

# 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
3.2	Les test du $\chi^2$ d'indépendance et d'homogénéité	19
3.2.1	Le $\chi^2$ d'indépendance	19
3.2.2	Test du $\chi^2$ d'indépendance	21
3.3	Le $\chi^2$ d'homogénéité	22
3.3.1	Pour deux échantillons	22
<b>4</b>	<b>Tests pour échantillons gaussiens</b>	<b>24</b>
4.1	Rappels du cours de statistiques mathématiques	24
4.2	Forme d'un test	25
4.3	Les test sur l'espérance	25
4.3.1	Test sur la moyenne pour 1 échantillon gaussien de variance inconnue. Test de students à 1 échantillon	25
4.3.2	Test sur des moyenne pour 2 échantillons gaussiens appariés	25
4.3.3	Test d'égalité des moyennes pour 2 échantillons gaussiens indépendant de variance connues	26
4.4	Test sur les variances	27
4.4.1	Test d'égalité des variances pour un échantillon gaussien de moyenne inconnus	27
4.4.2	Test de comparaison des variances de Fisher	27
4.4.3	Test de Student à 2 échantillons : Test de comparaison des moyenne de 2 échantillons gaussiens indépendants de variance égale	28
4.4.4	Test de Welch : Le test de Student se généralisant au cas des variances non égales	29
<b>5</b>	<b>Test de Mann-Whitney-Wilcoxon ou test de la somme des rangs</b>	<b>29</b>
<b>6</b>	<b>Test du signe et test du signe et rang de Wilcoxon</b>	<b>32</b>
6.1	Test du signe / test de la médiane	32
6.2	Le test des rangs et signe de Wilcoxon	34
<b>7</b>	<b>Remarques finales</b>	<b>36</b>
7.1	La table de fisher	36
7.2	Table de Mann-Whitney et Wilcoxon	36
<b>8</b>	<b>Test bonus</b>	<b>36</b>
8.1	Le test d'indépendance de Pearson	37
8.2	Comparaison asymptotique de proportion	38
<b>9</b>	<b>Comparaison de <math>K \geq 3</math> échantillons</b>	<b>38</b>
9.1	L'ANOVA à un facteur	38

# 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 :

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

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

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

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.  $(Y_i = \mathbb{1}_{X_i \leq t})_{i \in \mathbb{N}}$

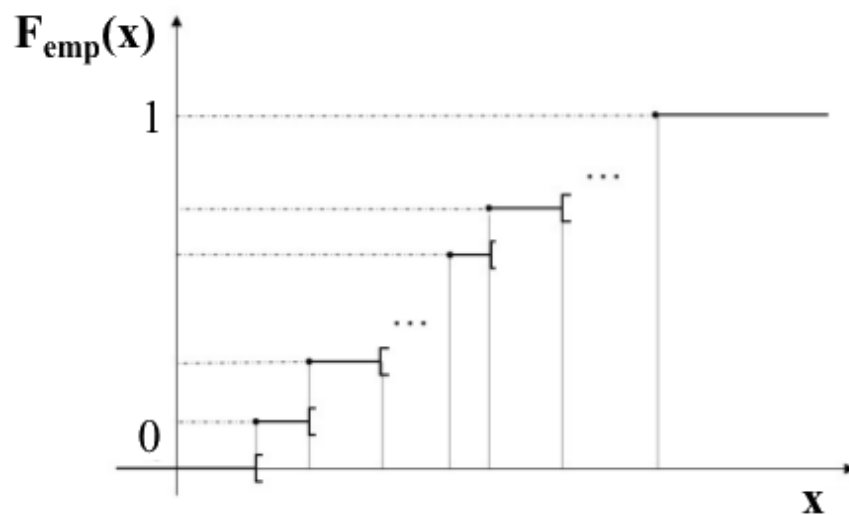


Figure 1 – Exemple de fonction de répartition empirique

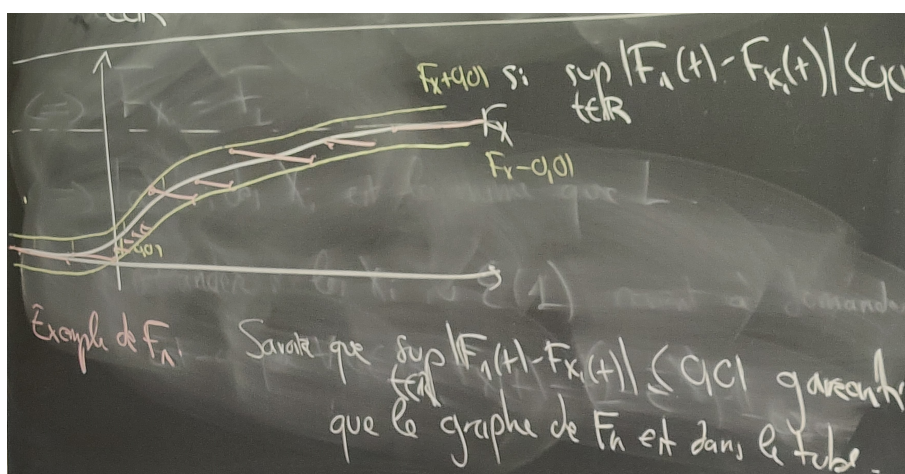


Figure 2 – Illustration graphique de Glivenko-Cantelli

**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  $h(F_n, F_U) = 0.22$

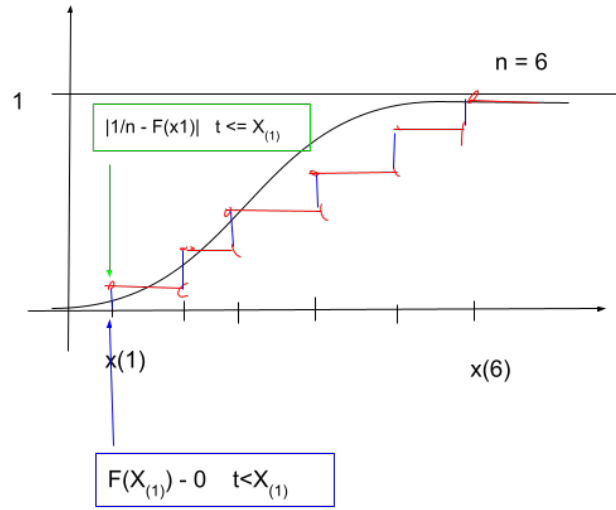


Figure 3 – Figure pour trouver la fonction  $h(F_n, F)$

$X_{(i)}$	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
$F(t) = 1 - e^{-x}$	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  $\chi^2$

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

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  $X_i$

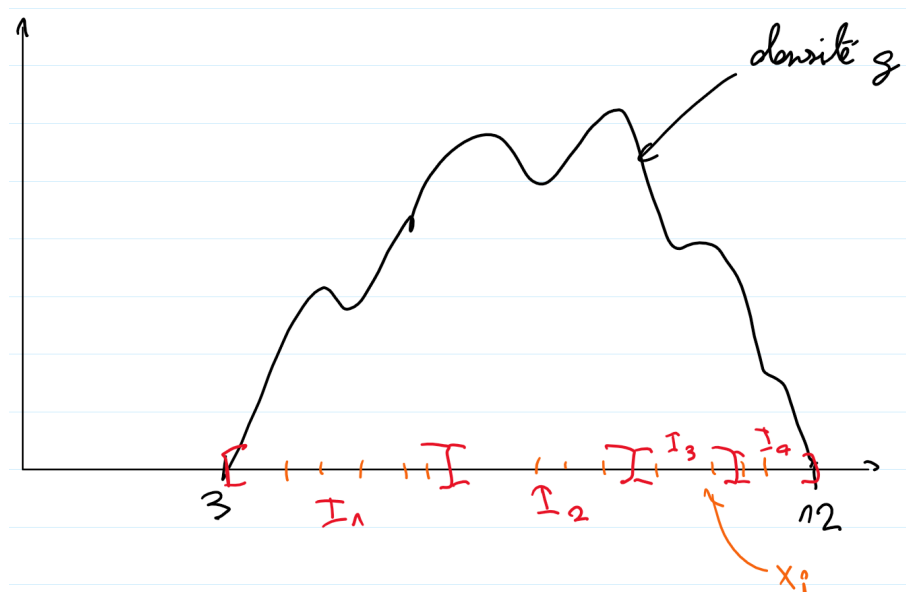


Figure 4 - Illustration de la partition de  $L$

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

On a alors :  $Y_1, \dots, Y_n$  variables à valeur dans  $\{1, \dots, d\}$ , avec comme proba cible :  $p = (p_1 = P(L \in I_1), \dots, p_d = P(L \in I_d))$ .

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 :

Intervalle	$I_1$	$I_2$	$I_3$	...	$I_d$
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

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

*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 :

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

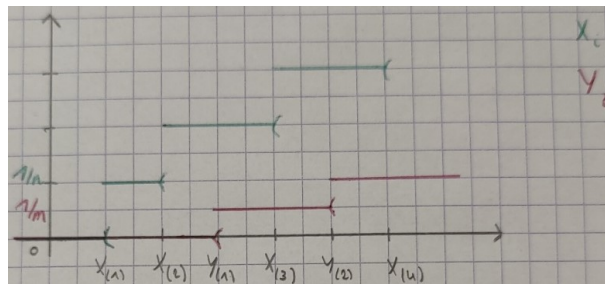


Figure 5 - <caption>

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 calcules 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	$\frac{1}{m}$	$\frac{1}{m}$	$\frac{2}{m}$	$\frac{2}{m}$	...

Nouveau cours du 03/03

## 3.2 Les test du $\chi^2$ d'indépendance et d'homogénéité

### 3.2.1 Le $\chi^2$ d'indépendance

**Rappel :** deux variables aléatoire réelle  $X, Y$  sont indépendante ssi

$$\forall A, B \subset \mathcal{B}(\mathbb{R}), P(X \in A \text{ et } Y \in B) = P(X \in A)P(Y \in B).$$

De manière informelle, la connaissance de  $X$  ne donne aucune information sur  $Y$

**Données :**

$(X_1, Y_1), \dots, (X_T, Y_T)$  données appariées **iid.** Cela mène à deux échantillons iid.  $X_1, \dots, X_T$  et  $Y_1, \dots, Y_T$ .

**Attention :**  $T$  = nombre totale de mesures

On veut déterminer si  $X_1$  est indépendant de  $Y_1 \Leftrightarrow X_1 \perp Y_1$ .

Ainsi on vas construire un test pour

—  $H_0 : X_1 \perp Y_1$

—  $H_0 : X_1 \not\perp Y_1$  ne sont pas indépendants

Quel genre de situation cela couvre-t-il? 2 exemples :

— Apparition d'effets secondaire pour un traitement

— Effet d'un facteur : réussite au bac en fonction du sexe?

