frac{i-1}{n} - F(X_{(i)})$	0.06	0.07	0.02	0.07	0

### 2.1.2.2 Comportement théorique de $h(F_n, F)$

$$h(F_n, F) = \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq t} - F(t) \right|.$$

est une variable aléatoire.

A priori, la loi de  $h(F_n, F)$  dépend

- de  $n$
- de la loi des  $X_i$

**Rappel :**  $H_0 : F = F_{X_0}$  contre  $H_1 : F \neq F_{X_i}$

Sous  $H_0$  quel est la loi de  $h(F_n, F)$ ?

$$h(F_n, F) = \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq t} - F_{X_1}(t) \right|.$$

Soit  $U_1, \dots, U_n$  iid. uniforme sur  $[0, 1]$

Soit  $F_{X_1}^{-1}$  l'inverse généralisé de  $F_{X_1}$

Alors  $F_{X_1}^{-1}(U_1), \dots, F_{X_1}^{-1}(U_n)$  ont même loi que  $X_1, \dots, X_n$ . Ainsi en loi

$$h(F_n, F) = \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{F_{X_1}^{-1}(U_i) \leq t} - F_{X_1}(t) \right|.$$

Or  $\{F_{X_1}^{-1} \leq t\} = \{U_i \leq F_{X_1}(t)\}$  donc  $\mathbb{1}_{F_{X_1}^{-1}(U_i) \leq t} = \mathbb{1}_{U_i \leq F_{X_1}(t)}$  et donc

$$h(F_n, F) = \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{U_i \leq F_{X_1}(t)} - F_{X_1}(t) \right|.$$



Si  $F_{X_1}$  est continue, alors  $]0, 1[ \subset F_{X_1}(\mathbb{R}) \subset [0, 1]$ . Ainsi en reparamétrant le sup on a

$$h(F_n, F) = \sup_{s \in ]0, 1[} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{U_i \leq s} - s \right|.$$

Dans cette formule, la loi de  $X$  (et sa fonction de répartition) n'apparaît pas!

**Bilan : La loi de  $h(F_n, F)$  ne dépend que de  $n$  sous  $H_0$**

La loi de  $h(F_n, F)$  est tabulée pour toutes les valeurs de  $n$ . On peut alors construire un test de niveau  $1 - \alpha$

### 2.1.2.3 Le test de Kolmogorov-Smirnov à 1 échantillon

Données :

- $X_1, \dots, X_n$
- $F$  une fonction de répartition continue
- $\alpha$  un niveau
- $H_0 : F_X = F$  contre  $H_1 : F_{X_1} \neq F$

Soit  $h_\alpha$  le quantile de niveau  $1 - \alpha$  de  $h(F_n, F)$

- Si  $h(F_n, F) > h_\alpha$ , on rejette  $H_0$
- Si  $h(F_n, F) \leq h_\alpha$ , on conserve  $H_0$

De manière formelle :  $\phi(\mathbb{X}) = \mathbb{1}_{h(F_n, F) > h_\alpha}$

**Exemple 2.4** (retour sur l'exemple). Dans le tableau, on avait lu  $h(F_n, F) = 0.22, n = 5$ .

Test de niveau 90% : la zone de rejet est  $h > 0.509$  (d'après la table). Dans l'exemple on conserve  $H_0$ , les  $X_i$  proviennent d'une  $\mathcal{U}([0, 1])$

**Exemple 2.5** (Autre exemple).  $X_1 = 1.67, X_2 = 1.3, X_3 = 0.01, X_4 = 2.48, X_5 = 0.11$  Est-ce que les  $X_i \sim \mathcal{E}(1)$ ? On applique le test de Kolmogorov-Smirnov.

$X_{(i)}$	0.01	0.11	1.3	1.67	2.48
$F_n$	0.2 + 1/n	0.4	0.6	0.8	1
$F(t) = 1 - e^{-t}$	0.01	0.1	0.72	0.81	0.91
Après le saut : $\frac{i}{n} - F(X_{(i)})$	0.19	0.3	0.12	0.01	0.09
Avant le saut : $\frac{i-1}{n} - F(X_{(i)})$	0.01	0.1	<b>0.32</b>	0.21	0.11

$$h_{F_5, F} = 0.32.$$

Test de niveau 99% : Rejet si  $h \leq 0.6689$  comme  $0.32 \leq 0.6685$  on conserve  $H_0$

Nouveau cours du 27/01

### Rappel du cours précédent

On a vu le test de Kolmogorov-Smirnov :  $X_1, \dots, X_n$  iid. de fdr.  $F_{X_1}$ .  
Fonction de répartition cible  $F$

$$H_0 = F_{X_1} = F \text{ contre } H_1 = F_{X_1} \neq F.$$

On calcule  $h(F_n, F) = \sup_{t \in \mathbb{R}} |F_n(t) - F(t)|$ .

La loi de  $h(F_n, F)$  est tabulée, il suffit alors pour un niveau  $\alpha$  donnée de vérifier si

$$h(F_n, F) > S_\alpha \text{ le seuil au niveau } \alpha.$$

### Début du cours

Si  $n$  est grand, on ne dispose pas de la table de  $h(F_n, F)$ . Solution : Utiliser un test asymptotique.

**Théorème 2.3.** Soit  $h_n = \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{U_i \leq t} - t \right|$  à  $U_1, \dots, U_n$  sont des va. iid. de loi uniforme sur  $[0, 1]$

$$\sqrt{n} h_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} W_\infty.$$

où  $P(W_\infty \leq t) = 1 - 2 \sum_{k=1}^{+\infty} (-1)^{k+1} e^{-2k^2 t^2}$ .

