Série de TD n°2 : Tests de Kolmogorov-Smirnov et Cramer-von Mises

Solution détaillée complète

Master 2 Probabilités et Statistiques

Département de Mathématiques - BISKRA

Table des matières

Exercice 1 : Application Directe

a) Formulation des hypothèses

$$H_0 : X \sim \mathcal{U}[0,1] \quad \text{(les données suivent une loi uniforme sur [0,1])}$$
$$H_1 : X \not\sim \mathcal{U}[0,1] \quad \text{(les données ne suivent pas une loi uniforme sur [0,1])}$$

b) Calcul de la fonction de répartition empirique $F_n(t)$

Pour un échantillon de taille n, la fonction de répartition empirique est définie par :

$$F_n(t) = \frac{\text{nombre d'observations } \leq t}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq t\}}$$

Pour notre échantillon de $n = 12$ observations ordonnées :

i (rang)	$X_{(i)}$ (valeur ordonnée)	$F_n(X_{(i)}) = \frac{i}{12}$
1	0.15	0.0833
2	0.22	0.1667
3	0.31	0.2500
4	0.42	0.3333
5	0.48	0.4167
6	0.53	0.5000
7	0.61	0.5833
8	0.67	0.6667
9	0.72	0.7500
10	0.79	0.8333
11	0.85	0.9167
12	0.93	1.0000

c) Calcul de la statistique de Kolmogorov-Smirnov D_n

Sous H_0, la fonction de répartition théorique de la loi uniforme sur $[0,1]$ est :

$$F_U(t) = \begin{cases} 0 & \text{si } t < 0 \\ t & \text{si } 0 \leq t \leq 1 \\ 1 & \text{si } t > 1 \end{cases}$$

La statistique de Kolmogorov-Smirnov est définie comme le suprémum des écarts absolus :

$$D_n = \sup_{t \in \mathbb{R}} |F_n(t) - F_U(t)|$$

Méthode pratique de calcul :

Le suprémum est atteint soit juste avant un saut de F_n, soit juste après. On calcule deux séries d'écarts :

1. $D_n^+ = \max\limits_{1 \leq i \leq n} \left(\frac{i}{n} - F_U(X_{(i)})\right)$ (écarts à droite des points)
2. $D_n^- = \max\limits_{1 \leq i \leq n} \left(F_U(X_{(i)}) - \frac{i-1}{n}\right)$ (écarts à gauche des points)

3. $D_n = \max(D_n^+, D_n^-)$

i	$X_{(i)}$	$F_n(X_{(i)}) = \frac{i}{12}$	$F_U(X_{(i)}) = X_{(i)}$	$\frac{i}{12} - X_{(i)}$	$X_{(i)} - \frac{i-1}{12}$
1	0.15	0.0833	0.15	-0.0667	0.1500
2	0.22	0.1667	0.22	-0.0533	0.1367
3	0.31	0.2500	0.31	-0.0600	0.1433
4	0.42	0.3333	0.42	-0.0867	0.1700
5	0.48	0.4167	0.48	-0.0633	0.1467
6	0.53	0.5000	0.53	-0.0300	0.1133
7	0.61	0.5833	0.61	-0.0267	0.1100
8	0.67	0.6667	0.67	-0.0033	0.0867
9	0.72	0.7500	0.72	0.0300	0.0533
10	0.79	0.8333	0.79	0.0433	0.0400
11	0.85	0.9167	0.85	0.0667	0.0167
12	0.93	1.0000	0.93	0.0700	0.0133

Calcul des maxima :

$$D_n^+ = \max(0, 0.0300, 0.0433, 0.0667, 0.0700) = 0.0700$$

$$D_n^- = \max(0.1500, 0.1367, 0.1433, 0.1700, 0.1467, 0.1133, 0.1100, 0.0867, 0.0533, 0.0400, 0.0167, 0.0133) =$$

$$D_n = \max(D_n^+, D_n^-) = 0.1700$$

Le suprémum est atteint à $x = 0.42$, juste avant le saut de F_n :

$$|F_n(0.42^-) - F_U(0.42)| = |0.2500 - 0.42| = 0.1700$$

d) Calcul de la statistique de Cramer-von Mises W_n^2

La statistique de Cramer-von Mises est définie par :

$$W_n^2 = n \int_{-\infty}^{+\infty} [F_n(t) - F_0(t)]^2 dF_0(t)$$

Formule pratique pour le calcul :

$$W_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left[F_0(X_{(i)}) - \frac{2i-1}{2n} \right]^2$$

Pour la loi uniforme, $F_U(X_{(i)}) = X_{(i)}$, donc :

$$W_n^2 = \frac{1}{12 \times 12} + \sum_{i=1}^{12} \left(X_{(i)} - \frac{2i-1}{24} \right)^2$$

i	$X_{(i)}$	$F_U(X_{(i)}) = X_{(i)}$	$\frac{2i-1}{24}$	$X_{(i)} - \frac{2i-1}{24}$	$\left(X_{(i)} - \frac{2i-1}{24}\right)^2$
1	0.15	0.15	0.0417	0.1083	0.01173
2	0.22	0.22	0.1250	0.0950	0.00903
3	0.31	0.31	0.2083	0.1017	0.01034
4	0.42	0.42	0.2917	0.1283	0.01647
5	0.48	0.48	0.3750	0.1050	0.01103
6	0.53	0.53	0.4583	0.0717	0.00514
7	0.61	0.61	0.5417	0.0683	0.00467
8	0.67	0.67	0.6250	0.0450	0.00203
9	0.72	0.72	0.7083	0.0117	0.00014
10	0.79	0.79	0.7917	-0.0017	0.00000
11	0.85	0.85	0.8750	-0.0250	0.00063
12	0.93	0.93	0.9583	-0.0283	0.00080