**Important :** Les données sont à valeurs dans un nombre **fini** de classes :

—  $X_1, \dots, X_T$  à valeurs dans  $A_1, \dots, A_M$

—  $Y_1, \dots, Y_T$  à valeurs dans  $B_1, \dots, B_N$

Si ce n'est pas le cas, on s'y ramène en **créant** des classes comme pour les autres test du  $\chi^2$

La loi de  $X_1$  est caractérisé par

$$p_m = P(X_1 \in A_m) \text{ pour } m \in \{1, \dots, M\}.$$

de même pour  $Y_1$

$$q_n = P(Y_1 \in B_n) \text{ pour } n \in \{1, \dots, N\}.$$

Si on a accès à ces probabilités, **l'indépendance** se lit

$$\forall m \in \{1, \dots, M\}, \forall n \in \{1, \dots, N\}.$$

$$p_{m,n} = P(X_1 \in A_m \text{ et } Y_1 \in B_n) = p_m * q_n \text{ par indépendance.}$$

Malheureusement : Ni les  $p_{m,n}$ , ni les  $p_m$ , ni les  $q_n$  ne sont connus!  $\rightarrow$  On vas estimer ces quantités et construire un test à partir de ces estimateur.

**Notation :** Pour  $m \in \{1, \dots, M\}$  et  $n \in \{1, \dots, N\}$ . On pose

$$\begin{aligned} N_{m,n} &= \sum_{i=1}^T \mathbb{1}_{X_i \in A_m, Y_i \in B_n} \\ &= \text{effectif observé sur la classe } A_m * B_n \\ N_{m,\cdot} &= \sum_{i=1}^T \mathbb{1}_{X_i \in A_m} \\ &= \text{effectif total de } A_m \\ N_{\cdot,n} &= \sum_{i=1}^T \mathbb{1}_{Y_i \in B_n} \\ &= \text{effectif totale de } B_n \end{aligned}$$

*Remarque.* On a immédiatement

$$T = \sum_{n=1}^N N_{\cdot,n} = \sum_{m=1}^M N_{m,\cdot} = \sum_{n=1}^N \sum_{m=1}^M N_{m,n}.$$

Estimateurs : Pour  $m \in \{1, \dots, M\}$  et  $n \in \{1, \dots, N\}$

$$\hat{p}_{m,n} = \frac{N_{m,n}}{T} \approx p_{m,n} \text{ si } T \text{ est grand.}$$

$$\hat{p}_m = \frac{N_{m,\cdot}}{T} \approx p_m \text{ si } T \text{ est grand.}$$

$$\hat{q}_n = \frac{N_{\cdot,n}}{T} \approx q_n \text{ si } T \text{ est grand.}$$

### 3.2.2 Test du $\chi^2$ d'indépendance

**Données :**  $(X_1, Y_1), \dots, (X_T, Y_T)$  iid appariés.

—  $X_1$  à valeur dans  $A_1, \dots, A_M$

—  $Y_1$  à valeur dans  $B_1, \dots, B_N$

**Hypothèse :**

—  $H_0 : X_1 \perp Y_1$

—  $H_1 : X_1 \not\perp Y_1$

**Statistique de test**

$$D = T * \sum_{m=1}^M \sum_{n=1}^N \frac{(\hat{p}_{m,n} - \hat{p}_m \hat{q}_n)^2}{\hat{p}_m \hat{q}_n}$$

$$= \sum_{m=1}^M \sum_{n=1}^N \frac{(N_{m,n} - \frac{N_{m,\cdot} N_{\cdot,n}}{T})^2}{\frac{N_{m,\cdot} N_{\cdot,n}}{T}}$$

**Condition :** Si  $\forall m \in \{1, \dots, M\}$  et  $\forall n \in \{1, \dots, N\}$

$$T \hat{p}_m \hat{q}_n \geq 5.$$

alors  $D$  suit approximativement une loi

$$\chi^2(MN - 1 - (M - 1) - (N - 1))$$

$$\Leftrightarrow \chi^2(MN - 1 - \text{estimation de } p_m - \text{Estimation des } q_n)$$

$$\Leftrightarrow \chi^2(MN - M - N + 1)$$

$$\Leftrightarrow \chi^2((M - 1)(N - 1))$$

**Seuil de rejet :** Au niveau  $\alpha$ .

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

Si  $D > h_\alpha$ , on rejette  $H_0$ . Sinon on conserve  $H_0$ .

**Que se passe-t-il sous  $H_1$**

Si  $X_1$  et  $Y_1$  ne sont pas indépendants, il existe  $m_0$  et  $n_0$  tels que

$$p_{m_0, n_0} \neq p_{m_0} q_{n_0}.$$

Ainsi,

$$\frac{(\hat{p}_{m_0, n_0} - \hat{p}_{m_0} \hat{q}_{n_0})^2}{\hat{p}_{m_0} \hat{q}_{n_0}} \xrightarrow{T \rightarrow +\infty} \frac{(p_{m_0, n_0} - p_{m_0} q_{n_0})^2}{p_{m_0} q_{n_0}}.$$

Donc  $D \rightarrow +\infty$  (on a multiplié par  $T$ )

**Exemple 3.2.** Indépendance de la couleur des yeux et des cheveux.

On a mesuré sur 1000 personnes leurs couleurs de yeux et cheveux qu'on a regroupé dans le tableau suivant.

yeux \ cheveux	Noirs ( $A_1$ )	Bruns	Blonds	Roux	Total
Marrons ( $B_1$ )	$N_{1,1} = 152$	$N_{2,1} = 247$	83	11	$N_{\cdot,1} = 152$
Vert ou Gris	73	114	37	8	232
Bleus	36	167	127	10	275
Total	$N_{1,\cdot} = 261$	463	247	29	1000

Condition :  $T * \hat{p}_m \hat{q}_n \geq 5 \Leftrightarrow T * \frac{N_{m,\cdot} * N_{\cdot,n}}{T * T} = \text{effectif attendu}$

→ Vous calculez les conditions comme vous voulez (?) Tableau des effectifs attendu + regarder si on respecte les conditions

yeux \ cheveux	Noirs ( $A_1$ )	Bruns	Blonds	Roux	Total
Marrons ( $B_1$ )	128.67	228.26	121.77	14.3	$N_{.,1} = 152$
Vert ou Gris	60.55	107.42	57.3	6.73	232
Bleus	71.78	127.32	67.93	7.98	275
Total	$N_{1,.} = 261$	463	247	29	1000

Tous les effectifs attendus sont  $\geq 5$  : On peut appliquer le test du  $\chi^2$  d'indépendance

$$\begin{aligned}
D &= T \sum_{m=1}^M \sum_{n=1}^N \frac{(\hat{p}_{m,n} - \hat{p}_m \hat{q}_n)^2}{\hat{p}_m \hat{q}_n} \\
&= \sum_{m=1}^M \sum_{n=1}^N \frac{(N_{m,n} - \frac{N_{m,.} N_{.,n}}{T})^2}{\frac{N_{m,.} N_{.,n}}{T}} \\
&= \frac{(152 - 128.67)^2}{128.67} + \frac{(247 - 228.26)^2}{228.26} + \dots + \frac{(10 - 7.98)^2}{7.98} \\
&= \text{Une somme à 12 termes} \\
&= 104.01
\end{aligned}$$

**Zone de rejet :** Sous  $H_0$ ,  $D \sim \chi^2(12 - 13 - 2) = \chi^2(6)$ . Pour un test au niveau 5%, on lit le quantile d'ordre 95% d'une  $\chi^2(6) = 12.6$

**Conclusion :**  $D > 12.6$  On rejette  $H_0$  couleur d'yeux et couleurs de cheveux ne sont pas indépendants

### 3.3 Le $\chi^2$ d'homogénéité

#### 3.3.1 Pour deux échantillons

**Données :**  $X_1, \dots, X_{n_1}$  et  $Y_1, \dots, Y_{n_2}$  deux échantillons iid indépendants entre eux (comme pour Kolmogorov-Smirnov).

Les variables sont toutes à valeurs dans les mêmes classes  $A_1, \dots, A_M$ .

**Hypothèse :** On veut tester l'homogénéité

- $H_0 = X_1$  et  $Y_1$  ont la même loi  $\Leftrightarrow \forall m \in \{1, \dots, M\}, P(X_1 \in A_m) = P(Y_1 \in A_m)$
- $H_1 = X_1$  et  $Y_1$  n'ont pas la même loi  $\Leftrightarrow \exists m \in \{1, \dots, M\}$  tel que  $P(X_1 \in A_m) \neq P(Y_1 \in A_m)$

**Remarque** (Lien entre test du  $\chi^2$  d'indépendance et d'homogénéité). On peut se ramener à un test d'indépendance en construisant l'échantillon apparié suivant  $i \leq n_1 + n_2 = T$

$$(W_i, Z_i) = \begin{cases} (X_i, 1) & \text{si } i \leq n_1 \\ (Y_{i-n_1}, 2) & \text{si } i > n_1 \end{cases}$$

On est passé de :

- $X_1, \dots, X_{n_1}$  à  $(X_1, 1), (X_2, 1), \dots, (X_{n_1}, 1)$
- et  $Y_1, \dots, Y_{n_2}$  à  $(Y_1, 2), (Y_2, 2), \dots, (Y_{n_2}, 2)$

On a :

$$W_1 \perp Z_1 \Leftrightarrow X_1 \text{ et } Y_1 \text{ ont la même loi.}$$

Pour tester l'homogénéité des deux populations, il suffit de tester l'indépendance de  $Z_1$  et  $W_1$

- $W_1$  est à valeur dans  $A_1, \dots, A_M$
- $Z_1$  est à valeur dans  $\{1, 2\}$
- Vu comme ça le test se généralise très bien ! On peut l'utiliser pour comparer beaucoup plus d'échantillon ! (exo 5 TD6)

**Exemple 3.3** (Mise en pratique sans l'indépendance). On va tester si  $P(X_1 \in A_m) = P(Y_1 \in A_m) \forall m \in \{1, \dots, M\}$ .

Sous  $H_0$  les populations sont homogènes. On estime  $p_m = P(X_1 \in A_m) = P(Y_1 \in A_m)$  par

$$\hat{p}_m = \frac{N_m^X + N_m^Y}{n_1 + n_2}.$$

Avec  $N_m^X = \sum_{i=1}^{n_1} \mathbb{1}_{X_i \in A_m}$  et  $N_m^Y = \sum_{j=1}^{n_2} \mathbb{1}_{Y_j \in A_m}$ .

On pose alors  $\hat{p}_m^X = \frac{N_m^X}{n_1}$  et  $\hat{p}_m^Y = \frac{N_m^Y}{n_2}$ .