Bonne nouvelle : La loi de  $W_\infty$  est tabulée!!

**Exemple 2.6** (Théorie de l'utilisation). Si  $n \geq 30$ . Pour avoir  $S_\alpha$  tel que  $P(h_n > S_\alpha) \approx 1 - \alpha$ . Si je prends  $k_\alpha$  tel que  $P(W_\infty > k_\alpha) = 1 - \alpha$  ( $k_\alpha$  est le quantile d'ordre  $1 - \alpha$  de  $W_\infty$ ). Alors, si on pose  $S_\alpha = \frac{k_\alpha}{\sqrt{n}}$  on a :

$$P(h_n \geq S_\alpha) = P(h_n \geq \frac{k_\alpha}{\sqrt{n}}) = P(\sqrt{n}h_n > k_\alpha) \approx P(W_\infty \geq k_\alpha).$$

Conclusion : Si  $n$  est grand (pas dans la table), on prend  $s_\alpha = \frac{k_\alpha}{\sqrt{n}}$  à  $h_\alpha$  est le quantile d'ordre  $1 - \alpha$  de  $W_\infty$

#### 2.1.2.4 Qu'est ce que $W_\infty$

$$\begin{aligned} \sqrt{n}h_n &= \sqrt{n} \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{U_i \leq t} - F(\mathbb{1}_{U_i \leq t}) \right| \\ &= \sup_{t \in \mathbb{R}} \left| \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{U_i \leq t} - F(\mathbb{1}_{U_i \leq t}) \right) \right| \end{aligned}$$

Cette quantité est approximativement une  $\mathcal{N}(0, t(1-t))$

$$Gt \rightarrow \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{U_i \leq t} - F(\mathbb{1}_{U_i \leq t}) \right).$$

Le graphe de  $G$  est aléatoire et est disponible sur moodle (ça ressemble à un cours de la bourse, dans notre cas on appelle ça un pont Brownien).

Pour la culture : un inégalité bien pratique

**Théorème 2.4** (Inégalité DKW). *Inégalité de Dvoretzky-Kiefer-Wolfowitz :  $X_i$  va. iid.*

$$\forall n \in \mathbb{N}, \forall \epsilon > 0, \mathbb{P}(\sup_{t \in \mathbb{R}} |F_n(t) - F_{x_1}(t)| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

Cette inégalité est

- Non asymptotique
- Pas génial si  $n$  petit

Mais elle permet aussi de construire une zone de rejet.

#### 2.1.2.5 Kolmogorov-Smirnov en pratique On fait ce test si

1. Les  $X_i$  semblent provenir d'une loi à fonction de répartition continue.  $\Rightarrow$  on n'a pas plusieurs fois la même valeur (sauf si celle-ci on était arrondi).  
Par exemple : si on voit 14 fois la même valeur  $\rightarrow$  on utilise pas KS. Mais si on voit 2 fois la même valeur  $\rightarrow$  c'est jouable
2. Fonctionne  $\forall n$  : même si  $n$  est petit, ce test est pertinent (alors qu'un test du khi-deux qu'on verra plus tard est exclusivement asymptotique)
3. Si  $n \geq 100$ , on fait le test asymptotique. Sinon on peut faire un test non asymptotique.

## 2.2 Ajustement à une famille de lois

On veut savoir si nos observations iid. proviennent d'une certaine famille de lois.

**Exemple 2.7.** — Est-ce que la loi  $X_i$  sont des  $\mathcal{E}(\lambda)$  pour  $\lambda > 0$ ?

- Est-ce que la loi  $X_i$  sont des  $\mathcal{N}(m, \sigma^2)$  pour  $m \in \mathbb{R}, \sigma^2 > 0$ ?
- Est-ce que la loi  $X_i$  sont des  $\mathcal{B}(n, p)$  pour  $m \in \mathbb{N}, p \in [0, 1]$ ?

Malheureusement, il est impossible de répondre à cette question en toute généralité.

Cependant il y a deux exemple important qu'on peut traiter.

### 2.2.0.1 Adéquation à une famille d'exponentielle Données : $X_1, \dots, X_n$ iid. loi inconnue

- $H_0$  : les  $X_i$  sont  $\mathcal{E}(\lambda)$  pour un certain  $\lambda \in \mathbb{R}_*^+$
- $H_1$  : les  $X_i$  ne sont pas exponentiels.

Idée : On utilise  $h(F_n, F_\lambda)$  pour un  $F_\lambda$  bien choisis :

$$F_\lambda = (1 - e^{-\lambda x}) \mathbb{1}_{x>0}.$$

Si on veut tester  $X_i \sim \mathcal{E}(\lambda)$ ,  $\lambda$  fixée, on regarde

$$h(F_n, F_\lambda) = \sup_{t \in \mathbb{R}} |F_n(t) - (1 - e^{-\lambda t}) \mathbb{1}_{t>0}|.$$

Problème :  $\lambda$  est inconnu  $\Rightarrow$  On l'estime !

$$\bar{\lambda}_n = \frac{n}{\sum_{i=1}^n X_i} \text{ estimateur Maximum Vraisemblance de } \lambda.$$

On regarde :  $X_i$  iid  $\mathcal{E}(\lambda)$

$$h(F_n, F_{\bar{\lambda}_n}) = \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq t} - (1 - e^{-\bar{\lambda}_n t}) \mathbb{1}_{t>0} \right|.$$

Miracle : La loi de  $h(F_n, F_{\bar{\lambda}_n})$  ne dépend pas de  $\lambda$ , mais uniquement de  $n$ .