Somme des carrés :

$$\sum_{i=1}^{12} \left(X_{(i)} - \frac{2i-1}{24}\right)^2 = 0.01173{+}0.00903{+}0.01034{+}0.01647{+}0.01103{+}0.00514{+}0.00467{+}0.00203{+}0.00$$

Calcul final :

$$\frac{1}{144} = 0.006944$$
$$W_n^2 = 0.006944 + 0.07200 = 0.078944$$

e) Valeurs critiques pour $\alpha = 5\%$

Pour le test de Kolmogorov-Smirnov :
La valeur critique peut être obtenue de plusieurs manières :

1. **Table exacte pour** $n = 12$: $D_{crit}(0.05) \approx 0.375$
2. **Approximation asymptotique :** $D_{crit} \approx \frac{1.36}{\sqrt{n}} = \frac{1.36}{\sqrt{12}} \approx 0.3927$
3. **Interpolation dans la table :**
   — Pour $n = 10$: $D_{crit} = 0.409$
   — Pour $n = 15$: $D_{crit} = 0.338$
   — Pour $n = 12$: $D_{crit} \approx 0.409 + \frac{12-10}{15-10} \times (0.338 - 0.409) = 0.3806$

On retient généralement : $D_{crit} \approx 0.3927$
Pour le test de Cramer-von Mises :
La valeur critique asymptotique est $W^2_{crit} \approx 0.461$

f) Conclusion et comparaison

Test de Kolmogorov-Smirnov :

$$D_n = 0.1700 < 0.3927 \quad \Rightarrow \quad \text{On ne rejette pas } H_0 \text{ au seuil de } 5\%$$

Test de Cramer-von Mises :

$$W_n^2 = 0.0789 < 0.461 \quad \Rightarrow \quad \text{On ne rejette pas } H_0 \text{ au seuil de } 5\%$$

Comparaison :

- Les deux tests conduisent à la même conclusion : non-rejet de H_0
- La statistique KS (0.1700) est plus proche de sa valeur critique (0.3927) que la statistique CvM (0.0789 vs 0.461)
- Le test CvM donne une marge plus confortable pour le non-rejet

Exercice 2 : Test de Normalité

a) Estimation des paramètres μ et σ^2

Les données représentent la taille (en cm) de 15 étudiants :

168.2	172.5	175.1	169.8	178.3	171.6	174.2	167.9
173.4	170.7	176.1	169.3	172.8	175.9	171.1	

Calcul de la moyenne $\bar{x}$

$$\bar{x} = \frac{1}{15}\sum_{i=1}^{15} x_i = \frac{168.2 + 172.5 + 175.1 + 169.8 + 178.3 + 171.6 + 174.2 + 167.9 + 173.4 + 170.7 + 176.1}{15}$$

$$\bar{x} = \frac{2592.9}{15} = 172.86$$

Calcul de la variance s^2 (estimateur non biaisé)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Calcul détaillé des écarts au carré :

$$\begin{aligned}
(168.2 - 172.86)^2 &= (-4.66)^2 = 21.7156 \\
(172.5 - 172.86)^2 &= (-0.36)^2 = 0.1296 \\
(175.1 - 172.86)^2 &= (2.24)^2 = 5.0176 \\
(169.8 - 172.86)^2 &= (-3.06)^2 = 9.3636 \\
(178.3 - 172.86)^2 &= (5.44)^2 = 29.5936 \\
(171.6 - 172.86)^2 &= (-1.26)^2 = 1.5876 \\
(174.2 - 172.86)^2 &= (1.34)^2 = 1.7956 \\
(167.9 - 172.86)^2 &= (-4.96)^2 = 24.6016 \\
(173.4 - 172.86)^2 &= (0.54)^2 = 0.2916 \\
(170.7 - 172.86)^2 &= (-2.16)^2 = 4.6656 \\
(176.1 - 172.86)^2 &= (3.24)^2 = 10.4976 \\
(169.3 - 172.86)^2 &= (-3.56)^2 = 12.6736 \\
(172.8 - 172.86)^2 &= (-0.06)^2 = 0.0036 \\
(175.9 - 172.86)^2 &= (3.04)^2 = 9.2416 \\
(171.1 - 172.86)^2 &= (-1.76)^2 = 3.0976
\end{aligned}$$

Somme des écarts au carré :

$$\sum_{i=1}^{15}(x_i - \bar{x})^2 = 134.28$$

Variance :

$$s^2 = \frac{134.28}{14} = 9.5914$$
$$s = \sqrt{9.5914} = 3.0968$$

b) Calcul de la statistique de Kolmogorov-Smirnov D_n pour le test de normalité

La fonction de répartition théorique estimée est :

$$F_0(x) = \Phi\left(\frac{x - \bar{x}}{s}\right)$$

où Φ est la fonction de répartition de la loi normale centrée réduite.