### Statistique de test :

$$D = n_1 \sum_{m=1}^M \frac{(\hat{p}_m^X - \hat{p}_m)^2}{\hat{p}_m} + n_2 \sum_{m=1}^M \frac{(\hat{p}_m^Y - \hat{p}_m)^2}{\hat{p}_m}$$

$$= \sum_{m=1}^M \frac{(N_m^X - n_1 \hat{p}_m)^2}{n_1 \hat{p}_m} + \sum_{m=1}^M \frac{(N_m^Y - n_2 \hat{p}_m)^2}{n_2 \hat{p}_m}$$

Si  $\forall m \in \{1, \dots, M\}, n_1 \hat{p}_m \geq 5$  et  $n_2 \hat{p}_m \geq 5$  alors  $D \sim \chi^2(M-1)$ .

**Seuil de rejet :** Au niveau  $\alpha$ , soit  $h_\alpha$  le quantile d'ordre  $1 - \alpha$  d'une  $\chi^2(M-1)$ . Si  $D > h_\alpha$ , on rejette  $H_0$ , sinon on conserve  $H_0$ .

Pop \ Groupe	O	A	B	AB	Total
Pop 1	121	120	79	33	353 = $n_1$
Pop 2	118	95	121	30	364 = $n_2$
Total	239	215	200	63	717

**Exemple 3.4** (Groupe sanguins dans 2 populations). Validité :  $n_1 \hat{p}_1, n_2 \hat{p}_1 \geq 5$ . Les calculs sont les mêmes !

$$D = \frac{(121 - \frac{353 \cdot 239}{717})^2}{\frac{353 \cdot 239}{717}} + \frac{(120 - \frac{353 \cdot 215}{717})^2}{\frac{353 \cdot 215}{717}} + \dots + \frac{(118 - \frac{364 \cdot 239}{717})^2}{\frac{364 \cdot 239}{717}} + \dots + \frac{(30 - \frac{364 \cdot 63}{717})^2}{\frac{364 \cdot 63}{717}}$$

$$\approx 11.75$$

On lit le quantile d'une loi  $\chi^2(3)$  et on décide. Faites le calcul et finissez.

**Exemple 3.5** (Pour un nombre quelconque de population). **Donnée :**

$$X_1^{(1)}, \dots, X_{n_1}^{(1)}$$

$$X_1^{(2)}, \dots, X_{n_1}^{(2)}$$

$$\dots$$

$$X_1^{(K)}, \dots, X_{n_1}^{(K)}$$

On a  $K$  échantillon indépendants de variable iid à valeur dans  $A_1, \dots, A_m$

#### Hypothèse :

- $H_0$  Tout les échantillons ont la même loi
- $H_1$  Il existe un échantillons qui diffère des autres

*Remarque.* On peut créer l'échantillon apparié fictif :

$$(W_i, Z_i) = (X_i^{(k)}, k).$$

et tester l'indépendance de  $Z_1$  et  $W_i$

Ou bien on utilise

$$N_m^{(k)} = \sum_{i=1}^{n_k} \mathbb{1}_{X_i^{(k)} \in A_m}$$

$$\hat{p}_m = \frac{N_m^{(1)} + \dots + N_m^{(K)}}{n_1 + \dots + n_K}$$

$$D = \sum_{k=1}^K \sum_{m=1}^M \frac{(N_m^{(k)} - n_k \hat{p}_m)^2}{n_k \hat{p}_m}$$

Condition : Si  $\forall k \leq K, \forall m \leq M : n_k \hat{p}_m \geq 5$  alors  $D \sim \chi^2((M-1)(K-1))$ .

Seuil de rejet : Quantile d'une  $\chi^2((M-1)(K-1))$ .

Sous  $H_1, D \rightarrow +\infty$  EXO (on avait plus le temps)

Nouveau cours du 10/03

## 4 Tests pour échantillons gaussiens

### 4.1 Rappels du cours de statistiques mathématiques

**Théorème 4.1** (Cochran).  $X_1, \dots, X_n$  v.a. iid. de loi  $\mathcal{N}(m, \sigma^2)$  alors

—  $\bar{X}_n$  et  $V_n$  sont indépendant à

$$\bar{X}_n \frac{1}{n} \sum_{i=1}^n X_i, V_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

—

$$\frac{(n-1)V_n}{\sigma^2} \sim \chi^2(n-1).$$

**Théorème 4.2** (Student).  $X_1, \dots, X_n$  v.a. iid.  $\mathcal{N}(m, \sigma^2)$  alors

$$\frac{\sqrt{n}}{\sqrt{V_n}}(\bar{X}_n - m) \sim \mathcal{T}(n-1).$$

#### Rappel

— La loi  $\chi^2(n)$  est la loi de

$$\sum_{i=1}^n Y_i^2 \text{ à } Y_i \text{ iid. } \mathcal{N}(0, 1).$$

— La loi  $\mathcal{T}(n)$  est la loi de

$$\frac{X}{\sqrt{V/n}} \text{ où } X \sim \mathcal{N}(0, 1), V \sim \chi^2(n).$$

— Opération sur les gaussiennes :  $X \sim \mathcal{N}(m_1, \sigma_1^2), Y \sim \mathcal{N}(m_2, \sigma_2^2), X \perp Y$  **indépendant** alors :

$$X + Y \sim \mathcal{N}(m_1 + m_2, \sigma_1^2 + \sigma_2^2)$$

$$X - Y \sim \mathcal{N}(m_1 - m_2, \sigma_1^2 + \sigma_2^2)$$

$$\lambda X \sim \mathcal{N}(\lambda m_1, \lambda^2 \sigma_1^2)$$

*Remarque.* Que se passe-t-il si  $X$  et  $Y$  ne sont pas indépendantes ?

$X - Y$  n'est a priori pas gaussienne. Cependant si  $(X, Y)$  est gaussien sur  $\mathbb{R}^2$  (vecteur gaussien) alors  $X - Y$  est gaussien d'espérance  $m_1 - m_2$  mais de variance **inconnue**.

$(X, Y)$  gaussien si sa densité est de la forme

$$g(x, y) = \frac{1}{c} e^{-ax^2 - by^2 - 2cxy}.$$

*Exemple 4.1* (Pas à savoir et un peu dur).

$$X \sim \mathcal{N}(0, 1)$$

$$Z \sim \mathcal{N}(0, 1) \perp X$$

$$B \sim \text{Ber}\left(\frac{1}{2}\right)$$

$$Y = BX + (1 - B)Z \sim \mathcal{N}(0, 1)$$

$$X - Y = \begin{cases} 0 & \text{si } B = 1 \\ X - Z & \text{si } B = 0 \end{cases}$$

$$X - Y \text{ n'est clairement pas gaussienne } P(Y \leq t) = P(Y \leq t \text{ et } B = 1) + P(Y \leq t \text{ et } B = 0)$$

$$= P(X \leq t) \frac{1}{2} + P(Z \leq t) \frac{1}{2} = P(X \leq 1)$$



## 4.2 Forme d'un test

- Nom du test / sa fonction
- Type de données / Condition d'utilisation
- $H_0, H_1$
- Statistique de test : sous  $H_0$  et  $H_1$
- Forme de la zone de rejet et le seuil de rejet au niveau  $\alpha$

*Remarque.* Je vous encourage fortement à revoir tous vos tests sous cette forme et à faire des fiche

## 4.3 Les test sur l'espérance

Dans le TD7, vous avez vu plusieurs test "élémentaire" sur les données gaussiennes. Ils sont à connaître et sont succinctement rappelé ici

### 4.3.1 Test sur la moyenne pour 1 échantillon gaussien de variance inconnue. Test de students à 1 échantillon

**Données**  $X_1, \dots, X_n$  iid.  $\mathcal{N}(m, \sigma^2)$  avec  $\sigma^2$  inconnu

#### Hypothèse

- $H_0 = m = m_0$
- $H_1 = m \neq m_0$  (**cas 1**) ou bien  $m > m_0$  (**cas 2**) ou bien  $m < m_0$  (**cas 3**)

#### Statistique de test

$$D = \frac{\sqrt{n}}{\sqrt{V_n}}(\bar{X}_n - m_0).$$

- Sous  $H_0, D \sim \mathcal{T}(n-1)$
- Sous  $H_1$

$$\begin{aligned} D &= \frac{\sqrt{n}}{\sqrt{V_n}}(\bar{X}_n - m) + \frac{\sqrt{n}}{\sqrt{V_n}} \\ &= \mathcal{T}(n-1) + \text{Biais? du signe de } m_1 - m_0 \end{aligned}$$

**Zone de rejet pour le cas 1** Soit  $h_{\alpha/2}$  le quantile d'ordre  $\frac{\alpha}{2}$  et  $h_{1-\alpha/2}$  le quantile d'ordre  $1 - \frac{\alpha}{2}$  de la loi  $\mathcal{T}(n-1)$

Si  $D > h_{1-\alpha/2}$  ou  $D < h_{\alpha/2}$ , on rejette  $H_0$

*Remarque (Attention).* Comme  $h_{\alpha/2} = -h_{1-\alpha/2}$  cela se ré-écrit  $|D| > h_{1-\alpha/2}$

#### Zone de rejet pour le cas 2

$$D = \mathcal{T}(n-1) + \frac{\sqrt{n}}{\sqrt{V_n}}(m - m_0) \text{ (biais } > 0 \text{)}.$$

Soit  $h_{1-\alpha}$  le quantile d'ordre  $1 - \alpha$  d'une loi  $\mathcal{T}(n-1)$ .

Si  $D > h_{1-\alpha}$  on rejette  $H_0$ . Sinon on conserve  $H_0$ .

**Zone de rejet pour le cas 3** Sous  $H_1 = m < m_0$ ,  $D$  prend des valeurs plutôt négatives. Soit  $h_\alpha$  le quantile d'ordre  $\alpha$  d'une  $\mathcal{T}(n-1)$ .

Si  $D < h_\alpha$  on rejette  $H_0$  sinon conserver  $H_0$ .