Si les  $(Y_i)_{i \in \mathbb{N}}$  sont iid. de loi  $\mathcal{E}(1)$ , les  $(\frac{1}{\lambda} Y_i)_{i \in \mathbb{N}}$  sont iid de loi  $\mathcal{E}(\lambda)$ . Pour comprendre la loi de  $h(F_n, F_{\bar{\lambda}_n})$ , je peux remplacer les  $X_i$  par  $\frac{1}{\lambda} Y_i$ .

$$\begin{aligned} h(F_n, F_{\bar{\lambda}_n}) &= \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\frac{Y_i}{\bar{\lambda}_n} \leq t} - (1 - e^{-\frac{n}{\sum_{i=1}^n Y_i} t}) \mathbb{1}_{t>0} \right| \\ &= \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{Y_i \leq \bar{\lambda}_n t} - (1 - e^{-\frac{n}{\sum_{i=1}^n Y_i} \bar{\lambda}_n t}) \mathbb{1}_{\bar{\lambda}_n t>0} \right| \text{ or } \mathbb{1}_{t>0} = \mathbb{1}_{\bar{\lambda}_n t>0} \\ &= \sup_{s \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{Y_i \leq s} - (1 - e^{-\frac{n}{\sum_{i=1}^n Y_i} s}) \mathbb{1}_{s>0} \right| \text{ avec } s = \bar{\lambda}_n t \end{aligned}$$

Cela ne dépend pas de  $\lambda$  mais seulement de  $n$ . On peut tabuler ! (Malheureusement elle n'a pas de nom) et construire un test de KS.

### 2.2.0.2 Adéquation à une loi normale On peut adapter le test précédent pour des gaussiennes en estimant $m$ et $\sigma^2$ avec $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ et $V_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ et construire un test.

Cela s'appelle le test de normalité de **Lilliefors** (voir exo de TD pour la suite)

## 2.3 Le test du $\chi^2$ d'ajustement

La lettre grecque  $\chi$  se prononce "khi".

On dispose de  $X_1, \dots, X_n$  va. iid.

On se place dans le cas particulier où les  $X_i$  sont à valeurs dans un ensemble fini  $\{x_1, \dots, x_d\}$ . La loi des  $X_i$  est donc entièrement déterminée par la donnée de  $p_k = P(X_1 = x_k)$  pour tout  $k$ . Le vecteur  $p = (p_1, \dots, p_d)$  caractérise la loi des  $X_i$

*Remarque.* On sait que  $p_1 + \dots + p_d = 1$

Hypothèse :  $\forall h \in \{1, \dots, d\}, p_h > 0$ . On ne s'est pas trompés dans le support, il faut prendre le plus petit  $d$ .

Ces restrictions ne sont pas si contraignantes dans beaucoup de cas pratiques, elles sont automatiquement vérifiées

**Exemple 2.8.** — Réponse à un questionnaire QCM : la réponse prend un nombre fini de valeurs

- Une notes sur 20 d'un examen
- Des variables qualitatives : fille/garçons, couleur des yeux

On a des observations  $X_1, \dots, X_n$  de loi inconnue  $p = (p_1, \dots, p_d)$ . On veut savoir si  $p = p^{ref}$  pour un vecteur  $p^{ref}$  fixé.

$$\begin{aligned} H_0 &= p = p^{ref} \text{ i.e. } \forall k \in \{1, \dots, d\}, p_k = p_k^{ref} \\ H_1 &= p \neq p^{ref} \text{ i.e. } \exists k \in \{1, \dots, d\} : p_k \neq p_k^{ref} \end{aligned}$$

**2.3.0.1 Préparatifs, introduction** Si on trie nos valeurs  $p^{ref} = (0; 3, 0.1, 0.2, 0.2, 0.2)$  On a envie de

	x1	x2	...	xs
Nombre d'observation	17	23	...	12

regarder  $\bar{p}_1 = \frac{\text{Nombre de } n_1}{n}, \dots, \bar{p}_s = \frac{\text{Nombre de } n_s}{n}$ . On a envie de construire quelque chose avec ces estimateurs

### Notation

$$\forall k \in \{1, \dots, d\}, N_{k,n} = \sum_{i=1}^n \mathbb{1}_{X_i=x_k}.$$

Les  $N_{k,n}$  sont les effectifs observés.

$$\forall k \in \{1, \dots, d\} \bar{p}_{k,n} = \frac{N_{k,n}}{n} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i=x_k}.$$

Les  $\bar{p}_{k,n}$  sont les proportions observés. On note  $\bar{p}_n = (\bar{p}_{1,n}, \dots, \bar{p}_{d,n})$

Sous  $H_0$ , les  $\bar{p}_{k,n}$  devraient être proches des  $p_k^{ref}$

*Note.* Comme dans KS, on va trouver une formule reliant les deux et pouvant être tabuler pour faire des tests. Mais elle est pas vraiment démontrable à notre niveau et utilise des vecteurs gaussiens

**Théorème 2.5.** Sous  $H_0$  on note

$$D(\bar{p}_n, p^{ref}) = n \sum_{k=1}^d \frac{(\bar{p}_{k,n} - p_k^{ref})^2}{p_k^{ref}}$$

$$D(\bar{p}_n, p^{ref}) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \chi^2(d-1)$$

Sous  $H_1$

$$D(\bar{p}_n, p^{ref}) \xrightarrow[n \rightarrow \infty]{p.s.} +\infty.$$

*Remarque* (Autre formulation, qu'on utilise en TD!). On peut aussi écrire

$$D(\bar{p}_n, p^{ref}) = \sum_{k=1}^d \frac{(N_{k,n} - np_k^{ref})^2}{np_k^{ref}}$$