Données ordonnées :

$$167.9, 168.2, 169.3, 169.8, 170.7, 171.1, 171.6, 172.5, 172.8, 173.4, 174.2, 175.1, 175.9, 176.1, 178.3$$

Tableau de calcul détaillé :

i	$X_{(i)}$	$\frac{i}{15}$	$z_i = \frac{X_{(i)} - \bar{x}}{s}$	$F_0(X_{(i)}) = \Phi(z_i)$	$\left\lvert \frac{i}{15} - F_0 \right\rvert$	$\left\lvert \frac{i-1}{15} - F_0 \right\rvert$
1	167.9	0.0667	-1.602	0.0545	0.0122	0.0545
2	168.2	0.1333	-1.506	0.0660	0.0673	0.0660
3	169.3	0.2000	-1.150	0.1251	0.0749	0.0584
4	169.8	0.2667	-0.988	0.1615	0.1052	0.0382
5	170.7	0.3333	-0.698	0.2425	0.0908	0.0908
6	171.1	0.4000	-0.569	0.2843	0.1157	0.1177
7	171.6	0.4667	-0.407	0.3421	0.1246	0.1246
8	172.5	0.5333	-0.116	0.4538	0.0795	0.0871
9	172.8	0.6000	-0.019	0.4924	0.1076	0.0924
10	173.4	0.6667	0.174	0.5691	0.0976	0.1024
11	174.2	0.7333	0.433	0.6673	0.0660	0.0660
12	175.1	0.8000	0.724	0.7657	0.0343	0.0990
13	175.9	0.8667	0.982	0.8365	0.0302	0.1032
14	176.1	0.9333	1.046	0.8524	0.0809	0.0857
15	178.3	1.0000	1.757	0.9605	0.0395	0.0272

Calcul de D_n^+ et D_n^- :

$$D_n^+ = \max(0.0122, 0.0673, 0.0749, 0.1052, 0.0908, 0.1157, 0.1246, 0.0795, 0.1076, 0.0976, 0.0660, 0.0343, 0.0$$
$$D_n^- = \max(0.0545, 0.0660, 0.0584, 0.0382, 0.0908, 0.1177, 0.1246, 0.0871, 0.0924, 0.1024, 0.0660, 0.0990, 0.1$$
$$D_n = \max(D_n^+, D_n^-) = 0.1246$$

Le maximum est atteint à $i = 7$ (valeur $x = 171.6$).

c) Détermination de la p-value avec la table de Lilliefors

La table de Lilliefors pour $n = 15$ donne les valeurs critiques suivantes :

D_n	p-value approximative
0.125	0.15
0.134	0.10
0.144	0.05
0.166	0.01

Pour $D_n = 0.1246$:

— Comparaison directe : $0.1246 < 0.125 \Rightarrow$ p-value > 0.15

— Interpolation pour estimation plus précise :

$$\text{p-value} \approx 0.15 - \frac{0.1246 - 0.125}{0.134 - 0.125} \times (0.15 - 0.10) \approx 0.15 + 0.0022 \approx 0.1522$$

Donc p-value ≈ 0.15 (15%).

d) Effectuer le test de normalité au niveau $\alpha = 5\%$

Valeur critique de Lilliefors pour $n = 15$ et $\alpha = 0.05$:

$$D_{crit} = 0.144$$

Comparaison :

$$D_n = 0.1246 < 0.144$$

Conclusion : On ne rejette pas H_0 au seuil de 5%. Les données sont compatibles avec une distribution normale.

e) Limitations du test KS lorsque les paramètres sont estimés

1. **Modification de la distribution de D_n** : La distribution de la statistique D_n n'est plus celle de Kolmogorov lorsqu'on estime les paramètres. Il faut utiliser la table spécifique de Lilliefors.
2. **Perte de puissance** : Le test est moins puissant (plus conservateur) car l'estimation des paramètres "ajuste" la distribution théorique aux données.
3. **Tables spécifiques nécessaires** : Pour chaque loi et chaque méthode d'estimation, il faut des tables spécifiques. La table de Lilliefors est spécifique au test de normalité avec paramètres estimés par maximum de vraisemblance.
4. **Approximations asymptotiques différentes** : La distribution asymptotique de $\sqrt{n}D_n$ change lorsqu'on estime les paramètres.
5. **Validité limitée pour petits échantillons** : Pour les très petits échantillons, les approximations peuvent être médiocres.
6. **Alternatives plus puissantes** : Pour le test de normalité, d'autres tests comme Shapiro-Wilk sont généralement plus puissants.

Exercice 3 : Test à Deux Échantillons

a) Formulation des hypothèses

Test d'homogénéité de Kolmogorov-Smirnov :

$$H_0 : F_A(x) = F_B(x) \quad \text{pour tout } x \quad \text{(les deux échantillons proviennent de la même distribution)}$$

$$H_1 : F_A(x) \neq F_B(x) \quad \text{pour au moins un } x \quad \text{(les distributions sont différentes)}$$

b) Calcul des fonctions de répartition empiriques $F_n(x)$ et $G_m(x)$

Groupe A (Méthode traditionnelle, $n = 8$) :

$$12.3, 14.7, 11.8, 13.5, 15.2, 12.9, 14.1, 13.8$$

Groupe B (Nouvelle méthode, $m = 9$) :

$$13.8, 15.6, 14.3, 16.1, 12.7, 15.9, 14.8, 16.3, 15.1$$

Étape 1 : Trier toutes les données ensemble

Valeur	Groupe	Rang global
11.8	A	1
12.3	A	2
12.7	B	3
12.9	A	4
13.5	A	5
13.8	A	6
13.8	B	6
14.1	A	8
14.3	B	9
14.7	A	10
14.8	B	11
15.1	B	12
15.2	A	13
15.6	B	14
15.9	B	15
16.1	B	16
16.3	B	17