### 4.3.2 Test sur des moyenne pour 2 échantillons gaussiens appariés

#### Données

- $X_1, \dots, X_n$  iid.  $\mathcal{N}(m_1, \sigma_1^2)$ ,  $\sigma_1^2$  inconnus
- $Y_1, \dots, Y_n$  iid.  $\mathcal{N}(m_2, \sigma_2^2)$ ,  $\sigma_2^2$  inconnus
- Échantillon apparié  $(X_i, Y_i)$  indépendant si  $i \neq j$  mais  $X_i$  n'est pas indépendant de  $Y_i$

### Hypothèse

- $H_0 = m_1 = m_2$
- $H_1 = m_1 \neq m_2$  ou  $m_1 > m_2$  ou  $m_1 < m_2$

**Statistique de test** Si on pose  $\bar{Z}_n = \frac{1}{n} \sum_{i=1}^n Z_i$  et  $V_n = \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z}_n)^2$

$$D = \frac{\sqrt{n}}{\sqrt{V_n}} \bar{Z}_n.$$

### Zone de rejet

- Sous  $H_0$ ,  $D \sim \mathcal{T}(n-1)$
- Sous  $H_1$  à compléter vous-même, c'est le même que le test précédent

**Vois exo2 du TD7 :**

**Correction de l'exo2 du TD7** On suppose que les  $Z_i$  sont iid gaussiennes! donc

$$\begin{aligned} E(Z_i) &= E(X_i) - E(Y_i) = m_1 - m_2 \\ \text{Var}(Z_i) &= \dots\dots\text{MANQUE DES TRUCS DEMANDER A JESS} \end{aligned}$$

Zone de rejet de niveau  $\alpha$  :

$$\mathcal{R} = \{|T| \geq t_{n-1}(1 - \alpha/2)\}.$$

Sous  $H_0$ ,  $T \sim \mathcal{T}(n-1)$ .

Application : au niveau 10%  $\mathcal{R} = \{T \geq 1.3\}$  et  $T = \sqrt{46} \frac{1.5}{\sqrt{8}} = 3.6$  Donc on rejette  $H_0$  : la note de stat math est plus grande que la note de simulation.

### 4.3.3 Test d'égalité des moyennes pour 2 échantillons gaussiens indépendant de variance connues

#### Données

- $X_1, \dots, X_{n_1}$  iid.  $\mathcal{N}(m_1, \sigma_1^2)$ ,  $\sigma_1^2$  connue
- $Y_1, \dots, Y_{n_2}$  iid.  $\mathcal{N}(m_2, \sigma_2^2)$ ,  $\sigma_2^2$  connue
- $(X_1, \dots, X_{n_1}) \perp (Y_1, \dots, Y_{n_2})$

### Hypothèse

- $H_0 = m_1 = m_2$
- $H_1 = m_1 \neq m_2$  ou  $m_1 > m_2$  ou  $m_1 < m_2$

**Statistique de test** Voir TD7 exo 3 : Correction exo 3 TD7

1.  $\bar{X}_{n_1} \sim \mathcal{N}(m_1, \frac{\sigma_1^2}{n_1})$  et  $\bar{Y}_{n_2} \sim \mathcal{N}(m_2, \frac{\sigma_2^2}{n_2})$
2.  $\bar{X}_{n_1} - \bar{Y}_{n_2} \sim \mathcal{N}(m_1 - m_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$
3.  $H_0 : m_1 = m_2, H_1 = m_1 \neq m_2, T = \bar{X}_{n_1} - \bar{Y}_{n_2}, \mathcal{R} = \{|T| \geq C_\alpha\}$  car sous  $H_0, T \sim \mathcal{N}(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

$$\begin{aligned} P(|\bar{X}_{n_1} - \bar{Y}_{n_2}| \geq C_\alpha) &= P\left(\frac{|\bar{X}_{n_1} - \bar{Y}_{n_2}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq \frac{C_\alpha}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) \\ &= \phi^{-1}(1 - \alpha/2) \end{aligned}$$

$$\text{donc } c_\alpha = \phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

4. Application :

- Lille : 10 mesures,  $\sigma^2 = 4$
- Sydney : 20 mesures,  $\sigma^2 = 9$
- $\bar{X}_{n_1} - \bar{Y}_{n_2} = 1$
- Sous  $H_0, \bar{X}_{n_1} - \bar{Y}_{n_2} \sim \mathcal{N}(0, \frac{4}{10} + \frac{9}{20} = 0.85)$

— Sous  $H_1$ ,  $\bar{X}_{n_1} - \bar{Y}_{n_2}$  est plus grand que sous  $H_0$

$$\mathcal{R} = \{\bar{X}_{n_1} - \bar{Y}_{n_2} \geq c_\alpha\}.$$

$$P_{H_0}(\bar{X}_{n_1} - \bar{Y}_{n_2} \geq C_\alpha) = P_{H_0}\left(\frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \geq \frac{c_\alpha}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

$$\text{donc } \frac{c_\alpha}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 2.06 \Leftrightarrow c_\alpha = 2.06\sqrt{0.85} = 1.89$$

Donc on conserve  $H_0$ , les deux composés peuvent avoir la même masse.

## Zone de rejet

### 4.4 Test sur les variances

#### 4.4.1 Test d'égalité des variances pour un échantillon gaussien de moyenne inconnue

**Données**  $X_1, \dots, X_n$  iid.  $\mathcal{N}(m, \sigma^2)$  avec  $m, \sigma^2$  inconnus

#### Hypothèse

- $H_0 = \sigma^2 = \sigma_0^2$
- $H_1 = \sigma^2 \neq \sigma_0^2$  ou  $\sigma^2 > \sigma_0^2$  ou  $\sigma^2 < \sigma_0^2$

#### Statistique de test

$$D = \frac{(n-1)}{\sigma_0^2} V_n.$$

**Zone de rejet** à compléter (attention  $\chi^2$  pas symétrique)

#### 4.4.2 Test de comparaison des variances de Fisher

**Définition 4.1** (Loi de Fisher). Soit  $V \sim \chi^2(d_1)$ ,  $W \sim \chi^2(d_2)$ ,  $V \perp W$ . La loi de  $\frac{V/d_1}{W/d_2}$  est appelée loi de Fisher à  $(d_1, d_2)$  degrés de liberté.

*Remarque.* Cette loi est tabulée pour  $d_1$  et  $d_2$  pas trop grands.

- Elle admet une densité
- Elle est importante car elle sert souvent (en ANOVA notamment)

On la note  $\mathcal{F}(d_1, d_2)$

#### Données

- $X_1, \dots, X_{n_1}$  iid.  $\mathcal{N}(m_1, \sigma_1^2)$   $m_1$  et  $\sigma_1^2$  inconnus
- $Y_1, \dots, Y_{n_2}$  iid.  $\mathcal{N}(m_2, \sigma_2^2)$   $m_2$  et  $\sigma_2^2$  inconnus
- $(X_1, \dots, X_{n_1}) \perp (Y_1, \dots, Y_{n_2})$  échantillons indépendants

#### Hypothèse

- $H_0 = \sigma_1^2 = \sigma_2^2$
- $H_1 = \sigma_1^2 \neq \sigma_2^2$  ou  $\sigma_1^2 > \sigma_2^2$  ou  $\sigma_1^2 < \sigma_2^2$

#### Statistique de test

$$D = \frac{V_{n_1}^X}{V_{n_2}^Y}.$$

avec  $V_{n_1}^X = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (X_i - \bar{X}_{n_1})^2$  et  $V_{n_2}^Y = \frac{1}{n_2-1} \sum_{i=1}^{n_2} (Y_i - \bar{Y}_{n_2})^2$

### Zone de rejet

- Sous  $H_0, \sigma^2 = \sigma_1^2 = \sigma_2^2$  :

$$D = \frac{\frac{V_{n_1}^X(n-1)}{\sigma^2} \frac{1}{n_1-1}}{\frac{V_{n_2}^Y(n_2-1)}{\sigma^2} \frac{1}{n_2-1}} \sim \mathcal{F}(n_1-1, n_2-1) \text{ par Cochran.}$$

- Sous  $H_1, \sigma_1^2 \neq \sigma_2^2$

$$D = \mathcal{F}(d_1, d_2) * \frac{\sigma_1^2}{\sigma_2^2}.$$

- **Cas 1** :  $H_1 : \sigma_1^2 \neq \sigma_2^2$  : Soit  $h_{\alpha/2}$  le quantile  $\frac{\alpha}{2}$  d'une  $\mathcal{F}(n_1-1, n_2-1)$  et  $h_{1-\alpha/2}$  le quantile d'ordre  $1 - \frac{\alpha}{2}$  d'une  $\mathcal{F}(n_1-1, n_2-1)$ .  
Si  $D > h_{1-\alpha/2}$  ou bien  $D < h_{\alpha/2}$  on rejette  $H_0$ . Sinon on conserve  $H_0$
- **Cas 2** :  $H_1 : \sigma_1^2 > \sigma_2^2$  : Soit  $h_{1-\alpha}$  le quantile  $1 - \alpha$  d'une  $\mathcal{F}(n_1-1, n_2-1)$   
Si  $D > h_{1-\alpha}$  on rejette  $H_0$ . Sinon on conserve  $H_0$ .
- **Cas 3** :  $H_1 : \sigma_1^2 < \sigma_2^2$  : Soit  $h_\alpha$  le quantile  $\alpha$  d'une  $\mathcal{F}(n_1-1, n_2-1)$   
Si  $D < h_\alpha$  on rejette  $H_0$ . Sinon on conserve  $H_0$ .

### 4.4.3 Test de Student à 2 échantillons : Test de comparaison des moyenne de 2 échantillons gaussiens indépendants de variance égale

#### Données

- $X_1, \dots, X_{n_1}$  iid.  $\mathcal{N}(m_1, \sigma^2)$ ,  $m_1$  inconnus
- $Y_1, \dots, Y_{n_2}$  iid.  $\mathcal{N}(m_2, \sigma^2)$ ,  $m_2$  inconnus (même variance)
- Échantillon indépendant  $(X_1, \dots, X_{n_1}) \perp (Y_1, \dots, Y_{n_2})$

#### Hypothèse

- $H_0 = m_1 = m_2$
- $H_1 = m_1 \neq m_2$  ou  $m_1 > m_2$  ou  $m_1 < m_2$

#### Statistique de test

$$D = \frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{\sqrt{W}}.$$

avec  $W = \frac{(n_1-1)V_{n_1}^X + (n_2-1)V_{n_2}^Y}{n_1+n_2-2}.$

#### Zone de rejet

- Sous  $H_0 : m_1 = m_2$

$$\bar{X}_{n_1} \sim \mathcal{N}(m_1, \frac{\sigma^2}{n_1})$$

$$\bar{Y}_{n_2} \sim \mathcal{N}(m_2, \frac{\sigma^2}{n_2})$$

$$\bar{X}_{n_1} - \bar{Y}_{n_2} \sim \mathcal{N}(0, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))$$

Donc

$$\frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sigma} (\bar{X}_{n_1} - \bar{Y}_{n_2}) \sim \mathcal{N}(0, 1).$$