Si on note  $N_k^{ref} = np_k^{ref}$  l'effectifs attendu, alors cela devient

$$D(\bar{p}_n, p^{ref}) = \sum_{k=1}^d \frac{(N_{k,n} - N_k^{ref})^2}{N_k^{ref}}.$$

$N_k^{ref}$  n'est pas un entier en général

### 2.3.0.2 Le test du $\chi^2$

- Données :  $X_1, \dots, X_n$  à valeur dans  $\{x_1, \dots, x_d\}$
- $p^{ref}$  qu'on veut tester
- Niveau  $\alpha$

Soit  $h_\alpha$  le quantile d'ordre  $1 - \alpha$  de la loi  $\chi^2(d-1)$  alors

- Si  $D(\bar{p}_n, p^{ref}) \geq h_\alpha$  on rejette  $H_0$
- Sinon  $D(\bar{p}_n, p^{ref}) < h_\alpha$  on conserve  $H_0$

**Attention :** Ce test est uniquement asymptotique!

Condition d'utilisation :

$$\forall k \in \{1, \dots, d\}, np_k^{ref}(1 - p_k^{ref}) \geq 5.$$

Cela implique  $n \geq 20$  mais en général il faut beaucoup plus

**Exemple 2.9** (dé truqué). On dispose d'un dé douteux, on relève les résultats de 100 lancés et on veut déterminer si il est pipé ou non.

**Condition :**  $100 * \frac{1}{6} * \frac{5}{6} = \frac{500}{36} = 13.88 > 5$  c'est bon le test du  $\chi^2$  est applicable.

	1	2	3	4	5	6
Effectifs	16	20	19	10	17	18
Proportions	0.16	0.2	0.19	0.1	0.17	0.18

- $H_0$  : dé non truqué  $\Leftrightarrow p^{ref} = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$
- $H_1$  : dé truqué  $p \neq p^{ref}$

On calcule

$$D = 100 * [\frac{(0.16 - \frac{1}{6})^2}{1/6} + \frac{(0.2 - \frac{1}{6})^2}{1/6} + \dots + \frac{(0.18 - \frac{1}{6})^2}{1/6}]$$

$$= 600 \sum_{k=1}^6 (\bar{p}_k - \frac{1}{6})^2 = 3.8$$

Pour faire un test à 90%, on doit comparer cette valeur avec le quantile d'ordre d'une loi  $\chi^2(6-1)$  degrés de liberté. Lecture de table :  $k = 9.24$ .

Ainsi comme  $D = 3.28 < 9.24$ , on conserve  $H_0$  le dé est équilibré

Nouveau cours du 03/02

### Bilan jusqu'à présent

Le test du  $\chi^2$  "basique" permet de tester l'adéquation de données iid  $X_1, \dots, X_n$  à valeurs dans  $\{x_1, \dots, x_d\}$  à une loi discrète sur  $\{x_1, \dots, x_d\}$  caractérisé par un vecteur de probabilité :  $p = (p_1, \dots, p_d)$

#### 2.3.0.3 Mise en place concrète :

1. Etape 0 : On vérifie les conditions

$$\forall k \in \{1, \dots, d\}, n * p_k \geq 5.$$

C'est la condition de Cochran (1954), il avait testé cas possible en observant l'approximation faites.

2. Etape 1 : On calcule les effectifs et proportions observées :  $N_{k,n}$  et  $\hat{p}_{k,n}$
3. Etape 2 : Calcul de la statistique de test

$$D = n \sum_{k=1}^d \frac{(\hat{p}_{k,n} - p_k)^2}{p_k}.$$

4. Etape 3 : Détermination de la zone de rejet au niveau  $\alpha$ . On lit  $h_\alpha$  le quantile d'ordre  $1 - \alpha$  de la loi  $\chi^2(d_1)$
5. Etape 4 : Décisions
  - si  $D > h_\alpha$ , on rejette  $H_0$  (au niveau  $\alpha$ ).
  - Si  $D \leq h_\alpha$  on conserve  $H_0$

**2.3.0.4 Test du  $\chi^2$  avec fusion des classes** Que fait-on si la condition  $np_k \geq 5$  n'est pas vérifiée? On fusionne des classes!

**Exemple 2.10.** On a observé des réponses à un questionnaire. On veut tester l'adéquation à la loi  $p = (\frac{1}{4}, \frac{1}{4}, \frac{7}{16}, \frac{1}{16})$  avec  $n = 40$

Modalité	1	2	3	4
Effectif	10	18	11	1

Vérification des conditions du test du  $\chi^2$

$$40 * p_1 = \frac{40}{4} = 10 > 5 \quad 40 * p_2 = \frac{40}{4} = 10 > 5 \quad 40 * p_3 = \frac{40 * 7}{16} = 17.5 > 5 \quad 40 * p_4 = \frac{40}{16} = 2.5 < 5 \text{ condition non vérifiée!}$$

On fusionne des colonnes de manière à remplir les conditions. On fusionne les colonnes 3 et 4 par exemple.

Modalité	1	2	3 ou 4
Effectif	10	18	12

La nouvelle probabilité de référence devient

$$p_{nouvelle}^{ref} = \left(\frac{1}{4}, \frac{1}{4}, \frac{7}{16} + \frac{1}{16}\right) = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right).$$

Nouvelle condition :

$$40 * p_1 = 10 > 540 * p_2 = 10 > 540 * p_3 = \frac{40}{2} = 20 > 5$$

Si on applique le test du  $\chi^2$  "de base", on obtient un test asymptotique de niveau  $\alpha$  pour le cas à 3 classes (fait avec un  $\chi^2(2)$ ), donc c'est aussi un test asymptotique de niveau  $\alpha$  pour le cas à 4 classes.