Étape 2 : Calculer $F_n(x)$ pour le groupe A

$$F_n(x) = \frac{\text{nombre d'observations dans A} \leq x}{8}$$

x	Nombre A $\leq x$	$F_n(x)$	Explication
11.8	1	0.125	Seulement 11.8
12.3	2	0.250	Ajout de 12.3
12.7	2	0.250	Aucun nouveau A
12.9	3	0.375	Ajout de 12.9
13.5	4	0.500	Ajout de 13.5
13.8	5	0.625	Ajout de 13.8 (celui du groupe A)
14.1	6	0.750	Ajout de 14.1
14.3	6	0.750	Aucun nouveau A
14.7	7	0.875	Ajout de 14.7
14.8	7	0.875	Aucun nouveau A
15.1	7	0.875	Aucun nouveau A
15.2	8	1.000	Ajout de 15.2 (dernier de A)
15.6	8	1.000	Tous les A $\leq$ 15.6
15.9	8	1.000	Tous les A $\leq$ 15.9
16.1	8	1.000	Tous les A $\leq$ 16.1
16.3	8	1.000	Tous les A $\leq$ 16.3

Étape 3 : Calculer $G_m(x)$ pour le groupe B

$$G_m(x) = \frac{\text{nombre d'observations dans B} \leq x}{9}$$

x	Nombre B $\leq x$	$G_m(x)$	Explication
11.8	0	0.000	Aucun B $\leq$ 11.8
12.3	0	0.000	Aucun B $\leq$ 12.3
12.7	1	0.111	Ajout de 12.7
12.9	1	0.111	Aucun nouveau B
13.5	1	0.111	Aucun nouveau B
13.8	2	0.222	Ajout de 13.8 (celui du groupe B)
14.1	2	0.222	Aucun nouveau B
14.3	3	0.333	Ajout de 14.3
14.7	3	0.333	Aucun nouveau B
14.8	4	0.444	Ajout de 14.8
15.1	5	0.556	Ajout de 15.1
15.2	5	0.556	Aucun nouveau B
15.6	6	0.667	Ajout de 15.6
15.9	7	0.778	Ajout de 15.9
16.1	8	0.889	Ajout de 16.1
16.3	9	1.000	Ajout de 16.3 (dernier de B)

c) Calcul de la statistique de Kolmogorov-Smirnov à deux échantillons

$$D_{n,m} = \sup_x |F_n(x) - G_m(x)|$$

Tableau complet des écarts :

x	$F_n(x)$	$G_m(x)$	$\|F_n(x) - G_m(x)\|$	Maximum local
11.8	0.125	0.000	0.125	
12.3	0.250	0.000	0.250	
12.7	0.250	0.111	0.139	
12.9	0.375	0.111	0.264	
13.5	0.500	0.111	0.389	
13.8	0.625	0.222	0.403	
14.1	0.750	0.222	**0.528**	← Maximum
14.3	0.750	0.333	0.417	
14.7	0.875	0.333	0.542	
14.8	0.875	0.444	0.431	
15.1	0.875	0.556	0.319	
15.2	1.000	0.556	0.444	
15.6	1.000	0.667	0.333	
15.9	1.000	0.778	0.222	
16.1	1.000	0.889	0.111	
16.3	1.000	1.000	0.000	

$$D_{n,m} = 0.528$$

Le maximum absolu est atteint à $x = 14.1$.

d) Détermination de la valeur critique pour un niveau $\alpha = 5\%$

Pour le test de Kolmogorov-Smirnov à deux échantillons, la valeur critique dépend de n, m et α.

Formule approximative pour $n, m \geq 4$:

$$D_{crit}(\alpha) = c(\alpha)\sqrt{\frac{n+m}{n \times m}}$$

avec $c(0.05) = 1.36$.

Application avec $n = 8$, $m = 9$, $\alpha = 0.05$:

$$D_{crit} = 1.36 \times \sqrt{\frac{8+9}{8 \times 9}} = 1.36 \times \sqrt{\frac{17}{72}} = 1.36 \times \sqrt{0.2361} = 1.36 \times 0.4859 = 0.661$$

Table exacte pour $n = 8$, $m = 9$: La table de Smirnov donne $D_{crit}(0.05) \approx 0.638$. On utilise la valeur de la table : $D_{crit} = 0.638$.

e) Effectuer le test et conclure

Comparaison :

$$D_{n,m} = 0.528 < 0.638 = D_{crit}$$

Décision : On ne rejette pas H_0 au seuil de 5%.

Conclusion : Il n'y a pas de preuve statistique suffisante pour affirmer que les deux méthodes d'apprentissage produisent des distributions de résultats différentes.

f) Proposer une interprétation pédagogique des résultats

1. **Résultat statistique :** Aucune différence significative n'a été détectée entre les distributions des résultats des deux méthodes au seuil de 5%.
2. **Analyse descriptive :**
   - La nouvelle méthode semble donner des résultats légèrement supérieurs en moyenne
   - Mais la variabilité est également plus grande dans le groupe B
   - Les écarts observés ne sont pas statistiquement significatifs avec ces effectifs
3. **Limitations :**
   - Effectifs modestes ($n = 8$, $m = 9$)
   - Test peu puissant avec petits échantillons
   - Risque d'erreur de type II (ne pas détecter une différence réelle)
4. **Recommandations pédagogiques :**
   - Répéter l'expérience avec un échantillon plus large (au moins 30 étudiants par groupe)
   - Considérer d'autres critères d'évaluation : coût, temps d'apprentissage, satisfaction des étudiants
   - Analyser si la nouvelle méthode profite davantage à certains types d'étudiants
5. **Conclusion pratique :** Avec les données actuelles, on ne peut pas recommander l'adoption généralisée de la nouvelle méthode basée sur des critères statistiques de performance.

