

# Liste des Tests du cours

Charles Vin

2022

## Table des matières

<b>1</b>	<b>Template</b>	<b>1</b>
<b>2</b>	<b>Test d'ajustement de Kolmogorov-Smirnov</b>	<b>1</b>
<b>3</b>	<b>Le test du <math>\chi^2</math> d'ajustement</b>	<b>2</b>
3.1	Le $\chi^2$ d'ajustement à une famille paramétrique de loi . . . . .	4
<b>4</b>	<b>Le test d'homogénéité de Kolmogorov-Smirnov</b>	<b>4</b>
<b>5</b>	<b>Test du <math>\chi^2</math> d'indépendance</b>	<b>5</b>
<b>6</b>	<b>Test du <math>\chi^2</math> d'homogénéité</b>	<b>6</b>
<b>7</b>	<b>Test sur les Gaussiennes</b>	<b>6</b>
7.1	Sur la moyenne . . . . .	6
7.2	Sur la variance . . . . .	6
<b>8</b>	<b>Test de la somme des rangs aka MWW</b>	<b>6</b>
<b>9</b>	<b>Test du signe</b>	<b>7</b>

## 1 Template

**Donnée**

**Conditions**

**Hypothèse**

**Statistique de test**

**Zone de Rejet**

**Méthode**

## 2 Test d'ajustement de Kolmogorov-Smirnov

**Conditions**

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).
2. Fonctionne  $\forall n$  : même si  $n$  est petit, ce test est pertinent
3. Si  $n \geq 100$ , on fait un test asymptotique.

## Hypothèse

- $H_0$  = les  $X_i$  ont pour fdr.  $F_X$
- $H_1$  = les  $X_i$  n'ont pas pour fdr.  $F_X$

## Statistique de test

$$\begin{aligned}h(F_n, F) &= \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \\&= \max_{1 \leq i \leq n} \left( \max \left( \left| \frac{i}{n} - F(X_{(i)}) \right|, \left| \frac{i-1}{n} - F(X_{(i)}) \right| \right) \right)\end{aligned}$$

## Zone de Rejet

### Si n est petit

La loi de  $h(F_n, F)$  est tabulé alors :

$$\mathcal{R} = \{h(F_n, F_X) \geq h_{1-\alpha}\}.$$

avec  $F_n$  fonction de répartition empirique,  $h_{1-\alpha}$  le quantile à aller chercher dans la table

### Si n est grand $n \geq 30$

Attention pas souvenir de l'avoir fait en TD.  
On a pas la table de  $h(F_n, F)$  mais on sait que

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

Donc on pose la zone de rejet

$$\mathcal{R} = \{h(F_n, F_X) \geq \frac{k_\alpha}{\sqrt{n}}\}.$$

avec  $F_n$  fonction de répartition empirique,  $k_\alpha$  le quantile de  $W_\infty$  à aller chercher dans sa table

## Méthode

Pour trouver la valeur de  $h(F_n, F_X)$  : Faire le grand tableau puis trouver le max. Exemple :

i	1	2	3	4	5
$X_{(i)}$	0.3	0.7	0.9	1.2	1.4
$X_{(i)} - 2$	-1.70	-1.30	-1.10	-0.80	-0.60
$F_0(X_{(i)})$	0.04	0.10	0.14	0.21	0.27
$\frac{i}{n}$	0.05	0.1	0.15	0.2	0.25
$ \frac{i}{n} - F_0(X_{(i)}) $	0.01	0.00	0.01	0.01	0.02
$ \frac{i-1}{n} - F_0(X_{(i)}) $	0.04	0.05	0.04	0.06	0.07

Table 1 – Ici le max c'est 0.07 à la dernière case

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

### Conditions

1. Les  $X_i$  sont à valeur dans un ensemble fini (loi discrète). Si a valeur dans  $\mathbb{N}$ , on fusionne les classes à partir d'un certain rang choisis
2. Test asymptotique :  $\forall k \in \{1, \dots, d\}, np_k^{ref}(1 - p_k^{ref}) \geq 5 \Leftrightarrow n \geq 20$

Si on ne remplit pas les conditions, on peut fusionner les classes

## Hypothèse

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

Avec  $p^{ref}$  un vecteur fixé à tester (par exemple pour un lancé de dé  $(\frac{1}{6}, \dots, \frac{1}{6})$ )

## Statistique de test

$$D(\bar{p}_n, p^{ref}) = n \sum_{k=1}^d \frac{(p_{k,n} - p_k^{ref})^2}{p_k^{ref}} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \chi^2(d-1)$$
$$= \sum_{k=1}^d \frac{(N_{k,n} - np_k^{ref})^2}{np_k^{ref}}$$

avec

- $N_{k,n} = \sum_{i=1}^n \mathbb{1}_{X_i = x_k}$  (ce qu'il y a dans le tableau de la consigne)
- $p_{k,n} = \frac{N_{k,n}}{n}$  les proportions observés

## Zone de Rejet

$$\mathcal{R} = \{D(\bar{p}_n, p^{ref}) \geq h_\alpha\}.$$

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

## Méthode

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  $H_0$

## 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.

$Z_{(i)}$	$F_n$	$G_n$	$h_{n_1, n_2}$
...	...	...	...