*Remarque.* Si on prend  $q = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$  qui appartient à  $H_1^4$  car  $q \neq p = (\frac{1}{4}, \frac{1}{4}, \frac{7}{16}, \frac{1}{16})$ . En fusionnant ce cas particulier, on se retrouve dans  $H_0^3$ .

$$q \in H_1^4 \rightarrow q^{reduit} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right) \in H_0^3.$$

On perd donc en information quand on fusionne des colonnes. La puissance du test se réduit car on se retrouve avec des cas dans  $H_1$  et dans  $H_0$ .

L'opération de fusion des colonnes permet toujours de construire un test de niveau asymptotique  $\alpha$  au détriment de la puissance.

## 2.4 Le test du $\chi^2$ pour une loi discrète

Données :  $X_1, \dots, X_n$  observation iid.

Loi cible à valeur dans  $\mathbb{N}$  caractérisée par

$$p = (p_k)_{k \in \mathbb{N}}.$$

Exemple pour une poisson

$$p_k = \frac{\lambda^k e^{-\lambda}}{k!}.$$

Est-ce que la loi des  $X_i$  est donnée par  $p$ ? C'est à dire

$$\forall k \in \mathbb{N}, P(X_1 = k) = p_k?$$

Valeur	0	1	2	3	4	5	6	...
Effectif	5	8	12	7	2	1	0	...

**Exemple 2.11.** "On ne peut pas faire un  $\chi^2$  avec une infinité de degrés de liberté" → On regroupe les classes à partir d'un certain rang On voudrait regarder  $np_0, np_1, np_2, np_3$  et pour la 4ème classe

Valeur	0	1	2	3	4 et plus
Effectif	5	8	12	7	4

$n(\sum_{k=4}^{+\infty} p_k)$ . Les classes sont déterminées afin que toutes les conditions soient satisfaites.

En pratique, on regarde à partir de quel indice la condition  $np_k < 5$  ne fonctionne plus, puis on regroupe à partir de la

### 2.4.0.1 En pratique Donnée : $X_1, \dots, X_n$

Loi cible :  $p = (p_k)_{k \in \mathbb{N}^*}$

1. Etape 0 : On détermine les classes en calculant  $np_1, np_2, \dots$  et ainsi de suite.
2. On regroupe les classes qui ne vérifient pas la condition
3. On calcule les effectifs de chaque classes  $N_{1,n}, \dots, N_{c-1,n}, N_{c,n}$  avec  $c$  l'effectif dans la classe agglomérée
4. On calcule les proportions observées  $\hat{p}_{k,n}$  et la stat de test  $D = n \sum_{k=1}^c \frac{(\hat{p}_{k,n} - p'_k)^2}{p'_k}$  où  $p'_k = p_k$  si  $k \leq c_1$  et  $p'_c = \sum_{k=c}^{+\infty} p_k$
5. On détermine la zone de rejet à l'aide du quantile d'ordre  $1 - \alpha$  d'une loi  $\chi^2(c-1)$ , noté  $h_\alpha$  Et on décide de conserver  $H_0$  si  $D \leq h_\alpha$ , on rejette sinon.

**2.4.0.2 Limite** Ce test permet de tester l'adéquation à n'importe quelle loi discrète au niveau  $\alpha$ . Cependant, dès lors qu'on regroupe des classes (ce qui est obligatoire ici) on perd la consistance du test.

## 2.5 Le test du $\chi^2$ pour une loi continue

Données :  $X_1, \dots, X_n$  iid.

Loi cible :  $L$  la loi d'une v.a.  $L$  (par exemple de densité  $g$ )

Idée : Transformer les données en les regroupant par paquets.

Soient  $I_1, \dots, I_d$  des intervalles qui forment une partition du support de  $L$ . (disjoints, dont l'union couvre toutes les valeurs de  $L$ ) Voir 4

Condition

$$\forall k \in \{1, \dots, d\} n * P(L \in I_k) \geq 5.$$

On crée de nouvelles variables  $Y_i$  : Numéro de l'intervalle dans lequel est  $X_i$

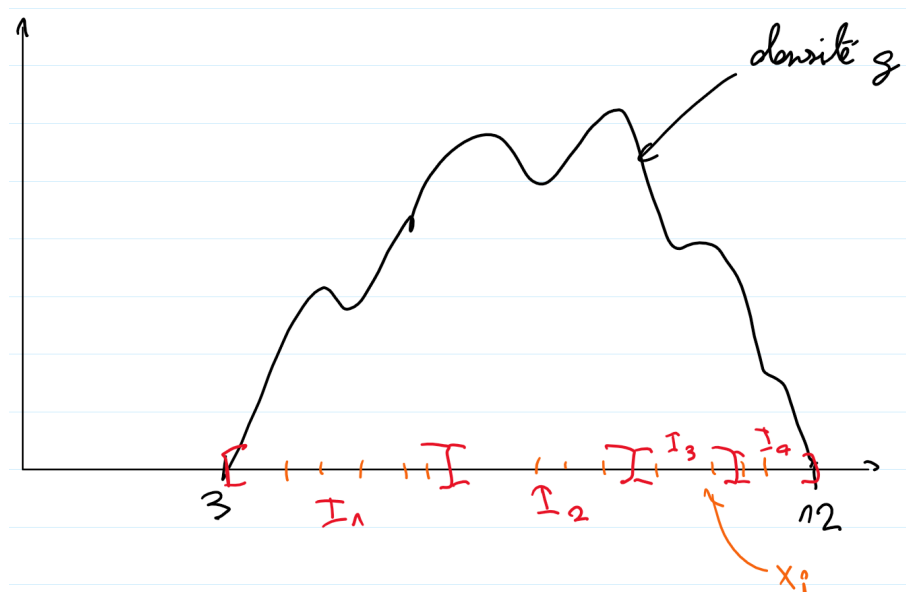


Figure 4 - Illustration de la partition de  $L$

$$P(Y_1 = k) = P(X_i \in I_k) = p_k \text{ (sous } H_0 \text{)}.$$

On a alors :  $Y_1, \dots, Y_n$  variables à valeur dans  $\{1, \dots, d\}$ , avec comme proba cible :  $p = (p_i = P(L \in I_i), \dots, p_d = P(L \in I_d))$ .