Exercice 4 : Test d'adéquation à la loi uniforme

a) Calcul de la fonction de répartition théorique $F_U(t)$ de la loi uniforme sur [0,1]

Pour la loi uniforme sur $[0, 1]$, la fonction de répartition est :

$$F_U(t) = \begin{cases} 0 & \text{si } t < 0 \\ t & \text{si } 0 \leq t \leq 1 \\ 1 & \text{si } t > 1 \end{cases}$$

b) Tableau contenant pour chaque observation

Les 20 observations ordonnées sont :

$$0.278, 0.452, 0.464, 0.494, 0.496, 0.505, 0.576, 0.592, 0.602, 0.608, 0.661, 0.683, 0.690, 0.696, 0.704, 0.728, 0.7$$

Tableau complet des calculs :

i	$x_{(i)}$	$F_n(x_{(i)}) = \frac{i}{20}$	$F_U(x_{(i)}) = x_{(i)}$	$\lvert F_n(x_{(i)}) - F_U(x_{(i)})\rvert$	$D^+ = \frac{i}{20} - x_{(i)}$	$D^- = x_{(i)} - \frac{i-1}{20}$
1	0.278	0.05	0.278	0.228	-0.228	0.228
2	0.452	0.10	0.452	0.352	-0.352	0.352
3	0.464	0.15	0.464	0.314	-0.314	0.314
4	0.494	0.20	0.494	0.294	-0.294	0.294
5	0.496	0.25	0.496	0.246	-0.246	0.246
6	0.505	0.30	0.505	0.205	-0.205	0.205
7	0.576	0.35	0.576	0.226	-0.226	0.226
8	0.592	0.40	0.592	0.192	-0.192	0.192
9	0.602	0.45	0.602	0.152	-0.152	0.152
10	0.608	0.50	0.608	0.108	-0.108	0.108
11	0.661	0.55	0.661	0.111	-0.111	0.111
12	0.683	0.60	0.683	0.083	-0.083	0.083
13	0.690	0.65	0.690	0.040	-0.040	0.040
14	0.696	0.70	0.696	0.004	-0.004	0.004
15	0.704	0.75	0.704	0.046	0.046	-0.046
16	0.728	0.80	0.728	0.072	0.072	-0.072
17	0.754	0.85	0.754	0.096	0.096	-0.096
18	0.850	0.90	0.850	0.050	0.050	-0.050
19	0.902	0.95	0.902	0.048	0.048	-0.048
20	0.949	1.00	0.949	0.051	0.051	-0.051

c) Détermination de la statistique $D_n = \sup \lvert F_n(t) - F_U(t)\rvert$

Calcul de D_n^+ et D_n^- :

$$D_n^+ = \max(0, 0.046, 0.072, 0.096, 0.050, 0.048, 0.051) = 0.096$$

$$D_n^- = \max(0.228, 0.352, 0.314, 0.294, 0.246, 0.205, 0.226, 0.192, 0.152, 0.108, 0.111, 0.083, 0.040, 0.004) = 0$$

$$D_n = \max(D_n^+, D_n^-) = 0.352$$

Le suprémum est atteint à $x = 0.452$ (vérification à gauche du point) :

$$\lvert F_n(0.452^-) - F_U(0.452)\rvert = \lvert 0.05 - 0.452\rvert = 0.402$$

Note : Dans notre tableau, nous avons 0.352, mais le calcul exact donne 0.402. Corrigeons :

À $i = 2$, $x_{(2)} = 0.452$: - $F_n(0.452^-) = 0.05$ - $F_U(0.452) = 0.452$ - Écart $= |0.05 - 0.452| = 0.402$

Donc $D_n = 0.402$.

d) Calcul de la statistique de Cramer-von Mises W_n^2

$$W_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(F_U(x_{(i)}) - \frac{2i-1}{2n} \right)^2$$
$$= \frac{1}{12 \times 20} + \sum_{i=1}^{20} \left(x_{(i)} - \frac{2i-1}{40} \right)^2$$

Tableau de calcul (extrait) :

i	$x_{(i)}$	$\frac{2i-1}{40}$	$x_{(i)} - \frac{2i-1}{40}$	$\left(x_{(i)} - \frac{2i-1}{40}\right)^2$
1	0.278	0.025	0.253	0.064009
2	0.452	0.075	0.377	0.142129
3	0.464	0.125	0.339	0.114921
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
18	0.850	0.875	-0.025	0.000625
19	0.902	0.925	-0.023	0.000529
20	0.949	0.975	-0.026	0.000676

Somme des carrés :

$$\sum_{i=1}^{20} \left(x_{(i)} - \frac{2i-1}{40} \right)^2 \approx 0.752616$$

Calcul final :

$$\frac{1}{240} = 0.004167$$
$$W_n^2 = 0.004167 + 0.752616 = 0.756783$$

e) Conclusion avec les valeurs critiques données

Valeurs critiques fournies : - KS : $c_{KS} \approx \frac{1.36}{\sqrt{n}} = \frac{1.36}{\sqrt{20}} \approx 0.304$ - CvM : $c_{CvM} \approx 0.461$

Comparaison :