On a déjà un terme de la stat de test, ce qu'il reste

$$\frac{(n_1-1)V_{n_1}^X}{\sigma^2} + \frac{(n_2-1)V_{n_2}^Y}{\sigma^2} \sim \chi^2(n_1+n_2-2)$$

$$\chi^2(n_1-1) \perp \chi^2(n_2-1)$$

$$\begin{aligned}\sqrt{W} &= \frac{\sigma \sqrt{(n_1 - 1)V_{n_1}^X + (n_2 - 1)V_{n_2}^Y}}{\sigma \sqrt{n_1 + n_2 - 2}} \\ &= \sigma \sqrt{\frac{\chi^2(n_1 + n_2 - 2)}{n_1 + n_2 - 2}}\end{aligned}$$

Ainsi sous  $H_0$

$$D \stackrel{\text{loi}}{=} \frac{\mathcal{N}(0, 1)}{\sqrt{\frac{\chi^2(n_1 + n_2 - 2)}{n_1 + n_2 - 2}}} \sim \mathcal{T}(n_1 + n_2 - 2).$$

CCL à retenir : sous  $H_0$

$$D \sim \mathcal{T}(n_1 + n_2 - 2).$$

- Sous  $H_1$  :
  - Si  $m_1 > m_2$ ,  $D$  prend des valeurs plus grandes qu'une Student à  $n_1 + n_2 - 2$  degrés de libertés
  - Si  $m_1 < m_2$ ,  $D$  prend des valeurs négatives
- **Cas 1** :  $H_1 : m_1 \neq m_2$  : Soit  $h_{\alpha/2}$  le quantile  $\frac{\alpha}{2}$  d'une  $\mathcal{T}(n_1 + n_2 - 2)$  et  $h_{1-\alpha/2}$  le quantile d'ordre  $1 - \frac{\alpha}{2}$  d'une  $\mathcal{T}(n_1 + n_2 - 2)$ . **Attention**  $h_{\alpha/2} = -h_{1-\alpha/2}$   
 Si  $|D| > h_{1-\alpha/2}$  on rejette  $H_0$ . Sinon on conserve  $H_0$
- **Cas 2** :  $H_1 : m_1 > m_2$  : Soit  $h_{1-\alpha}$  le quantile  $1 - \alpha$  d'une  $\mathcal{T}(n_1 + n_2 - 2)$   
 Si  $D > h_{1-\alpha}$  on rejette  $H_0$ . Sinon on conserve  $H_0$ .
- **Cas 3** :  $H_1 : m_1 < m_2$  : Soit  $h_{\alpha}$  le quantile  $\alpha$  d'une  $\mathcal{T}(n_1 + n_2 - 2)$   
 Si  $D < h_{\alpha} = -h_{1-\alpha}$  on rejette  $H_0$ . Sinon on conserve  $H_0$ .

#### 4.4.4 Test de Welch : Le test de Student se généralisant au cas des variances non égales

##### Données

- $X_1, \dots, X_n$  iid.  $\mathcal{N}(m_1, \sigma_1^2)$ ,  $m_1$  et  $\sigma_1^2$  inconnus
- $Y_1, \dots, Y_n$  iid.  $\mathcal{N}(m_2, \sigma_2^2)$ ,  $m_2$  et  $\sigma_2^2$  inconnus
- Échantillons indépendants

##### Hypothèse

- $H_0 = m_1 = m_2$
- $H_1 = m_1 \neq m_2$  ou  $m_1 > m_2$  ou  $m_1 < m_2$

##### Statistique de test

$$D = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{\sqrt{\frac{V_{n_1}^X}{n_1} + \frac{V_{n_2}^Y}{n_2}}}.$$

##### Zone de rejet

- Sous  $H_0$ ,  $D$  suit **approximativement** une loi  $\mathcal{T}(\mu)$ .  $\mu$  n'est pas connu et est approximé par des formules horribles
  - Sous  $H_1$
- Nouveau cours du 17/03

## 5 Test de Mann-Whitney-Wilcoxon ou test de la somme des rangs

**Définition 5.1** (Ordre Stochastique). Soit  $X$  et  $Y$  deux variables aléatoires, on dit que  $Y$  domine stochastiquement  $X$  si

$$\forall t \in \mathbb{R}, F_Y(t) \leq F_X(t).$$

Cela équivaut à

$$\forall t \in \mathbb{R}, P(X > t) \leq P(Y > t).$$

Si  $Y \succ X$  et  $Y \neq X$  alors

$$E(Y) > E(X).$$

et si on note  $m_X$  et  $m_Y$  les médianes de  $X$  et  $Y$  on a également

$$m_X \leq m_Y.$$

Ce test est proche de KS à deux échantillons, en pratique il est même mieux.

### Données

- $X_1, \dots, X_{n_1}$  iid.
- $Y_1, \dots, Y_{n_2}$  iid.
- Échantillons indépendants
- On suppose que  $F_X$  et  $F_Y$  sont **continues**.

### Hypothèse

- $H_0 = X_1$  et  $Y_1$  ont la même loi.  $F_{X_1} = F_{Y_1}$
- $H_0 = X_1$  et  $Y_1$  n'ont pas la même loi.  $F_{X_1} \neq F_{Y_1}$ 
  - Ou  $X_1 \succ Y_1$  C'est à dire  $F_{X_1} \neq F_{Y_1}$  et  $\forall t \in \mathbb{R}, F_{Y_1}(t) \leq F_{X_1}(t)$
  - Ou  $Y_1 \succ X_1$  C'est à dire  $F_{X_1} \neq F_{Y_1}$  et  $\forall t \in \mathbb{R}, F_{X_1}(t) \leq F_{Y_1}(t)$

**Statistique de test** On note  $n = n_1 + n_2$ . On crée le vecteur

$$Z = (X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}) = (Z_1, \dots, Z_n).$$

$Z$  est la concaténation des deux échantillons. On ordonne  $Z$  par ordre croissant  $Z' = (Z_{(1)}, \dots, Z_{(n)})$  et on pose  $\forall i \in \{1, \dots, n_1\}$

$$\begin{aligned} R(i) &= \text{Rang de } X_i \text{ dans } Z' \\ &= \sum_{j=1}^n \mathbb{1}_{X_i \leq Z_j} \end{aligned}$$

On pose finalement la stat de test suivant

$$U = \sum_{i=1}^{n_1} R(i) = \text{la somme des rangs des } X_i \text{ dans } Z'.$$

*Remarque.* En cas d'ex-æquo, on leur attribue le rang moyen des rangs. Voir exemple.

C'est ici la puissance de ce test par rapport à KS, il départage les ex-æquo d'une manière mathématique, contrairement à KS

En général  $U$  est à valeurs entre

$$1 + 2 + \dots + n_1 = \frac{n_1(n_1 + 1)}{2} \text{ et } (n_2 + 1) + (n_2 + 2) + \dots + n = n_1(n_2 + \frac{n_1 + 1}{2}).$$

(réfléchir au cas les plus extremes)

### Zone de rejet

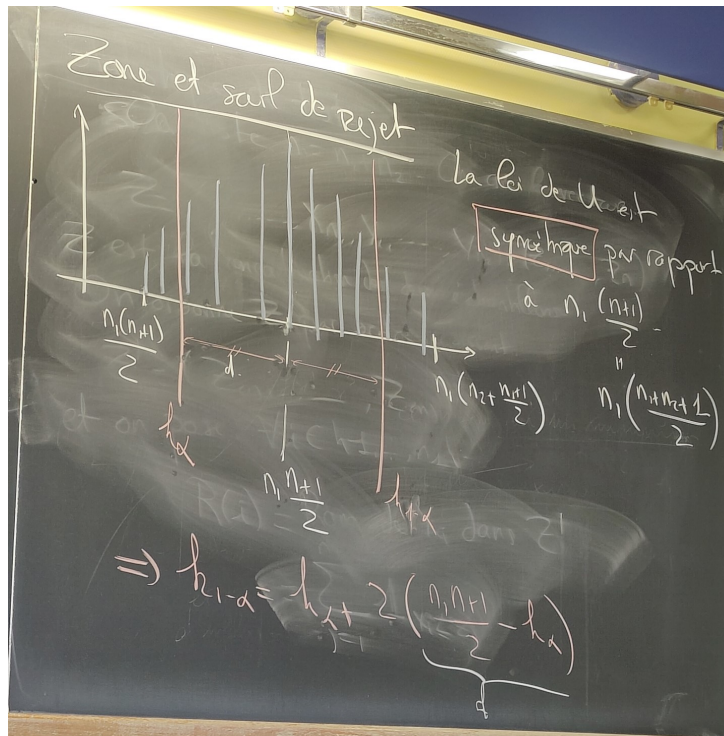
- Sous  $H_0$  : Quel est le rang de  $X_i$  ?  $R(1)$  est une variable. uniforme sur  $\{1, \dots, n\}$ . Sous  $H_0$ , les  $X_i$  sont uniformément répartis dans le vecteur  $Z'$ , et cela indépendamment de leur loi. Tout ce qui compte, c'est que le vecteur  $Z$  sont un vecteur de variables iid.

**CCL** : Ainsi, sous  $H_0$ , la loi de  $U$  ne dépend pas de la loi de  $X_1$  et  $Y_1$ . Elle ne dépend que de  $n_1$  et  $n_2$ . On peut alors tabuler la variable  $U$ .

- Sous  $H_1 = Y_1 \succ X_1$ , les  $X_i$  sont plutôt au début du vecteur  $Z'$ ,  $U$  prend donc de petites valeurs.
- Si  $H_1 = X_1 \succ Y_1$ , les  $X_i$  sont plutôt à la fin du vecteur  $Z'$ .  $U$  prend donc des grandes valeurs.
- Si  $H_1 : F_{X_1} \neq F_{Y_1}$ ,  $U$  va prendre des valeurs extremes, mais on ne sait pas de quel côté.

La loi de  $U$  est **symétrique** par rapport à  $n_1 \frac{n+1}{2}$

- **Cas 1**  $H_1 : Y_1 \succ X_1$  au niveau  $\alpha$ , on pose  $H_\alpha$  le quantile d'ordre  $\alpha$  de la loi  $U(n_1, n_2)$ . Si  $U < H_\alpha$  on rejette  $H_0$  sinon on conserve  $H_0$ .
- **Cas 2**  $H_1 : X_1 \succ Y_1$  au niveau  $\alpha$  on pose  $h_{1-\alpha}$  le quantile d'ordre  $1 - \alpha$  de la loi de  $U(n_1, n_2)$  Si  $U > h_{1-\alpha}$ , on rejette  $H_0$  sinon on conserve  $H_0$ .



- **Cas 3**  $H_1 : F_{X_1} \neq F_{Y_1}$  au niveau  $\alpha$  on pose  $h_{\alpha/2}$  et  $h_{1-\alpha/2}$  les quantiles d'ordre  $\alpha/2$  et  $1 - \alpha/2$  de la loi de  $U(n_1, n_2)$  Si  $U < h_{\alpha/2}$  ou  $U > h_{1-\alpha/2}$ , on rejette  $H_0$  sinon on conserve  $H_0$ .