On applique alors le test du  $\chi^2$  "basique" aux variables  $Y_i$ . Cela fournit un test asymptotique de niveau  $\alpha$ . Le tableau à considérer est :

Intervale	$I_1$	$I_2$	$I_3$	...	$I_d$
Effectif	.	.	.	...	.

### 2.5.0.1 Bilan de la méthode Aspects positifs :

— **Fonctionne pour toutes les lois**

— Facile à faire

Aspects négatifs :

— Problème de consistance. Regrouper les variables par intervalle ruiner l'erreur de seconde espèce.

— Asymptotique

— Dépendant du choix des intervalles. Ce qui n'est pas canonique.

## 2.6 Le $\chi^2$ d'ajustement à une famille paramétrique de loi

On dispose d'observation iid.  $X_1, \dots, X_n$ .

On veut savoir si la loi des  $X_i$  fait partie d'une famille paramétrique  $\mathcal{F} = (P_\theta)_{\theta \in \Theta}$  à  $\Theta \subset \mathbb{R}^M$ .

Par exemple

- Lois de Poisson  $(Pois(\lambda))_{\lambda \in \mathbb{R}_*^+}$ ,  $M = 1$
- Lois Exponentielles :  $(\mathcal{E}(\lambda))_{\lambda \in \mathbb{R}_*^+}$ ,  $M = 1$
- Lois géométrique :  $(Geom(p))_{p \in ]0,1[}$ ,  $M = 1$
- Lois normales :  $(\mathcal{N}(m, \sigma^2))_{m \in \mathbb{R}, \sigma^2 \in \mathbb{R}_*^+}$ ,  $M = 2$

Les hypothèses :

- $H_0$  = la loi des  $X_i$  appartient à  $\mathcal{F}$
- $H_1$  = la loi des  $X_i$  n'appartient pas à  $\mathcal{F}$

1. Etape 1 : Soit  $\hat{\theta}_n$  l'estimateur du maximum de vraisemblance de  $\theta$  (pour  $P_\theta$ ). On estime **tous** les paramètres de la loi  $(p_1^{\hat{\theta}_n}, \dots, p_d^{\hat{\theta}_n})$
2. Etape 2 : On va tester l'ajustement de  $X_1, \dots, X_n$  à  $P_{\hat{\theta}_n}$ . On calcule les fréquences observées  $\hat{p}_{k,n}$ .

**Erreur à ne pas commettre** : il est faux de dire que

$$D = n \sum_{k=1}^d \frac{(\hat{p}_{k,n} - p_k^{\hat{\theta}_n})^2}{p_k^{\hat{\theta}_n}} \rightarrow \chi^2(d-1).$$

**Théorème 2.6.** Sous  $H_0$ ,

$$D = n \sum_{k=1}^d \frac{(\hat{p}_{k,n} - p_k^{\hat{\theta}_n})^2}{p_k^{\hat{\theta}_n}} \rightarrow \chi^2(d-1-M).$$

Avec

- $d$  = Nombre de classes à la fin, après regroupement éventuel
- $M$  = nombre de paramètre

### 2.6.0.1 En pratique

1. Etape 1 : Soit  $\hat{\theta}_n$  l'estimateur du maximum de vraisemblance de  $\theta$  (pour  $P_\theta$ ). On estime **tous** les paramètres de la loi  $(p_1^{\hat{\theta}_n}, \dots, p_d^{\hat{\theta}_n})$
2. Etape 2 : On va tester l'ajustement de  $X_1, \dots, X_n$  à  $P_{\hat{\theta}_n}$ . On calcule les fréquences observées  $\hat{p}_{k,n}$ .
3. Etape 3 : Vérification des conditions  $np_k^{\hat{\theta}_n}$  et possible regroupement en classes
4. Etape 4 : Calcul de la stat de test  $D$
5. Etape 5 : Zone de rejet : lecture de  $H_\alpha$  le quantile d'ordre  $1 - \alpha$  d'une  $\chi^2(d-1-M)$
6. Etape 6 : Décision
  - $D > h_\alpha$  on rejette  $H_0$
  - $D \leq h_\alpha$  on conserve  $H_0$

Nouveau cours du 10/02

**Exemple 2.12** (Test d'ajustement à une loi de Poisson). On dispose d'observation  $X_1, \dots, X_n$  iid. (représentant le nombre d'heure entre 2 pannes de métro). On veut tester pour savoir si les données proviennent d'une loi de Poisson.