Test de Kolmogorov-Smirnov :

$$D_n = 0.402 > 0.304 \quad \Rightarrow \quad \text{Rejet de } H_0$$

Test de Cramer-von Mises :

$$W_n^2 = 0.7568 > 0.461 \quad \Rightarrow \quad \text{Rejet de } H_0$$

Conclusion : Les deux tests rejettent l'hypothèse que les données proviennent d'une loi uniforme sur [0,1] au seuil de 5%.

f) Comparer les sensibilités des deux tests pour cet échantillon

1. **Sensibilité du test de Kolmogorov-Smirnov :**
   — Détecte le suprémum des écarts absolus
   — Ici : $D_n = 0.402$, valeur critique = 0.304
   — Rapport : $0.402/0.304 = 1.32$ (32% au-dessus du seuil)
   — Sensible aux écarts localisés importants
2. **Sensibilité du test de Cramer-von Mises :**
   — Intègre tous les écarts au carré
   — Ici : $W_n^2 = 0.7568$, valeur critique = 0.461
   — Rapport : $0.7568/0.461 = 1.64$ (64% au-dessus du seuil)
   — Sensible aux écarts répartis sur tout l'intervalle
3. **Analyse comparative :**
   — Le test CvM montre une déviation plus forte par rapport à sa valeur critique
   — Cela suggère que les écarts sont répartis sur tout l'intervalle plutôt que concentrés en un point
   — Les données semblent montrer une concentration vers le milieu de l'intervalle [0,1]
   — Manque relatif de données près des extrêmes (0 et 1)
4. **Conclusion sur la sensibilité :** Pour cet échantillon, le test de Cramer-von Mises est plus sensible que le test de Kolmogorov-Smirnov, indiquant que les écarts à la loi uniforme sont de nature globale plutôt que locale.

Exercice 5 : Synthèse Théorique

a) Propriétés d'invariance

i) Démontrer que la statistique de Kolmogorov-Smirnov est invariante par transformation monotone croissante

Démonstration :

Soient : - $X_1, X_2, ..., X_n$ un échantillon i.i.d. de loi F_X - $g : \mathbb{R} \to \mathbb{R}$ une fonction strictement croissante et continue - $Y_i = g(X_i)$ pour $i = 1, ..., n$

1. **Fonction de répartition de Y :**

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

2. **Fonction de répartition empirique de Y :** Pour les Y_i ordonnés $y_{(1)} \leq ... \leq y_{(n)}$:

$$F_n^Y(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Y_i \leq y\}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq g^{-1}(y)\}} = F_n^X(g^{-1}(y))$$

3. **Statistique KS pour Y :**

$$D_n^Y = \sup_{y\in\mathbb{R}} |F_n^Y(y) - F_Y(y)| = \sup_{y\in\mathbb{R}} |F_n^X(g^{-1}(y)) - F_X(g^{-1}(y))|$$

4. **Changement de variable :** Posons $t = g^{-1}(y)$. Comme g est strictement croissante, la transformation $y \mapsto t$ est bijective et préserve l'ordre. De plus, quand y parcourt $\mathbb{R}$, t parcourt $\mathbb{R}$ aussi.

$$D_n^Y = \sup_{t\in\mathbb{R}} |F_n^X(t) - F_X(t)| = D_n^X$$

Conclusion : La statistique de Kolmogorov-Smirnov est invariante par transformation monotone croissante.

ii) Montrer que cette propriété s'applique également au test de Cramer-von Mises

Démonstration :

$$W_n^{2,Y} = n\int_{-\infty}^{+\infty} [F_n^Y(y) - F_Y(y)]^2 dF_Y(y)$$

Par changement de variable $y = g(t)$, on a $dy = g'(t)dt$ et :

$$\begin{aligned} dF_Y(y) &= f_Y(y)dy = f_X(g^{-1}(y)) \cdot |(g^{-1})'(y)|dy \\ &= f_X(t) \cdot \frac{1}{g'(t)} \cdot g'(t)dt = f_X(t)dt \end{aligned}$$

Donc :

$$W_n^{2,Y} = n\int_{-\infty}^{+\infty} [F_n^X(t) - F_X(t)]^2 f_X(t)dt = W_n^{2,X}$$

Conclusion : Le test de Cramer-von Mises est également invariant par transformation monotone croissante.

iii) Expliquer l'importance pratique de cette invariance

1. **Flexibilité d'application :** Permet de tester l'adéquation sur différentes échelles sans changer la conclusion.
   — Exemple : Tester la log-normalité en testant la normalité des données transformées par logarithme
   — Exemple : Transformation pour stabiliser la variance avant test
2. **Utilisation des tables :** Les mêmes tables de valeurs critiques s'appliquent aux données transformées.
3. **Robustesse :** Le test reste valide après transformation monotone des données.
4. **Applications courantes :**
   — En économétrie : tests sur échelles logarithmiques
   — En fiabilité : tests sur données transformées par fonction de hasard
   — En biostatistique : transformations pour normaliser les données
5. **Simplicité :** On peut choisir l'échelle la plus pratique pour les calculs ou l'interprétation.

b) Consistance des tests

i) Démontrer que le test de Kolmogorov-Smirnov est consistant contre toute alternative fixe

Démonstration :
Soit $H_1 : F \neq F_0$. Il existe alors au moins un point t_0 tel que :

$$|F(t_0) - F_0(t_0)| = \delta > 0$$

Par le théorème de Glivenko-Cantelli :

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow{\text{p.s.}} 0 \quad \text{quand } n \to \infty$$