**Exemple 5.1.** On a 2 échantillons :

- 8 étudiants qui viennent en amphi notes : 15, 16.4, 5.6, 16.4, 18.8, 15.6, 15.2, 12.8
- 9 étudiants qui ne viennent pas en amphi : 15.6, 11, 12.6, 7.4, 9.6, 14.8, 13, 15.4, 12.6

Est-ce que la présence en amphi a un impact sur les notes?

- $H_0$  pas d'impact = même loi
- $H_1$  impact : lois différentes

On choisit d'effectuer le test de Mann-Whitney (On n'utilise pas KS car on a plusieurs notes répété).

On trie les données :

Obs	<b>5.6</b>	7.4	9.6	11	12.6	12.6	<b>12.8</b>	13	
Rang	1	2	3	4	5.5	5.5	7	8	
Obs	14.8	<b>15</b>	<b>15.2</b>	15.4	<b>15.6</b>	15.6	<b>16.4</b>	<b>16.4</b>	<b>18.8</b>
Rang	9	10	11	12	13.5	13.5	15.5	15.5	17

On calcule

$$U = 1 + 7 + 10 + 11 + 13.5 + 15.5 + 15.5 + 17$$

$$= \text{Somme des rangs de } X_i = 90.5$$

Ici  $n_1 = 8$  et  $n_2 = 9$ , niveau 5%, on retrouve

$$h_{0.025} \approx 51$$

$$h_{0.975} = 51 + 2(72 - 51)$$

$$= 93$$

Revoir le graphique précédent.

Zone de rejet : Comme  $U \geq 51$  et  $U \leq 93$ , on conserve  $H_0$ .

**Bilan de ce test** Ce test est une alternative au test d'homogénéité de KS.

- KS détecte n'importe quelle différence de loi.
- MW est plus sensible à des changements de médiane, plutôt des translations.

- MW est plus utilisé et apprécié.
- MW gère les ex-æquo. Alors que KS déteste les ex-æquo.
- MW pas ouf si juste un changement de variance et pas de médiane.
- Si  $n_1, n_2$  sont grands. On n'a pas la table, on utilise alors le test asymptotique.

$$\frac{U - E(U)}{\sqrt{Var(U)}} = \frac{U - n \frac{n_1+1}{2}}{\sqrt{\frac{n_1 n_2 (n+1)}{12}}} \rightarrow Z \sim \mathcal{N}(0, 1).$$

## 6 Test du signe et test du signe et rang de Wilcoxon

### Données

- $X_1, \dots, X_{n_1}$  iid.
- $Y_1, \dots, Y_{n_2}$  iid.
- Échantillon **appariées**  $(X_1, Y_1), \dots, (X_n, Y_n)$  sont iid  $X_1 \perp Y_1$

On note  $Z_i = Y_i - X_i$ . On suppose que  $Z_i$  a une fonction de répartition continue donc aucun des  $Z_i$  ne vaut 0.

### 6.1 Test du signe / test de la médiane

#### Hypothèse

- $H_0$  La médiane de  $Z$  vaut 0.  $m_Z = 0$ . C'est à dire que  $P(Y_1 < X_1) = 1/2$
- $H_1 = m_Z \neq 0$  ou  $m_Z > 0 \Leftrightarrow P(Z \leq 0) > 1/2 \Leftrightarrow P(Y_1 > X_1) > 1/2$  ou  $m_Z < 0$

#### Statistique de test

$$S_n = \sum_{i=1}^n \mathbb{1}_{Z_i \leq 0}$$

= Nombre de  $Y_i > X_i$

- Sous  $H_0 : P(Z_i > 0) = P(Y_i > X_i) = \frac{1}{2}$

$$\mathbb{1}_{Z_i > 0} \sim \text{Ber}\left(\frac{1}{2}\right).$$

donc

$$S_n \sim \text{Bin}\left(n, \frac{1}{2}\right).$$

- Sous  $H_1$ 
  - Si  $m_Z > 0, P(Y_i > X_i) > \frac{1}{2}, S_n \sim \text{Bin}(n, p), p > \frac{1}{2}$  donc  $S_n$  est "grand"
  - Si  $m_Z < 0, P(Y_i < X_i) > \frac{1}{2}$  donc  $S_n$  est petit.
  - Si  $m_Z \neq 0, S_n$  a un comportement proche des extremes (petit/grand).

#### Zone de rejet Au niveau $\alpha$

- Si  $H_1 : m_Z > 0$ . Soit  $h_{1-\alpha}$  le quantile d'ordre  $1 - \alpha$  de la loi  $\text{Bin}(n, \frac{1}{2})$ . Si  $S_n > h_{1-\alpha}$  on rejette  $H_0$  sinon on conserve  $H_0$
- Si  $H_1 : m_Z < 0$ . Soit  $h_\alpha$  le quantile d'ordre  $\alpha$  de la loi  $\text{Bin}(n, \frac{1}{2})$ . Si  $S_n < h_\alpha$  on rejette  $H_0$  sinon on conserve  $H_0$
- Si  $H_1 : m_Z \neq 0$ . Soit  $h_{\alpha/2}$  et  $h_{1-\alpha/2}$  le quantile d'ordre  $\alpha/2$  et  $1 - \alpha/2$  de la loi  $\text{Bin}(n, \frac{1}{2})$ . Si  $S_n < h_{\alpha/2}$  ou  $S_n > h_{1-\alpha/2}$  on rejette  $H_0$  sinon on conserve  $H_0$

*Remarque* (En cas d'égalité). Si il existe  $i$  tel que  $Z_i = 0$ . On exclut ces données et on recommence avec le reste  $n \rightarrow n - 1$



## CCL sur le test

- Ce test ne regarde **que** le signe et pas les amplitudes. (exemple : on peut avoir 5 fois  $-100$  dans nos données et 5 fois  $+1$  c'est la même chose). En général on va appliquer ce test quand **on ne connaît pas** les amplitudes mais juste les signes ("Est ce que la situation c'est améliorer?")
- Si  $n$  est grand on fait un test asymptotique

$$\frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} \rightarrow \mathcal{N}(0, 1).$$

Sous  $H_0$ ,  $E(S_n) = n/2$ ,  $Var(S_n) = n/4$

Nouveau cours du 31/03

## Rappel Résumé du test du signe

- $X_1, \dots, X_n, Y_1, \dots, Y_n$  2 échantillon appariés
- Hypothèse : la loi des  $Z_i = Y_i - X_i$  est continue
- Statistique de test :  $S_n = \sum_{i=1}^n \mathbb{1}_{Z_i > 0}$
- HP :  $H_0$  la médiane des  $Z_i$  vaut 0 :  $m_z = 0$  contre  $H_1 = m_z > 0$  ou  $m_z < 0$  ou  $m_z \neq 0$
- $m_z > 0$  signifie  $P(Y_i > X_i) > \frac{1}{2}$
- Sous  $H_0$  :  $S_n \sim \mathcal{B}(n, \frac{1}{2})$
- Sous  $H_1$  : si  $m_z > 0$  alors

$$p = P(z_i = 1) = P(Y_i > X_i) > \frac{1}{2}.$$

Ainsi sous  $H_1$ ,  $S_n \sim \mathcal{B}(n, p)$ . Sous  $H_1$ ,  $S_n$  prend de "grande valeur".

- Zone et seuil de rejet pour  $H_1 = m_z > 0$ . Soit  $h_{1-\alpha}$  le quantile d'ordre  $1 - \alpha$  de la  $\mathcal{B}(n, \frac{1}{2})$ . Si  $S_n > h_{1-\alpha}$  on rejette  $H_0$  sinon on conserve  $H_0$
- Remarque :

1. S'il existe un  $i$  tel que  $Z_i = 0$ . On est alors obligé d'exclure les données correspondantes. On adapte alors la valeur de  $n$ .
2. Si  $n$  est petit, la table de la loi binomiale ne fournit pas de quantile exact.

**Exemple 6.1.** Si  $n = 10$  :

- $P(\mathcal{B}(10, \frac{1}{2}) \leq 7) = 0.945$
- $P(\mathcal{B}(10, \frac{1}{2}) \leq 8) = 0.989$

Si on veut faire un test à 5% que faire? On peut être tenté de regarder

(a)  $\mathcal{R} = \{S \geq 8\}$

(b)  $\mathcal{R} = \{S \geq 9\}$

Laquelle est la bonne pour avoir 5%?

(a) Pour 1 :  $P_{H_0}(\mathcal{B}(10, \frac{1}{2}) \in \mathcal{R}) = 0.055 > 0.05$  ce choix n'est pas acceptable pour un niveau  $\alpha$

(b) On choisit la zone 2

En fait tout dépend de si on peut discuter avec les personnes, si c'est un texte de loi, pas le choix pour bouger  $\alpha$  de quelque dixième. Sinon penser à travailler avec la p valeur (qu'on va calculer plus tard).

**Exemple 6.2** (propre d'un test du signe). Donnée de poids de patients après un changement d'alimentation

i	1	2	3	4	5	6	7	8	9
Avant $X_i$	80	82	75	90	90	87	100	107	103
Après $Y_i$	79	83	75	81	95	86	101	105	100
Différence $Z_i = Y_i - X_i$	-1	1	0	-9	-3	-1	1	-2	-3

- $H_0$  Pas d'effet du changement sur le poids, c'est à dire  $m_z = 0$
- $H_1$  Le changement induit une diminution du poids  $m_z < 0$

Ici on dispose de **deux échantillons appariés**. Pour une valeur de  $i$ ,  $X_i$  et  $Y_i$  représentent le poids d'une personne avant et après le changement d'alimentation.

Comme  $Z_3 = 0$ , on doit exclure  $i = 3$

$$S = \sum_{i=1, i \neq 3}^9 \mathbb{1}_{Z_i > 0} = 2.$$

Zone de rejet :  $\alpha = 5\%$

Sous  $H_1$ ,  $S$  prend des plus petite valeur. Lecture de table :  $n = 8$

$$- P(\mathcal{B}(2, \frac{1}{2}) \leq 1) = 0.035$$

$$- P(\mathcal{B}(2, \frac{1}{2}) \leq 2) = 0.145$$

Zone de rejet au niveau 5% :  $\mathcal{R} = \{S \leq 1\}$ . CCL : On conserve  $H_0$ .

**CCL** : Le test du signe ne prend en compte **que** le signe des  $Z_i$ . C'est à la fois sa force **et** sa faiblesse.

Dans de nombreux cas, il est délicat d'obtenir les amplitudes d'évolution. En revanche, dès qu'on les a et qu'elles ont du sens, il serait malavisé de ne pas utiliser.