*Remarque (Rappel).*  $Z \sim Pois(\lambda)$ ,  $P(Z = k) = \frac{\lambda^k e^{-\lambda}}{k!}$ ,  $\lambda \in \mathbb{R}^*$

- $H_0$  La loi des observation est  $Pois(\lambda)$  pour un certain  $\lambda > 0$
- $H_1$  la loi des  $X_i$  n'est pas une loi de Poisson

1. Estimer les paramètres (par un maximum de vraisemblance) :  
On rappelle (1er semestre) que l'EMV pour  $\lambda$  est

$$\bar{\lambda}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Données : Sur ces 100 données, on calcule  $\bar{\lambda}_n$  :

$$\bar{\lambda} = \frac{1}{100} (14 * 0 + 22 * 1 + \dots + 1 * 7) = 2.29.$$



Valeurs	0	1	2	3	4	5	6	7	8	9	...
Effectif	14	22	20	25	7	9	2	1	0	0	0

2. Calcul de la statistique de test comme si on faisait un  $\chi^2$  d'ajustement à une  $\mathcal{P}(2.29)$ .

Si  $Z \sim \mathcal{P}(2.29)$ ,  $P(X = k) = (2.29)^k \frac{e^{-2.29}}{k!} = p_k$ . On calcule les  $np_k$ . On détermine les classes à

k	0	1	2	3	4	5	6	7	...
$100p_k$	10.13	23.19	26.55	20.27	11.6	5.3	2.03	0.66	...

regrouper pour avoir  $np_k \geq 5$ . On se rend compte rapidement qu'il faut regrouper 5 à  $+\infty$ . Calcul

k	0	1	2	3	4	5 et +
$100p_k$	10.13	23.19	26.55	20.27	11.6	$100 - \sum \text{autres} = 8.26$

de la statistique de test :

$$D = 100 \sum_{k=0}^5 \frac{(p_{k,n} - p_k)^2}{p_k} = \sum_{k=0}^5 \frac{(N_{k,n} - 100p_k)^2}{100p_k} = 7.78.$$

3. Zone de rejet au niveau  $\alpha = 5\%$

On lit dans la table le quantile d'ordre  $1 - \alpha = 0.95$  de la loi  $\chi^2(6 - \text{Nombre de classe} - \text{Nombre de paramètre estimé})$

$\chi^2(6 - 1 - 1)$ . Ici  $k_{0.95} = 9.48$ .

CCL : Comme  $D = 7.78 \leq k_{0.95} = 9.48$  on conserve  $H_0$ .

*Remarque* (Chapitre 1). — Remarque sur le  $\chi^2$  d'ajustement à une formule paramétrique de lois :

Principe :

1. On estime

2. On calcule comme si on faisait un  $\chi^2$  d'ajustement à une seule loi

3. Attention au degrés de liberté dans la zone de rejet!

— Remarque sur la consistance : Si le nombre de classes utilisées tend vers  $+\infty$  quand  $n \rightarrow +\infty$ , le test du  $\chi^2$  est consistant.

## 2.7 Bilan du chapitre

On a deux tests d'ajustement : Kolmogorov-Smirnov et  $\chi^2$

— KS : ajustement à une loi de fdr. continue. Fonctionne pour toutes valeurs de  $n$ . Si  $n$  grand, on prend  $\frac{1}{\sqrt{n}}$  quantile de  $W_\infty$ .

**Attention :** Si  $n$  est grand sur des données réelles, KS est très sensible au bruit et rejette très souvent. Une erreur de 0.01 sur la fdr. des données mène vite à un ... si  $n \geq 10^5$ .

—  $\chi^2$  : Test asymptotique,  $n \geq 50$  au minimum + Condition. Fonctionne dans tous les cas.

## 3 Loi de comparaison

Dans ce chapitre, on dispose de deux jeux de données

—  $X_1, \dots, X_n$  avec  $n > 0$  iid.

—  $Y_1, \dots, Y_n$  avec  $n > 0$  iid.

On cherche à comparer les lois sous-jacentes.

1. Est-ce que les  $X_i$  et  $Y_j$  ont la même loi? (homogénéité)

2. Est-ce que les  $X_i$  sont indépendants des  $Y_j$  (indépendance)

3. Est-ce que les lois de  $X_i$  et  $Y_j$  ont la même moyenne ou la même médiane?

Deux cas de figure :

— Échantillon appariés : les  $X_i$  et  $Y_j$  proviennent d'une même mesure / tirage  $(X_i, Y_j)$ . Cela implique  $n = m$ .

**Exemple 3.1.** On mesure la taille et le poids de pluviomètres à Roubaix et à Croix

— Échantillons indépendants : si  $(X_1, \dots, X_n)$  est indépendant de  $(Y_1, \dots, Y_n)$ , on dira que les échantillons sont indépendants.

### 3.1 Le test d'homogénéité de Kolmogorov-Smirnov

On dispose des données iid.  $(X_1, \dots, X_n)$  et  $(Y_1, \dots, Y_n)$ . Les échantillons sont indépendants. On veut tester

- $H_0$  : les  $X_i$  et  $Y_i$  ont la même loi, c'est à dire  $F_{X_1} = F_{Y_1}$  où  $F_{X_1}, F_{Y_1}$  sont continues.
- $H_1$  les lois sont différentes

Comme pour le test d'ajustement de KS, on va construire un test non asymptotique se basant sur les fdr. empirique.

Notation :

$$\begin{aligned} F_n : \mathbb{R} &\rightarrow [0, 1] & G_n : \mathbb{R} &\rightarrow [0, 1] \\ t &\mapsto \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq t} & t &\mapsto \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{Y_j \leq t} \end{aligned}$$

#### Théorème 3.1. .

1. Si  $F_{X_1} = F_{Y_1}$  alors la variable

$$h(F_n, G_n) = \sup_{t \in \mathbb{R}} |F_n(t) - G_n(t)|.$$

a même loi que la variable

$$h_{n,m} = \sup_{s \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{U_i \leq s} - \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{V_j \leq s} \right|.$$

avec  $(U_1, \dots, U_n)$  et  $(V_1, \dots, V_n)$  sont deux échantillons indépendants de variable iid. uniformes sur  $[0, 1]$ .