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

Pratiquement comme avant, pas encore fait en TD, mais copier collé du cours quand même

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  $\chi^2(d - 1 - M)$  avec  $M$  nombre de paramètre.
6. Etape 6 : Décision
  - $D > h_\alpha$  on rejette  $H_0$
  - $D \leq h_\alpha$  on conserve  $H_0$

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

### Conditions

- Deux échantillons indépendants de variable iid.
- De fdr. continue  $F_X, F_Y$

### Hypothèse

- $H_0$  : 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

### Statistique de test

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

### Zone de Rejet

- Ce test est de taille  $\alpha$ , si on utilise la table de  $h_{n,m} = \sup_{s \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{U_i \leq s} - \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{V_j \leq s} \right|$ .
- Si  $n$  et  $m$  sont trop grands, on utilise le résultat suivant :  
Sous  $H_0$

$$h_{n_1, n_2} = \sqrt{\frac{nm}{n+m}} h(F_n, G_n) \xrightarrow[n, m \rightarrow +\infty]{\alpha} W_\infty \text{ voir KS asymptotique.}$$

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

### Méthode

Même qu'un khi deux classique!

$$Z_{(i)} = (X_i, Y_i)$$

## 5 Test du $\chi^2$ d'indépendance

### Donnée

- $(X_1, Y_1), \dots, (X_T, Y_T)$  iid appariés.
- $X_1$  à valeur dans  $A_1, \dots, A_M$
- $Y_1$  à valeur dans  $B_1, \dots, B_N$

### Conditions

- Loi discrète
- $n$  ou  $T$  plutôt grand
- $\forall i < M, j < N : T * \hat{p}_m \hat{q}_m \geq 5$  ou avec la notation en TD :  $E_{i,j} \geq 5$

### Hypothèse

- $H_0 : X_1 \perp Y_1$
- $H_1 : X_1 \not\perp Y_1$

### Statistique de test

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

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

On utilise la deuxième en TD, la fraction est équivalente à  $E_{i,j}$  aka le produit en croix à l'intérieur du tableau durant les TD (groupe 2)

### Zone de Rejet

- Sous  $H_0$ ,  $D \rightarrow \chi^2((M-1)(N-1))$
- Sous  $H_1$ ,  $D \rightarrow +\infty$

$$\mathcal{R} = \{D \geq h_\alpha\}.$$

### Méthode

$E_{1,1} * 1000 = 152 * 261$   
 $\Leftrightarrow E_{1,1} = \frac{152 * 261}{1000} = 128.67$

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

  

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

Puis calculer la stat de test

$$D = \sum_{\text{chaque case du tableau}} \frac{N_{1,1} - E_{1,1}}{E_{1,1}}.$$

## 6 Test du $\chi^2$ d'homogénéité

### Donnée

- $X_1, \dots, X_{n_1}$  échantillons iid
- $Y_1, \dots, Y_{n_2}$  échantillons iid
- Échantillons indépendant entre eux

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

### Conditions

### 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  $P(X_1 \in A_m) \neq P(Y_1 \in A_m)$

### Statistique de test

### Zone de Rejet

### Méthode

## 7 Test sur les Gaussiennes

### 7.1 Sur la moyenne

- Test sur 1 échantillon : Loi de Student
  - Variance connu : on l'utilise à la place de  $V_n$
  - Variance inconnu : On utilise  $\bar{X}_n$  dans  $V_n$
- Test sur 2 échantillons indépendants :
  - Variances connus :
    - Même variance inconnu :  $\bar{X}_{n_1} - \bar{Y}_{n_2} \sim \mathcal{N}(m_1 - m_2, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))$  Same stat de test sauf qu'on estime la variance avec  $W = \frac{(n_1-1)V_{n_1}^X + (n_2-1)V_{n_2}^Y}{n_1+n_2-2}$
  - Variance inconnu : Test de welch :  $D = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{\sqrt{\frac{V_{n_1}^X}{n_1} + \frac{V_{n_2}^Y}{n_2}}} \sim_{H_0} \mathcal{T}(\mu)$  avec  $\mu$  Formule horrible
- Test sur 2 échantillons apparié :  $Z_n$  into student (pas trouvé de raison yet)

### 7.2 Sur la variance

- Test sur 1 échantillon : Comme le semestre d'avant
  - Moyenne connu : On l'utilise dans le calcul de  $V_n$
  - Moyenne inconnu : On utilise  $\bar{X}_n$  dans le calcul de  $V_n$
- Test sur 2 échantillons indépendants :
  - Moyennes connus : L'utiliser dans les calcul des  $V_n$
  - Même Moyenne inconnu : X Pas de solution so do same as before
  - Moyenne inconnu :  $D = \frac{X_{n_1}^X}{V_{n_2}^Y}$  qui suit  $\mathcal{F}(n_1 - 1, n_2 - 1)$  sans besoin de transformation.
- Test sur 2 échantillons apparié : X

## 8 Test de la somme des rangs aka MWW

C'est le test de sur l'ordre stochastique.

## Donnée

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

## Conditions

- On suppose que  $F_X$  et  $F_Y$  sont **continues**.
- Mieux qu'un KS à deux échantillons!

## 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

$$U = \sum_{i=1}^{n_1} R(i) = \sum_{i=1}^{n_1} \sum_{j=1}^n \mathbb{1}_{X_i \leq Z_j}.$$

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

## Zone de Rejet

La loi est symétrique. On a uniquement la table d'un côté, il faut calculer l'autre coté  $h_{1-\alpha} = h_\alpha + 2(\frac{n_1 n_2 + 1}{2} - h_\alpha) =$ .

Si  $n$  est grand, on utilise le TCL suivant

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

## Méthode

On trie les données : On calcule

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

$$\begin{aligned} U &= 1 + 7 + 10 + 11 + 13.5 + 15.5 + 15.5 + 17 \\ &= \text{Somme des rangs de } X_i = 90.5 \end{aligned}$$

## 9 Test du signe

### Donnée

- $X_1, \dots, X_{n_1}$  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.

## Conditions

Fonction de répartition continue.

## 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  $Y_i > X_i$

## Zone de Rejet

- 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).
- On utilise donc une table de la loi binomiale. Si  $n$  est grand, on utilise le TCL.

## Méthode