Donc, pour n assez grand, avec probabilité tendant vers 1 :

$$|F_n(t_0) - F_0(t_0)| \approx |F(t_0) - F_0(t_0)| = \delta$$

Ainsi :

$$D_n = \sup_t |F_n(t) - F_0(t)| \geq \delta > 0$$

La puissance du test :

$$\pi_n = P_{H_1}(D_n > c_\alpha) \to 1 \quad \text{quand } n \to \infty$$

Conclusion : Le test de Kolmogorov-Smirnov est consistant contre toute alternative fixe.

ii) Montrer que le test de Cramer-von Mises est également consistant

Démonstration :
Sous $H_1 : F \neq F_0$, on a :

$$\int [F(t) - F_0(t)]^2 dF_0(t) = \Delta > 0$$

Par convergence de F_n vers F (Glivenko-Cantelli) :

$$W_n^2 = n \int [F_n(t) - F_0(t)]^2 dF_0(t) \xrightarrow{\text{p.s.}} \infty$$

Donc :

$$P_{H_1}(W_n^2 > c_\alpha) \to 1 \quad \text{quand } n \to \infty$$

Conclusion : Le test de Cramer-von Mises est également consistant.

iii) Comparer la puissance asymptotique des deux tests

1. **Test de Kolmogorov-Smirnov :**
   - Plus sensible aux alternatives présentant des écarts localisés importants
   - Moins sensible aux alternatives avec des écarts faibles mais répartis
   - Puissance élevée contre les alternatives avec discontinuités
2. **Test de Cramer-von Mises :**
   - Plus sensible aux alternatives présentant des écarts faibles mais répartis
   - Intègre tous les écarts, pas seulement le maximum
   - Généralement plus puissant contre les alternatives régulières
3. **Comparaison théorique :**
   - Pour des alternatives de la forme $F(t) = t + \epsilon h(t)$ avec h régulière :
     - Le test CvM est localement plus puissant
     - L'efficacité relative de Pitman est plus élevée pour CvM
   - Pour des alternatives avec sauts localisés :
     - Le test KS peut être plus puissant
     - KS détecte mieux les discontinuités
4. **Résultats pratiques :**
   - En général, CvM est plus puissant que KS contre la plupart des alternatives
   - Mais KS est plus simple à comprendre et à mettre en œuvre
   - Le choix dépend de la nature des alternatives attendues

c) Distribution asymptotique

i) Établir que sous H_0 :

$$\sqrt{n} D_n \xrightarrow{\mathcal{L}} \sup_{0 \leq t \leq 1} |B(t)|$$

où $B(t)$ est un pont brownien.
Démonstration :

1. **Théorème de Donsker (principe d'invariance) :** Si $X_1, ..., X_n \sim F_0$ continue, alors :
$$\sqrt{n}(F_n(t) - F_0(t)) \xrightarrow{\mathcal{L}} B(F_0(t))$$
où $B(\cdot)$ est un pont brownien sur [0,1].
2. **Cas particulier $F_0(t) = t$ (loi uniforme) :**
$$\sqrt{n}(F_n(t) - t) \xrightarrow{\mathcal{L}} B(t)$$
3. **Application au suprémum :** Par continuité de la fonctionnelle "suprémum" :
$$\sqrt{n}D_n = \sqrt{n}\sup_t |F_n(t) - t| \xrightarrow{\mathcal{L}} \sup_{0\leq t\leq 1} |B(t)|$$
4. **Cas général :** Pour une loi F_0 continue quelconque, par transformation de la variable :
$$\sqrt{n}D_n \xrightarrow{\mathcal{L}} \sup_{0\leq t\leq 1} |B(t)|$$

ii) Donner l'expression de la fonction de répartition limite de $\sqrt{n}D_n$

Pour $K = \sup_{0\leq t\leq 1} |B(t)|$, la fonction de répartition est donnée par :
$$P(K \leq x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2} \quad \text{pour } x > 0$$
Développement alternatif :
$$P(K \leq x) = \frac{\sqrt{2\pi}}{x}\sum_{k=1}^{\infty} e^{-(2k-1)^2\pi^2/(8x^2)}$$
Approximation pratique :
Pour les calculs, on utilise souvent :
$$P(\sqrt{n}D_n \leq x) \approx 1 - 2e^{-2x^2} + 2e^{-8x^2} - 2e^{-18x^2} + \cdots$$
Cette série converge rapidement, surtout pour x modéré.

iii) Expliquer comment cette distribution est utilisée pour construire les tables du test KS

1. **Pour les grands échantillons ($n > 50$) :**
   — On utilise directement la distribution asymptotique
   — Les quantiles x_α vérifiant $P(K \leq x_\alpha) = 1 - \alpha$ sont tabulés
   — Puis on déduit $D_{crit} = x_\alpha/\sqrt{n}$
2. **Valeurs critiques asymptotiques courantes :**

α	x_α
0.20	1.073
0.10	1.224
0.05	1.358
0.025	1.480
0.01	1.628

3. **Pour les petits échantillons :**
   - La distribution exacte de D_n pour n fini est connue
   - Formule de Birnbaum et Tingey (1951) :

$$P(D_n \leq d) = 1 - d \sum_{j=0}^{\lfloor n(1-d) \rfloor} \binom{n}{j} \left(\frac{j}{n} + d \right)^{j-1} \left(1 - d - \frac{j}{n} \right)^{n-j}$$

   - Ces valeurs exactes sont tabulées pour différents n et α
4. **Méthode de construction des tables :**
   (a) Pour chaque n (de 1 à 50 généralement)
   (b) Calculer la distribution exacte de D_n sous H_0
   (c) Déterminer les quantiles correspondant aux seuils α usuels
   (d) Présenter sous forme de tableau
5. **Approximation pour n intermédiaire :**

$$D_{crit}(\alpha) \approx \frac{x_\alpha}{\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}}$$

   Cette formule de Stephens (1974) donne une bonne approximation.

d) Extensions et généralisations

i) Proposer une extension du test KS pour des données censurées

Problème : Avec des données censurées à droite, la fonction de répartition empirique F_n n'est pas directement observable.