2. De plus si  $F_{X_1} \neq F_{Y_1}$  alors

$$h(F_n, G_n) \xrightarrow{n, m \rightarrow \infty} \|F_{X_1} - F_{Y_1}\|_{\infty} = \sup_{t \in \mathbb{R}} |F_{X_1}(t) - F_{Y_1}(t)| > 0.$$

*Preuve :* (a) D'après le théorème de simulation par inversion de la fdr. si  $U_1, \dots, U_n$  sont des variables aléatoire iid. uniforme sur  $[0, 1]$ ,  $(F_{X_1}^{-1}(U_1), \dots, F_{X_1}^{-1}(U_n))$  a même loi que  $(X_1, \dots, X_n)$ . Si  $U_1, \dots, U_n$  sont des variables aléatoire iid. uniforme sur  $[0, 1]$ ,  $(F_{X_1}^{-1}(V_1), \dots, F_{X_1}^{-1}(V_n))$  a même loi que  $(Y_1, \dots, Y_n)$ .

Ainsi,  $h(F_n, G_n)$  a même loi que  $\sup_{s \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{F_{X_1}^{-1}(U_i) \leq s} - \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{F_{X_1}^{-1}(V_j) \leq s} \right|$

Par les propriétés classique de l'inverse généralisée,

$$\begin{aligned} F_{X_1}^{-1}(U_i) \leq t &\Leftrightarrow U_i \leq F_{X_1}(t) \\ F_{X_1}^{-1}(V_j) \leq t &\Leftrightarrow V_j \leq F_{X_1}(t) \end{aligned}$$

$$\text{Ainsi } A = \sup_{s \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{U_i \leq F_{X_1}(s)} - \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{V_j \leq F_{X_1}(s)} \right|$$

(b) Conséquence immédiate du théorème de Glivenko Cantelli.

□

#### 3.1.1 Test d'homogénéité de Kolmogorov-Smirnov

Données :  $(X_1, \dots, X_n)$  et  $(Y_1, \dots, Y_n)$  iid deux échantillon indépendants.  $H_0 : F_{X_1} = F_{Y_1}$  contre  $H_1 : F_{X_1} \neq F_{Y_1}$ .

Statistique de test : on calcule

$$\sup_{s \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq s} - \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{Y_j \leq s} \right|.$$

Zone de rejet au niveau  $\alpha$  :

Soit  $k_\alpha$  le quantile d'ordre  $1 - \alpha$  de la loi de

$$\sup_{s \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{U_i \leq s} - \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{V_j \leq s} \right|.$$

où  $(U_1, \dots, U_n) \perp (V_1, \dots, V_n) \text{ iid. } \sim U([0, 1])$

CCL : Si  $h(F_n, G_n) \leq k_\alpha$ , on conserve  $H_0$  au niveau  $\alpha$ . Sinon on rejette  $H_0$

Remarque. .

1. Ce test est de taille  $\alpha$ , si on utilise la table de  $h_{n,m}$ .
2. Si  $n$  et  $m$  sont trop grands, on utilise le résultat suivant :  
Sous  $H_0$

$$\sqrt{\frac{nm}{n+m}} h(F_n, G_n) \xrightarrow[n, m \rightarrow +\infty]{\alpha} W_\infty \text{ voir KS asymptotique.}$$

On utilise alors comme zone de rejet  $\sqrt{\frac{n+m}{nm}} W_\infty$  avec  $W_\infty$  le quantile d'ordre  $1 - \alpha$  de  $W_\infty$ .

### 3.1.1.1 En pratique (cas $n$ et $m$ grand) :

En R :  $X, Y$  vecteur, `ks.test(X,Y)`

A la main :

$$F_n(t) - G_n(t) = \frac{\text{nb de } X_i \leq t}{n} - \frac{\text{nb de } Y_i \leq t}{m}.$$

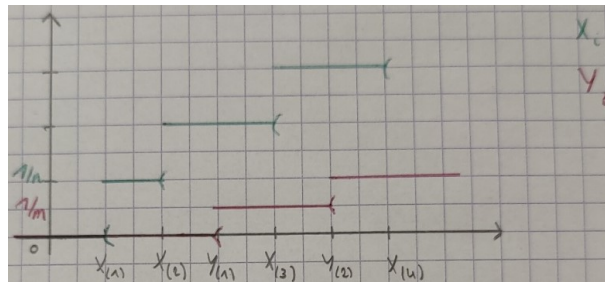


Figure 5 – <caption>

On range par ordre croissant :

$X_{(1)}$	$X_{(2)}$	$Y_{(1)}$	$X_{(3)}$	$Y_{(2)}$	$X_{(4)}$	...
$\frac{1}{n} - \frac{0}{m}$	$\frac{2}{n}$	$\frac{2}{n} - \frac{1}{m}$	$\frac{3}{n} - \frac{1}{m}$	$\frac{3}{n} - \frac{2}{m}$	$\frac{4}{n} - \frac{2}{m}$	...

Je calcule les  $n + m$  quantités et je garde la plus grande valeur.

Méthode inefficace :

$X_i$	$X_{(1)}$	$X_{(2)}$		$X_{(3)}$		$X_{(4)}$	
$Y_i$			$Y_{(1)}$		$Y_{(2)}$		
$F_n$	$\frac{1}{n}$	$\frac{2}{n}$	$\frac{2}{n}$	$\frac{3}{n}$	$\frac{3}{n}$	$\frac{4}{n}$	...
$G_m$	0	0	$\frac{1}{m}$	$\frac{1}{m}$	$\frac{2}{m}$	$\frac{2}{m}$	...