## 6.2 Le test des rangs et signe de Wilcoxon

**Données** : 2 échantillons appariés :  $X_1, \dots, X_n$  et  $Y_1, \dots, Y_n$ . On pose  $Z_i = Y_i - X_i$  iid.

*Remarque.* — Moyen rapide de voir si échantillons appariés : on a le même nombre  $n$ .

— Attention le  $n$  dans les formules est le nombre de couple  $(X_i, Y_i)$

— En réalité si on regarde sur wikipedia, le test demande même pas iid

**Condition** :

— La loi des  $Z_i$  est continue

— Les  $Z_i$  sont symétriques par rapport à leur médiane  $m$ .

*Remarque (Attention).* .

— si  $E(|Z_i|) < +\infty$  alors  $m = E(Z_i)$ .

— La condition sur la symétrie est délicate à vérifier. On se contentera de la supposer vrais dès que c'est raisonnable (typiquement en faisant un histogramme)

*Remarque (Importante).*  $|Z_i|$  et  $\text{Signe}(Z_i) = \frac{Z_i}{|Z_i|} = \begin{cases} 1 & \text{si } Z_i > 0 \\ -1 & \text{si } Z_i < 0 \end{cases}$

*Mini preuve :*

$$\begin{aligned} \forall t > 0, P(\text{sgn}(Z) = 1 \text{ et } |Z| > t) &= P(Z > t) \\ &= \frac{1}{2}P(Z > t) + P(Z < -t) \\ &= \frac{1}{2}P(|Z| > t) \\ &= P(\text{sgn}(Z) = 1)P(|Z| > t) \end{aligned}$$

idem si  $\text{sgn}(Z) = -1$ . Bref :  $\text{sgn}(Z)$  et  $|Z|$  sont indépendantes □

**Hypothèse** :

—  $H_0 = m = 0$

—  $H_0 = m \neq 0$  ou  $m > 0$  ou  $m < 0$

Si  $m = 0$  alors  $P(Y_i > X_i) = P(X_i < Y_i) = \frac{1}{2}$

**Statistique de test** Soit  $R_i$  le rang de  $|Z_i|$  dans l'échantillon ordonnée issus de  $|Z_1|, \dots, |Z_n|$ .

On pose

$$W_n^+ = \sum_{i=1}^n \mathbb{1}_{Z_i > 0} R_i \text{ la somme des rangs positifs.}$$

et

$$W_n^- = \sum_{i=1}^n \mathbb{1}_{Z_i < 0} R_i.$$

On remarque que

$$W_n^+ + W_n^- = \sum_{i=1}^n R_i = 1 + 2 + \dots + n = \frac{n(n+1)}{2}.$$

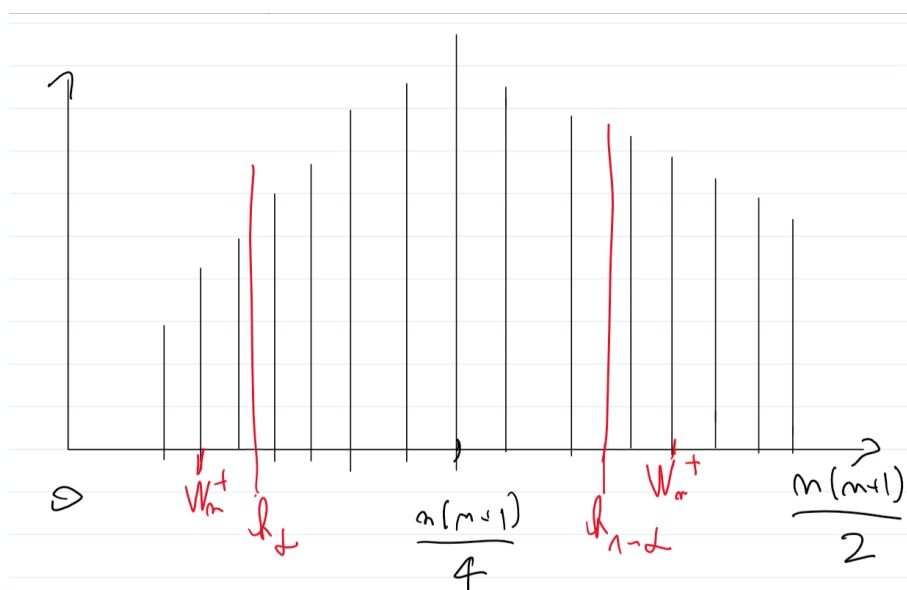
Sous  $H_0$ ,  $W_n^+$  et  $W_n^-$  ont la même loi, et celle-ci ne dépend pas de la loi des  $Z_i$ . On l'appelle la loi de Wilcoxon. Elle est tabulée pour  $n$  entre 5 et 20.

Sous  $H_0$ , le vecteur  $(R_1, \dots, R_n)$  est une permutation aléatoire uniforme. Les  $\mathbb{1}_{Z_i > 0}$  sont des variables de Bernoulli  $\frac{1}{2}$  indépendantes des  $|Z_i|$  donc des  $R_i$ .

Ainsi,  $W_n^+$  a même loi que  $\sum_{i=1}^n b_i R_i$  avec  $b_i$  iid.  $\text{Ber}(\frac{1}{2})$  et  $R_i \perp b_i$  et  $(R_1, \dots, R_n)$  permutation uniforme.

Sous  $H_0$ , la loi de  $W_n^+$  et  $W_n^-$  est symétrique par rapport à leur moyenne  $\frac{n(n+1)}{4}$ .

Figure 6 – Loi de Wilcoxon



**Zone et seuil de rejet** Sous  $H_0$ ,  $W_n^+$  et  $W_n^-$  sont tabulées

Sous  $H_1 = m > 0$  alors  $W_n^+$  prend de grandes valeurs et  $W_n^-$  de petites valeurs.

Soit  $h_\alpha$  le quantile d'ordre  $\alpha$  de la loi de Wilcoxon et  $h_{1-\alpha}$  celui d'ordre  $1 - \alpha$

On peut utiliser comme zone de rejet  $W_n^+ \geq h_{1-\alpha}$  ou  $W_n^- \leq h_\alpha$ . En fait, il s'agit de la **même** condition, car, grâce à la symétrie de la loi de Wilcoxon.

Pour le test bilatéral :

Si  $H_1 = m \neq 0$  alors on utilise comme zone de rejet

$$W_n^+ \geq h_{1-\frac{\alpha}{2}} \text{ ou } W_n^+ \leq h_{\alpha/2} \text{ (tout avec } W_n^+).$$

Ce qui peut se réécrire

$$W_n^- \leq h_{\alpha/2} \text{ ou bien } W_n^+ \leq h_{\alpha/2} \text{ (tout avec un seul quantile).}$$

avec les  $h_{\alpha/2}$  et  $h_{1-\alpha/2}$  sont les quantiles de la loi de Wilcoxon.

Il n'y a pas de différence entre les deux présentations c'est une affaire de goût personnel.

**Exemple 6.3** (Retour sur l'exemple). A partir de l'exemple donnée dans le test du signe, on remplit le tableau.  $n = 8$

$ Z_i $	1	1	0	9	3	1	1	2	3
Signe	-	+	X	-	-	-	+	-	-
Rang	2.5	2.5	X	8	6.5	2.5	2.5	5	6.5

En cas d'égalité on attribue le rang moyen.

Dans l'exemple, les données sont appariées et

- $H_0 = m = 0$
- $H_1 = m \neq 0$

Ici, sous  $H_1$  C'est  $W_n^+$  qui prend des grandes valeurs. Calculons  $W_n^+ = 2.5 + 2.5 = 5$  (automatiquement  $W_n^- = \frac{8*9}{2} - 5 = 36 - 5 = 31$ )

**Zone et seuil de rejet** : Dans la table pour  $n = 8$ , on lit que  $P(W_n^+ \leq 5) \leq 0.05$  et  $P(W_n^+ \leq 6) \geq 0.05$ . Ici on rejette  $H_0$  si  $W_n^+ \leq 5$ . Or  $W_n^+ = 5$  on rejette  $H_0$

**Bilan sur le test de Wilcoxon** Les avantages :

- Ce test est une alternative **non paramétrique** au test de Student pour échantillons appariées (on ne suppose pas que c'est gaussien)
- Il fonctionne **toujours** mieux que le test du signe
- Si les données sont gaussiennes, le test de Wilcoxon est **à peine** moins bon que Students.

Bref : c'est un des meilleurs tests du cours pour les échantillons appariées

Les désavantages :

- Hypothèse dure à vérifier, on se contente souvent de supposer que c'est applicable.
- Version asymptotique :

$$\frac{W_n^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \xrightarrow{n \rightarrow +\infty} \mathcal{N}(0, 1).$$

## 7 Remarques finales

### 7.1 La table de fisher

La loi de Fisher à  $(d_1, d_2)$  degrés de liberté est la loi de  $\mathcal{F}_{d_1, d_2} = \frac{U_1/d_1}{U_2/d_2}$  où  $U_1 \sim \chi^2(d_1), U_2 \sim \chi^2(d_2), U_1 \perp U_2$ .

Ainsi  $F \sim \mathcal{F} \Rightarrow \nabla(d_1, d_2), \frac{1}{F} \sim \mathcal{F} \Rightarrow \nabla(d_2, d_1)$ .

Cela implique : si on note  $h_{1-\alpha, d_1, d_2}$  le quantile d'ordre  $1 - \alpha$  de la loi  $\mathcal{F} \Rightarrow \nabla(d_1, d_2)$  alors

$$h_{\alpha, d_1, d_2} = \frac{1}{h_{1-\alpha, d_2, d_1}}.$$

**Exemple 7.1.**

$$h_{0.95, 10, 5} = 4.75 \Leftrightarrow h_{0.05, 10, 5} = \frac{1}{3.33} = \frac{1}{h_{0.95, 5, 10}}.$$

### 7.2 Table de Mann-Whitney et Wilcoxon

Dans les deux cas, on utilise la symétrie par rapport à la médiane

## 8 Test bonus

Il est impossible de couvrir tous les tests de comparaison d'échantillons.

But : Devenir autonome et savoir aller chercher de nouveau test dans la littérature/sur internet.

A chaque fois que vous verrez un test, j'aimerais que vous vous posiez les questions suivantes :

- Combien d'échantillon? Indépendant? Liée?
- Quelle loi sur les données? Donnée gaussiennes? Bernoulli? Poisson?
- Si vous n'avez aucune information sur la loi, regardez du côté des tests non paramétriques.
- Quelle taille d'échantillon? Si les échantillons sont assez grands, vous pouvez faire des tests asymptotiques.
- Comment traduire ma question en hypothèses? "Les échantillons sont-ils différents?" C'est la manière de comprendre cette question qui vas **guider votre choix de test**.
- Comparaison de moyennes? De variances? De médianes? De fonction de répartition?