Solution : Utiliser l'estimateur de Kaplan-Meier $\hat{S}(t)$:

1. **Estimateur de Kaplan-Meier :**

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i} \right)$$

   où d_i est le nombre d'événements au temps t_i et n_i est le nombre d'individus à risque juste avant t_i.
2. **Fonction de répartition estimée :**

$$\hat{F}_n(t) = 1 - \hat{S}(t)$$

3. **Statistique KS modifiée :**

$$D_n^{cens} = \sup_t |\hat{F}_n(t) - F_0(t)|$$

4. **Difficultés :**
   - La distribution de D_n^{cens} sous H_0 est complexe
   - Doit être déterminée par simulation
   - Dépend du schéma de censure
5. **Variantes :**
   - Test de Fleming-Harrington
   - Test de log-rank modifié
   - Tests basés sur les processus de martingales

ii) Discuter des adaptations nécessaires pour le test CvM dans le cas de données groupées

Problème : Pour des données groupées en k classes, F_n est une fonction en escalier.
Adaptations :

1. **Formulation discrète :**

$$W^2 = n \sum_{j=1}^{k} p_j \left(\frac{F_n(x_j) - F_0(x_j)}{p_j} \right)^2$$

   où :
   - $p_j = F_0(b_j) - F_0(a_j)$ est la probabilité théorique de la classe j
   - $F_n(x_j)$ est la fréquence cumulée observée jusqu'à la classe j
   - $F_0(x_j)$ est la probabilité cumulée théorique jusqu'à la classe j

2. **Formulation alternative :**

$$W^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j} + \text{termes correctifs}$$

   où O_j et E_j sont les effectifs observés et théoriques.

3. **Modifications nécessaires :**
   - Utiliser les centres de classe pour le calcul
   - Adapter la formule d'intégration 4cm→ Remplacer l'intégrale par une somme discrète
   - Ajuster les valeurs critiques 4cm→ Tables spécifiques ou simulation

4. **Limitations :**
   - Perte d'information due au regroupement
   - Puissance réduite
   - Dépendance au choix des classes

5. **Alternative :** Test du chi-deux, plus adapté aux données groupées.

iii) Mentionner d'autres tests d'adéquation non paramétriques et situer KS et CvM dans ce paysage

1. **Principaux tests d'adéquation non paramétriques :**
   - **Test d'Anderson-Darling :**

$$A_n^2 = n \int_{-\infty}^{+\infty} \frac{[F_n(t) - F_0(t)]^2}{F_0(t)[1 - F_0(t)]} dF_0(t)$$

     - Pondère les queues de distribution
     - Plus puissant que KS et CvM pour détecter des écarts dans les queues
     - Particulièrement utile pour les tests de normalité
   - **Test de Shapiro-Wilk :**
     - Spécifique au test de normalité
     - Basé sur les statistiques d'ordre et les coefficients optimaux

— Généralement le plus puissant pour détecter la non-normalité
— Limitée à la normalité

— **Test du chi-deux :**

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-p-1}$$

— Pour données discrètes ou groupées
— Nécessite des effectifs suffisants par classe
— Moins puissant que les tests basés sur la fonction de répartition pour données continues

— **Test de Kuiper :**

$$V_n = D_n^+ + D_n^-$$

— Invariant aux transformations circulaires
— Utile pour les données angulaires ou périodiques
— Moins sensible aux écarts au centre de la distribution

— **Test de Watson :**

$$U_n^2 = n \int_{-\infty}^{+\infty} \left[F_n(t) - F_0(t) - \int (F_n(u) - F_0(u)) dF_0(u) \right]^2 dF_0(t)$$

— Modifié pour être invariant par changement d'origine
— Utile pour les données circulaires

2. **Position des tests KS et CvM :**

Test	Simplicité	Puissance	Généralité	Usage courant
Kolmogorov-Smirnov				Tests généraux
Cramer-von Mises				Alternatives lisses
Anderson-Darling				Tests de normalité
Shapiro-Wilk				Normalité seulement
Chi-deux				Données groupées

3. **Recommandations pour le choix :**
   — **Pour un test général simple :** KS
   — **Pour plus de puissance contre alternatives lisses :** CvM ou Anderson-Darling
   — **Pour tester spécifiquement la normalité :** Shapiro-Wilk ou Anderson-Darling
   — **Pour données groupées :** Chi-deux
   — **Pour données circulaires :** Kuiper ou Watson

4. **Avantages de KS et CvM :**
   — Applicables à toute loi continue
   — Invariants par transformation monotone
   — Tables largement disponibles

— Interprétation graphique intuitive (surtout pour KS)
— Consistants contre toute alternative

5. **Limitations de KS et CvM :**
   — Moins puissants que des tests spécialisés
   — Sensibles aux arrondis ou discrétisations
   — Pour KS : moins sensible aux écarts dans les queues
   — Pour CvM : calcul plus complexe que KS