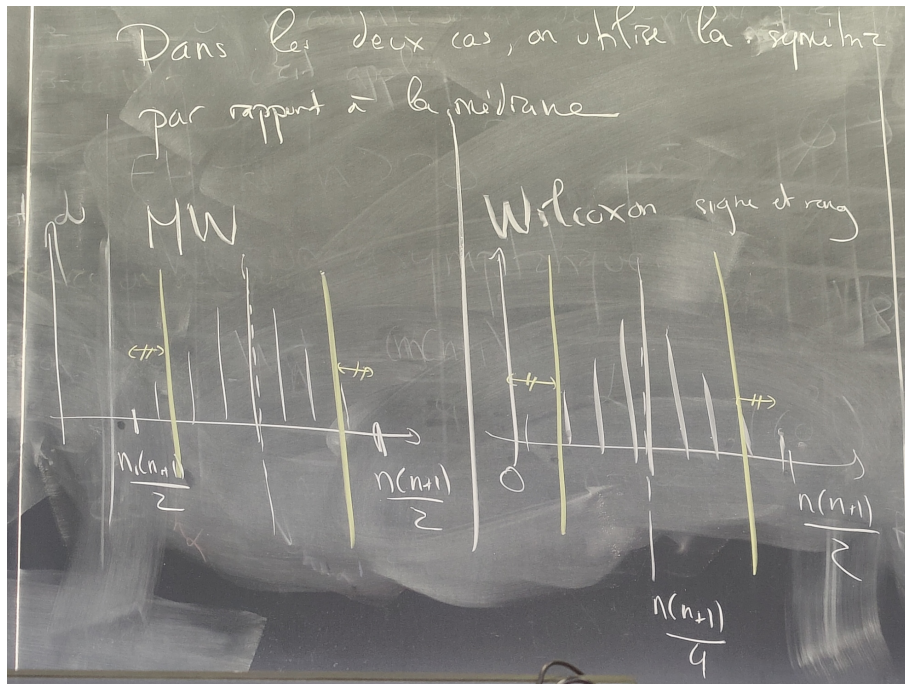


Figure 7 – Loi de Mann-Whitney et Wilcoxon

## 8.1 Le test d'indépendance de Pearson

### Données :

- $X_1, \dots, X_n$  iid  $\mathcal{N}(m, \sigma^2)$
- $Y_1, \dots, Y_n$  iid  $\mathcal{N}(m, \sigma^2)$
- Échantillon appariés et  $(X_i, Y_i)$  vecteur gaussien

### Hypothèse

- $H_0 = X_i \perp Y_i$  ( $\text{cor}(X, Y) = 0$ )
- $H_0 = X_i \not\perp Y_i$  ( $\text{cor}(X, Y) \neq 0$ )

Ici dans le cas particulier des gaussiens, le lien entre indépendance et corrélation est une équivalent.

**Statistique de test** Soit  $R$  la corrélation empirique :

$$R = \frac{\text{cov}_n((X_1, \dots, X_n), (Y_1, \dots, Y_n))}{\sqrt{V_n^X * V_n^Y}}$$

$$R_n = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{(\sum_{i=1}^n (X_i - \bar{X}_n)^2) \sum_{i=1}^n (Y_i - \bar{Y}_n)^2}}$$

$R_n$  est un estimateur fortement consistant de  $\text{cor}(X_i, Y_i)$

$$D = \frac{R_n}{\sqrt{1 - R_n^2}} \sqrt{n-1}.$$

- Sous  $H_0$ ,  $D \sim \mathcal{T}(n-2)$
- Sous  $H_1$ ,  $D$  est grand en valeur absolue

**Zone et seuil de rejet** On utilise des quantiles de la loi  $\mathcal{T}(n-2)$

$$\mathcal{R} = \{D < h_{\alpha/2}\} \cup \{D > h_{1-\alpha/2}\}.$$

*Remarque.* Il existe aussi des tests d'indépendance non-paramétriques : Ce sont les tests de Spearman et Kendall

## 8.2 Comparaison asymptotique de proportion

**Motivation** On a plus jamais reparlé de variable de Bernoulli depuis le semestre dernier alors que c'est ce qu'on risque de rencontrer le plus souvent.

**Données :**

- $X_1, \dots, X_n$  iid  $\text{Ber}(p_1)$
- $Y_1, \dots, Y_n$  iid  $\text{Ber}(p_2)$
- $(X_1, \dots, X_n) \perp (Y_1, \dots, Y_n)$

**Hypothèse**

- $H_0 = p_1 = p_2$
- $H_1 = p_1 \neq p_2$  ou  $p_1 < p_2$  ou  $p_1 > p_2$

**Statistique de test**

$$S_{n,m} = \frac{\bar{X}_n - \bar{Y}_n}{\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n} + \frac{\bar{Y}_n(1-\bar{Y}_n)}{n}}}.$$

- Sous  $H_0$ ,  $S_{n,m} \xrightarrow[n, m \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, 1)$
- Sous  $H_1$ 
  - Si  $p_1 > p_2$ ,  $S_{n,m} \xrightarrow[n, m \rightarrow \infty]{} +\infty$
  - Si  $p_1 < p_2$ ,  $S_{n,m} \xrightarrow[n, m \rightarrow \infty]{} -\infty$
  - Si  $p_1 \neq p_2$ ,  $S_{n,m} \xrightarrow[n, m \rightarrow \infty]{} +\infty$

**Zone et seuil de rejet**

- $H_1 : p_1 > p_2$ ,  $\mathcal{R} = \{S > h_{1-\alpha}\}$
- $H_1 : p_1 < p_2$ ,  $\mathcal{R} = \{S < h_\alpha\}$
- $H_1 : p_1 \neq p_2$ ,  $\mathcal{R} = \{S > h_{\alpha/2}\} \cup \{S < h_{1-\alpha/2}\}$

Où les  $h_\alpha$  sont des quantiles de la loi normale.

## 9 Comparaison de $K \geq 3$ échantillons

On dispose de  $K \geq 3$  échantillons indépendants. On cherche à déterminer s'il existe une différence entre ces échantillons.

Cette "différence" peut se caractériser de plusieurs manières : différences de moyenne, médianes, variances, fonction de répartition

- $H_0$  les échantillons ont la même caractéristiques
- $H_1$  Il existe au moins 2 populations qui diffèrent :  $\exists i, j \in \{1, \dots, K\}$  tq  $m_i \neq m_j$

Jusqu'à présent vous avez vu un seul test qui entre dans ce cadre : le  $\chi^2$  d'homogénéité. Le problème avec ce test est qu'il faut vraiment beaucoup de données.

### 9.1 L'ANOVA à un facteur

Vous avez vu dans le cours de modélisation statistique les aspects théorique du modèle linéaire gaussien. C'est le modèle le plus important en statistiques.

Ici la présentation ne traitera **que** des aspects pratiques.

**Données**

- $K$  échantillons **indépendants**
- $X_1^{(1)}, \dots, X_{n_1}^{(1)}$  va iid  $\mathcal{N}(m, \sigma_1^2)$
- $X_1^{(2)}, \dots, X_{n_2}^{(2)}$  va iid  $\mathcal{N}(m, \sigma_2^2)$
- $X_1^{(K)}, \dots, X_{n_K}^{(K)}$  va iid  $\mathcal{N}(m, \sigma_K^2)$
- On suppose de plus l'**homoscédasticité** :  $\sigma_1^2 = \dots = \sigma_K^2$

*Remarque.* Si on pose  $n = n_1 + \dots + n_K$ , cela correspond à la vouloir expliquer les variables continues  $(Y_1, \dots, Y_n) = (X_1^{(1)}, \dots, X_{n_1}^{(1)}, \dots, X_{n+K}^{(K)})$  par la variable catégorielle à valeur dans  $\{1, \dots, K\}$

Cette situation correspond à la situation classique : effet de "catégorie" sur "score continue". Exemple :

- Effet du département de résidence sur la taille
- Effet de la mention au bac sur le temps au 100m

### ANOVA en pratique

- $H_0 = m_1 = \dots = m_k$
- $H_1 = \exists i, j \text{ tq } m_i \neq m_j$

La somme des carrés des écarts ( $\approx$  variance sans diviser )

$$SCE^{(p)} = \sum_{i=1}^{n_p} (X_i^{(p)} - \bar{X}^{(p)})^2.$$

$$\frac{SCE^{(p)}}{n_p - 1} = V^{(p)} \text{ variance empirique de l'échantillon } p.$$

*Remarque.* Sous  $H_0$ ,  $\frac{SCE^{(p)}}{\sigma^2} \sim \chi^2(n_p - 1)$   
Sous  $H_1$  aussi

On calcule alors SCE ou SCT total ou intra

$$SCE_{intra}^{totale} = \sum_{p=1}^K SCE^{(p)}.$$

*Remarque.* sous  $H_0$  comme sous  $H_1$

$$\begin{aligned} \frac{SCE_{tot}}{\sigma^2} &= \chi^2(n_1 - 1) + \dots + \chi^2(n_K - 1) \text{ somme de khi deux indépendante} \\ &= \chi^2(n_1 + n_2 + \dots + n_K - K) \\ &= \chi^2(n - K) \end{aligned}$$

Enfin, on calcule

$$\bar{X} = \text{moyenne totale} = \frac{1}{n} \sum_{p=1}^K n_p \bar{X}^{(p)}.$$

On calcule

$$SCE_{inter} = \sum_{p=1}^K n_p (\bar{X}^{(p)} - \bar{X})^2.$$

Sous  $H_0$  :  $\frac{SCE_{inter}}{\sigma^2} \sim \chi^2(K - 1)$   
Sous  $H_1$ ,  $SCE_{inter}$  est **grand**

### Stat de test

$$F = \frac{SCE_{inter}/(K - 1)}{SCE_{intra}^{totale}/(n - K)}.$$

Sous  $H_0$  :  $F \sim \mathcal{F}(K - 1, n - K)$   
Sous  $H_1$ ,  $F$  prend de grande valeurs.

**Zone et seuil de rejet** Au niveau  $\alpha$ , soit  $h_{1-\alpha}$  le quantile d'ordre  $1 - \alpha$  de la loi  $\mathcal{F}(K - 1, n - K)$ . Si  $F > h_{1-\alpha}$  on rejette  $H_0$ .

*Remarque.* L'ANOVA porte ce nom à cause de la décomposition :

$$\begin{aligned} \sum_{p=1}^K \sum_{i=1}^{n_p} (X_i^{(p)} - \bar{X})^2 &= SCE_{inter} + SCE_{intra}^{totale} \\ &= \text{Variance totale} = \text{entre les groupes} + \text{interne à chaque groupe} \end{aligned}$$

Analyse (de la décomposition) de la variance.